

**SAMPLING
TECHNIQUES**
second edition

**William
G. Cochran**
Professor
of Statistics
Harvard
University

JOHN WILEY & SONS, INC.
NEW YORK — LONDON
1963

Уильям Кокрен

**МЕТОДЫ
ВЫБОРОЧНОГО
ИССЛЕДОВАНИЯ**

Перевод с английского Н. М. СОННИНА
Под редакцией А. Г. ВОЛКОВА
Предисловие к русскому переводу Н. К. ДРУЖНИНА



МОСКВА «СТАТИСТИКА» 1976

Кокрен У.

К59 Методы выборочного исследования.

Пер. с англ. И. М. Сонина. Под ред. А. Г. Волкова. Вступит. статья Н. К. Дружинина. М., «Статистика», 1976. 440 с. с ил.

Книга содержит систематическое изложение методов отбора, оценивания и вычисления ошибок выборки при выборочных исследованиях. Вывод формул оценок и их выборочных дисперсий сопровождается многочисленными примерами из практики английской и американской статистики. Рассмотрены преимущества, недостатки и условия применения различных способов отбора. Излагаются результаты новейших исследований в области методики выборочных исследований. Каждая из 13 глав книги сопровождается упражнениями для самостоятельной работы и списком литературы.

Книга представляет интерес для широкого круга исследователей. Она полезна экономистам, статистикам, социологам, а также всем, кто проводит выборочные исследования или пользуется их результатами.

К $\frac{10805^*-079}{008(01)-76}$ 114-76

517.8

* Второй индекс 10803.

ПРЕДИСЛОВИЕ К РУССКОМУ ПЕРЕВОДУ

Автор предлагаемой советскому читателю книги Уильям Кокрен — известный американский ученый-статистик, профессор Гарвардского университета. Кроме настоящего фундаментального труда, перу профессора Кокрена принадлежит ряд статей на темы, относящиеся к теории выборочного метода; Кокрен занимался также близкими к теории выборочного метода вопросами теории эксперимента. Вместе с Гертрудой Кокс — директором Института статистики при университете штата Северная Каролина — он написал книгу о планировании экспериментов («Experimental Designs»), вышедшую в свет вторым изданием в 1957 г.

Содержание настоящей книги профессора Кокрена несомненно привлечет внимание специалистов-статистиков. Излишне говорить о том, какое важное место занимает выборочный метод в статистической науке. Вся математическая ветвь этой науки развивалась как учение о вероятностной оценке данных, получаемых в результате выборочного статистического наблюдения или эксперимента. Практическое значение выборочного метода также очень велико. Возможность сократить расходы на проведение статистических работ представляет собой немаловажный довод в пользу применения выборочного наблюдения вместо наблюдения сплошного. К этому нередко присоединяется отсутствие возможности вообще, по тем или иным причинам, провести сплошное статистическое обследование.

Роль выборочного метода остается весьма значительной и в государственной статистике СССР, где наряду со статистической информацией, получаемой в порядке функционирования единой системы народнохозяйственного учета или проведения специальных сплошных переписей, перед статистическими органами возникают и такие задачи, которые возможно и желательно разрешать лишь путем выборочного наблюдения.

Заглавие книги Кокрена свидетельствует о том, что в ней не излагаются теоретические основы выборочного метода. Хотя в предисловии к книге автор говорит, что «цель этой книги — как можно более полно

изложить теорию выборочного метода», но собственно теории выборочного метода, в нашем ее понимании, он посвящает лишь немногие страницы, сосредоточивая внимание на математической стороне этого метода, на вычислении оценок по данным наблюдения и их ошибок, на организации самого выборочного наблюдения. Это отнюдь не умаляет достоинств книги. Общие теоретические основания выборочного метода, заключающиеся в положениях теории вероятностей, достаточно хорошо известны и едва ли требуется лишний раз к ним обращаться. Гораздо важнее осветить многочисленные и нередко сложные вопросы методики выборочных обследований. Автор, говоря о той или иной форме отбора, обычно дает математическую модель и вывод формул, что также представляет большой интерес для читателя-специалиста. Книга Кокрена, несомненно, одна из лучших работ о выборочном методе в западной литературе. Поэтому издание ее перевода представляется весьма желательным.

Кокрен в своем изложении обращается прежде всего к классическому образцу применения выборочного метода — к простому случайному отбору. В такой выборке с наибольшей отчетливостью проявляются законы случая. Здесь не налагается никакого ограничения на действие механизма важнейшего из них — закона больших чисел, определяющего всю логическую конструкцию выборочного метода. Теория выборочного метода создавалась, исходя именно из представления о случайном отборе, а практика выборочного статистического наблюдения во всех своих видах почти всегда имеет в виду такой отбор как основу своей организации. Таким образом, изложение вопросов, касающихся главных положений теории случайного отбора и математического выражения оценки его результатов, играет в книге Кокрена роль необходимого введения читателя в круг проблем, составляющих основное содержание этой книги.

Знакомясь в работе Кокрена со всем многообразием приемов выборочного статистического наблюдения, читатель, естественно, желает задать вопрос: а каковы же характерные черты современной практики выборочного метода? Конечно, исследователь, применяющий этот метод, всегда стремится к тому, чтобы повысить точность результатов своей работы, т. е. уменьшить случайную ошибку выборки и таким образом иметь возможность с большей вероятностью указать меньшие пределы, в которых может находиться неизвестная характеристика исходной совокупности (так называемый доверительный интервал). Не стоит добавлять к этому, что сама ошибка должна иметь действительно случайный характер, не отражая какие-либо систематические влияния. Каков же путь, ведущий к достижению этой цели, исключая увеличение численности выборки, позволяющее закону больших чисел проявиться полнее? Этот путь — ограничения, налагаемые на «игру случая», сужение области его действия, позволяющее исключить из вариации полученных данных также и ту ее часть, которая не формирует случайную ошибку выборки. Современная методика выборочного статистического наблюдения располагает для этого рядом средств, с которыми книга Кокрена позволяет ознакомиться с достаточной полнотой.

Первое и, пожалуй, самое важное из этих средств это — расслоение, или стратификация (*stratum* — слой, пласт) исходной совокупности перед отбором из нее некоторого числа объектов. В нашей литературе показывается, почему такой отбор, называемый обычно «типическим» или «районированным», приводит к уменьшению случайной ошибки. Для того чтобы понять логику этой операции полезно, может быть, провести некоторую аналогию с экспериментальной работой. Очень важной задачей в экспериментальной работе является преодоление тех влияний, которые идут со стороны неоднородности условий проведения опыта. Источниками этой неоднородности могут быть в агрономических экспериментах, например, различия в плодородии почвы, в технологических опытах — разные партии сырья и т. д. Понятно, что эта неоднородность условий способна исказить выводы исследования, поскольку она присоединяет свое влияние к действию всех прочих, в том числе и образующих ошибку опыта, случайных причин. Экспериментатор выходит из положения таким образом, что уравнивает, по возможности, условия проведения опыта лишь в каждом «блоке» или повторении его. Обеспечивая такое сходство условий в пределах сравнительно небольшого блока, экспериментатор уже может сравнивать без помех, обусловленных различием между блоками, действие различных факторов или способов обработки, почти свободное от посторонних влияний. Что же касается различий между блоками или повторениями опыта, создающих, так сказать, общий «пестрый» фон, то вариация, обязавшая им, в итоге, при статистической обработке результатов опыта, подлежит исключению. Так налагается в опыте ограничение на случайное размещение его вариантов. Оно сохраняется лишь внутри блоков (повторений), где создается, по возможности, однородная среда.

При организации расслоенного отбора статистик также имеет перед собой неоднородную совокупность, иначе не было бы смысла расчленять ее на слои. Изучение семейных бюджетов, когда речь идет о всей неоднородной массе потребителей, может служить здесь примером. Другой пример — выборочное обследование различных торговых предприятий города, которые по характеру товарооборота в разных районах различны. Таким образом, слои, которые создаются так, чтобы они по возможности представляли собой однородные в известном отношении группы единиц со случайной вариацией значений признака, могут быть уподоблены блокам в эксперименте.

Подобно статистической обработке результатов эксперимента, при определении случайной ошибки выборки только вариация внутри слоев будет принята во внимание, а вариация, обязавшая различиям между слоями, исключена. Следовательно, и при расслоенном отборе со случайным извлечением определенного числа единиц из слоев ограничивается, подобно эксперименту с блоками, область «игры случая». Она остается источником вариации значений признака лишь внутри слоев. Так, в свою очередь, преодолевает статистик неоднородность условий, в которых производится отбор. Вместе с тем он уменьшает величину случайной ошибки, иными словами, повышает точность своего исследования.

Однако нужно подчеркнуть, что эффективность расслоенной выборки в высокой степени зависит от выбора признака расслоения, а также от порядка «размещения» отбираемых из слоев единиц. Что касается выбора признака расслоения, то можно провести аналогию между ним и выбором группировочного признака в статистической таблице. Если только при правильном выборе группировочного признака может с достаточной отчетливостью выступить в статистической таблице связь между изучаемыми явлениями, то, точно так же, лишь при удачном выборе признака расслоения может оказаться достаточно большой та доля общей вариации, которая обязана различиям между слоями и которая в дальнейшем подлежит исключению. Как и вариация средних в сказуемом групповой таблицы, эта вариация будет, очевидно, тем больше, чем теснее связь между изучаемым признаком и признаком расслоения. Надо полагать, что Кокрен и имеет это в виду, когда он говорит о регрессионной модели, в связи с вопросом о наиболее целесообразном числе слоев. Действительно, вариация средних по слоям, как и вариация средних значений признака в сказуемом статистической таблицы, может рассматриваться как изменчивость признака «по линии» эмпирической регрессии. Как же лучше разместить отбираемые единицы по слоям, чтобы добиться большей точности результатов отбора? Штудирова книгу Кокрена, читатель может познакомиться с двумя способами решения этой задачи: с пропорциональным размещением и с размещением оптимальным. В нашей литературе, посвященной проблемам выборочного метода, оптимальное размещение, приводящее в общем к большей точности выборки по сравнению не только с простой случайной выборкой, но и с пропорциональным размещением, не получило должного освещения, если не считать забытой работы А. Г. Ковалевского «Основы теории выборочного метода», опубликованной в Саратове в 1924 г., главы, написанной В. Н. Старовским в монографии «Теория математической статистики» (авторы А. Я. Боярский, В. Н. Старовский, В. И. Хотимский, Б. С. Ястремский. М.—Л., Соцэкгиз, 1931, гл. 5), да одного примера, приведенного в работе Дж. Юла и М. Кендэла «Теория статистики», переведенной на русский язык (М., Госстатиздат, 1960). Обычно в западной литературе способ оптимального размещения связывается с именем Е. Неймана, который предложил этот способ в 1934 г. Однако известный русский статистик А. А. Чупров еще в 1923 г. изложил этот способ в одной из своих статей, опубликованной на английском языке. Между тем идея оптимального размещения заслуживает, разумеется, особого внимания, так как распределение отбираемых единиц по слоям в соответствии с величинами средних квадратических отклонений вполне отвечает общему логическому принципу выборочного метода: чем «пестрее» исходная совокупность, тем многочисленней должна быть выборка. Именно за счет вариации величин средних квадратических отклонений, которая исключается из общей вариации вместе с вариацией средних по слоям, и создается преимущество выборки с оптимальным размещением перед выборкой, имеющей размещение пропорциональное. Разумеется, организация выборочного обследования как с пропорциональным размещением, так и с размещением оптимальным

может встретиться с большими трудностями, поскольку в обоих случаях требуется иметь какую-то информацию о «весе» выделяемых из исходной совокупности слоев. При оптимальном размещении эти трудности усугубляются еще необходимостью иметь представление о величинах средних квадратических отклонений. Но во всех случаях, когда эти затруднения могут быть преодолены, исследователь не должен пренебрегать возможностью организовать свое выборочное обследование на этих принципах, обещающих при правильном их применении большую эффективность результатов.

Какие же еще средства повышения точности наблюдения путем ограничения «игры случая» выработала теория и практика выборочного метода? Эти средства — метод отношения средних, метод регрессии и систематический отбор, который в нашей литературе именуется механическим. Все эти средства, их преимущества и недостатки Кокрен также подробно описывает в своей книге.

В основе метода отношения средних и метода регрессии лежит одна и та же идея: связать данные наблюдения с помощью отношения средних или, еще лучше, коэффициента регрессии с результатами каких-либо обследований, производившихся предварительно, чтобы затем исключить из вариации наблюдаемых данных ту ее часть, которая обусловлена этой связью. Метод отношения средних можно считать лишь известным приближением к методу регрессии: при строгой пропорциональности в изменениях изучаемых признаков оба метода приведут к одинаковым результатам. В книге Кокрена можно найти пример с оценкой среднего числа жителей в 49 городах США в 1930 г., которая произведена так, что данные 1930 г. были связаны с данными 1920 г. отношением средних. Логика обоих методов ясна. Она заключается в том, что вариация, обязанная регрессии, не может уже рассматриваться как вариация случайная. Она уже «объяснена» этой регрессией, установлена предшествующими наблюдениями. Таким образом, только колебания «около» линии регрессии выражают ничем не связанную «игру случая». Несомненно, здесь имеется сходство с тем положением вещей, которое складывается при расслоенной выборке. Как уже говорилось, вариация средних по слоям также может представляться как вариация «по линии» регрессии.

Если при расслоенном отборе или методе отношения средних и методе регрессии позволено все же проявляться законам случая хотя бы и в суженной области, то практика выборочного метода в поисках способов получения более репрезентативной выборочной средней выдвинула такую форму отбора, которая уже почти совсем не оставляет места для действия этих законов. Это — систематический отбор. Когда первая единица отбирается из исходной совокупности случайно, например, при помощи таблицы случайных чисел, то систематический отбор называют также «псевдо-случайным» отбором. Но, по существу, это не меняет дела.

Кокрен описывает в своей книге, какие опасности таятся для исследователя, например, при линейном тренде в исходной совокупности или при периодических колебаниях значений изучаемого признака. Хронологический порядок расположения членов совокупности также

может присоединить к случайной вариации вариацию, обязанную действию фактора времени. Список работников может отразить частичное сходство рабочих групп и вообще может обнаружиться стремление соседних членов совокупности быть похожими друг на друга или одинаково отклоняться от средней величины. В почтовом списке членов какого-либо общества, подписчиков или клиентов некоторого обслуживающего их учреждения может проявиться географический порядок и т. д. Та или иная тенденция, существующая в исходной совокупности, может быть нейтрализована соответствующим порядком отбора. Так, например, линейная тенденция нейтрализуется отбором объектов из середины интервалов. Но как же обстоит дело с вычислением ошибки выборки? Кокрен не останавливается подробно на этом вопросе, а между тем последний должен привлечь к себе особое внимание именно потому, что систематическая выборка широко вошла в практику.

При систематическом отборе неизбежно возникает корреляция между членами выборок. Если она положительна, то ошибка, исчисленная как случайная, преуменьшит действительную величину ошибки, при отрицательной же корреляции она преувеличит последнюю. В статистической литературе можно найти утверждения, что рендомизация исходной совокупности при систематической выборке способна поправить дело. Так, например, Йейтс пишет: «Систематическая выборка была бы полностью равносильна случайной выборке, если бы элементы в списке располагались совершенно случайно» (Ф. Йейтс. Выборочный метод в переписях и обследованиях. М., «Статистика», 1965, с. 56). Однако это справедливо лишь для математического ожидания ошибки при систематическом отборе из полностью рендомизированной бесконечной совокупности. У Кокрена можно найти доказательство правильности этого положения.

Наряду с вполне понятным желанием сделать результаты выборочного наблюдения более точными, более репрезентативными, в организации этого наблюдения наметилась и другая линия: линия упрощения всей работы, сокращения затрат на нее времени и труда, а следовательно, и уменьшения связанных с ней издержек. Наиболее ясное свое выражение эта линия получила в организации серийного отбора (по английской терминологии, cluster sampling — группового отбора; в нашей сельскохозяйственной статистике она получила название «гнездового» отбора). Разумеется, легче организовать отбор и изучение нескольких десятков серий или групп единиц, чем сотен отдельных единиц. Практические преимущества серийного отбора особенно очевидны, например, в сельскохозяйственной статистике, где обследование нескольких групп хозяйств, расположенных в непосредственной близости друг к другу, менее затруднительно, чем обследование такого же числа отдельных хозяйств, разбросанных по всей территории района. Но организационные выгоды, полученные при отборе сериями, могут быть перекрыты потерей в точности получаемых результатов, так как отбор групп близлежащих объектов создает благоприятные условия для возникновения внутриклассовой корреляции, т. е. связи между единицами, входящими в отбираемую серию. Поскольку обычно между ними устанавливается положительная корреляция,

ошибка серийного отбора оказывается, как правило, больше ошибки простой случайной выборки. Кокрен представляет весь механизм образования ошибки при серийном отборе в терминах внутриклассовой корреляции и это делает его более понятным.

Таковы основные направления в современной организации выборочного статистического наблюдения, и читатель получает о них вместе с суммой других сведений отчетливое представление, штудировав книгу Кокрена. Нельзя при этом думать, что эта книга — нечто вроде элементарного курса практики выборочного статистического обследования. Напротив, успешное освоение положений, рассмотренных в книге, предполагает, что читатель уже подготовлен к этому. Во всяком случае для этого необходимо знание основ теории выборочного метода и известное представление о проблемах корреляции и регрессии, а также умение пользоваться соответствующими методами математики. На это указывает сам автор. Читатель должен всегда помнить, что вопросы организации выборочного наблюдения и статистическая обработка полученных результатов тесно друг с другом связаны, представляя собой, по сути дела, две стороны одной и той же проблемы. Определение же формы выборочного наблюдения, в свою очередь, диктуется той познавательной задачей, которая выдвигается в данном исследовании, и характером изучаемого явления. В общем, книга Кокрена написана не для начинающего статистика. Ее надо рассматривать как труд, который предназначен для углубления и расширения знаний.

В заключение хотелось бы подчеркнуть, что работы, посвященные проблемам выборочного статистического метода, вообще имеют большое научное значение. Роль этого метода в изучении социально-экономических явлений, может быть, не столь велика, как роль опирающегося на те же философско-методологические и математические основания эксперимента в науках о природе. Но когда перед экономистом или социологом стоят задачи испытания гипотезы о характеристиках и составе некоторой еще не изученной большой совокупности фактов, выборочный метод служит единственным средством решения этой задачи. При этом то обстоятельство, что исследователь в своих умозаключениях опирается на вероятностную логику, не только не умаляет научной ценности получаемых выводов, но придает им строгий научный смысл. Важно только помнить, что математико-статистические оценки должны сочетаться с пониманием сущности реальных внутренних отношений в изучаемых явлениях. В связи с этим нужно сделать упрек автору: в скупых его пояснениях относительно обстоятельств, при которых применяется та или иная форма отбора, а также в математических моделях недостаточно раскрывается логика самого метода. Но и в таком изложении книга Кокрена должна принести несомненную пользу.

ПРЕДИСЛОВИЕ

Цель этой книги — как можно более полно изложить теорию выборочного метода в том виде, в каком она разработана для применения к выборочным обследованиям, с примерами, показывающими, как эта теория применяется на практике, и с упражнениями для самостоятельной работы. Я надеюсь, что эта книга может служить как учебным пособием по курсу выборочных обследований, в котором основное внимание уделяется теории, так и для индивидуального изучения выборочного метода теми, кому недоступен формальный учебный процесс.

Минимальная математическая подготовка, необходимая для свободного понимания доказательств, включает знание дифференциального исчисления в объеме, достаточном для нахождения максимумов и минимумов (с применением там, где это необходимо, множителей Лагранжа), а также знакомство с элементарной алгеброй, и особенно умение обращаться со сравнительно сложными алгебраическими суммами. Особенно желательно знание основ теории вероятностей для конечных выборочных пространств, включая комбинаторные вероятности, свойства математических ожиданий и понятие условной вероятности. Что касается статистики, то автор предполагает, что читатель прошел вводный курс, охватывающий такие разделы, как средние значения и средние квадратичные отклонения, нормальное, биномиальное и полиномиальное распределения, доверительные границы, t -распределение Стьюдента, линейная регрессия и простейшие приемы дисперсионного анализа. Поскольку я везде пытался подчеркнуть связь теории выборочных обследований с главным направлением статистической теории, в отдельных случаях применяются и более новые результаты из статистики. В первых главах книги каждый шаг доказательства должен с очевидностью вытекать для читателя из предыдущих рассуждений. Ближе к концу, где доказательства более насыщены, для полного понимания отдельных их этапов читателю может потребоваться проделать небольшие выкладки на бумаге.

Порядок изложения тем в этом издании в основном совпадает с порядком изложения в первом издании и большинство теорем сохраняет свои прежние номера. Однако гл. 5, посвященная расслоению, которая уже в первом издании была излишне длинной, разбита на две главы. Нынешняя гл. 5 содержит обычные, более известные результаты. Новая гл. 5А посвящена многочисленным специальным вопросам, рассмотрение которых необходимо для более результативного применения расслоения. Еще одно изменение порядка изложения состоит в том, что основные сведения об оценках по отношению даются теперь в гл. 2 и 3, а не откладываются до шестой, как было в первом издании. Эта перестановка была сделана потому, что оценки по отношению встречаются на практике, часто в скрытом виде, даже в наиболее простых обследованиях, так что, по-видимому, целесообразно ознакомиться с ними раньше. Разумеется, преподаватель, предпочитающий оставлять эту тему до гл. 6, может сохранить старый порядок изложения.

Для того чтобы охватить результаты исследований, опубликованные после подготовки в 1951—1952 гг. первого издания, во многих местах были введены дополнительные параграфы. Материал, содержащийся в них, имеет довольно разнородный характер. Перечислим некоторые наиболее важные из таких разделов. Несколько параграфов посвящено статистическим методам, применяющимся в тех случаях, когда результаты обследования должны быть представлены отдельно для некоторых определенных подразделений совокупности (например, для людей разного возраста или для домовладельцев и квартиронанимателей), а также, когда сравнение отдельных групп необходимо для анализа. В отношении расслоенного отбора новые разделы посвящены формированию слоев и определению числа слоев, нахождению оптимальных объемов выборки в слоях при определенных уровнях точности по каждой из некоторого числа переменных и расслоению по двум признакам при небольшом объеме выборки. Приводится краткое изложение новых исследований по отбору без возвращения, когда исходные единицы отбираются с неодинаковыми вероятностями. Изучение ошибок, не обусловленных отбором, дало возможность разработать методы для исследования эффективности повторных обращений, новые по сравнению с другими приемами уменьшения смещения, вызванного неполучением ответа, и новые методы сбора данных о роли ошибок наблюдения в общих ошибках оценок, получаемых в обследованиях.

Некоторые новые параграфы должны восполнить пробелы в изложении, на которые мне указали преподаватели, пользовавшиеся этой книгой, а также подсказал мой собственный опыт. Для двухступенчатого отбора, например, формулы дисперсии приводятся отдельно для каждого из основных методов получения выборки и оценивания, и более подробно рассматривается результативность равновзвешенных оценок. Хотя эти формулы можно получить на основе одной или двух общих теорем, удобнее иметь готовые их точные выражения, по-видимому, даже ценой некоторого увеличения объема книги. Многие старые параграфы были переделаны для того, чтобы включить новые результаты или сделать изложение более четким. Более чем вдвое увеличено число упражнений.

Настоящее издание примерно на треть больше первого. Это увеличение вызывает у меня смешанные чувства. Даже материал первого издания книги было трудно уложить в полугодовой курс лекций, оставляя при этом время для изучения примеров реальных обследований, интересующих самих учащихся. Поэтому материал разбит по параграфам таким образом, что многие из них можно опустить или свести к краткому упоминанию результатов без ущерба для понимания последующих глав книги. Хотя выбор разделов для изучения определяется взглядами преподавателя, а также уровнем подготовки и кругом интересов учащихся, для вводного курса можно рекомендовать опустить или сократить следующие параграфы: 2.8, 2.13, 2.14; 3.9, 3.11; 4.6, 4.7; 5.8, 5.9; 5A.1, 5A.3, 5A.4, 5A.5, 5A.10, 5A.12; 6.4, 6.5, 6.9, 6.13, 6.14, 6.15, 6.17; 7.4, 7.5, 7.7, 7.8, 7.9; 8.5, 8.6, 8.8, 8.11, 8.12; 9.5, 9.6, 9.11, 9.12, 9.13; 10.7, 10.8, 10.9, 10.10; 11.7, 11.9, 11.16; 12.5, 12.6, 12.7, 12.8; 13.5, 13.7, 13.15, 13.16.

Значительную часть лекционных записей, по которым было составлено первое издание, подготовили д-р Элва Л. Финкнер и д-р Эмиль Х. Джебе. Д-р Ф. К. Корнелл, д-р Дж. А. Доулл и д-р Финкнер любезно разрешили сослаться на данные их обследований. Некоторые исследования как теоретического, так и прикладного характера, послужившие исходным материалом для книги, стали возможными благодаря исследовательскому контракту с Управлением военно-морских исследований Военно-морского министерства США. При подготовке настоящего издания перепечатку рукописи превосходно выполнили сотрудники канцелярии отделения статистики Гарвардского университета. Помощь в проверке доказательств великодушно оказали г-жа Клео Ятц и все те студенты старших курсов, которые, на их беду, проходили мимо дверей моего кабинета в решающий период подготовки книги. Авторский указатель составила г-жа Сюзан Роджерс, а ответы к упражнениям проверил г-н Самбашива Рао. Я многим обязан своим коллегам Лилу Д. Кэлвину, Р. М. Кайерту, У. Эдвардсу Демингу, Торе Далениусу, Моррису Х. Хансену, Херману О. Хартли, Уильяму Н. Хервицу, Лесли Кишу, Уильяму Дж. Мэдоу и Фредерику Ф. Стивану за плодотворное обсуждение новых направлений выборочного метода. Всем этим лицам, как названным по имени, так и не названным, я хотел бы выразить свою благодарность.

Кембридж, Массачусетс,
ноябрь 1962 г.

УИЛЬЯМ ДЖ. КОКРЕН

ГЛАВА I

ВВЕДЕНИЕ

1.1. ПРЕИМУЩЕСТВА ВЫБОРОЧНОГО МЕТОДА

Наши знания, суждения и поступки в очень большой мере основаны на выборочных данных. Это утверждение одинаково справедливо как для повседневной жизни, так и для научных исследований. Впечатление об учреждении, в котором ежедневно производятся тысячи различных операций, складывается часто на основании лишь одного или двух посещений этого учреждения за несколько лет. Путешественник, проведя десять дней в чужой стране, собирается написать книгу и в ней посоветовать жителям этой страны, как оживить промышленность, преобразовать политическую систему, сбалансировать бюджет и улучшить питание в гостиницах. Это — персонаж анекдотический. Но на самом деле, от ученого-обществоведа, который прожил 20 лет в этой стране, изучая ее, он отличается лишь тем, что основывает свои выводы на гораздо меньшем числе наблюдений, да еще, вероятно, меньше осведомлен о степени своего невежества. И в науке и в житейских делах нам доступен для изучения лишь фрагмент той общей картины, которая должна расширить наши знания.

Тому, как правильно получить выборку и как сделать по ее данным обоснованные выводы, еще лет 30 назад не уделяли внимания. Эти проблемы не играли бы особой роли, если бы материал, из которого мы производим отбор, был однороден, так что любая выборка дала бы приблизительно одинаковые результаты. Заключение о состоянии нашего здоровья делается по нескольким каплям крови, проанализированным в лаборатории. Такой метод основан на предположении, что циркулирующая кровь всегда хорошо перемешана и каждая ее капля несет одинаковую информацию, — предположении, в правильность которого мы, будучи неспециалистами, свято верим. Однако, когда изучаемый материал далеко не однороден, как это часто и бывает, способ получения выборки приобретает решающее значение, а изучение методов, позволяющих получить достоверные сведения, становится весьма важным.

В этой книге излагаются основы теории, созданной для обоснования методов правильного отбора. На практике, в большинстве случаев, для которых эта теория была разработана, совокупность, о которой мы хотим получить сведения, конечно и имеет четкие границы — жи-

тели города, станки на заводе, рыбы в озере. Иногда удобнее, казалось бы, получить нужные сведения, произведя сплошное обследование или перепись этой совокупности. Практические работники, привыкшие к сплошным переписям, сначала недоверчиво относились к выборочному методу и пользовались им неохотно. Хотя такого предубеждения более не существует, имеет смысл перечислить основные преимущества выборочного метода по сравнению со сплошной переписью.

Меньше стоимость

Затраты на получение данных лишь относительно небольшой части всей совокупности меньше, чем при сплошной переписи. Для большой совокупности достаточно точные данные можно получить по выборке, составляющей лишь очень небольшую долю этой совокупности. В США наиболее важные периодические обследования, предпринимаемые правительством, основаны на выборках, охватывающих около 100 000 человек, т. е. обследуется приблизительно один из каждых 1800 жителей страны*. Обследования для сбора сведений, касающихся торговли и рекламной политики при изучении рынка, могут основываться на выборках объемом всего в несколько тысяч единиц.

Короче сроки

По тем же причинам данные выборочного обследования можно собрать и обобщить быстрее, чем при сплошной переписи. Это особенно важно, когда сведения нужны срочно.

Шире область применения

При некоторых видах обследований для сбора данных необходимо привлечь высококвалифицированный персонал или воспользоваться специальным оборудованием; как правило, и то и другое ограничено. В этих случаях сплошное обследование невозможно: приходится либо получать сведения выборочным путем, либо не получать их совсем. Таким образом, выборочные обследования имеют более широкую область применения и дают большую возможность получать сведения самого разнообразного характера. С другой стороны, если желательно получить точную информацию о мелких подразделениях исходной совокупности, то нужный для этого объем выборки может оказаться столь большим, что предпочтительнее окажется сплошная перепись.

Больше достоверность

Если общий объем работы меньше, то можно привлечь более квалифицированный персонал, лучше его подготовить, более тщательно контролировать проведение обследования и обработку его результатов. Поэтому выборочное обследование может дать более достоверные сведения, чем соответствующее сплошное обследование.

* Данные относятся к началу 60-х годов. — Примеч. ред.

1.2. ПРИМЕРЫ ПРИМЕНЕНИЯ ВЫБОРОЧНОГО МЕТОДА

Если проследить за развитием выборочного метода за последние 10 лет, то наибольшее впечатление производит быстрое увеличение числа и видов проведенных выборочных обследований. Статистическое бюро ООН время от времени публикует сообщения о выборочных обследованиях, проводимых в странах — членах ООН, в издании «Sample Surveys of Current Interest». В сообщении за 1960 г. перечисляются обследования, проведенные в 52 странах. Многие из этих обследований преследовали цель получить несомненно важные для национального планирования сведения в таких областях, как сельскохозяйственное производство и землепользование, безработица и трудовые ресурсы, промышленное производство, оптовые и розничные цены, состояние здоровья населения, доходы и расходы семей. Проводились обследования и по более частным темам: были исследованы, например, жилищные и социальные проблемы пожилых людей (Австрия), задолженность арендаторов (Цейлон), стоимость жилищного строительства (Чехословакия), возраст учеников начальных школ (Италия), влияние телевидения на школьников (Голландия), условия домашней работы домохозяек (Швеция), состав женщин, берущих детей на воспитание (Великобритания), использование технической информации на мелких предприятиях (Великобритания), занятость ученых и инженеров в промышленности (США).

Выборочный метод стал играть значительную роль в национальных переписях населения, проводимых каждые десять лет. В США 5%-ная выборка была впервые применена в переписи 1940 г., когда дополнительные вопросы о роде занятий, происхождении, числе детей и т. д. задавали лицам, чьи фамилии попадали на две из каждых 40 строк на лицевой и на оборотной сторонах переписного листа. При переписи 1950 г. выборочный метод применялся гораздо шире. По 20%-ной выборке (каждая пятая строка переписного листа) были получены сведения по таким признакам, как доход, число лет обучения, миграция, служба в вооруженных силах. Путем отбора в этой 20%-ной выборке каждого шестого человека дополнительно была взята 3 1/3 %-ная выборка для получения сведений о браках и числе рожденных детей. Кроме того, группа вопросов, касающихся срока службы и состояния жилища, была разбита на пять подгрупп и ответы на вопросы соответствующей подгруппы были получены в каждом пятом доме. Выборочный метод применялся также для ускорения публикации результатов переписи. Предварительные результаты по многим важным показателям, полученные путем выборочной разработки, появились более чем за полтора года до опубликования окончательных итогов.

Выборочный метод широко применялся и в переписи населения 1960 г. За исключением некоторых основных данных, требуемых по конституции или согласно закону от каждого человека, полная перепись была проведена на 25%-ной выборочной основе: только одно из каждых четырех домохозяйств получало полный переписной лист. Это изменение наряду с существенным повышением механизации об-

работки материалов переписи значительно ускорило публикацию результатов и удешевило перепись.

На более низком уровне местные власти — городов, штатов и графств — стали шире пользоваться выборочными обследованиями, чтобы получать сведения, необходимые для перспективного планирования и решения неотложных проблем. В США в большинстве крупных городов существуют коммерческие агентства, которые по заказам планируют и проводят выборочные обследования.

В значительной степени требует выборочного подхода и так называемое исследование рынка. Постоянно необходимы сведения о числе радиослушателей, телезрителей по различным программам, а также о читательских аудиториях газет и журналов (включая читающих рекламу). Промышленники и торговцы интересуются реакцией населения на новые продукты или новые методы упаковки, жалобами на ранее выпущенные продукты и причинами предпочтения одного продукта другому.

В промышленности, торговле и обслуживании часто пользуются выборочным методом, пытаясь повысить результативность работы предприятий. Такие важные области применения выборочного метода, как контроль качества и выборочная приемка продукции, находятся за рамками настоящей книги. Очевидно, однако, что решения, касающиеся уровня качества партии изделий или его изменения или же принятия или отклонения такой партии, могут быть хорошо обоснованными только в том случае, когда результаты, полученные по выборочным данным, справедливы (с достаточной точностью) для партии изделий в целом. Выборка деловых документов (отчетов, платежных ведомостей, акций, личных дел), получить которую обычно гораздо проще, чем провести выборочный опрос людей, может дать нужные сведения быстро и экономично. Применяв выборочный метод, можно сэкономить средства и время также при оценке запасов, при изучении условий и продолжительности службы оборудования, при оценке качества и эффективности канцелярской работы, при исследовании того, как руководящие работники распределяют свое время на решение различных вопросов и вообще в новой области управления, называемой «исследование операций». В книгах Деминга (Deming, 1960)* и Слонима (Slonim, 1960) содержится много интересных примеров, демонстрирующих диапазон применения выборочного метода в торгово-промышленной деятельности.

Опросы общественного мнения и предвыборные опросы, которые сыграли большую роль в ознакомлении общественности с методами выборочного исследования, продолжают привлекать внимание газет. В счетоводстве и финансовом контроле, где выборочный метод применяется уже в течение многих лет, растет интерес к приложению его современных достижений при решении конкретных задач. Предметом оживленной дискуссии служит возможность полагаться на результаты выборочных обследований в ходе судебного процесса.

* Фамилия и год издания в скобках указывают на источник в списке литературы, помещенном в конце главы. — *Примеч. ред.*

Выборочные обследования можно условно разделить на два вида: *описательные* и *аналитические*. Цель описательного обследования состоит просто в том, чтобы получить сведения о некоторых больших группах: например, о числе мужчин, женщин и детей, смотрящих ту или иную телевизионную программу. При аналитическом обследовании сравниваются различные подгруппы совокупности для того, чтобы установить, существуют ли между ними такие различия, которые позволили бы нам построить или проверить гипотезы о природе сил, действующих в данной совокупности. Например, обследование рождаемости в Индианаполисе было предпринято с целью выяснить, в какой степени супружеские пары планируют число и время появления детей, отношение мужей и жен к такому планированию, причины того или иного отношения и в какой мере супруги достигают успеха в своих действиях (Kiser and Whelpton, 1953).

Разумеется, между описательными и аналитическими обследованиями нельзя провести четкой границы. Многие обследования предоставляют данные, пригодные для обеих целей. Наряду с ростом числа описательных обследований наблюдается, однако, и значительное увеличение числа обследований, предпринятых главным образом с аналитическими целями, особенно для изучения поведения и здоровья людей. В качестве примеров можно назвать обследования состояния зубов у школьников до и после фторизации воды, уровня и причин смертности курильщиков в зависимости от интенсивности курения и обширное обследование эффективности противополиомиелитной вакцины Солка.

Успешное проведение выборочных обследований привело к их применению для оценки довольно необычных величин: например, длины сигаретных окурков, числа мух в городе, числа подписей под петицией, в действительности не поставленных указанными людьми, и даже числа людей, умеющих складывать язык «трубочкой». Эти величины имели отношение к изучению соответственно связи между курением и раком легких, эффективности борьбы с мухами, юридической силы петиций и наследования умения складывать язык «трубочкой», хотя последнее, на мой взгляд, не может служить объектом большого обследования.

1.3. ОСНОВНЫЕ ПРОБЛЕМЫ ВЫБОРОЧНОГО ОБСЛЕДОВАНИЯ

Прежде чем рассматривать роль, которую играет теория в выборочном обследовании, полезно вкратце охарактеризовать основные проблемы, связанные с планированием и проведением обследования. Обследования могут сильно различаться по их сложности. Взять выборку из 5000 карточек, пронумерованных и аккуратно расставленных в картотеке, нетрудно. Совсем другое дело получить выборку жителей района, где средством сообщения служат реки, протекающие в лесах, где карты отсутствуют, жители говорят на 15 разных диалектах и весьма подозрительно относятся к любопытным незнакомцам. Проблемы, вызывающие затруднения в одном обследовании, могут оказаться несущественными или совсем не возникнуть в другом.

Основные проблемы, связанные с обследованием, сгруппированы более или менее произвольно в следующие 11 пунктов.

Цели обследования

Чрезвычайно полезна четкая формулировка целей обследования. Без нее, погружаясь в детали планирования сложного обследования, легко забыть о его общих целях и принять решения, расходящиеся с ними.

Совокупность, из которой производится отбор

Словом *совокупность* пользуются для обозначения множества объектов, из которого извлекается выборка. Определение совокупности может не представлять никакой трудности, как, например, в случае, когда отбирается партия электрических лампочек для оценки среднего времени их горения. Напротив, при выборочном исследовании совокупности ферм необходимо сформулировать правила, позволяющие выделить ферму и отграничить одну из них от другой. Такие правила должны быть практичными: нужно, чтобы в ходе работы исследователь был в состоянии без особых колебаний определять, принадлежит ли сомнительный объект к совокупности или нет.

Совокупность, из которой производится отбор (*обследуемая совокупность*), должна совпадать с совокупностью, о которой мы хотим собрать сведения (*изучаемая совокупность*). Иногда по практическим соображениям или ради удобства исследуемая совокупность суживается по сравнению с изучаемой. В этом случае следует помнить, что выводы, сделанные по выборке, относятся лишь к исследуемой совокупности. Суждение о степени применимости этих выводов также и к изучаемой совокупности должно основываться на других источниках сведений. Может оказаться полезной любая доступная дополнительная информация о характере различий между исследуемой и изучаемой совокупностями.

Собираемые данные

Необходимо убедиться в том, что все собираемые данные соответствуют целям обследования и никаких важных данных не пропущено. Существует распространенная тенденция, особенно при обследовании совокупностей людей, задавать слишком много вопросов, часть которых впоследствии вовсе не анализируется. Перегруженный опросный лист ухудшает качество ответов как на важные, так и на второстепенные вопросы.

Желательная степень точности

Результаты выборочных обследований всегда отчасти неопределенны. Это происходит потому, что исследуется только часть всей совокупности и измерения производятся с ошибками. Эту неопределенность можно уменьшить, извлекая выборки большего объема и производя более точные измерения. Но это обычно увеличивает затраты времени и средств. Следовательно, важный момент состоит в определении же-

лательной степени точности результатов. Ответственность за это несет лицо, которое будет пользоваться собранными данными. Принятие решения относительно желательной точности может оказаться затруднительным, поскольку многие практические работники не привыкли мыслить в терминах величины погрешностей, допустимой при получении оценок и тем не менее дающей возможность принять правильное решение. Статистик часто может оказать им помощь на этом этапе.

Способы наблюдения

Существует большой выбор средств и методов изучения совокупности. Данные о состоянии здоровья человека могут быть получены либо с его слов, либо по результатам медицинского исследования. При обследовании можно предоставить опрашиваемому самому заполнять опросный лист или поручить исследователям задавать стандартный набор вопросов в определенной форме или же вести опрос в виде беседы, при которой вопросы задают в различной форме и в произвольном порядке. Обследование можно произвести по почте, по телефону, путем личного посещения или же так или иначе сочетая эти способы. Методы опроса и связанные с ними проблемы подробно изучались [см., например, (Human, 1954) и (Payne, 1951)].

Значительная часть предварительной работы состоит в разработке форм документов, в которых будут содержаться вопросы и куда нужно будет записывать ответы. Если опросные листы сравнительно просты, то возможные ответы можно иногда заранее закодировать, т. е. записать в таком виде, чтобы потом их легко было преобразовать для машинной обработки. Для разработки удачного инструментария обследования необходимо отчетливо представлять структуру таблиц с итоговыми данными, которыми будут пользоваться при анализе материалов обследования.

Основа выборки

Прежде чем производить отбор, необходимо разбить совокупность на части, которые называются *единицами отбора* или просто *единицами*. Эти единицы должны вместе исчерпывать всю совокупность и не должны перекрывать одна другую, т. е. каждый элемент совокупности должен принадлежать одной и только одной единице. Иногда единицы отбора выделяются очевидным образом, как, например, в совокупности электрических лампочек, где единицей отбора служит отдельная лампочка. Иногда приходится выбирать из нескольких возможных единиц отбора. Например, при обследовании людей в городе единицей отбора может быть отдельный человек, члены одной семьи или же все жители городского квартала. При выборочном изучении урожая сельскохозяйственных культур единицами отбора могут служить поля, фермы или же участки земли, форма и размеры которых заранее известны.

Построение такого перечня единиц отбора, называемого *основой выборки*, на практике часто бывает одной из главных задач. Наученные горьким опытом организаторы обследований с недоверием относятся к спискам, составленным ранее для других целей. Несмотря на заверения в обратном, такие списки часто оказываются неполными

или неудобочитаемыми или содержат неизвестное число повторений. Хорошую основу выборки иногда трудно создать и для специфических совокупностей, например, для букмекеров или для совокупности людей, разводящих индюшек. Джессен (Jessen, 1955) приводит интересный метод построения основы выборки, в которой единицами отбора служат ветки плодового дерева.

Извлечение выборки

Существует целый ряд способов, которыми можно извлечь выборку. Для каждого такого способа, зная желательный уровень точности, можно приближенно оценить на этом основании объем выборки. Кроме того, при принятии решения о способе отбора учитываются относительная стоимость и время, необходимые для осуществления того или иного способа.

Пробное исследование

Оказалось полезным проверять опросный лист и методику собственно исследования в пробном исследовании небольшого масштаба. Это почти всегда приводит к улучшению опросного листа и может выявить другие недочеты, которые станут серьезной проблемой при большом масштабе работы, как, например, стоимость исследования, если она окажется значительно больше ожидаемой.

Организация собственно исследования

В больших исследованиях встречаются многие проблемы, связанные с их организацией. Следует обучить персонал в соответствии с целями и применяемыми методами наблюдения и обеспечить соответствующий контроль его работы. Неоценимую роль играет методика как можно более ранней проверки качества ответов. Следует разработать план действий на случай неполучения ответа, т. е. при невозможности для исследователя получить информацию об определенных единицах в выборке.

Сводка и анализ данных

Первый этап состоит в просмотре заполненных опросных листов с тем, чтобы исправить ошибки регистрации или, по крайней мере, исключить заведомо неправильные данные. Необходимо принять решение о порядке сводки и группировки данных в случае, когда опрашиваемые не дали ответа на некоторые вопросы или эти ответы были исключены при просмотре. После этого данные сводятся и группируются для получения оценок. К одним и тем же данным могут быть применены различные способы оценивания.

При изложении результатов исследования желательно указывать величину ожидаемой ошибки для наиболее важных оценок. Одно из

преимуществ вероятностного отбора заключается в возможности сделать такие указания на величину ошибки, хотя ценность их несколько снижается, если значительно число неответивших.

Информация для будущих исследований

Чем больше информации о совокупности мы имеем первоначально, тем легче получить выборку, обеспечивающую точные оценки. Каждое завершённое выборочное исследование представляет собой потенциальное средство улучшения будущего отбора, поскольку оно содержит данные о средних значениях, средних квадратичных отклонениях и о природе изменчивости основных характеристик, а также о стоимости получения необходимых данных. Техника проведения выборочных исследований развивается быстрее, если приняты меры для накопления и обобщения информации такого рода.

Существует еще одна важная причина, по которой каждое завершённое исследование облегчает их проведение в будущем. В сложных исследованиях ход работы никогда в точности не соответствует запланированному. Внимательный исследователь учится распознавать ошибки, допущенные при проведении исследования, и предотвращать их в будущих исследованиях.

1.4. РОЛЬ ТЕОРИИ ВЫБОРОЧНОГО МЕТОДА

Приведенный в предыдущем параграфе перечень вопросов должен был подчеркнуть, что организация выборочного исследования — это дело, требующее знаний и навыков разнообразного характера. В некоторых из этих вопросов — таких, как определение совокупности, установление круга данных, подлежащих сбору, и методов наблюдения, организация практической работы — теория выборочного метода по большей части не играет значительной роли. Хотя эти вопросы не будут в дальнейшем рассматриваться в этой книге, необходимо ясно представлять себе их важность. Выборочное исследование требует внимания ко всем этапам работы: плохое выполнение одного из них может погубить исследование, в котором все остальное проделано правильно.

Цель теории выборочного метода состоит в повышении результативности выборочного исследования. В ней разрабатываются методы извлечения выборки и оценивания, позволяющие при минимальных затратах получать оценки с достаточной для нашей цели точностью. К принципу достижения определенной точности при минимальных затратах постоянно обращаются при изложении теории выборочного метода.

Для того чтобы применить этот принцип, мы должны уметь предвидеть для любого рассматриваемого метода отбора ожидаемую точность и ожидаемую стоимость. Что касается точности, то мы не можем точно предсказать, какую ошибку будет содержать оценка в каждом конкретном случае, поскольку для этого нужно знать истинное значение оцениваемой величины для всей совокупности. Вместо этого точность

оценки, получаемой с помощью того или иного способа отбора, определяется на основании распределения частот этой оценки, которое получается, если соответствующий способ многократно применять к одной и той же совокупности. Это, конечно, обычный прием суждения о точности в статистической теории.

Можно ввести еще одно упрощение. Для выборок того объема, который обычно встречается на практике, часто есть все основания полагать, что выборочные оценки имеют приблизительно нормальное распределение. Для нормально распределенных оценок вид распределения частот полностью известен, если известны среднее значение и среднее квадратичное отклонение (или дисперсия). Значительная часть теории выборочного метода посвящена нахождению формул для таких средних и дисперсий.

Существует некоторое различие между теорией выборочных обследований и классической теорией выборочного метода, заключающееся в том, что при обследовании совокупности состоят из *конечного* числа единиц. Когда отбор производится из конечной, а не из бесконечной совокупности, методы доказательства теорем иные и результаты несколько более сложны. Для практических целей эти различия в результатах для конечных и для бесконечных совокупностей обычно не имеют значения. Если объем выборки (по числу первичных единиц отбора) мал по сравнению с объемом всей совокупности, то вполне применимы результаты, полученные для бесконечной совокупности. В основном в этой книге излагаются результаты, относящиеся к конечным совокупностям. В некоторых, более сложных вопросах, чтобы упростить изложение, мы будем пользоваться теорией для бесконечных совокупностей.

1.5. ВЕРОЯТНОСТНЫЙ ОТБОР

Все методы отбора, для которых будет излагаться соответствующая теория, должны обладать следующими общими математическими свойствами.

1. Должна существовать возможность указать множество различных выборок S_1, S_2, \dots, S_r , которые могут быть получены при применении данного метода отбора к некоторой конкретной совокупности. Это значит, что мы можем точно указать, какие единицы отбора принадлежат к S_1 , к S_2 и т. д. Предположим, например, что совокупность состоит из шести единиц, пронумерованных числами от 1 до 6. При извлечении выборки объемом в две единицы принятый способ отбора дает три возможных исхода: $S_1 \sim (1,4)$; $S_2 \sim (2,5)$; $S_3 \sim (3,6)$. Заметим, что при этом не обязательно перечислять все возможные выборки объема 2.

2. Для каждой из возможных выборок S_i задана известная нам вероятность ее извлечения π_i .

3. Мы извлекаем одну из выборок S_i с помощью некоторого процесса, при котором вероятность извлечения каждой выборки принимает соответствующее значение π_i . В рассматриваемом примере мы можем приписать трем указанным выборкам равные вероятности. Тогда

само извлечение может быть произведено с помощью равновероятного выбора целого числа от 1 до 3. Если таким числом оказалось j , то считается извлеченной выборка S_j .

4. Должен быть установлен метод вычисления оценки по выборке и для каждой конкретной выборки он должен приводить к единственному значению. Мы можем принять, например, в качестве оценки среднее значение результатов наблюдений отдельных единиц в выборке.

Для каждого способа отбора, удовлетворяющего этим свойствам, мы можем вычислить распределение частот значений соответствующей оценки, которое получилось бы в результате многократного применения этого способа отбора к рассматриваемой совокупности. Действительно, мы знаем, с какой частотой будет извлечена любая отдельная выборка S_i и как вычислить оценку по данным этой выборки. Таким образом, для любого метода отбора рассмотренного типа можно развивать дальнейшую теорию, хотя конкретные детали могут и оказаться довольно сложными.

К методу отбора такого типа применяется термин *вероятностный отбор*. Это, конечно, не единственный способ, которым можно извлечь выборку. Далее указаны распространенные способы отбора, не имеющие вероятностного характера.

1. Отбор ограничивается легко доступной частью совокупности. Например, выборка угля из открытого вагона берется лишь с глубины от 6 до 9 дюймов.

2. Отбор производится беспорядочно. Исследователь, выбирая десять кроликов из большой клетки в лаборатории, может делать это без продуманного плана, забирая тех, до которых он может дотянуться.

3. Имеется небольшая, но неоднородная совокупность. Исследователь просматривает всю совокупность и отбирает небольшое число «типичных» единиц, т. е. единиц, отвечающих его представлению о среднем для совокупности. Такой метод называют иногда *предвзятым* или *направленным отбором*.

4. Выборка состоит преимущественно из добровольцев в исследованиях, где процесс измерения неприятен или опасен для обследуемого.

При надлежащих условиях каждый из этих способов может дать полезные сведения. Однако развитие теории выборочного метода не связано с этими способами, поскольку в них отсутствует элемент случайного отбора. Что касается проверки ценности того или иного из этих методов, то единственно возможный путь состоит в отыскании ситуации, при которой известны результаты или по всей совокупности или по данным вероятностной выборки, и в сопоставлении с ними результатов отбора. Однако даже если при одном таком сравнении метод и окажется удовлетворительным, это не гарантирует, что так же будет и в других обстоятельствах.

На практике мы редко получаем вероятностную выборку, записывая S_i и π_i , как было указано ранее. Для больших совокупностей, где принятый порядок отбора дал бы миллиарды возможных выборок, это оказалось бы немыслимо трудоемкой работой. Обычно отбор производится путем указания вероятностей включения в выборку от-

дельных единиц и затем извлечения единиц по одной или по несколько сразу до тех пор, пока не будет образована выборка нужного объема и типа. Для теоретических же целей достаточно знать, что при желании, располагая достаточным временем, мы можем выписать все S_i и p_i .

1.6. ПРИМЕНЕНИЕ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Как уже упоминалось, при обследованиях выборки часто достаточно велики, так что получаемые по ним оценки имеют приблизительно нормальное распределение. Для вероятностного отбора, кроме того, существуют формулы среднего значения и дисперсии оценки. Рассмотрим сначала *несмещенные* оценки. Оценка $\hat{\mu}$, получаемая согласно некоторой схеме отбора, называется *несмещенной* оценкой некоторой характеристики совокупности, μ , если среднее значение* $\hat{\mu}$, взятое по всем возможным выборкам, равно μ . В обозначениях параграфа 1.5 это условие можно записать в виде

$$E(\hat{\mu}) = \sum_{i=1}^N p_i \hat{\mu}_i = \mu,$$

где $\hat{\mu}_i$ — оценка, получаемая по i -й выборке. Символ E , заменяющий выражение «математическое ожидание величины» (the expected value of), общепринят.

Предположим, что мы получили выборку методом, обеспечивающим несмещенную оценку, и вычислили соответствующее значение выборочной оценки $\hat{\mu}$ и ее среднее квадратичное отклонение $\sigma_{\hat{\mu}}$ (часто называемое иначе ее стандартной ошибкой**). Насколько хороша наша оценка? Мы не знаем точную величину ошибки оценки ($\hat{\mu} - \mu$), но из свойств нормального распределения вытекает, что с вероятностями:

0,32 (или приблизительно в одном случае из трех) абсолютное значение ошибки $|\hat{\mu} - \mu|$ превосходит $\sigma_{\hat{\mu}}$;

0,05 (или в одном случае из двадцати) абсолютное значение ошибки $|\hat{\mu} - \mu|$ превосходит $1,96 \sigma_{\hat{\mu}} \approx 2 \sigma_{\hat{\mu}}$;

0,01 (или в одном случае из ста) абсолютное значение ошибки $|\hat{\mu} - \mu|$ превосходит $2,58 \sigma_{\hat{\mu}}$.

Например, если при определении срока службы некоторых приборов на большом предприятии при обычной нагрузке вероятностная

* Далее слово «значение» будет иногда опускаться. — *Примеч. ред.*

** Английский термин *standard error*, который здесь употребляет автор, передается далее термином *стандартная ошибка*, чтобы подчеркнуть и терминологически отличить среднее квадратичное отклонение выборочной оценки (средней квадратичной ошибки) от среднего квадратичного отклонения вообще. — *Примеч. пер.*

выборка показала, что среднее время работы прибора для этой выборки $\hat{\mu} = 394$ дням при среднем квадратичном отклонении (стандартной ошибке) $\sigma_{\hat{\mu}} = 4,6$ дня, то среднее время работы приборов для всей их совокупности в 99 случаях из ста заключено между

$$\hat{\mu}_L = 394 - (2,58) \cdot (4,6) = 382 \text{ дням}$$

и

$$\hat{\mu}_U = 394 + (2,58) \cdot (4,6) = 406 \text{ дням.}$$

Эти границы, 382 дня и 406 дней, называются нижней (lower) и верхней (upper) *доверительными границами*. Для отдельной оценки, сделанной по однократному обследованию, утверждение « μ заключено между 382 и 406 днями» не является абсолютно правильным. «99%-ный доверительный уровень» означает, что если бы та же схема отбора многократно применялась к рассматриваемой совокупности и утверждение о доверительных границах делалось по каждой выборке, то приблизительно в 99% случаев оно было бы правильным и в 1% случаев ошибочным. Если выборочный метод применяется там, где ранее производились сплошные переписи, то это свойство иногда можно продемонстрировать, извлекая повторно выборки предлагаемого типа из совокупности, по которой имеются полные данные, так что μ известно [см., например, (Trueblood and Cyert, 1957)]. Практические работники лучше и глубже понимают природу выборочного метода, убедившись на деле в том, что за небольшим исключением заранее установленная доля утверждений оказывается правильной. Подобным же образом, если извлекается однократная выборка из каждой совокупности, принадлежащей некоторому ряду различных совокупностей, то окажутся правильными приблизительно 95% утверждений, сделанных на 95%-ном доверительном уровне.

Ранее предполагалось, что $\sigma_{\hat{\mu}}$, вычисленное по выборке, определяет-ся точно. В действительности $\sigma_{\hat{\mu}}$, как и $\hat{\mu}$, подвержено ошибкам выборки. Если случайная переменная распределена нормально, то при малом объеме выборки для нахождения доверительных границ для μ вместо таблиц нормального распределения применяются таблицы t -распределения Стьюдента. Замена таблиц нормального распределения таблицами t -распределения почти не играет роли, если число степеней свободы при вычислении $\sigma_{\hat{\mu}}$ превосходит 60. При некоторых видах рас-слоенного отбора и применении метода дублированного отбора (см. параграф 13.14) число степеней свободы невелико и необходимо пользоваться таблицами t -распределения.

1.7. СМЕЩЕНИЕ И ЕГО РОЛЬ

В теории выборочных обследований приходится рассматривать смещенные оценки. Это нужно делать по двум причинам.

1. В некоторых, часто встречающихся случаях, особенно при оценивании отношений двух величин, оценки, которые по другим соображениям удобны и целесообразны, оказываются смещенными.

2. Даже если оценки при вероятностном отборе и будут несмещенными, ошибки наблюдения и неполучение ответа могут привести к смещению в окончательных результатах обследования. Так произойдет, например, если среди ответивших на вопросы обследования о расходовании общественных фондов на некоторые цели половина настроена «за» и половина «против», а среди отказавшихся отвечать настроены «против» почти все.

Для того чтобы исследовать эффект смещения, предположим, что оценка $\hat{\mu}$ распределена нормально со средним значением m , которое находится на расстоянии B от истинного значения для совокупности μ ,

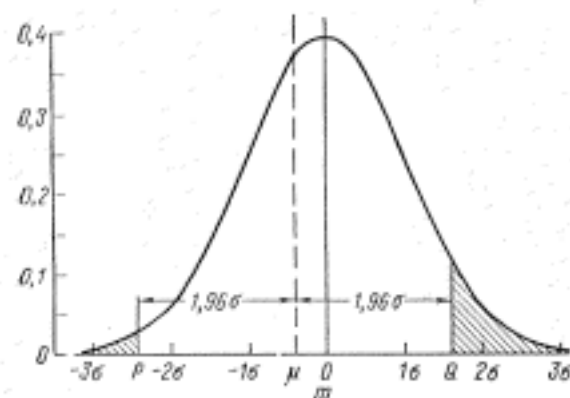


Рис. 1.1. Влияние смещения на ошибки оценивания

как показано на рис. 1.1. Величина смещения $B = m - \mu$. Предположим, что мы не знаем о существовании смещения. Мы вычисляем среднее квадратичное отклонение σ распределения частот оценки — оно будет, конечно, средним квадратичным отклонением от среднего значения m для распределения, а не от истинного среднего значения μ . Вместо $\sigma_{\hat{\mu}}$ мы принимаем σ . В качестве утверждения относительно достоверности оценки мы заявляем, что лишь с вероятностью 0,05 оценка $\hat{\mu}$ отклонится от своего истинного значения на величину, большую чем $1,96 \sigma$.

Рассмотрим теперь, как искажается эта вероятность при наличии смещения. Для этого мы вычислим истинную вероятность того, что ошибка оценки превысит $1,96 \sigma$, при этом ошибка измеряется относительно истинного значения μ . Два «хвоста» распределения нужно исследовать отдельно. Для правого «хвоста» вероятность того, что ошибка превысит $+1,96 \sigma$, равна площади заштрихованной области на рис. 1.1 справа от точки Q. Эта площадь равна

$$\frac{1}{\sigma \sqrt{2\pi}} \int_{\mu + 1,96 \sigma}^{\infty} e^{-(\hat{\mu} - m)^2 / 2\sigma^2} d\hat{\mu}.$$

Положим $\hat{\mu} - m = t\sigma$. Нижний предел интегрирования по t равен:

$$\frac{\mu - m}{\sigma} + 1,96 = 1,96 - \frac{B}{\sigma}.$$

Следовательно, эта площадь равна

$$\frac{1}{\sqrt{2\pi}} \int_{1,96 - (B/\sigma)}^{\infty} e^{-t^2/2} dt.$$

Аналогично для левого «хвоста» площадь заштрихованной области слева от точки P равна

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1,96 - (B/\sigma)} e^{-t^2/2} dt.$$

Из вида интегралов ясно, что величина искажения вероятностей зависит только от отношения смещения к среднему квадратичному отклонению. Результаты вычислений приведены в табл. 1.1.

Таблица 1.1

ВЛИЯНИЕ СМЕЩЕНИЯ B НА ВЕРОЯТНОСТЬ ТОГО, ЧТО ОШИБКА ПРЕВЫСИТ ВЕЛИЧИНУ $1,96 \sigma$

B/σ	Вероятность того, что ошибка		Общая вероятность
	$< -1,96 \sigma$	$> 1,96 \sigma$	
0,02	0,0238	0,0262	0,0500
0,04	0,0228	0,0274	0,0502
0,06	0,0217	0,0287	0,0504
0,08	0,0207	0,0301	0,0508
0,10	0,0197	0,0314	0,0511
0,20	0,0154	0,0392	0,0546
0,40	0,0091	0,0694	0,0685
0,60	0,0052	0,0869	0,0921
0,80	0,0029	0,1230	0,1259
1,00	0,0015	0,1685	0,1700
1,50	0,0003	0,3228	0,3231

На общую вероятность того, что ошибка превысит величину $1,96 \sigma$, смещение влияет очень мало при условии, что оно составляет менее одной десятой среднего квадратичного отклонения. В этом случае общая вероятность составляет 0,0511 вместо предполагавшихся 0,05. По мере того как смещение увеличивается, искажение вероятности становится более значительным. Для $B = \sigma$ общая вероятность равна 0,17, что более чем в три раза превышает предполагавшуюся величину.

На «хвосты» распределения смещение влияет по-разному. При положительном смещении, как в нашем примере, вероятность преуменьшить истинное значение быстро уменьшается от предполагавшихся 0,025 до пренебрежимо малой величины при $B = \sigma$. Вероятность соответствующего преувеличения постепенно растет. Для большинства

приложений интерес, главным образом, представляет совокупная ошибка, но иногда интересуются ошибкой, имеющей определенный знак.

В качестве рабочего правила принимается, что влиянием смещения на достоверность оценки можно пренебречь, если смещение составляет менее одной десятой среднего квадратичного отклонения оценки. Если имеется смещенный метод оценивания, для которого $B/\sigma < 0,1$, где B — абсолютная величина смещения, то можно считать, что смещение не будет существенным недостатком этого метода. Даже при $B/\sigma = 0,2$ искажение вероятности ошибки довольно незначительно.

Пользуясь этим правилом, следует различать два источника смещения, упомянутые в начале этого параграфа. Для смещений, аналогичных тем, что возникают при оценивании отношений двух величин, верхняя граница для отношения B/σ может быть найдена теоретически. Если выборка достаточно велика, то мы можем быть уверены в том, что B/σ не превышает 0,1. С другой стороны, для смещений, вызванных ошибками наблюдения или неполучением ответа, обычно невозможно найти надежную и достаточно малую верхнюю границу для B/σ . Этот трудный вопрос рассматривается в гл. 13.

1.8. СРЕДНИЙ КВАДРАТ ОШИБКИ

При сравнении смещенной оценки с несмещенной или двух оценок с разными величинами смещения полезным критерием служит средний квадрат ошибки (СКО) оценки, здесь ошибка измеряется относительно оцениваемого параметра для совокупности. Формально

$$\begin{aligned} \text{СКО}(\hat{\mu}) &= E(\hat{\mu} - \mu)^2 = E[(\hat{\mu} - m) + (m - \mu)]^2 = \\ &= E(\hat{\mu} - m)^2 + 2(m - \mu)E(\hat{\mu} - m) + (m - \mu)^2 = \\ &= (\text{дисперсия } \hat{\mu}) + (\text{смещение})^2 \end{aligned}$$

(члены удвоенного произведения исчезают, так как $E(\hat{\mu} - m) = 0$).

Применение СКО в качестве критерия достоверности оценки равносильно рассмотрению двух оценок, имеющих одинаковый СКО, как эквивалентных. Это не вполне строгое заключение, потому что распределения частот ошибок $(\hat{\mu} - \mu)$ разной величины для двух оценок не будут одинаковы, если у них разные величины смещения. Однако, как показали Хансен, Хервиц и Мэдоу (Hansen, Hurwitz and Madow, 1953), если B/σ меньше чем приблизительно 1/2, то распределения частот абсолютных величин ошибок $|\hat{\mu} - \mu|$ почти одинаковы. Табл. 1.2 иллюстрирует это утверждение.

Даже при $B/\sigma = 0,6$ соответствующие вероятности меняются незначительно по сравнению со случаем $B/\sigma = 0$.

Поскольку трудно проследить за тем, чтобы в оценках не присутствовало никаких незаподозренных смещений, мы будем говорить обычно о *точности* (precision) оценки, а не о ее *достоверности* (accuracy). Термин *достоверность* относится к величине отклонений от истинного среднего значения μ , в то время как термин *точность* относится к величине отклонений от среднего значения m , получаемого в результате многократного применения одного и того же способа отбора.

Таблица 1.2
ВЕРОЯТНОСТЬ ТОГО, ЧТО АБСОЛЮТНАЯ ВЕЛИЧИНА ОШИБКИ БОЛЬШЕ ИЛИ РАВНА:
 $1\sqrt{\text{СКО}}$; $1,96\sqrt{\text{СКО}}$; $2,576\sqrt{\text{СКО}}$

B/σ	Вероятность		
	$1\sqrt{\text{СКО}}$	$1,96\sqrt{\text{СКО}}$	$2,576\sqrt{\text{СКО}}$
0	0,317	0,0500	0,0100
0,2	0,317	0,0499	0,0100
0,4	0,319	0,0496	0,0095
0,6	0,324	0,0479	0,0083

Упражнения

1.1. Предположим, что вы собираетесь применить выборочный метод, чтобы оценить общее число слов в какой-либо книге с иллюстрациями: (а) Возникает ли здесь проблема определения совокупности? (б) Приведите доводы «за» и «против» применения в качестве единицы отбора: (1) страницы, (2) строки.

1.2. Нужно извлечь выборку из перечня фамилий. Фамилии записаны на пронумерованных и последовательно расположенных в ящике карточках, по одной фамилия на каждой карточке. Каждая фамилия должна иметь одинаковый шанс попасть в выборку. Какие проблемы возникают в следующих типичных ситуациях? (а) Некоторые фамилии не принадлежат изучаемой совокупности, хотя обнаружить, относится ли к ней та или иная фамилия можно только после извлечения карточки. (б) Некоторые фамилии встречаются на нескольких карточках. Все карточки с одной и той же фамилией имеют последовательные номера и, следовательно, расположены в картотеке рядом. (в) Некоторые фамилии встречаются на нескольких карточках, но карточки с одной и той же фамилией могут быть размещены в картотеке где угодно.

1.3. Сложной проблемой часто оказывается нахождение основы выборки, из которой можно произвести отбор со сравнительно небольшими затратами. Какого вида основы выборки подошли бы для следующих обследований? Имеют ли эти основы серьезные недостатки? (а) Обследование в большом городе магазинов, где продаются сафьяны и тому подобные товары. (б) Обследование вещей, забытых в метро или автобусах. (в) Обследование лиц, ужаленных в прошлом году змеями.

1.4. В городском справочнике, выпущенном четыре года назад, перечислены адреса домов в порядке их расположения вдоль улиц и указаны фамилии проживающих по каждому адресу. Каковы недостатки такой основы при обследовании жителей города, производимом сейчас и связанным с их опросом? Могут ли исследователи устранить эти недостатки в ходе самого обследования? Пользуясь справочником, будете ли вы извлекать выборку адресов (жилых помещений) или выборку лиц?

1.5. При оценивании выборочным путем фактической стоимости мелких товаров, числящихся в описях крупной фирмы, по каждой позиции, попавшей в выборку, фиксировалась стоимость фактическая и по торговым книгам. Для всей выборки отношение фактической стоимости к числящейся по книгам составило 1,021, причем эта оценка распределена приблизительно нормально со стандартной ошибкой, равной 0,0082. Вычислите 95%-ные доверительные границы для фактической стоимости, если общая стоимость всех этих товаров по торговым книгам составляет 80 000 долл.

1.6. Имеющиеся данные часто можно рассматривать как данные некоторой выборки, хотя, на первый взгляд, они кажутся результатом сплошного учета. Владелец стоянки для автомобилей обнаружил, что в воскресенье по утрам он получает низкий доход. После 26 воскресений средняя выручка за воскресенье

утро составила ровно 10 долл. Стандартная ошибка этой величины, вычисленная по изменениям от недели к неделе, равна 1,2 долл. Содержание стоянки в воскресенье обходится в 7 долл. Владелец намерен держать стоянку открытой в это время, если ожидаемый им доход в расчете на одно воскресное утро будет составлять в среднем 5 долл. Какова доверительная вероятность того, что средний доход за большой период времени составит не менее 5 долл.? Какие еще предположения нужно сделать, чтобы получить ответ на этот вопрос?

1.7. Что произойдет с приведенными в табл. 1.2 вероятностями того, что абсолютное значение ошибки превысит $1/\sqrt{СКО}$; $1,96/\sqrt{СКО}$ и $2,576/\sqrt{СКО}$, если B/σ будет стремиться к бесконечности, т. е. если СКО будет полностью определяться смещением? Согласуется ли ваш ответ с характером изменения данных в табл. 1.2 по мере того, как B/σ меняется от 0 до 0,6?

1.8. Если необходимо сравнить две оценки, имеющие различные распределения частот ошибок ($\mu - \mu$), то в некоторых специальных задачах иногда можно подсчитать расходы или потери, которые вызовет ошибка ($\mu - \mu$) той или иной заданной величины. При прочих равных условиях предпочтительнее оценки с меньшими ожидаемыми потерями. Покажите, что если потери представляют собой квадратичную функцию ошибки вида $\lambda(\mu - \mu)^2$, то следует выбрать оценку с меньшим средним квадратом ошибки.

ЛИТЕРАТУРА

- Deming W. E. (1960). *Sample design in business research*. John Wiley and Sons. New York.
- Hansen M. H., Hurwitz W. N. and Madow W. G. (1953). *Sample survey methods and theory*. John Wiley and Sons. New York, Vol. I, p. 58.
- Hyman H. H. (1954). *Interviewing in social research*. University of Chicago Press.
- Jessen R. J. (1955). Determining the fruit count on a tree by randomized branch sampling. *Biometrics*, 11, 99—109.
- Kiser C. V. and Whelpton P. K. (1953). Résumé of the Indianapolis study of social and psychological factors affecting fertility. *Population Studies*, 7, 95—110.
- Payne S. L. (1951). *The art of asking questions*. Princeton University Press.
- Slonim M. J. (1960). *Sampling in a nutshell*. Simon & Schuster. New York.
- Trueblood R. M. and Cyert R. M. (1957). *Sampling techniques in accounting*. Prentice-Hall, Englewood Cliffs, N. J.
- U. N. Statistical Office (1960). *Sample surveys of current interest*. Eighth Report.

ГЛАВА 2

ПРОСТОЙ СЛУЧАЙНЫЙ ОТБОР

2.1. ПРОСТОЙ СЛУЧАЙНЫЙ ОТБОР

В выборочных обследованиях рассматриваются выборки, извлеченные из совокупностей, содержащих некоторое конечное число N единиц. Если эти единицы различимы между собой, то число различных выборок объема n , которые могут быть извлечены из N единиц, задается комбинаторной формулой

$$C_N^n = \frac{N!}{n!(N-n)!} \quad (2.1)$$

Например, если совокупность содержит пять единиц, обозначаемых A, B, C, D и E , то существует десять различных выборок объема 3, а именно:

$ABC \ ABD \ ABE \ ACD \ ACE$
 $ADE \ BCD \ BCE \ BDE \ CDE$

Заметим, что одна и та же буква не может встречаться в выборке дважды. Порядок, в котором буквы расположены в выборке, не играет роли, так что шесть выборок ABC, ACB, BAC, BCA, CAB и CBA рассматриваются как тождественные.

Простым случайным отбором называется способ извлечения n единиц из N , при котором каждая из C_N^n выборок имеет равную вероятность быть отобранной. Иногда этот способ отбора называют просто случайным отбором. Поскольку слово случайный употребляется в литературе во многих различных смыслах, целесообразно воспользоваться дополнительным уточняющим прилагательным. Некоторые авторы предпочитают выражение неограниченный случайный отбор*.

На практике простую случайную выборку получают, отбирая последовательно единицу за единицей. Единицы в совокупности нумеруются числами от 1 до N , после чего выбирается последовательность n случайных чисел, заключенных между 1 и N . Ее можно выбрать, либо пользуясь таблицей случайных чисел, либо помещая жетоны с номерами от 1 до N в урну, тщательно перемешивая их и отбирая последо-

* Здесь и далее термин случайный употребляется как синоним термина равновероятный, если не оговорено иное. — Примеч. пер.

вательно и жетонов. Единицы совокупности, имеющие эти номера, составляют выборку. На каждом этапе отбора такой процесс обеспечивает для всех еще не выбранных номеров равную вероятность быть отобранными. Легко проверить, что равную вероятность быть отобранными имеют все C_N^n возможных выборок.

Когда жетоны с номерами извлекаются из урны, они не возвращаются в нее, поскольку иначе одна и та же единица могла бы попасть в выборку более одного раза. Поэтому такой отбор называется *отбором без возвращения*. Аналогично, если применяется таблица случайных чисел, то отбрасываются номера, уже полученные ранее. Отбор с возвращением легко осуществить, но им, за исключением особых случаев, пользуются редко, поскольку нет особых оснований допускать, чтобы одна и та же единица встречалась в выборке дважды*.

Другие методы отбора часто оказываются предпочтительнее простого случайного отбора по соображениям удобства или повышения точности. Однако для введения в теорию выборочного метода простой случайный отбор наиболее удобен.

2.2. ОПРЕДЕЛЕНИЯ И ОБОЗНАЧЕНИЯ

При выборочном обследовании мы сосредоточиваем внимание на определенных свойствах единиц совокупности, которые мы пытаемся измерить и зафиксировать для каждой единицы, попавшей в выборку. Эти свойства мы будем называть *характеристиками* или просто *признаками*.

Численные значения, полученные для какого-либо признака для N единиц, составляющих совокупность, обозначаются через y_1, y_2, \dots, y_N . Соответствующие значения для единиц, попавших в выборку, обозначаются через y_1, y_2, \dots, y_n или, если мы хотим указать общий член выборки, через y_i ($i = 1, 2, \dots, n$). Заметим, что выборка не будет состоять из *первых* n единиц совокупности, за исключением того редкого случая, когда именно эти единицы окажутся отобранными. Если помнить это обстоятельство, то, как показывает мой опыт, никакой путаницы не возникает.

Заглавные буквы относятся к характеристикам *совокупности*, строчные — к соответствующим характеристикам *выборки*. Для суммарных и средних значений мы примем следующие определения:

	Совокупность	Выборка
Суммарное значение	$Y = \sum_{i=1}^N y_i = y_1 + y_2 + \dots + y_N$	$\sum_{i=1}^n y_i = y_1 + y_2 + \dots + y_n$
Среднее значение	$\bar{Y} = \frac{y_1 + y_2 + \dots + y_N}{N} = \frac{\sum_{i=1}^N y_i}{N}$	$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{\sum_{i=1}^n y_i}{n}$

* В нашей литературе соответствующие приемы отбора часто называют *бесповторным* и *повторным*. Термины *без возвращения* и *с возвращением* более предпочтительны, поскольку основное различие между этими видами отбора состоит не в возможности повторного извлечения одной и той же единицы, а в порядке отбора (с возвращением или без возвращения отобранной единицы), создающем такую возможность. — *Примеч. ред.*

Хотя отбор применяется для многих целей, наибольший интерес обычно представляют четыре характеристики совокупности:

1. Среднее значение \bar{Y} (например, среднее число детей на одну школу).
2. Суммарное значение Y (например, общее число акров под пшеницей в некотором районе).
3. Отношение двух суммарных или средних значений $R = Y/X = \bar{Y}/\bar{X}$ (например, отношение стоимости ликвидного имущества к общей стоимости имущества у группы семей).
4. Доля единиц, попадающих в некоторую определенную группу (например, доля людей, имеющих искусственные зубы).

В этой главе рассматривается оценивание первых трех из этих величин.

Символ $\hat{}$ обозначает оценку значения некоторого признака для совокупности, сделанную по выборке. В этой главе рассматриваются только наиболее простые оценки:

	Оценка
Среднее значение для совокупности \bar{Y}	$\hat{\bar{Y}} = \bar{y}$ — выборочное среднее
Суммарное значение для совокупности Y	$\hat{Y} = N\bar{y} = N \sum_{i=1}^n y_i / n$
Отношение для совокупности R	$\hat{R} = \bar{y}/\bar{x} = \sum_{i=1}^n y_i / \sum_{i=1}^n x_i$

Входящий в \hat{Y} множитель N/n , на который умножается суммарное значение для совокупности, называют иногда *множителем распространения* или *множителем повышения* или *инфляции*. Обратная ему величина n/N — отношение объема (числа единиц) выборки к объему совокупности — называется *долей отбора* и обозначается буквой f .

2.3. СВОЙСТВА ОЦЕНОК

Точность какой-либо оценки, полученной по выборке, зависит от двух факторов: от способа, которым оценка вычисляется по данным выборки, и от способа отбора. Для экономии места мы иногда будем писать «точность выборочного среднего» или «точность простого случайного отбора», не упоминая особо другой существенный фактор. Мы надеемся, что в таких случаях из контекста всегда будет ясно, какой из двух факторов пропущен. Знакомая с какой-либо новой формулой, читатель должен быть уверен, что он понимает, для какого конкретного способа отбора и способа оценивания эта формула справедлива.

В этой книге способ оценивания называется *состоятельным*, если оценка становится в точности равной оцениваемому значению для совокупности при $n = N$, т. е. когда выборку составляет вся совокупность. Очевидно, что при простом случайном отборе \bar{y} и $N\bar{y}$ представляют собой состоятельные оценки соответственно среднего и суммарного значений для совокупности. Состоятельность — желательное свойство оценок. Но может найти себе применение и несостоятельная оцен-

ка, если только она обеспечивает удовлетворительную точность, когда n мало по сравнению с N . По-видимому, ее полезность этим случаем и ограничивается. Другое определение состоятельности для случая конечной совокупности дают Хансен, Хервиц и Мэдоу (Hansen, Hurwitz and Madow, 1953).

Как мы уже говорили, метод оценивания называется *несмещенным*, если среднее значение оценки, взятое по всем возможным выборкам данного объема n , в точности равно истинному значению для совокупности. Если метод называют несмещенным без всяких оговорок, то это утверждение должно быть справедливым для любой совокупности конечных значений y_i и для любого n . Чтобы установить, будет ли \bar{y} несмещенной оценкой при простом случайном отборе, вычислим значение \bar{y} для всех C_N^n выборок и найдем их среднее. Символ E означает, что среднее берется по всем возможным выборкам.

Теорема 2.1. Выборочное среднее \bar{y} есть несмещенная оценка \bar{Y} .

Доказательство. По определению

$$E\bar{y} = \frac{\sum \bar{y}}{C_N^n} = \frac{\sum (y_1 + y_2 + \dots + y_n)}{n \{N!/n! (N-n)!\}}, \quad (2.2)$$

где суммирование распространяется на все C_N^n выборок. Для подсчета этой суммы найдем, во скольких выборках участвует каждое конкретное значение y_i . Поскольку на остальных $(n-1)$ местах в выборке могут стоять любые из $(N-1)$ единиц, остающихся для отбора, число выборок, содержащих y_i , равно:

$$C_{N-1}^{n-1} = \frac{(N-1)!}{(n-1)! (N-n)!}. \quad (2.3)$$

Следовательно,

$$\sum (y_1 + y_2 + \dots + y_n) = \frac{(N-1)!}{(n-1)! (N-n)!} (y_1 + y_2 + \dots + y_N).$$

Отсюда, учитывая (2.2), получаем

$$\begin{aligned} E\bar{y} &= \frac{(N-1)!}{(n-1)! (N-n)!} \frac{n! (N-n)!}{nN!} (y_1 + y_2 + \dots + y_N) = \\ &= \frac{(y_1 + y_2 + \dots + y_N)}{N} = \bar{Y}. \end{aligned} \quad (2.4)$$

Следствие. $\hat{Y} = N\bar{y}$ есть несмещенная оценка суммарного значения для совокупности, Y .

Менее громоздкое доказательство теоремы 2.1 можно получить следующим образом. Поскольку каждая единица участвует в одном и том же числе выборок, ясно, что

$$\frac{E(y_1 + y_2 + \dots + y_n)}{y_1 + y_2 + \dots + y_N} \quad (2.5)$$

Соответствующим множителем должно быть n/N , так как первое выражение содержит n членов, а второе выражение — N членов. Это и дает искомый результат.

2.4. ДИСПЕРСИИ ОЦЕНОК

Дисперсия y_i для конечной совокупности обычно определяется по формуле

$$\sigma^2 = \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N}. \quad (2.6)$$

Мы будем пользоваться несколько иным выражением, в котором в знаменателе вместо N стоит $(N-1)$. Положим

$$S^2 = \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1}. \quad (2.7)$$

В таком виде формула применяется теми, кто рассматривает теорию выборочного метода с точки зрения дисперсионного анализа. Преимущество такой записи в том, что большинство результатов принимает несколько более простой вид. При условии, что одни и те же обозначения применяются постоянно, все результаты эквивалентны в любой из двух форм записи.

Рассмотрим теперь дисперсию \bar{y} . Под ней мы подразумеваем $E(\bar{y} - \bar{Y})^2$, взятое по всем C_N^n возможным выборкам.

Теорема 2.2. Дисперсия среднего \bar{y} для простой случайной выборки равна:

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \frac{S^2}{n} \frac{(N-n)}{N} = \frac{S^2}{n} (1-f), \quad (2.8)$$

где $f = n/N$ есть доля отбора.

Доказательство

$$n(\bar{y} - \bar{Y}) = (y_1 - \bar{Y}) + (y_2 - \bar{Y}) + \dots + (y_n - \bar{Y}). \quad (2.9)$$

Из соображений симметрии, высказанных при рассмотрении выражения (2.5), следует, что

$$E[(y_1 - \bar{Y})^2 + \dots + (y_n - \bar{Y})^2] = \frac{n}{N} [(y_1 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2], \quad (2.10)$$

а также, что

$$\begin{aligned} E[(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \dots + (y_{n-1} - \bar{Y})(y_n - \bar{Y})] = \\ = \frac{n(n-1)}{N(N-1)} [(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \\ + \dots + (y_{n-1} - \bar{Y})(y_n - \bar{Y})]. \end{aligned} \quad (2.11)$$

В (2.11) суммирование произведений распространяется на все пары единиц соответственно в выборке и в совокупности. Сумма в левой части содержит $n(n-1)/2$ членов, а сумма в правой $N(N-1)/2$ членов.

Возведем теперь (2.9) в квадрат и вычислим среднее по всем возможным простым случайным выборкам. Пользуясь (2.10) и (2.11), получаем

$$n^2 E(\bar{y} - \bar{Y})^2 = \frac{n}{N} \left\{ (y_1 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2 + \right. \\ \left. + \frac{2(n-1)}{N-1} [(y_1 - \bar{Y})(y_2 - \bar{Y}) + \dots + (y_{N-1} - \bar{Y})(y_N - \bar{Y})] \right\}.$$

Дополняя до квадрата суммы выражение в квадратных скобках, имеем

$$n^2 E(\bar{y} - \bar{Y})^2 = \frac{n}{N} \left\{ \left(1 - \frac{n-1}{N-1}\right) [(y_1 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2] + \right. \\ \left. + \frac{n-1}{N-1} [(y_1 - \bar{Y}) + \dots + (y_N - \bar{Y})]^2 \right\}.$$

Второй член внутри фигурных скобок исчезает, так как сумма всех y_i равна $N\bar{Y}$. Деление на n^2 дает

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \frac{N-n}{nN(N-1)} \sum_{i=1}^N (y_i - \bar{Y})^2 = \frac{S^2}{n} \frac{(N-n)}{N}.$$

Следствие 1. Стандартная ошибка \bar{y} равна:

$$\sigma_{\bar{y}} = \frac{S}{\sqrt{n}} \sqrt{(N-n)/N} = \frac{S}{\sqrt{n}} \sqrt{1-f}. \quad (2.12)$$

Следствие 2. Дисперсия величины $\hat{Y} = N\bar{y}$, применяемой в качестве оценки суммарного значения для совокупности, Y , равна:

$$V(\hat{Y}) = E(\hat{Y} - Y)^2 = \frac{N^2 S^2}{n} \frac{(N-n)}{N} = \frac{N^2 S^2}{n} (1-f). \quad (2.13)$$

Следствие 3. Стандартная ошибка \hat{Y} равна:

$$\sigma_{\hat{Y}} = \frac{NS}{\sqrt{n}} \sqrt{(N-n)/N} = \frac{NS}{\sqrt{n}} \sqrt{1-f}. \quad (2.14)$$

2.5. ПОПРАВКА НА КОНЕЧНОСТЬ СОВОКУПНОСТИ

Хорошо известно, что для случайной выборки объема n из бесконечной совокупности дисперсия среднего равна σ^2/n . Если совокупность конечна, то единственное изменение в этом результате состоит в том, что нужно ввести множитель $(N-n)/N$. Множители $(N-n)/N$ для дисперсии и $\sqrt{(N-n)/N}$ для стандартной ошибки называются поправками на конечность совокупности (пкс). Авторы, которые ве-

дут изложение в терминах σ , приводят их с делителем $(N-1)$ вместо N . Если доля отбора n/N остается низкой, то эти множители близки к единице и объем совокупности сам по себе не оказывает непосредственного влияния на стандартную ошибку выборочного среднего. Например, если для двух совокупностей S одинаково, то выборка объемом в 500 единиц из совокупности, насчитывающей 200 000, обеспечивает почти ту же точность оценки среднего для совокупности, что и выборка в 500 единиц из 10 000. Лица, незнакомые с выборочным методом, с трудом воспринимают этот, действительно замечательный, результат. Им кажется интуитивно очевидным, что выборочное среднее не может быть достаточно достоверной оценкой, если оно получено на основе сведений об очень небольшой части совокупности. Читателю стоит подумать над тем, почему такое представление ошибочно.

На практике пкс можно не учитывать, если доля отбора не превышает 5%, а для многих целей даже если она достигает 10%. Если поправка не учитывается, то это приводит к некоторому преувеличению стандартной ошибки оценки \bar{y} .

Следующая теорема, обобщающая теорему 2.2, в этой главе не применяется, но мы будем ссылаться на нее далее.

Теорема 2.3. Пусть y_i, x_i — пара переменных, определенных для каждой единицы совокупности, и \bar{y}, \bar{x} — соответствующие средние для простой случайной выборки объема n . Тогда их ковариация имеет вид

$$E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}). \quad (2.15)$$

Теорема сводится к теореме 2.2, если значения переменных y_i, x_i совпадают для каждой единицы.

Доказательство. Применим теорему 2.2 к переменной $u_i = y_i + x_i$. Среднее значение u_i для совокупности есть $\bar{U} = \bar{Y} + \bar{X}$ и по теореме 2.2 получаем

$$E(\bar{u} - \bar{U})^2 = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N (u_i - \bar{U})^2$$

или

$$E[(\bar{y} - \bar{Y}) + (\bar{x} - \bar{X})]^2 = \\ = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N [(y_i - \bar{Y}) + (x_i - \bar{X})]^2. \quad (2.16)$$

Раскроем квадратные скобки в обеих частях равенства. По теореме 2.2

$$E(\bar{y} - \bar{Y})^2 = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2.$$

Аналогичное равенство справедливо для $E(\bar{x} - \bar{X})^2$.

Следовательно, эти два члена в левой и правой частях (2.16) сокращаются. Утверждение теоремы [формула (2.15)] следует из равенства удвоенных произведений.

2.6. ОЦЕНИВАНИЕ СТАНДАРТНОЙ ОШИБКИ ПО ВЫБОРКЕ

Формулы стандартных ошибок оценок средних и суммарных значений для совокупности применяются в основном для трех целей: (1) сравнить точность, которую дает простой случайный отбор, с точностью других способов отбора, (2) оценить объем выборки, необходимый для предполагаемого обследования, (3) оценить точность, действительно достигнутую в проведенном обследовании. В эти формулы входит S^2 , дисперсия для совокупности. В действительности она заранее не известна, но ее можно оценить по данным выборки. Соответствующее утверждение сформулировано в теореме 2.4.

Теорема 2.4. Для простой случайной выборки

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

есть несмещенная оценка

$$S^2 = \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1}.$$

Доказательство. Мы можем записать

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2 = \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right]. \end{aligned}$$

Возьмем теперь среднее по всем случайным выборкам объема n . По соображениям симметрии, высказанным в теореме 2.2,

$$E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 \right] = \frac{n}{N} \sum_{i=1}^N (y_i - \bar{Y})^2 = \frac{n(N-1)}{N} S^2,$$

согласно определению S^2 . Далее, по теореме 2.2,

$$E [n(\bar{y} - \bar{Y})^2] = \frac{N-n}{N} S^2.$$

Следовательно,

$$E(s^2) = \frac{S^2}{(n-1)N} [n(N-1) - (N-n)] = S^2. \quad (2.17)$$

Следствие. Несмещенные оценки дисперсий \hat{y} и $\hat{Y} = N\hat{y}$ есть

$$v(\hat{y}) = s_{\hat{y}}^2 = \frac{s^2}{n} \left(\frac{N-n}{N} \right) = \frac{s^2}{n} (1-f); \quad (2.18)$$

$$v(\hat{Y}) = s_{\hat{Y}}^2 = \frac{N^2 s^2}{n} \left(\frac{N-n}{N} \right) = \frac{N^2 s^2}{n} (1-f). \quad (2.19)$$

Для стандартных ошибок положим

$$s_{\hat{y}} = \frac{s}{\sqrt{n}} \sqrt{1-f}, \quad s_{\hat{Y}} = \frac{Ns}{\sqrt{n}} \sqrt{1-f}. \quad (2.20)$$

Эти оценки несколько смещены. Для большинства приложений их смещение не играет роли.

Читателю следует обратить внимание на символы, применяемые для обозначения истинной и оцениваемой по данным выборки дисперсий оценок. Так, для \hat{y} мы пишем:

истинная дисперсия: $V(\hat{y}) = \sigma_{\hat{y}}^2$,

оцениваемая дисперсия: $v(\hat{y}) = s_{\hat{y}}^2$.

2.7. ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ

Обычно предполагается, что оценки \hat{y} и \hat{Y} нормально распределены относительно соответствующих значений для совокупности. Основания для такого предположения и область его применимости рассматриваются в параграфе 2.13. Если это предположение справедливо, то нижняя и верхняя доверительные границы среднего и суммарного значений для совокупности имеют вид:

для среднего значения:

$$\hat{Y}_L = \bar{y} - \frac{ts}{\sqrt{n}} \sqrt{1-f}; \quad \hat{Y}_U = \bar{y} + \frac{ts}{\sqrt{n}} \sqrt{1-f}, \quad (2.21)$$

для суммарного значения:

$$\hat{Y}_L = N\bar{y} - \frac{tNs}{\sqrt{n}} \sqrt{1-f}; \quad \hat{Y}_U = N\bar{y} + \frac{tNs}{\sqrt{n}} \sqrt{1-f}. \quad (2.22)$$

Здесь t — квантиль нормального распределения, соответствующий желательной доверительной вероятности. Его наиболее употребительные значения таковы:

Доверительная вероятность (в %)	50	80	90	95	99
t	0,67	1,28	1,64	1,96	2,58

Если объем выборки меньше 60, то процентные значения квантилей можно взять из таблиц t -распределения Стьюдента с $(n-1)$ степенями свободы, поскольку таково число степеней свободы выборочной дисперсии s^2 . Полное соответствие t -распределению имеет место только тогда, когда сами наблюдения y_i нормально распределены и N беско-

нечно. Умеренные отклонения от нормальности не оказывают большого влияния. Для небольших выборок в случае очень асимметричных распределений необходимы специальные методы.

Пример. Подписи под некоторой петицией были собраны на 676 листах. На каждом листе могут разместиться 42 подписи, но многие листы содержат меньшее число подписей. Для простой случайной выборки объемом в 50 листов (приблизительно 7%-ная выборка) было подсчитано число подписей на каждом листе. Полученные данные приведены в табл. 2.1.

Таблица 2.1
ДАННЫЕ ВЫБОРКИ, СОСТОЯЩЕЙ ИЗ 50 ЛИСТОВ С ПОДПИСЯМИ

Число подписей	Частота	Число подписей	Частота
x_i	f_i	x_i	f_i
42	23	14	1
41	4	11	1
36	1	10	1
32	1	9	1
29	1	7	1
27	2	6	3
23	1	5	2
19	1	4	1
16	2	3	1
15	2		
Итого 50			

Нужно оценить общее число подписей под петицией и найти 80%-ные доверительные границы.

Единицей отбора служит лист, а наблюдения y_i представляют собой число подписей на каждом листе. Поскольку почти половина листов содержит максимальное число подписей — 42, данные представлены в виде распределения частот. Заметим, что исходное распределение кажется весьма далеким от нормального, наибольшая частота приходится на верхний конец. Тем не менее опыт позволяет полагать, что средние значения выборок объемом в 50 листов распределены приблизительно нормально. Находим

$$n = \sum f_i = 50; \quad y = \sum f_i y_i = 1471; \quad \sum f_i y_i^2 = 54\,497.$$

Следовательно, оценка суммарного числа подписей равна:

$$\hat{Y} = N\bar{y} = \frac{676 \cdot 1471}{50} = 19\,888.$$

Для оценки дисперсии по выборке, s^2 , имеем

$$s^2 = \frac{1}{n-1} [\sum f_i (y_i - \bar{y})^2] = \frac{1}{n-1} \left[\sum f_i y_i^2 - \frac{(\sum f_i y_i)^2}{\sum f_i} \right] =$$

$$= \frac{1}{49} \left[54\,497 - \frac{(1471)^2}{50} \right] = 229,0.$$

Согласно (2.22) 80%-ные доверительные границы вычисляются как

$$19\,888 \pm \frac{t_{Ns}}{\sqrt{n}} \sqrt{1-f} = 19\,888 \pm \frac{1,28 \cdot 676 \cdot 15,13 \sqrt{1-0,0740}}{\sqrt{50}}.$$

Это дает значения 80%-ных границ, равные 18 107 и 21 669. При сплошном подсчете общее число подписей оказалось равным 21 045.

2.8. ДРУГОЙ МЕТОД ДОКАЗАТЕЛЬСТВА

Корифилд (Cornfield, 1944) предложил метод доказательства основных результатов для простого случайного отбора без возвращения, позволяющий пользоваться обычными результатами, относящимися к бесконечным совокупностям. Пусть a_i — случайная переменная, принимающая значение 1, если i -я единица попала в выборку, и 0 в противном случае. Выборочное среднее \bar{y} можно записать как

$$\bar{y} = \frac{1}{n} \sum_{i=1}^N a_i y_i, \quad (2.23)$$

где суммирование распространяется на все N единиц совокупности. В этом выражении a_i — случайные переменные и y_i — некоторый набор неизменных чисел.

Очевидно, что соответствующие вероятности (Pr — от английского «probability» — вероятность) будут равны:

$$\text{Pr}(a_i = 1) = \frac{n}{N}; \quad \text{Pr}(a_i = 0) = 1 - \frac{n}{N}.$$

Таким образом, a_i имеет биномиальное распределение для одного испытания с $P = n/N$.

Следовательно,

$$E(a_i) = P = \frac{n}{N}; \quad V(a_i) = PQ = \frac{n}{N} \left(1 - \frac{n}{N} \right). \quad (2.24)$$

Для того чтобы найти $V(\bar{y})$, нужно знать, кроме того, ковариацию a_i и a_j . Произведение $a_i a_j$ равно 1, если i -я и j -я единицы обе попали в выборку, и равно 0 в противном случае. Вероятность того, что две конкретные единицы попали в выборку, как легко показать, равна $n(n-1)/N(N-1)$. Отсюда (cov — от английского «covariance» — ковариация)

$$\text{Cov}(a_i a_j) = E(a_i a_j) - E(a_i) E(a_j) = \frac{n(n-1)}{N(N-1)} -$$

$$- \left(\frac{n}{N} \right)^2 = - \frac{n}{N(N-1)} \left(1 - \frac{n}{N} \right). \quad (2.25)$$

Применяя этот подход для нахождения $V(\bar{y})$ из (2.23) и пользуясь также (2.24) и (2.25), получаем

$$V(\bar{y}) = \frac{1}{n^2} \left[\sum_{i=1}^N y_i^2 V(a_i) + 2 \sum_{i < j}^N y_i y_j \text{Cov}(a_i a_j) \right] =$$

$$= \frac{1-f}{nN} \left(\sum y_i^2 - \frac{2}{N-1} \sum y_i y_j \right).$$

Дополняя до квадрата суммы второй член в круглых скобках, имеем

$$V(\bar{y}) = \frac{1-f}{nN} \left(\frac{N}{N-1} \sum y_i^2 - \frac{1}{N-1} Y^2 \right) = \\ = \frac{1-f}{n(N-1)} \sum (y_i - \bar{Y})^2 = \frac{(1-f) S^2}{n}$$

Рассмотренный метод позволяет получить простые доказательства теорем 2.3 и 2.4. Им можно пользоваться для нахождения моментов распределения y высших порядков, хотя для этой цели существует более мощный метод, предложенный Тьюки (Tukey, 1950) и в дальнейшем усовершенствованный Уишартом (Wishart, 1952).

Аналогичный подход применяется в случае отбора с возвращением. В этом случае i -я единица может попасть в выборку 0, 1, 2, ..., n раз. Пусть t_i — число попаданий i -й единицы в выборку. Тогда

$$\bar{y} = \frac{1}{n} \sum_{i=1}^N t_i y_i \quad (2.26)$$

Поскольку i -я единица попадает в выборку при каждом извлечении с вероятностью $1/N$, то величина t_i распределена как число успехов в n испытаниях, т. е. по биномиальному закону с $P = 1/N$. Следовательно,

$$E(t_i) = \frac{n}{N}; \quad V(t_i) = n \left(\frac{1}{N} \right) \left(1 - \frac{1}{N} \right). \quad (2.27)$$

Совместно величины t_i имеют полиномиальное распределение. Для него

$$\text{Cov}(t_i, t_j) = -\frac{n}{N^2}. \quad (2.28)$$

Пользуясь (2.26), (2.27) и (2.28), получаем, что для отбора с возвращением

$$V(\bar{y}) = \frac{1}{n^2} \left[\sum_{i=1}^N y_i^2 \frac{n(N-1)}{N^2} - 2 \sum_{i < j} y_i y_j \frac{n}{N^2} \right] = \\ = \frac{1}{nN} \sum_{i=1}^N (y_i - \bar{Y})^2 = \frac{\sigma^2}{N} = \frac{N-1}{N} \frac{S^2}{n}.$$

2.9. ОЦЕНИВАНИЕ ОТНОШЕНИЯ

Часто величиной, которую нужно оценить по простой случайной выборке, служит отношение двух переменных, значения каждой из которых меняются от единицы к единице. При обследовании домохозяйств примерами могут служить среднее число костюмов, приходящееся на одного взрослого мужчину, средние расходы на косметику на одну женщину и среднее число часов в неделю, проводимых перед телевизором, на одного ребенка в возрасте от 10 до 15 лет. Для того чтобы оценить первую из этих характеристик, мы должны для каждого

i -го домохозяйства ($i = 1, 2, \dots, n$) зарегистрировать число взрослых мужчин x_i , живущих в нем, и общее число принадлежащих им костюмов y_i . Параметром совокупности, который нужно оценить, служит отношение

$$R = \frac{\text{общее число костюмов}}{\text{общее число взрослых мужчин}} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}.$$

Соответствующей выборочной оценкой будет

$$\hat{R} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{\bar{y}}{\bar{x}}.$$

Примеры такого рода часто возникают, когда единица отбора (домохозяйство) представляет собой группу или гнездо элементов (взрослых мужчин), а нас интересует среднее значение для совокупности на один элемент. Отношения могут встретиться и во многих других случаях, например, отношение банковских ссуд на строительство к общему объему банковских ссуд или отношение площади, занятой под пшеницей, к общей площади земли на ферме.

Выборочное распределение \hat{R} более сложно, чем распределение \bar{y} , потому что и числитель \bar{y} и знаменатель \bar{x} меняются от выборки к выборке. Для небольших выборок распределение \hat{R} асимметрично и \hat{R} в качестве оценки R имеет обычно небольшое смещение. При увеличении объема выборки распределение \hat{R} стремится к нормальному и смещение становится пренебрежимо малым. Для большинства практических целей достаточно следующего приближенного результата; более детально распределение \hat{R} рассматривается в гл. 6.

Теорема 2.5. Если переменные y_i, x_i наблюдаются для каждой единицы простой случайной выборки объема n (предполагается, что он велик), то дисперсия $\hat{R} = \bar{y}/\bar{x}$ составляет приблизительно

$$V(\hat{R}) \approx \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^n (y_i - R x_i)^2}{N-1}, \quad (2.29)$$

где $R = \bar{Y}/\bar{X}$ есть отношение средних для совокупности и $f = n/N$.
Доказательство

$$\hat{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}}, \quad (2.30)$$

Если l велико, то \bar{x} не должно сильно отличаться от \bar{X} . Аппроксимация состоит в том, что в знаменателе (2.30) вместо \bar{x} помещается \bar{X} . Это дает

$$\hat{R} - R \approx \frac{\bar{y} - R\bar{x}}{\bar{X}}. \quad (2.31)$$

Возьмем теперь среднее по всем простым случайным выборкам объема n .

$$E(\hat{R} - R) \approx \frac{E(\bar{y} - R\bar{x})}{\bar{X}} = \frac{\bar{Y} - R\bar{X}}{\bar{X}} = 0, \quad (2.32)$$

так как $R = \bar{Y}/\bar{X}$. Отсюда следует, что с точностью до порядка применяемой здесь аппроксимации \hat{R} есть несмещенная оценка R .

Из (2.31) получаем также

$$V(\hat{R}) = E(\hat{R} - R)^2 \approx \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2.$$

Величина $\bar{y} - R\bar{x}$ представляет собой выборочное среднее переменной $d_i = y_i - Rx_i$, среднее значение которой для совокупности $\bar{D} = \bar{Y} - R\bar{X} = 0$. Следовательно, мы можем найти $V(\hat{R})$, применяя к переменной d_i теорему 2.2 о дисперсии среднего для простой случайной выборки и деля результат на \bar{X}^2 . Это дает

$$\begin{aligned} V(\hat{R}) &\approx \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2 = \frac{1}{\bar{X}^2} \frac{S_d^2}{n} (1-f) = \\ &= \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^N (d_i - \bar{D})^2}{(N-1)} = \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^N (y_i - Rx_i)^2}{N-1}. \end{aligned}$$

Доказательство закончено.

Стоит отметить способ, которым была доказана теорема 2.5. Было показано, что формула приближенной дисперсии отношения \bar{y}/\bar{x} получается из формулы теоремы 2.2 для дисперсии выборочного среднего \bar{y} , если заменить переменную y_i переменной $(y_i - Rx_i)/\bar{X}$. Такое же утверждение, или его естественное обобщение, справедливо также и в более сложных случаях отбора и будет неоднократно применяться далее в этой книге.

В качестве выборочной оценки величины

$$\frac{\sum_{i=1}^N (y_i - Rx_i)^2}{N-1}$$

естественно взять

$$\frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1}.$$

Можно показать, что эта оценка имеет смещение порядка $1/n$. Для оценки стандартной ошибки \hat{R} по выборке имеем

$$s(\hat{R}) = \frac{\sqrt{1-f}}{\sqrt{n}\bar{X}} \sqrt{\frac{\sum (y_i - \hat{R}x_i)^2}{n-1}}. \quad (2.33)$$

Если \bar{X} неизвестно, то в знаменатель подставляется его выборочная оценка \bar{x} . Для более быстрого подсчета $s(\hat{R})$ на малых вычислительных машинах нужно записать эту оценку в виде

$$s(\hat{R}) = \frac{\sqrt{1-f}}{\sqrt{n}\bar{X}} \sqrt{\frac{\sum y_i^2 - 2\hat{R} \sum y_i x_i + \hat{R}^2 \sum x_i^2}{n-1}}. \quad (2.34)$$

Пример. В табл. 2.2. указаны число членов семьи (x_1), недельный доход семьи в долларах (x_2) и недельные расходы на питание в долларах (y) для простой случайной выборки, содержащей 33 семьи с низким уровнем дохода. Поскольку выборка мала, данные приводятся только для иллюстрации вычислений.

Оценим по выборке: (а) средние недельные расходы на питание на одну семью, (б) средние недельные расходы на питание на одного человека, (в) процент дохода, затрачиваемый на питание, и найдем стандартные ошибки этих оценок.

Недельные расходы на питание на одну семью. Это обычное выборочное среднее

$$\bar{y} = \frac{907,2}{33} = 27,49 \text{ долл.}$$

По теореме 2.2 (опуская пкс) стандартная ошибка этого среднего равна:

$$\begin{aligned} s_{\bar{y}} &= \frac{1}{\sqrt{n}} \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \frac{1}{\sqrt{n(n-1)}} \sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}} = \\ &= \frac{1}{\sqrt{33 \cdot 32}} \sqrt{28\,224 - (907,2)^2/33} = 1,76 \text{ долл.} \end{aligned}$$

(Сумма квадратов без поправки, равная 28 224, указана под табл. 2.2.)

Недельные расходы на питание на одного человека. Поскольку величина семьи меняется, эта оценка представляет собой отношение двух переменных:

$$\hat{R}_1 = \frac{\sum y}{\sum x_1} = \frac{907,2}{123} = 7,38 \text{ долл. на одного человека.}$$

Суммы квадратов и произведений, нужные для вычисления $S(\hat{R})$ по формуле (2.34), указаны под табл. 2.2. Кроме того, необходимо знать

$$2\hat{R}_1 = 14,7512; \hat{R}_1^2 = 54,3996; \bar{x}_1 = 3,7273.$$

Значения \hat{R}_1 , $2\hat{R}_1$, \hat{R}_1^2 приводятся с дополнительными десятичными знаками, чтобы сохранить точность.

Таблица 2.2

ЧИСЛО ЧЛЕНОВ СЕМЬИ, НЕДЕЛЬНЫЙ ДОХОД И РАСХОДЫ НА ПИТАНИЕ ДЛЯ 33 СЕМЕЙ

Номер семьи	Число членов семьи x_1	Доход x_2	Расходы на питание y	Номер семьи	Число членов семьи x_1	Доход x_2	Расходы на питание y
1	2	62	14,3	18	4	83	36,0
2	3	62	20,8	19	2	85	20,6
3	3	87	22,7	20	4	73	27,7
4	5	65	30,5	21	2	66	25,9
5	4	58	41,2	22	5	58	23,3
6	7	92	28,2	23	3	77	39,8
7	2	88	24,2	24	4	69	16,8
8	4	79	30,0	25	7	65	37,8
9	2	83	24,2	26	3	77	34,8
10	5	62	44,4	27	3	69	28,7
11	3	63	13,4	28	6	95	63,0
12	6	62	19,8	29	2	77	19,5
13	4	60	29,4	30	2	69	21,6
14	4	75	27,1	31	6	69	18,2
15	2	90	22,2	32	4	67	20,1
16	5	75	37,7	33	2	63	20,7
17	3	69	22,6				
Итого				123		2394	907,2

$$\Sigma x_1^2 = 533; \Sigma x_2^2 = 177\,254; \Sigma y^2 = 28\,224;$$

$$\Sigma x_1 y = 3595,5; \Sigma x_2 y = 66\,678.$$

Следовательно, из (2.34) имеем

$$s(\hat{R}_1) = \frac{1}{\sqrt{33 \cdot 3,7273}} \sqrt{\frac{28\,224 - 14,7512 \cdot 3595,5 + 54,3996 \cdot 533}{32}} = 0,534 \text{ долл.}$$

Процент дохода, затрачиваемый на питание. Это опять отношение двух переменных

$$\hat{R}_2 = 100 \frac{\Sigma y}{\Sigma x_2} = \frac{100 \cdot 907,2}{2394} = 37,9\%.$$

По формуле (2.34) читатель может проверить, что стандартная ошибка равна 2,38%.

2.10. ОЦЕНКИ СРЕДНИХ ЗНАЧЕНИЙ ДЛЯ ПОДСОВОКУПНОСТЕЙ

Во многих обследованиях получают отдельные оценки для каждой из некоторого числа групп, на которые подразделена совокупность. При обследовании домохозяйств может оказаться желательным иметь отдельные оценки для семей с 0, 1, 2, ... детьми, для домовладельцев и квартиронанимателей или для семей по различным группам занятий. Для таких подсовокупностей Подкомиссией ООН по статистической выборке (U. N. Statistical Office, 1950) был рекомендован термин *области изучения*.

В простейшем случае каждая единица совокупности попадает в одну из таких областей. Пусть j -я область содержит N_j единиц и пусть n_j — число единиц простой случайной выборки объема n , принадлежащих к этой области. Если y_{jk} ($k = 1, 2, \dots, n_j$) — значение наблюдений по этим единицам, то среднее значение для совокупности по j -й области, \bar{y}_j , оценивается по формуле

$$\bar{y}_j = \sum_{k=1}^{n_j} \frac{y_{jk}}{n_j}. \quad (2.35)$$

На первый взгляд \bar{y}_j кажется такой же оценкой отношения, как в параграфе 2.9, поскольку, хотя n неизменно, n_j будет меняться от одной выборки объема n к другой. Связанного с этим осложнения можно избежать, рассматривая распределение \bar{y}_j по выборкам, у которых неизменны как n , так и n_j .

На множестве выборок с заданными n и n_j вероятность того, что в выборку попадет некоторый конкретный набор n_j единиц, принадлежащих области j , из общего числа N_j таких единиц, составляет

$$\frac{C_{N-N_j}^{n-n_j}}{C_{N-N_j}^{n-n_j} \cdot C_{N_j}^{n_j}} = \frac{1}{C_{N_j}^{n_j}}.$$

Действительно, каждый конкретный набор n_j единиц, принадлежащих области j , может попасть в выборку вместе с любыми $(n - n_j)$ единицами из общего числа $(N - N_j)$ единиц, не принадлежащих области j . Поэтому числитель предыдущей формулы представляет собой число выборок, содержащих конкретный набор n_j единиц, а знаменатель — общее число выборок с заданными n и n_j . Отсюда следует, что к y_{jk} можно применить теоремы 2.1, 2.2 и 2.4, если мы возьмем n_j вместо n и N_j вместо N .

По теореме 2.1: \bar{y}_j есть несмещенная оценка \bar{Y}_j . (2.36)

По теореме 2.2: стандартная ошибка \bar{y}_j равна

$$\frac{s_j}{\sqrt{n_j}} \sqrt{1 - (n_j/N_j)}, \quad (2.37)$$

где

$$s_j^2 = \sum_{k=1}^{n_j} \frac{(y_{jk} - \bar{y}_j)^2}{n_j - 1}. \quad (2.38)$$

По теореме 2.4: оценкой стандартной ошибки \bar{y}_j служит

$$\frac{s_j}{\sqrt{n_j}} \sqrt{1 - (n_j/N_j)}, \quad (2.39)$$

где

$$s_j^2 = \sum_{k=1}^{n_j} \frac{(y_{jk} - \bar{y}_j)^2}{n_j - 1}. \quad (2.40)$$

Если значение N_j неизвестно, то при вычислении пкс вместо n_j/N_j можно воспользоваться величиной n/N . (При простом случайном отборе n_j/N_j есть несмещенная оценка n/N).

2.11. ОЦЕНКИ СУММАРНЫХ ЗНАЧЕНИЙ ДЛЯ ПОДСОВОКУПНОСТЕЙ

Предположим, что у некоторой фирмы имеется список счетов, в котором часть счетов оплачена, а часть нет, и мы хотели бы оценить по выборке общую сумму неоплаченных счетов. Если N_j (число всех неоплаченных счетов фирмы) известно, то не возникает никакой проблемы. Выборочной оценкой будет $N_j \bar{y}_j$, а ее условной стандартной ошибкой — выражение (2.37), умноженное на N_j .

В другом случае, если известна общая сумма счетов в списке, можно применить оценку отношения. По выборке получаем оценку отношения (общая сумма неоплаченных счетов, деленная на общую сумму всех счетов). После этого умножаем ее на известную общую сумму всех счетов.

Если неизвестны ни N_j , ни общая сумма счетов, то эти оценки применить нельзя. Вместо них умножим выборочное суммарное значение всех y -ов по единицам, принадлежащим области j , на множитель распространения N/n . Это даст оценку

$$\hat{Y}_j = \frac{N}{n} \sum_{k=1}^{n_j} y_{jk}. \quad (2.41)$$

Мы покажем, что \hat{Y}_j — несмещенная оценка, и найдем ее стандартную ошибку по всем многократным выборкам объема n . Прием, заключающийся в том, чтобы считать неизменными как n_j , так и n , в данном случае не помогает.

При изложении доказательства мы вернемся к прежним обозначениям, в которых y_i — значение наблюдения у i -й единицы в совокупности. Для каждой единицы совокупности определим новую переменную y'_i , считая

$$y'_i = \begin{cases} y_i, & \text{если единица принадлежит } j\text{-й области изучения} \\ 0 & \text{в противном случае.} \end{cases}$$

Суммарное значение y'_i по совокупности

$$\sum_{i=1}^N y'_i = \sum_{j \in \text{обл}} y_i = Y_j.$$

Для простой случайной выборки объема n $y'_i = y_i$ для каждой из n_j единиц, принадлежащих j -й области; $y'_i = 0$ для каждой из оставшихся $n - n_j$ единиц. Если \bar{y}' — обычное выборочное среднее y'_i , то справедливо равенство

$$N\bar{y}' = \frac{N}{n} \sum_{i=1}^n y'_i = \frac{N}{n} \sum_{k=1}^{n_j} y_{jk} = \hat{Y}_j.$$

Полученный результат показывает, что оценка \hat{Y}_j , определяемая равенством (2.41), в N раз больше выборочного среднего y'_i .

Для многократных выборок объема n мы можем, очевидно, применить к переменной y'_i теоремы 2.1, 2.2 и 2.4. Из них следует, что \hat{Y}_j представляет собой несмещенную оценку Y_j со стандартной ошибкой

$$\sigma(\hat{Y}_j) = \frac{NS'}{\sqrt{n}} \sqrt{1 - (n/N)}, \quad (2.42)$$

где S' — среднее квадратичное отклонение y'_i для совокупности. Для того чтобы вычислить S' , мы считаем, что совокупность состоит из N_j значений y_i , лежащих в j -й области, и $N - N_j$ значений, равных нулю. Таким образом,

$$S'^2 = \frac{1}{N-1} \left(\sum_{j \in \text{обл}} y_i^2 - \frac{Y_j^2}{N} \right). \quad (2.43)$$

По теореме 2.4 выборочной оценкой стандартной ошибки \hat{Y}_j служит

$$s(\hat{Y}_j) = \frac{Ns'}{\sqrt{n}} \sqrt{1 - (n/N)}. \quad (2.44)$$

При вычислении s' значение каждой единицы, не принадлежащей j -й области изучения, считается равным нулю. Изучающим выборочный метод бывает, по-видимому, трудно внутренне согласиться с этим правилом, но тем не менее оно верно.

Методы этого и предшествующего параграфов применимы также к обследованиям, в которых основа выборки содержит единицы, не принадлежащие совокупности согласно ее определению. Приведем пример, иллюстрирующий это замечание.

Пример. Из общего списка 2422 мелких статей домашних расходов была получена выборка 180 статей, предназначенная для оценки общей суммы расходов на ведение домашнего хозяйства. Некоторые виды расходов (на одежду и на содержание автомобиля) не были признаны мелкими расходами и из 180 единиц выборки были оставлены 152. Сумма расходов по оставленным статьям (в долларах) и сумма квадратов (без поправки) оказались следующими:

$$\sum y'_i = 343,5; \quad \sum y_i'^2 = 1491,38.$$

Оценим общую сумму расходов на ведение домашнего хозяйства и найдем стандартную ошибку этой оценки.

$$\hat{Y}_J = \frac{N}{n} \sum_{i=1}^n y_i' = \frac{2422 \cdot 343,5}{180} = 4622 \text{ долл.}$$

Из (2.44)

$$s(\hat{Y}_J) = \frac{Ns'}{\sqrt{n}} \sqrt{1 - (n/N)}.$$

При вычислении s' мы считаем, что наша выборка 180 статей имеет 28 нулевых значений. Следовательно,

$$\begin{aligned} s'^2 &= \frac{1}{179} \left(\sum y_i'^2 - \frac{(\sum y_i')^2}{180} \right) = \\ &= \frac{1}{179} \left(1491,38 - \frac{(343,5)^2}{180} \right) = 4,670. \end{aligned}$$

Окончательно

$$s_{\hat{Y}_J} = 2422 \sqrt{\frac{4,670}{180} \left(1 - \frac{180}{2422} \right)} = 375 \text{ долл.}$$

Оценка не очень точна, ее коэффициент вариации составляет 375/4622, или приблизительно 8%.

В этом примере расходы на содержание автомобиля и на одежду были исключены как несоответствующие определению совокупности, следовательно, рассматривались в выборке как нулевые значения. В некоторых случаях заранее известно, что отдельные единицы совокупности не участвуют в образовании оцениваемого суммарного значения. Например, при обследовании магазинов с целью оценить общую выручку от продажи саквояжей оказывается, что некоторые магазины не продают саквояжей; на отдельных участках территории, принятых в качестве единиц отбора для исследования ферм, фермы отсутствуют. Иногда можно, затратив некоторые усилия, определить такие пустые единицы и подсчитать их число, так что, в наших обозначениях, станет известным $(N - N_J)$, а следовательно, и N_J .

Поэтому имеет смысл выяснить, насколько уменьшается $V(\hat{Y}_J)$, если известно N_J . Если N_J не известно, то согласно (2.42)

$$V(\hat{Y}_J) = \frac{N^2 S'^2}{n} \left(1 - \frac{n}{N} \right).$$

Если \bar{Y}_J и S_J — среднее значение и среднее квадратичное отклонение для рассматриваемой области изучения (т. е. среди ненулевых единиц), то, как может проверить читатель,

$$(N-1)S'^2 = (N_J-1)S_J^2 + N_J\bar{Y}_J^2 \left(1 - \frac{N_J}{N} \right).$$

Поскольку членами при $1/N_J$ и $1/N$ почти всегда можно пренебречь,

$$S'^2 \approx P_J S_J^2 + P_J Q_J \bar{Y}_J^2, \quad (2.45)$$

где $P_J = N_J/N$ и $Q_J = 1 - P_J$. Отсюда получаем

$$V(\hat{Y}_J) \approx \frac{N^2}{n} (P_J S_J^2 + P_J Q_J \bar{Y}_J^2) \left(1 - \frac{n}{N} \right). \quad (2.46)$$

Если ненулевые единицы определены, то мы извлекаем выборку объема n_J из их числа. Оценкой суммарного значения по области изучения служит $N_J \bar{y}_J$ с дисперсией

$$V(N_J \bar{y}_J) = \frac{N_J^2}{n_J} S_J^2 \left(1 - \frac{n_J}{N_J} \right) = \frac{N^2}{n_J} P_J^2 S_J^2 \left(1 - \frac{n_J}{N_J} \right). \quad (2.47)$$

Сравним дисперсии (2.46) и (2.47). Для (2.46) среднее число ненулевых единиц в выборке объема n равно $n P_J$. Если в (2.47) мы положим $n_J = n P_J$, так что число наблюдаемых ненулевых единиц станет приблизительно одинаковым для обоих методов, то (2.47) примет вид

$$V(N_J \bar{y}_J) = \frac{N^2}{n} P_J S_J^2 \left(1 - \frac{n}{N} \right). \quad (2.48)$$

Отношение дисперсий (2.48) и (2.46) будет

$$\frac{V(N_J \text{ известно})}{V(N_J \text{ не известно})} = \frac{S_J^2}{S_J^2 + Q_J \bar{Y}_J^2} = \frac{C_J^2}{C_J^2 + Q_J},$$

где $C_J = S_J/\bar{Y}_J$ есть коэффициент вариации среди ненулевых единиц. Как и следовало ожидать, уменьшение дисперсии, связанное с тем, что N_J известно, больше, если доля нулевых единиц велика и если y_J среди ненулевых единиц меняется сравнительно мало. Более подробный анализ этой проблемы читатель найдет в работе Джессена и Хауземана (Jessen and Houseman, 1944).

2.12. СРАВНЕНИЕ СРЕДНИХ ЗНАЧЕНИЙ ДЛЯ ОБЛАСТЕЙ ИЗУЧЕНИЯ

Пусть \bar{y}_j, \bar{y}_k — значения выборочных средних для j -й и k -й из числа областей изучения, к которым отнесены единицы простой случайной выборки. Дисперсия их разности равна:

$$V(\bar{y}_j - \bar{y}_k) = V(\bar{y}_j) + V(\bar{y}_k).$$

Эта формула применима также и к разности двух отношений: \hat{R}_j и \hat{R}_k .

Необходимо отметить одно обстоятельство. Проверка равенства $\bar{Y}_j = \bar{Y}_k$ редко представляет научный интерес, потому что за исключением отдельных случаев в конечной совокупности эти средние не будут в точности равны, даже если данные по обоим областям получены путем случайного отбора из одной и той же бесконечной совокуп-

ности. Вместо этого нулевой гипотезой обычно служит гипотеза о том, что две области изучения получены из бесконечных совокупностей, имеющих одинаковые средние значения. Поэтому при вычислении $V(\bar{y}_j)$ и $V(\bar{y}_h)$ мы опускаем пкс, пользуясь формулой

$$V(\bar{y}_j - \bar{y}_h) = \frac{S_j^2}{n_j} + \frac{S_h^2}{n_h}.$$

2.13. ОБОСНОВАННОСТЬ АППРОКСИМАЦИИ НОРМАЛЬНЫМ РАСПРЕДЕЛЕНИЕМ

Уверенность в том, что для большинства практических приложений аппроксимация нормальным распределением оправдана, происходит из многих источников. В теории вероятностей большое внимание уделялось распределению средних значений случайных выборок. Было доказано, что для любой совокупности с конечным средним квадратичным отклонением распределение выборочного среднего стремится к нормальному при увеличении n [см., например, книгу Феллера (Feller, 1957)]. Эти исследования относятся к бесконечным совокупностям.

Для отбора без возвращения из конечных совокупностей Гаек (Hájek, 1960), следуя работам Эрдеша и Реньи (Erdős and Rényi, 1959) и Мэдоу (Madow, 1948), указал необходимые и достаточные условия, при которых распределение выборочного среднего стремится к нормальному. Гаек исходил из последовательности значений n_v, N_v , стремящихся к бесконечности таким образом, что $(N_v - n_v)$ также стремится к бесконечности. Наблюдения в v -й совокупности обозначаются через y_{vi} ($i = 1, 2, \dots, N_v$). Пусть для этой совокупности S_{vt} будет набором единиц совокупности, для которых выполняется условие

$$|y_{vi} - \bar{Y}_v| > \tau \sqrt{n_v(1-f_v)} S_v,$$

где \bar{Y}_v, S_v, f_v — соответственно среднее значение, среднее квадратичное отклонение и пкс для этой совокупности, а τ — некоторое число > 0 . Тогда условие типа условия Линдеберга

$$\lim_{v \rightarrow \infty} \frac{\sum_{i \in S_{vt}} (y_{vi} - \bar{Y}_v)^2}{(N_v - 1) S_v^2} = 0$$

будет необходимым и достаточным для того, чтобы распределение \bar{y}_v стремилось к нормальному со средним и дисперсией, определяемыми теоремами 2.1 и 2.2.

Эти внушительные результаты оставляют кое-что недосказанным. Все-таки нелегко ответить на прямой вопрос: насколько большим должно быть n для данной совокупности, чтобы аппроксимация нормальным распределением была достаточно верной? Распределения, отличные от нормального, сильно разнятся одно от другого как по своей приро-

де, так и по степени их отклонения от нормального распределения. В практике выборочных обследований нельзя предполагать, что все распределения частот будут достаточно близки к нормальному. Распределения для многих видов экономических объектов (магазины, птицеводческие фермы, города) обнаруживают заметную положительную асимметрию, когда имеется несколько крупных единиц и большое

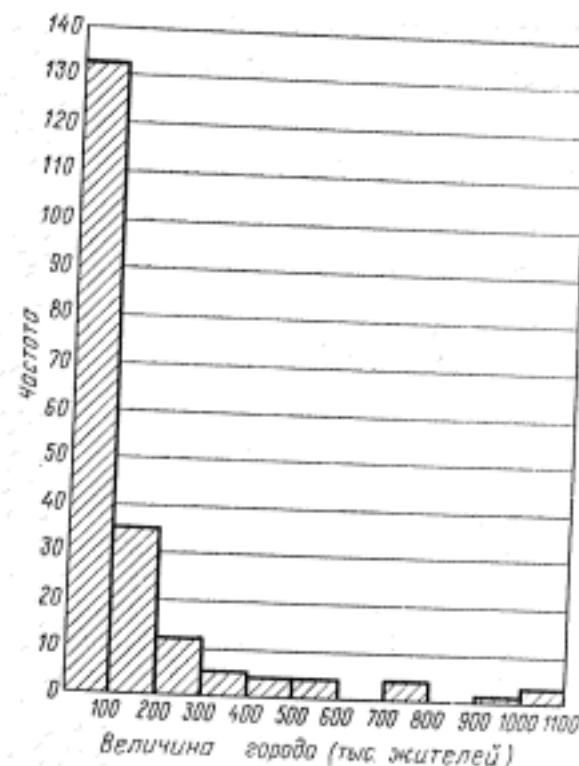


Рис. 2.1. Распределение частот величины 196 городов США в 1920 г.

число малых. Асимметрию того же типа проявляют и некоторые биологические популяции (например, число крыс или мух в городском квартале).

В качестве иллюстрации положительно асимметричного распределения на рис. 2.1 изображено распределение частот числа жителей в 196 больших городах США в 1920 г. (Четыре крупнейших города — Нью-Йорк, Чикаго, Филадельфия и Детройт — исключены. Их включение удлинит бы горизонтальную шкалу более чем в пять раз по сравнению с приведенной и, конечно, еще больше подчеркнуло бы асимметрию.) На рис. 2.2 изображено распределение частот суммарного числа жителей 200 простых случайных выборок объемом в $n = 49$ городам, полученных из этой совокупности. Распределение сум-

марных и соответственно средних выборочных значений значительно более походит на нормальную кривую, хотя и обнаруживает еще некоторую положительную асимметрию.

При любом обсуждении обоснованности аппроксимации нормальным распределением мы должны уточнить, что подразумевается под словами, что аппроксимация нормальным распределением «достаточно верна». При выборочных обследованиях аппроксимация нормальным распределением применяется в основном для вычисления доверительных границ. Если для среднего значения совокупности, \bar{Y} , вычисляются с помощью нормальной аппроксимации 95%-ные доверительные границы, то мы делаем следующее утверждение:

$$\bar{y} - 1,96s_{\bar{y}} < \bar{Y} < \bar{y} + 1,96s_{\bar{y}}. \quad (2.49)$$

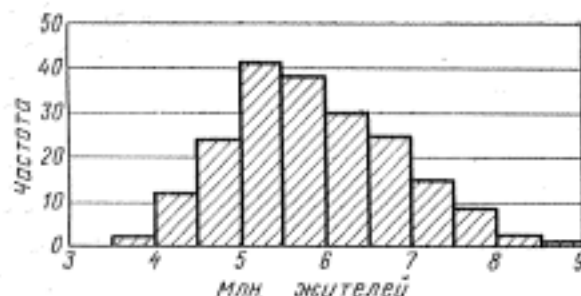


Рис. 2.2. Распределение частот суммарного значения 200 простых случайных выборок объема $n=49$

Мы заявляем, что при многократном отборе подобные утверждения будут ошибочны только в 5% случаев. Следовательно, мы могли бы сказать, что нормальная аппроксимация достаточно верна, если такие утверждения в действительности ошибочны в 4—6% случаев. Выбор чисел 4 и 6 довольно произволен: некоторых исследователей удовлетворяют и более широкие пределы.

Изучение теоретических распределений, обладающих асимметрией, и результаты выборочных экспериментов с реальными асимметричными совокупностями дают возможность сделать некоторые утверждения относительно того, что обычно происходит с доверительными вероятностями при отборе из положительно асимметричных совокупностей. Объем выборки предполагается достаточно большим, чтобы распределение \bar{y} обнаруживало некоторое сходство с нормальным, как на рис. 2.2. Эти утверждения таковы:

1. Частота, с которой ошибочно утверждение

$$\bar{y} - 1,96s_{\bar{y}} < \bar{Y} < \bar{y} + 1,96s_{\bar{y}},$$

обычно больше 5%.

2. Частота, с которой

$$\bar{Y} > \bar{y} + 1,96s_{\bar{y}},$$

больше 2,5%.

3. Частота, с которой

$$\bar{Y} < \bar{y} - 1,96s_{\bar{y}},$$

меньше 2,5%.

В качестве иллюстрации рассмотрим переменную y , распределенную строго по биномиальному закону, так что точное распределение \bar{y} можно получить из таблиц биномиального распределения. Переменная y принимает только два значения: значение h с вероятностью P и значение 0 с вероятностью Q . Среднее для совокупности $\bar{Y} = Ph$. В простой случайной выборке объема n оказалось a единиц, имеющих значение h и $n - a$ единиц со значением 0. Для этой выборки

$$\Sigma y = ah; \quad \bar{y} = \frac{ah}{n};$$

$$(n-1)s^2 = \Sigma y^2 - n\bar{y}^2 = ah^2 - \frac{a^2 h^2}{n};$$

$$s_{\bar{y}}^2 = \frac{s^2}{n} = \frac{h^2}{n^2} \frac{a(n-a)}{n-1}.$$

Следовательно, 95%-ные доверительные границы для \bar{Y} равны

$$\bar{y} \pm 1,96s_{\bar{y}} = \frac{h}{n} \left[a \pm 1,96 \sqrt{\frac{a(n-a)}{n-1}} \right]. \quad (2.50)$$

Пусть $n = 400$, $P = 0,1$. Тогда $\bar{Y} = 0,1h$. Подбирая разные a , находим, что при $a = 29$ выражение (2.50) дает значение верхней границы, равное $39,18h/400 = 0,098h$, в то время как при $a = 30$ получаем значение $40,34h/400 = 0,101h$. Следовательно, при любых $a \leq 29$ верхняя граница оказывается слишком низкой. Аналогично устанавливаем, что при $a \geq 54$ нижняя граница слишком высока.

Переменная a распределена по биномиальному закону при $n = 400$, $P = 0,1$. Из таблиц (Harvard Computation Laboratory, 1955) следует, что вероятности высказанных ранее утверждений будут следующими:

$$\Pr \{ \text{установленная верхняя граница слишком низка} \} = \Pr \{ a < 29 \} = 0,0357$$

$$\Pr \{ \text{установленная нижняя граница слишком высока} \} = \Pr \{ a > 54 \} = 0,0217$$

$$\Pr \{ \text{доверительное утверждение ошибочно} \} = 0,0574$$

Общая вероятность ошибиться ненамного превышает 0,05. В более чем 60% ошибочных утверждений истинное среднее значение будет больше, чем вычисленная верхняя граница.

Надежного общего правила о том, насколько велико должно быть n , чтобы при вычислении доверительных границ можно было применять нормальную аппроксимацию, не существует. Для совокупностей, в которых основное отклонение от нормальности заключается в замет-

ной положительной асимметрии, в некоторых случаях оказывается полезным грубое правило

$$n > 25G_1^2,$$

где G_1 — мера асимметрии Фишера (Fisher, 1932),

$$G_1 = \frac{E(y_i - \bar{Y})^3}{\sigma^3} = \frac{1}{N\sigma^3} \sum_{i=1}^N (y_i - \bar{Y})^3.$$

Это правило построено таким образом, что вероятностное утверждение на 95%-ном доверительном уровне ошибочно не более чем в 6% случаев. Математически оно выведено в предположении, что влиянием моментов распределения \bar{y} выше третьего можно пренебречь. Предложенное правило учитывает только общую частоту ошибочных утверждений, направление ошибки оценки во внимание не принимается.

Вычислив для конкретной совокупности G_1 или ее оценку, мы можем получить приближенное представление об объеме выборки, необходимом для того, чтобы при вычислении доверительных границ можно было применить нормальную аппроксимацию. Полученные результаты нужно проверить, если это возможно, путем экспериментального отбора.

Пример. В табл. 2.3 приведены данные о числе акров посевной площади на 556 фермах в графстве Сенека штата Нью-Йорк. Данные взяты из серии исследований, проведенных Уэстом (West, 1951), который многократно извлекал выборки объема 100 из этой совокупности и исследовал распределения частот \bar{y} , s и t -распределение Стюдента для нескольких признаков, изучаемых при сельскохозяйственных обследованиях.

Таблица 2.3

РАСПРЕДЕЛЕНИЕ ЧАСТОТ ЧИСЛА АКРОВ ПОСЕВНОЙ ПЛОЩАДИ НА 556 ФЕРМАХ

Группы по величине посевной площади (в акрах)	Условная шкала y_i	Частота f_i	$f_i y_i$	$f_i y_i^2$	$f_i y_i^3$
0—29	—0,9	47	—42,3	38,1	—34,3
30—63	0	143	0	0	0
64—97	1	154	154	154	154
98—131	2	82	164	328	656
132—165	3	62	186	558	1 674
166—199	4	33	132	528	2 112
200—233	5	13	65	325	1 625
234—267	6	6	36	216	1 296
268—301	7	4	28	196	1 372
302—335	8	6	48	384	3 072
336—369	9	2	18	162	1 458
370—403	10	0	0	0	0
404—437	11	2	22	242	2 662
438—471	12	0	0	0	0
472—505	13	2	26	338	4 394
Итого		556	836,7	3469,1	20 440,7

$$E(y_i) = \bar{Y} = \frac{836,7}{556} = 1,50486;$$

$$E(y_i^2) = \frac{3469,1}{556} = 6,23939;$$

$$E(y_i^3) = \frac{20440,7}{556} = 36,76385;$$

$$\sigma^2 = E(y_i^2) - \bar{Y}^2 = 3,97479;$$

$$\kappa_3 = E(y_i - \bar{Y})^3 = E(y_i^3) - 3E(y_i^2)\bar{Y} + 2\bar{Y}^3 = 15,411;$$

$$G_1 = \frac{\kappa_3}{\sigma^3} = \frac{15,411}{7,925} = 1,9.$$

Порядок вычислений G_1 показан после таблицы. Все вычисления сделаны по условной шкале и, так как G_1 — отвлеченное число, возвращаться к исходной шкале не обязательно. Отметим, что величина первого интервала несколько отличается от остальных.

Поскольку $G_1 = 1,9$, в качестве предполагаемого минимального объема выборки мы принимаем

$$n = 25 \cdot (1,9)^2 = 90.$$

Для выборок объема 100 Уэст установил, что для рассматриваемого признака (акры посевной площади) ни распределение \bar{y} , ни t -распределение Стюдента не отличаются значительно от соответствующих теоретических нормальных распределений.

На практике в образцовых выборочных исследованиях стараются сделать аппроксимацию нормальным распределением более обоснованной. Неправомерность такой аппроксимации, как правило, связана с тем, что совокупность содержит элементы, сильно отличающиеся от остальных, и они определяют выборочное среднее, если попадают в выборку. Еще сильнее влияние этих элементов сказывается в увеличении дисперсии выборки и в уменьшении точности. Поэтому весьма разумно отделить их от остальных и исследовать отдельно, возможно путем сплошного наблюдения, если эти элементы не слишком многочисленны. Такое выделение крайних элементов из основного состава совокупности уменьшает асимметрию и повышает обоснованность нормальной аппроксимации. Этот прием служит примером расслоенного отбора, который рассматривается в гл. 5.

2.14. ВЛИЯНИЕ ОТКЛОНЕНИЯ РАСПРЕДЕЛЕНИЯ ОТ НОРМАЛЬНОГО НА ВЫБОРОЧНУЮ ДИСПЕРСИЮ

Одно из последствий отклонения распределения генеральной совокупности (parent distribution) от нормального заключается в том, что выборочная дисперсия, s^2 , может значительно сильнее меняться от выборки к выборке, чем можно было бы ожидать, предполагая, что отбор производится из нормально распределенной совокупности. Для любой бесконечной совокупности дисперсия s^2 для случайных выборок объема n равна (Fisher, 1932):

$$V(s^2) = \frac{2\sigma^4}{n-1} + \frac{\kappa_4}{n}. \quad (2.51)$$

Первое слагаемое в этом выражении представляет собой значение дисперсии s^2 для случая, когда распределение генеральной совокупности нормально. Второй член характеризует отклонение от нормальности. Величина κ_4 представляет собой четвертый кумулянт Фишера (Fisher, 1932) и задается формулой

$$\kappa_4 = E(y_i - \bar{Y})^4 - 3\sigma^4.$$

Заметим, что асимметрия исходного распределения, измеряемая с помощью G_1 , не влияет на устойчивость s^2 : решающее значение имеет четвертый момент генеральной совокупности.

Для нормального распределения кумулянт κ_4 равен нулю. Для других распределений он может принимать положительные и отрицательные значения, но для распределений, встречающихся в практике выборочных исследований, κ_4 , по-видимому, гораздо чаще оказывается положительным, а для некоторых из них может принимать довольно большие значения.

Выражение (2.51) можно записать в виде

$$V(s^2) = \frac{2\sigma^4}{n-1} \left(1 + \frac{n-1}{2n} \frac{\kappa_4}{\sigma^4} \right) = \frac{2\sigma^4}{n-1} \left(1 + \frac{n-1}{2n} G_2 \right),$$

где $G_2 = \kappa_4/\sigma^4$ есть фишера мера эксцесса (*loc. cit.*). Величина в скобках выражает степень влияния на дисперсию s^2 , оказываемого отклонением от нормальности. Отметим, что этот множитель почти не зависит от n , так что указанное влияние сохраняется даже при больших выборках.

По данным Уэста о посевных площадях на фермах (табл. 2.3) величина G_2 составляет приблизительно 6. Таким образом, величина $V(s^2)$ почти в четыре раза больше того значения, которое получилось бы, если бы исходное распределение числа акров под посевами считалось нормальным. В своих исследованиях Уэст установил, что аналогичное изменение происходит с дисперсией среднего квадратичного отклонения s по трем признакам, которые он рассматривал. Отношение $V(s)$ к теоретической дисперсии s для нормальной совокупности равнялось 3,7 для числа акров под посевами, 2,1 — для общей обрабатываемой площади и 13,7 — для условных единиц производительности труда. (Согласно теории это отношение должно быть приблизительно одинаковым для s и для s^2 .)

Значение этих результатов для практического применения выборочного метода связано с тем, что мы иногда пользуемся значениями s^2 , чтобы сравнить точность одного метода отбора с точностью другого или оценить объем выборки, необходимый для достижения желательного уровня точности \bar{y} (см. гл. 4). Для этих целей полезно иметь некоторое представление о точности оценки s^2 , особенно если она была вычислена по весьма скудным данным. Как показывают полученные ранее результаты, применение «нормальных» формул для нахождения дисперсии s^2 может создать очень обманчивое впечатление об устойчивости s^2 .

Упражнения

2.1. Для совокупности с $N = 6$ значения y_i равны 8, 3, 1, 11, 4 и 7. Вычислите выборочное среднее \bar{y} для всех возможных простых случайных выборок объема 2. Проверьте, что \bar{y} представляет собой несмещенную оценку \bar{Y} и что ее дисперсия совпадает с указанной в теореме 2.2.

2.2. Для той же совокупности вычислите s^2 для всех простых случайных выборок объема 3 и проверьте, что $E(s^2) = S^2$.

2.3. Если из той же совокупности путем отбора с возвращением извлекаются выборки объема 2, покажите, найдя все возможные выборки, что $V(\bar{y})$ удовлетворяет уравнению

$$V(\bar{y}) = \frac{\sigma^2}{n} = \frac{S^2}{n} \frac{(N-1)}{N}.$$

2.4. Простая случайная выборка объемом в 30 домохозяйств была отобрана из городского района, содержащего 14 848 домохозяйств. Числа лиц для каждого домохозяйства в выборке следующие:

5, 6, 3, 3, 2, 3, 3, 3, 4, 4, 3, 2, 7, 4, 3, 5, 4, 4, 3, 3, 4, 3, 3, 1, 2, 4, 3, 4, 2, 4. Оцените общее число людей в районе и вычислите вероятность того, что эта оценка находится в пределах $\pm 10\%$ истинного значения.

2.5. При изучении возможности применения выборочного метода для учета товаров на складе произвели подсчет стоимости товаров на каждой из 36 полок складского помещения. Получены следующие округленные до целых долларов значения:

29, 38, 42, 44, 45, 47, 51, 53, 53, 54, 56, 56, 56, 58, 58, 59, 60, 60, 60, 60, 61, 61, 61, 62, 64, 65, 65, 67, 67, 68, 69, 71, 74, 77, 82, 85.

Оценка суммарного значения, сделанная по выборке, должна быть точной в пределах 200 долл., за исключением одного случая из 20. Эксперт считает, что для этого достаточно взять простую случайную выборку объемом в 12 полок. Согласны ли вы с ним?

$$\Sigma y = 2138; \Sigma y^2 = 131\,682.$$

2.6. После того как выборка, приведенная в табл. 2.1 (с. 42), была получена, подсчитали число полностью заполненных листов (каждый с 42 подписями), оно оказалось равным 326. Пользуясь этими сведениями, получите улучшенную оценку общего числа подписей и найдите стандартную ошибку вашей оценки.

2.7. Из списка 468 небольших двухгодичных колледжей была взята простая случайная выборка объемом в 100 колледжей. В выборке оказалось 54 государственных и 46 частных колледжей. Данные о числе студентов (y) и числе преподавателей (x) показаны в следующей таблице:

Колледжи	n	Σy	Σx
Государственные	54	31 281	2 024
Частные	46	13 707	1 075
Колледжи	Σy^2	Σyx	Σx^2
Государственные	29 881 219	1 729 349	111 090
Частные	6 366 785	431 041	33 119

(а) Для колледжей каждого типа оцените отношение числа студентов к числу преподавателей. (б) Вычислите стандартные ошибки ваших оценок. (в) По данным для государственных колледжей найдите 90%-ные доверительные границы для отношения студенты/преподаватели во всей совокупности.

2.8. В предыдущем примере проверьте на 5%-ном уровне значимости, будет ли отношение студенты/преподаватели существенно различаться для двух типов колледжей.

2.9. Для государственных колледжей оцените общее число преподавателей: (а) если известно, что общее число государственных колледжей в совокупности равно 251; (б) не зная этого числа. В обоих случаях вычислите стандартную ошибку вашей оценки.

2.10. Далее в таблице указано число жителей в каждом из 197 американских городов, насчитывавших в 1940 г. более 50 000 населения. Вычислите стандартную ошибку оценки общего числа жителей во всех 197 городах для следующих вариантов отбора: (а) простая случайная выборка объемом в 50 городов; (б) выборка, включающая пять крупнейших городов и кроме того простую случайную выборку объемом 45 из оставшихся 192 городов; (в) выборка, включающая девять крупнейших городов и простую случайную выборку объемом 41 из оставшихся городов.

РАСПРЕДЕЛЕНИЕ ЧАСТОТ ВЕЛИЧИНЫ ГОРODOB

Группа по величине города (тыс. жителей)	Частота f	Группа по величине города (тыс. жителей)	Частота f	Группа по величине города (тыс. жителей)	Частота f
50—100	105	550—600	2
100—150	36	600—650	1	1500—1550	1
150—200	13	650—700	2
200—250	6	700—750	0	1600—1650	1
250—300	7	750—800	1
300—350	8	800—850	1	1900—1950	1
350—400	4	850—900	2
400—450	1	900—950	0	3350—3400	1
450—500	3	950—1000	0
500—550	0	1 000—1 050	0	7450—7500	1

... означает пропуск интервала.

2.11. Вычислите коэффициент асимметрии G_1 для исходной совокупности и для совокупности, остающейся после удаления (а) пяти крупнейших городов, (б) девяти крупнейших городов.

2.12. Нужно провести небольшое обследование для сравнения домовладельцев с квартиронанимателями. В совокупности приблизительно 75% домовладельцев и 25% съемщиков. Предполагают, что для некоторого признака дисперсия составляет приблизительно 15 как для домовладельцев, так и для нанимателей. Стандартная ошибка разности средних для этих двух областей изучения не должна превышать 1. Какого объема должна быть выборка: (а) если число домовладельцев и нанимателей известно заранее, до извлечения выборки, (б) если оно не известно заранее. Для (б) достаточно получить приближенный ответ; для получения точного ответа нужны таблицы биномиального распределения.

2.13. Из совокупности объема N путем возвратного отбора получена простая случайная выборка объема 3. Покажите, что вероятности того, что выборка содержит 1, 2 и 3 различных единицы (например, aaa , aab , abc соответственно), равны:

$$P_1 = \frac{1}{N^3}; \quad P_2 = \frac{3(N-1)}{N^3}; \quad P_3 = \frac{(N-1)(N-2)}{N^3}.$$

В качестве оценки \bar{y} мы принимаем \bar{y}' — невзвешенное среднее по различным единицам выборки. Покажите, что среднее значение дисперсии \bar{y}' равно:

$$V(\bar{y}') = \frac{(2N-1)(N-1)S^2}{6N^2}.$$

Для доказательства этого утверждения достаточно, например, показать, что

$$V(\bar{y}') = S^2 \left(\frac{N-1}{N} P_1 + \frac{N-2}{2N} P_2 + \frac{N-3}{3N} P_3 \right).$$

Выведите отсюда, что $V(\bar{y}') < V(\bar{y})$, где \bar{y} — обычное среднее по n наблюдениям в выборке. Неравенство $V(\bar{y}') < V(\bar{y})$ для любого $n > 2$ было доказано Раджем и Хамисом (Raj and Khamsis, 1958).

2.14. Два зубных врача A и B проводят обследование состояния зубов

у 200 детей некоторой деревни. Д-р A отобрал простую случайную выборку, включающую 20 детей, и подсчитал число испорченных зубов у каждого ребенка, получив следующие результаты:

Число испорченных зубов на одного ребенка	0	1	2	3	4	5	6	7	8	9	10
Число детей	8	4	2	2	1	1	0	0	0	1	1

Д-р B , пользуясь теми же стоматологическими методами, осмотрел всех 200 детей, просто отмечая тех, у кого не было испорченных зубов. Число таких детей оказалось равным 60.

Оцените суммарное число испорченных зубов у детей этой деревни: (а) пользуясь только результатами A ; (б) пользуясь результатами A и B ; (в) будут ли эти оценки несмещенными? (г) От какой оценки вы ожидаете большей точности?

2.15. Компания собирается провести опрос, получив простую случайную выборку лиц, работающих на ее предприятиях более пяти лет. На проведение опроса ассигновано 1000 долл., стоимость опроса одного человека составляет 10 долл. Отдельного списка работающих более пяти лет не существует. Стоимость составления такого списка 200 долл. Компания может: (а) составить список работающих более пяти лет и опросить всех попавших в простую случайную выборку из этого списка, (б) получить простую случайную выборку всех работающих и опросить только работающих более пяти лет. Стойкостью неоправданного включения в выборку работающих менее пяти лет можно пренебречь.

Покажите, что при вычислении оценки суммарного значения некоторого признака для совокупности интересующих компанию работников план (а) обеспечивает меньшую дисперсию, чем план (б) только при $V_j < 2\sqrt{Q_j}$, где V_j — коэффициент вариации изучаемого признака среди интересующих компанию работников и Q_j — доля работающих менее пяти лет. Пикс не учитывайте.

ЛИТЕРАТУРА

- Cornfield J. (1944). On samples from finite populations. *Jour. Amer. Stat. Assoc.*, 39, 236—239.
- Erdős P. and Rényi A. (1959). On the central limit theorem for samples from a finite population. *Pub. Math. Inst. Hungarian Acad. Sci.*, 4, 49—57.
- Feller W. (1957). *An introduction to probability theory and its applications*. John Wiley and Sons, New York, second edition. Есть русский перевод: Феллер В. Введение в теорию вероятностей и ее приложения, т. 1, М., «Мир», 1964; т. 2, М., «Мир», 1967.
- Fisher R. A. (1932). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, fourth edition.
- Hájek J. (1960). Limiting distributions in simple random sampling from a finite population. *Pub. Math. Inst. Hungarian Acad. Sci.*, 5, 361—374.
- Hansen M. H., Hurwitz W. N. and Madow W. G. (1953). *Sample survey methods and theory*. John Wiley and Sons, New York, Vol. II.
- Harvard Computation Laboratory (1955). *Tables of the cumulative binomial probability distribution*, Harvard University Press, Cambridge, Mass.
- Jessen R. J. and Houseman E. E. (1944). *Statistical investigations of farm sample surveys taken in Iowa, Florida and California*. *Iowa Agr. Exp. Sta. Res. Bull.* 329.
- Madow W. G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Ann. Math. Stat.*, 19, 535—545.
- Raj Des and Khamsis S. H. (1958). Some remarks on sampling with replacement. *Ann. Math. Stat.*, 29, 550—557.
- Tukey J. W. (1950). Some sampling simplified. *Jour. Amer. Stat. Assoc.*, 45, 501—519.
- U. N. Statistical Office (1950). *The preparation of sample survey reports*. Stat. Papers series C no. 1.
- West Q. M. (1951). *The results of applying a simple random sampling process to farm management data*. Agricultural Experiment Station, Cornell University.
- Wishart J. (1952). Moment-coefficients of the k -statistics in samples from a finite population. *Biometrika*, 39, 1—13.

ОТБОР ДЛЯ ОЦЕНИВАНИЯ
ДОЛЕЙ И ПРОЦЕНТОВ

3.1. КАЧЕСТВЕННЫЕ ПРИЗНАКИ

Иногда мы хотим оценить общее число, долю или процент единиц совокупности, обладающих некоторым признаком или свойством или же относящихся к некоторому определенному классу единиц. В таком виде обычно публикуется большая часть результатов переписей и обследований, например число безработных, процент жителей — уроженцев данной местности и т. д. Такая классификация может быть введена непосредственно в опросный лист путем постановки вопросов, на которые нужно отвечать простыми «да» или «нет». В других случаях исходные результаты наблюдения имеют более или менее непрерывный характер и классификация производится при сводке и группировке данных обследования. Так, мы можем регистрировать возраст опрашиваемых с точностью до года, но опубликовать процент населения в возрасте 60 лет и старше.

Обозначения. Предположим, что все единицы в совокупности относятся к одному из двух классов C и C' . Обозначения имеют вид:

Число единиц класса C		Доля единиц класса C	
в совокупности	в выборке	в совокупности	в выборке
A	a	$P = A/N$	$p = a/n$

Выборочной оценкой P служит p , а выборочной оценкой A служит Np или Na/n . В статистической практике при рассмотрении оценок типа a и p часто применяется биномиальное распределение. Как мы увидим, в случае конечных совокупностей нужно применять гипергеометрическое распределение, хотя биномиальное распределение дает обычно удовлетворительное приближение.

3.2. ДИСПЕРСИИ ВЫБОРОЧНЫХ ОЦЕНОК

Для того чтобы применить в рассматриваемой ситуации теоремы, доказанные в гл. 2, можно воспользоваться следующим простым приемом. Для каждой единицы в выборке или совокупности положим y_i

равным 1, если единица относится к классу C , и 0, если она относится к C' . Для такой совокупности значений y_i очевидно, что

$$Y = \sum_{i=1}^N y_i = A; \quad (3.1)$$

$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N} = \frac{A}{N} = P. \quad (3.2)$$

Аналогично для выборки

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{a}{n} = p. \quad (3.3)$$

Следовательно, задачу оценивания A и P можно рассматривать как задачу оценивания суммарного и среднего значений для совокупности, в которой каждое y_i равно либо 1, либо 0. Для того чтобы применить теоремы гл. 2, выразим сначала S^2 и s^2 через P и p . Заметим, что

$$\sum_{i=1}^N y_i^2 = A = NP; \quad \sum_{i=1}^n y_i^2 = a = np.$$

Следовательно,

$$S^2 = \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1} = \frac{\sum_{i=1}^N y_i^2 - N\bar{Y}^2}{N-1} = \frac{1}{N-1} (NP - NP^2) = \frac{N}{N-1} PQ, \quad (3.4)$$

где $Q = 1 - P$. Аналогично

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{n}{n-1} pq. \quad (3.5)$$

При простом случайном отборе единиц, классифицированных указанным образом, применение теорем 2.1, 2.2 и 2.4 к такой совокупности дает следующие результаты.

Теорема 3.1. Выборочная доля $p = a/n$ есть несмещенная оценка доли для совокупности $P = A/N$.

Теорема 3.2. Дисперсия p равна:

$$V(p) = E(p - P)^2 = \frac{S^2}{n} \left(\frac{N-n}{N} \right) = \frac{PQ}{n} \left(\frac{N-n}{N-1} \right). \quad (3.6)$$

Последнее равенство получаем на основании (3.4).

Следствие 1. Если p и P есть соответственно для выборки и для совокупности процент единиц, относящихся к классу C , то для дисперсии p продолжает выполняться равенство (3.6).

Следствие 2. Дисперсия $A = Np$, оценки общего числа единиц класса C , равна:

$$V(\hat{A}) = \frac{N^2 PQ}{n} \left(\frac{N-n}{N-1} \right). \quad (3.7)$$

Теорема 3.3. Несмещенная оценка дисперсии p по данным выборки есть

$$v(p) = s_p^2 = \frac{N-n}{(n-1)N} pq. \quad (3.8)$$

Доказательство. В следствии из теоремы 2.4 было показано, что для непрерывной переменной y , несмещенной оценкой дисперсии выборочного среднего \bar{y} служит

$$v(\bar{y}) = \frac{s^2}{n} \frac{(N-n)}{N}. \quad (3.9)$$

Для оценивания долей роль \bar{y} играет p , и в (3.5) мы показали, что

$$s^2 = \frac{n}{n-1} pq. \quad (3.10)$$

Следовательно,

$$v(p) = s_p^2 = \frac{N-n}{(n-1)N} pq.$$

Отсюда вытекает, что если N очень велико по отношению к n , так что пкс можно пренебречь, то несмещенной оценкой дисперсии p будет

$$\frac{pq}{n-1}.$$

Некоторых читателей этот результат может смутить, потому что на практике для оценки дисперсии почти всегда применяется выражение pq/n . В действительности pq/n не будет несмещенной оценкой даже для бесконечной совокупности.

Следствие. Несмещенной оценкой дисперсии $\hat{A} = Np$, оценки общего числа единиц в совокупности, относящихся к классу C , служит

$$v(\hat{A}) = s_{Np}^2 = \frac{N(N-n)}{n-1} pq. \quad (3.11)$$

Пример. Простая случайная выборка объемом в 200 фамилий из некоторого списка 3042 фамилий и адресов показала при исследовании, что 38 адресов неверны. Оценим общее число адресов списка, нуждающихся в исправлении, и найдем стандартную ошибку этой оценки. Имеем

$$N = 3042; n = 200; a = 38; p = 0,19.$$

Оценка общего числа неправильных адресов равна:

$$\hat{A} = Np = 3042 \cdot 0,19 = 578,$$

$$s_{\hat{A}} = \sqrt{3042 \cdot 2842 \cdot 0,19 \cdot 0,81 / 199} = \sqrt{6685} = 81,8.$$

Поскольку доля отбора составляет менее 7%, пкс влияет незначительно. Для того чтобы исключить ее, подставим N вместо $N - n$. Если, кроме того, мы подставим n вместо $n - 1$, то получим более простую формулу

$$s_{Np} = N \sqrt{pq/n} = 3042 \sqrt{0,19 \cdot 0,81 / 200} = 84,4.$$

Это значение довольно хорошо согласуется с предыдущим результатом 81,8.

Полученные для дисперсии и для оценки дисперсии p формулы справедливы только для случая, когда по классам C и C' распределяются сами единицы, так что p представляет собой отношение числа единиц класса C в выборке к общему числу единиц в выборке. Во многих обследованиях каждая единица отбора образована некоторой группой элементов, и классифицируются именно эти элементы. Приведем несколько примеров.

Единица отбора	Элементы
Семья	Члены семьи
Ресторан	Работники ресторана
Корзина с яйцами	Отдельные яйца
Персиковое дерево	Отдельные персики

Для случая, когда извлекается простая случайная выборка единиц, чтобы оценить долю P элементов совокупности, принадлежащих классу C , предыдущие формулы не применимы. Соответствующие методы рассматриваются в параграфе 3.12.

3.3. ВЛИЯНИЕ P НА СТАНДАРТНЫЕ ОШИБКИ

Формула (3.6) показывает, как меняется дисперсия оценки процента с изменением P при неизменных n и N . Если пренебречь пкс, то имеем

$$V(p) = \frac{PQ}{n}.$$

Значения функции PQ и квадратного корня из нее приведены в табл. 3.1. Эти функции можно рассматривать соответственно как дисперсию и среднее квадратичное отклонение для выборки объема 1.

Таблица 3.1

ЗНАЧЕНИЯ PQ И \sqrt{PQ}

P — процент единиц совокупности, относящихся к классу C

P	0	10	20	30	40	50	60	70	80	90	100
PQ	0	900	1 600	2 100	2 400	2 500	2 400	2 100	1 600	900	0
\sqrt{PQ}	0	30	40	46	49	50	49	46	40	30	0

Обе функции принимают наибольшее значение, когда совокупность поровну разделена на два класса, и симметричны относительно такого разбиения. Стандартная ошибка p меняется сравнительно мало, если P заключено между 30 и 70%. При максимальном значении \sqrt{PQ} , равном 50, для того чтобы уменьшить стандартную ошибку выборки до 5%, нужна выборка объема 100. Для того чтобы получить стандартную ошибку 1%, нужна выборка объема 2500.

Если мы заинтересованы в том, чтобы оценить общее число единиц совокупности, относящихся к классу C , то такой подход непригоден. В этом случае более естественно задаться вопросом: будет ли оценка отклоняться от истинного числа единиц, скажем, не более чем на 7%? Поэтому лучше рассматривать стандартную ошибку, выразив ее как долю или процент истинного значения, NP . Такая доля равна:

$$\frac{\sigma_{NP}}{NP} = \frac{N\sqrt{PQ}}{\sqrt{n}NP} \sqrt{\frac{N-n}{N-1}} = \frac{1}{\sqrt{n}} \sqrt{\frac{Q}{P}} \sqrt{\frac{N-n}{N-1}}. \quad (3.12)$$

Эту величину называют обычно коэффициентом вариации оценки. Если пренебречь п.к.с, то этот коэффициент будет равен $\sqrt{Q/P}$. Значения отношения $\sqrt{Q/P}$, которые можно рассматривать как значения коэффициента вариации для выборки объема 1, приведены в табл. 3.2.

Таблица 3.2

значения $\sqrt{Q/P}$ для разных значений P
 P — процент единиц совокупности, относящихся к классу C

P	0	0.1	0.5	1	5	10	20
$\sqrt{Q/P}$	∞	31.6	14.1	9.9	4.4	3.0	2.0
P	30	40	50	60	70	80	90
$\sqrt{Q/P}$	1.5	1.2	1.0	0.8	0.7	0.5	0.3

При неизменном объеме выборки коэффициент вариации оценки общего числа единиц, относящихся к классу C , монотонно убывает по мере того, как истинный процент единиц класса C увеличивается. Когда P составляет менее 5%, коэффициент этот довольно велик. Для того чтобы точно оценить число единиц совокупности, обладающих каким-либо редким свойством, нужны очень большие выборки. При $P = 1\%$, чтобы уменьшить коэффициент вариации оценки до 0.1, или 10%, мы должны иметь $\sqrt{n} = 99$. Это требует объема выборки, равного 9801. Простой случайный отбор или какой-либо другой способ отбора общего характера будет слишком дорогостоящим методом оценки общего числа единиц, обладающих каким-либо редким признаком.

3.4. БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Поскольку в рассматриваемом случае совокупность имеет очень простой характер, все y_i равны либо 1, либо 0, мы можем найти не только среднее значение и дисперсию оценки p , но и фактическое распределение частот этой оценки.

Совокупность содержит A единиц класса C и $N - A$ единиц класса C' , причем $P = A/N$. Если первая извлеченная единица относится к C , то в совокупности останется $A - 1$ единиц класса C и $N - A$ единиц класса C' . Таким образом, доля единиц класса C после первого извлечения несколько изменится и станет равной $(A - 1)/(N - 1)$. Если же первая извлеченная единица относится к C' , то доля единиц класса C станет равной $A/(N - 1)$. При отборе без возвращения эта доля меняется сходным образом в течение всей последовательности извлечений. В данном параграфе такие изменения не учитываются, т. е. P считается постоянным. Это эквивалентно предположению о том, что как A , так и $N - A$ велики по сравнению с объемом выборки n .

При таком предположении процесс извлечения выборки состоит из последовательности n испытаний, в каждом из которых вероятность того, что извлеченная единица относится к классу C , равна P . Отсюда вытекает, что число единиц класса C в выборке подчиняется знакомому нам биномиальному распределению. Вероятность того, что выборка содержит a единиц класса C , равна

$$\Pr(a) = \frac{n!}{a!(n-a)!} P^a Q^{n-a}. \quad (3.13)$$

По этой формуле можно составить таблицы распределения частот a , или $p = a/n$, или оценки общего числа единиц, Np .

Существует три подходящих сборника таблиц*. Во всех таблицах P приведено через интервал 0.01. Таблицы составлены для следующих значений n .

Таблицы Бюро Стандартов США (U.S. Bureau of Standards, 1950):
 $n = 1$ (1) 49, т. е. пробегает значения от 1 до 49 через интервал 1.

Таблицы Ромига (Romig, 1952):

$n = 50$ (5) 100.

Таблицы Гарвардской вычислительной лаборатории (Harvard Computation Laboratory, 1955):

$n = 1(1)50(2)100(10)200(20)500(50)1000$.

* Таблицы биномиального распределения имеются, в частности, в следующих изданиях на русском языке: Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М., ВЦ АН СССР, 1968, с. 346, 347.

$n = 5(5) 30$.

$P = 0.01; 0.02 (0.02) 0.10 (0.10) 0.50$.

Мостеллер Ф., Рурке Р., Томас Дж. Вероятность. М., «Мир», 1969.

$n = 2(1) 25$.

$P = 0.01; 0.05; 0.10 (0.10) 0.90; 0.95; 0.99$ — Примеч. ред.

3.5. ГИПЕРГЕОМЕТРИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ

Распределение p можно найти и без предположения о том, что совокупность велика по сравнению с выборкой. Пусть числа единиц в совокупности, относящихся к классам C и C' , составляют соответственно A и A' . Вычислим вероятность того, что соответствующие числа единиц в выборке составят a и a' , где

$$a + a' = n; A + A' = N.$$

При простом случайном отборе каждый из C_N^a различных наборов по a единиц из N имеет одинаковую вероятность быть отобранным. Для того чтобы найти искомую вероятность, подсчитаем, сколько таких выборок содержит ровно a единиц класса C и a' единиц класса C' . Число различных наборов a единиц из A , принадлежащих классу C , равно C_A^a , а число различных наборов a' единиц из A' равно $C_{A'}^{a'}$. Каждый набор первого типа может сочетаться с любым набором второго типа, образуя различные выборки нужного типа. Общее число таких выборок равно $C_A^a \cdot C_{A'}^{a'}$.

Следовательно, если извлекается простая случайная выборка объема n , то вероятность того, что она окажется выборкой нужного типа, равна:

$$\Pr(a, a' | A, A') = C_A^a C_{A'}^{a'} / C_N^n. \quad (3.14)$$

Мы получили распределение частот a или np , из которого сразу следует распределение p . Такое распределение называется *гипергеометрическим*. Для вычислительных целей гипергеометрические вероятности (3.14) можно записать следующим образом:

$$\frac{n!}{a!(n-a)!} \cdot \frac{A(A-1)\dots(A-a+1)(A')(A'-1)\dots(A'-a'+1)}{N(N-1)\dots(N-n+1)}. \quad (3.14')$$

Пример. Семья из восьми человек состоит из трех мужчин и пяти женщин. Найдем распределение частот числа мужчин в простой случайной выборке объема 4. Для нашего случая

$$A = 3; A' = 5; N = 8; n = 4.$$

По формуле (3.14') распределение числа мужчин a имеет вид

a	Вероятность
0	$\frac{4!}{0!4!} \cdot \frac{5 \cdot 4 \cdot 3 \cdot 2}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{1}{14}$
1	$\frac{4!}{1!3!} \cdot \frac{3 \cdot 5 \cdot 4 \cdot 3}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{6}{14}$
2	$\frac{4!}{2!2!} \cdot \frac{3 \cdot 2 \cdot 5 \cdot 4}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{6}{14}$
3	$\frac{4!}{3!1!} \cdot \frac{3 \cdot 2 \cdot 1 \cdot 5}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{1}{14}$
4	$= 0$ (частота невозможна)

Читатель может проверить, что среднее число мужчин составляет $3/2$, а дисперсия есть $15/28$. Эти результаты согласуются с формулами, установленными ранее в параграфе 3.2, по которым

$$E(np) = nP = \frac{nA}{N} = \frac{4 \cdot 3}{8} = \frac{3}{2};$$

$$V(np) = nPQ \frac{N-n}{N-1} = 4 \cdot \frac{3}{8} \cdot \frac{5}{8} \cdot \frac{4}{7} = \frac{15}{28}.$$

3.6. ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ

Сначала мы выясним смысл доверительных границ в случае качественных признаков. Пусть a из n единиц выборки относятся к классу C . Предположим, что нужно сделать заключение о числе A единиц совокупности, относящихся к классу C . В качестве верхней доверительной границы для A возьмем такое значение \hat{A}_U , при котором вероятность получения в выборке числа единиц класса C , меньшего или равного a , принимала бы некоторое малое значение α_U , например 0,025. Формально \hat{A}_U удовлетворяет уравнению

$$\sum_{j=0}^a \Pr(j, n-j | \hat{A}_U, N - \hat{A}_U) = \alpha_U, \quad (3.15)$$

где \Pr — значение вероятности для гипергеометрического распределения, определенное формулой (3.14).

Если α_U выбрано заранее, то в общем случае (3.15) может иметь решением не целое \hat{A}_U , в то время как по смыслу \hat{A}_U должно быть целым. На практике мы выбираем в качестве \hat{A}_U наименьшее целое число A , такое, что левая часть (3.15) меньше или равна α_U . Аналогично нижней доверительной границей \hat{A}_L служит наибольшее целое число, такое, что

$$\sum_{j=a}^n \Pr(j, n-j | \hat{A}_L, N - \hat{A}_L) \leq \alpha_L. \quad (3.16)$$

Теперь мы находим доверительные границы для P , полагая $\hat{P}_U = \hat{A}_U/N$, $\hat{P}_L = \hat{A}_L/N$.

Для вычисления доверительных границ существуют многочисленные способы.

Точные методы

Чанг и Делюри (Chung and De Lury, 1950) составили номограммы 90, 95 и 99%-ных доверительных границ для P при $N = 500, 2500$ и 10 000. Значения для промежуточных объемов совокупностей можно получить путем интерполяции. Либерман и Оуэн (Lieberman and

Owen, 1961) приводят таблицы для отдельных членов и для суммы членов гипергеометрического распределения при N , не превосходящих 100.

Нормальная аппроксимация

Из формулы (3.8) для дисперсии p следует, что одним из видов нормальной аппроксимации доверительных границ для P будет

$$p \pm \left[t \sqrt{1-f} \sqrt{pq/(n-1)} + \frac{1}{2n} \right], \quad (3.17)$$

где $f = n/N$ и t — квантиль нормального распределения, отвечающий заданной доверительной вероятности. Для тех, кто привык пользоваться более простым выражением $\sqrt{pq/n}$, заметим, что обе формулы дают мало отличающиеся значения. Последний член в правой части выражения представляет собой поправку на непрерывность. Он лишь незначительно улучшает аппроксимацию. Однако без этой поправки нормальная аппроксимация обычно дает слишком узкий доверительный интервал.

Ошибка нормальной аппроксимации зависит от всех величин: n , p , N , α_U и α_L . Наиболее чувствительна она к величине np , или более определенно, к числу наблюдений в меньшем классе. Табл. 3.3. дает некоторые практические ориентиры для решения вопроса о том, когда можно применять нормальную аппроксимацию (3.17).

Таблица 3.3
НАИМЕНЬШИЕ ЗНАЧЕНИЯ np , ПРИ КОТОРЫХ МОЖНО ПРИМЕНЯТЬ
НОРМАЛЬНУЮ АППРОКСИМАЦИЮ

p	Число наблюдений в меньшем классе np	Объем выборки n
0,5	15	30
0,4	20	50
0,3	24	80
0,2	40	200
0,1	60	600
0,05	70	1400
~0*	80	сб

* Это означает, что p очень мало, так что np подчиняется распределению Пуассона.

Значения в табл. 3.3 даны таким образом, что для 95%-ных доверительных границ истинная частота, с которой P не попадает в эти границы, не превышает 5,5%. Кроме того, вероятность того, что верхняя граница меньше P , заключена между 2,5 и 3,5%, а вероятность того, что нижняя граница превышает P , — между 2,5 и 1,5%.

Пример 1. В простой случайной выборке объема 100, извлеченной из совокупности объема 500, содержится 37 единиц класса C . Найдем 95%-ные доверительные границы для доли и общего числа единиц класса C в совокупности. В нашем примере

$$n = 100; N = 500; p = 0,37.$$

Пример относится к тому случаю, когда рекомендуется нормальная аппроксимация. Оценка стандартной ошибки p равна:

$$\sqrt{(1-f)pq/(n-1)} = \sqrt{0,8 \cdot 0,37 \cdot 0,63/99} = 0,0434.$$

Поправка на непрерывность $1/2 n$ равна 0,005. Следовательно, в качестве 95%-ных доверительных границ для P принимаем

$$0,37 \pm (1,96 \cdot 0,0434 + 0,005) = 0,37 \pm 0,090;$$

$$\hat{P}_L = 0,280; \hat{P}_U = 0,460.$$

Значения, полученные из таблиц Чанга и Делюри, составляют соответственно 0,285 и 0,462.

Для нахождения границ для общего числа единиц класса C в совокупности умножаем полученные значения на N и получаем соответственно 140 и 230.

Биномиальная аппроксимация

В том случае, когда нельзя применить нормальную аппроксимацию, доверительные границы для P можно найти по биномиальным таблицам (параграф 3.4), приняв во внимание, если это необходимо, пкс. В табл. VIII,1 из *Статистических таблиц* Фишера и Йейтса (Fisher and Yates, 1957) в отличие от обычных биномиальных таблиц указаны биномиальные доверительные границы для P при любом значении n . В примере 2 показано, как вычисляются границы в случае биномиальной аппроксимации.

Пример 2. Для другого признака в выборке из примера 1 девять из 100 единиц выборки принадлежат классу C . По таблицам Ромига при $n = 100$ 95%-ные доверительные границы для P оказываются 0,041 и 0,165. (Таблицы Фишера и Йейтса дают значения 0,042 и 0,164). Если доля отбора, f , составляет менее 5%, то границы, найденные таким образом, для большинства целей достаточно узки. В нашем примере $f = 0,2$ и необходима поправка.

Для того чтобы внести эту поправку, мы уменьшаем интервал между p и каждой границей в $\sqrt{1-f}$ раз; $\sqrt{1-f} = \sqrt{0,8} = 0,894$. После поправки границы равны:

$$\hat{P}_L = 0,090 - 0,894(0,090 - 0,041) = 0,046;$$

$$\hat{P}_U = 0,090 + 0,894(0,165 - 0,090) = 0,157.$$

Границы, полученные по таблицам Чанга и Делюри, равны соответственно 0,045 и 0,157.

Пример 3. При контроле долгоиграющих пластинок, когда допустим весьма низкий уровень брака, наибольший интерес представляет верхняя доверительная граница для A . Предположим, что проверяется 200 пластинок и партия в 1000 изделий принимается, если брака не обнаружено. Существуют специальные таблицы значений верхней доверительной границы для числа бракованных изделий в партии. Хоро-

шее приближение дает следующее соотношение. Для гипергеометрического распределения вероятность того, что среди n изделий нет бракованных, если среди N изделий A бракованных, равна:

$$\frac{(N-A)(N-A-1)\dots(N-A-n+1)}{N(N-1)\dots(N-n+1)} = \left(\frac{N-A-u}{N-u}\right)^n,$$

где $u = (n-1)/2$. Например, при $n = 200$, $A = 10$, $N = 1000$ это приближение дает $(890,5/900,5)^{200}$, откуда по таблицам логарифмов получаем значение 0,107. Таким образом, 90%-ной верхней доверительной границей для числа бракованных изделий в партии служит приближенно $A = 10$ (1%-ный уровень брака).

3.7. КЛАССИФИКАЦИЯ ПО НЕСКОЛЬКИМ ПРИЗНАКАМ

Часто при изложении результатов единицы распределяются более чем по двум классам. Так выборка из совокупности людей может быть распределена по 15 пятилетним возрастным группам. Даже если предполагается, что на вопрос отвечают простыми «да» или «нет», результаты могут распасться на четыре класса: «да», «нет», «не знаю» и «ответа нет». Мы проиллюстрируем обобщение теории на такие случаи, рассмотрев ситуацию, когда имеется три класса.

Предположим, что число единиц i -го класса равно A_i для совокупности и a_i для выборки, причем

$$N = \sum A_i; \quad n = \sum a_i; \quad P_i = \frac{A_i}{N}; \quad p_i = \frac{a_i}{n}.$$

Если объем выборки n мал по сравнению со всеми A_i , то вероятности P_i можно считать практически постоянными для всей последовательности извлечений. Вероятность получения выборки с заданными a_i задается полиномиальным выражением

$$\text{Pr}(a_i) = \frac{n!}{a_1! a_2! a_3!} P_1^{a_1} P_2^{a_2} P_3^{a_3}. \quad (3.18)$$

Оно представляет собой обобщение биномиального распределения и служит хорошим приближением, если доля отбора мала.

Правильное выражение для вероятности получения выборки с заданными a_i будет

$$\text{Pr}(a_i | A_i) = C_{A_1}^{a_1} C_{A_2}^{a_2} C_{A_3}^{a_3} / C_N^n. \quad (3.19)$$

Это выражение представляет собой естественное обобщение формулы (3.14) из параграфа 3.5 для гипергеометрического распределения. В числителе стоит число различных выборок объема n , содержащих a_1 единиц класса 1, a_2 — класса 2 и a_3 — класса 3.

3.8. ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ ПРИ КЛАССИФИКАЦИИ ПО НЕСКОЛЬКИМ ПРИЗНАКАМ

Необходимо различать два разных случая.

Случай 1. Мы вычисляем

$$p = \frac{\text{число единиц какого-либо одного класса в выборке}}{n} = \frac{a_1}{n}$$

или

$$p = \frac{\text{общее число единиц группы классов}}{n} = \frac{a_1 + a_2 + a_3}{n}.$$

В любой из этих ситуаций, хотя исходная классификация содержит более чем два класса, само p получается при разделении n единиц только на два класса. К этому случаю применима ранее изложенная теория. Доверительные границы вычисляются так, как описано в параграфе 3.6.

Случай 2. Иногда некоторые классы исключаются и p вычисляется после разбиения на две части единиц из оставшихся классов. Например, мы можем исключить всех лиц, ответивших «не знаю» или не давших ответа, и рассмотреть отношение числа ответов «да» к общему числу ответов «да» и «нет». В выборочных обследованиях часто представляют интерес отношения, структура которых имеет такой характер. В знаменателе этого отношения стоит не n , а некоторое меньшее число n' .

Хотя n' меняется от выборки к выборке, можно воспользоваться полученными ранее результатами, рассматривая условное распределение p для тех выборок, в которых неизменны как n , так и n' . Такой прием уже применялся в параграфе 2.10. Положим

$$p = \frac{a_1}{a_1 + a_2}; \quad n' = a_1 + a_2; \quad n = a_1 + a_2 + a_3,$$

так что a_3 представляет собой число единиц выборки, относящихся к классу, не интересующему нас в данный момент. Тогда, как показано в следующем параграфе, условным распределением a_1 и a_2 будет гипергеометрическое распределение для случая, когда объем выборки равен n' и объем совокупности $N' = A_1 + A_2$. Следовательно, согласно (3.17) нормальная аппроксимация условных доверительных границ для $P = A_1/(A_1 + A_2)$ дает значения

$$p \pm \left[t \sqrt{\left(1 - \frac{n'}{N'}\right) \frac{pq}{(n'-1)} + \frac{1}{2n'}} \right]. \quad (3.20)$$

Если значение N' неизвестно, то в член пкс в (3.20) вместо n'/N' можно подставить значение n/N .

3.9. УСЛОВНОЕ РАСПРЕДЕЛЕНИЕ p

Для того чтобы найти это распределение, рассмотрим только те выборки объема n , у которых $n' = a_1 + a_2$ единиц относятся к классам 1 и 2. Число различных выборок такого вида равно:

$$C_{N'}^{n'} C_{N-N'}^{n-n'} = C_{A_1+A_2}^{a_1+a_2} C_{A_3}^{a_3}. \quad (3.21)$$

Среди таких выборок число выборок, содержащих a_1 единиц класса 1 и a_2 единиц класса 2, уже было приведено в числителе формулы (3.19)

из параграфа 3.7. Деля этот числитель на (3.21), получаем

$$P(a_1 | A_1, A_2, n, n') = C_{A_1}^{a_1} C_{A_2}^{n-a_1} / C_{A_1+A_2}^{n'} \quad (3.22)$$

Это обычное гипергеометрическое распределение для выборки объема n' из совокупности объема $N' = A_1 + A_2$.

Пример. Рассмотрим совокупность, состоящую из пяти единиц b, c, d, e, f , относящихся к трем классам.

Класс	A_i	Обозначения единиц
1	1	b
2	2	c, d
3	2	e, f

По простой случайной выборке объема 3 мы хотим оценить $P = A_1 / (A_1 + A_2)$, равное в данном случае $1/3$. Имеем $N = 5$ и $N' = 3$.

Существует 10 возможных выборок объема 3 и все они имеют равные исходные вероятности, соответствующие $1/3$ и $2/3$, согласуются с общей формулой (3.22). Далее они рассматриваются в соответствии со значениями n' .

$n' = 1$					
Выборка	a_1	a_2	p	Условная вероятность	$(p - P)$
bef	1	0	1	1/3	2/3
cef или def	0	1	0	2/3	-1/3

Если выборки различаются только значениями a_1, a_2 , то можно получить только два вида выборок: $a_1 = 1, a_2 = 0$; $a_1 = 0, a_2 = 1$. Их условные вероятности, соответственно $1/3$ и $2/3$, согласуются с общей формулой (3.22). Далее,

$$E(p) = \frac{1}{3};$$

$$\sigma_p^2 = \frac{1}{3} \cdot \frac{4}{9} + \frac{2}{3} \cdot \frac{1}{9} = \frac{6}{27} = \frac{2}{9}.$$

Оценка p несмещенная, и ее дисперсия согласуется с общей формулой

$$\sigma_p^2 = \left(\frac{N' - n'}{N' - 1} \right) \frac{PQ}{n'} = \frac{3-1}{3-1} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}.$$

При $n' = 2$ существует шесть возможных выборок, которые дают только два набора значений a_1, a_2 .

$n' = 2$					
Выборка	a_1	a_2	p	Условная вероятность	$(p - P)$
bce, bcf, bde или bdf	1	1	1/2	2/3	1/6
cde или cdf	0	2	0	1/3	-1/3

Оценка опять несмещенная и ее дисперсия равна:

$$\sigma_p^2 = \frac{2}{3} \cdot \frac{1}{36} + \frac{1}{3} \cdot \frac{1}{9} = \frac{1}{18}.$$

что можно получить и по общей формуле. Заметим, что значение дисперсии составляет только одну четвертую ее значения при $n' = 1$. В случае применения условного распределения дисперсия меняется в зависимости от вида полученной выборки.

При $n' = 3$ существует только одна возможная выборка bcd . Она дает правильную долю для совокупности $1/3$. Условная дисперсия p равна нулю, как показывает и общая формула, которая обращается в нуль при $N' = n'$.

3.10. ДОЛИ И СУММАРНЫЕ ЗНАЧЕНИЯ ДЛЯ ПОДСОВОКУПНОСТЕЙ

Если нужно получить отдельные оценки для каждой из некоторого числа подсовокупностей или областей изучения, к которым относятся единицы выборки, то можно применить результаты параграфов 3.8. и 3.9. Данные выборки можно записать следующим образом:

Класс	Область 1		Область 2		...		Область k		Всего
	C	C'	C	C'	...		C	C'	
Число единиц	a_1	a'_1	a_2	a'_2	...		a_k	a'_k	n

Из n единиц ($a_1 + a'_1$) относятся к области 1 и среди них a_1 относятся к классу C. Доля единиц класса C в области 1 оценивается с помощью $p_1 = a_1 / (a_1 + a'_1)$. Распределение частот и доверительные границы для p_1 рассматривались в параграфах 3.8 и 3.9 (случай 2).

Оценить общее число A_1 единиц класса C в области 1 можно двойным образом. Если N_1 , общее число единиц совокупности, принадлежащих области 1, известно, то можно применить условную оценку

$$\hat{A}_1 = N_1 p_1 = \frac{N_1 a_1}{a_1 + a'_1}.$$

Ее стандартная ошибка вычисляется по формуле

$$s(\hat{A}_1) = N_1 \sqrt{1 - (n_1/N_1)} \sqrt{p_1 q_1 / (n_1 - 1)},$$

где $n_1 = a_1 + a'_1$.

Если N_1 неизвестно, то оценкой служит

$$\hat{A}_1 = \frac{Na_1}{n}$$

с оценкой стандартной ошибки

$$s(\hat{A}_1) = N \sqrt{1 - (n/N)} \sqrt{pq/(n-1)},$$

где $p = a_1/n$.

3.11. СРАВНЕНИЕ МЕЖДУ РАЗЛИЧНЫМИ ОБЛАСТЯМИ

Поскольку доли в различных областях оцениваются независимо, сравнение таких долей проводится обычными элементарными методами. Например, чтобы проверить, будет ли доля $p_1 = a_1/(a_1 + a'_1)$ значительно отличаться от доли $p_2 = a_2/(a_2 + a'_2)$, мы образуем обычную таблицу 2×2 .

Класс	Область	
	1	2
C	a_1	a_2
C'	a'_1	a'_2
Итого	n_1	n'_1

К распределению $(p_1 - p_2)$ можно применять обычный критерий χ^2 (Fisher, 1958) или нормальную аппроксимацию. Аналогично сравнение долей для более чем двух областей проводится с помощью таблиц сопряженности $2 \times k$.

Иногда желательно оценить, будет ли a_1 значительно отличаться от a_2 ; например, превышает ли число республиканцев, одобряющих некоторое предложение, число одобряющих его демократов. Согласно нулевой гипотезе, что эти два числа в совокупности совпадают, общее число единиц в рассматриваемых группах $n' = a_1 + a_2$ должно распределяться с равными вероятностями между двумя этими группами. Следовательно, мы можем рассматривать a_1 как число успехов в биномиальной схеме с n' испытаниями с вероятностью успеха $1/2$ согласно нулевой гипотезе. Можно проверить, что квантиль нормального распределения (с поправкой на непрерывность) имеет вид

$$\frac{2 \left(\left| a_1 - \frac{1}{2} n' \right| - \frac{1}{2} \right)}{\sqrt{n'}}$$

3.12. ОЦЕНИВАНИЕ ДОЛЕЙ ПРИ ГНЕЗДОВОМ ОТБОРЕ

Как уже упоминалось в параграфе 3.2, изложенные методы не применимы, если каждая единица состоит из гнезда элементов и мы хотим оценить долю элементов, относящихся к классу C.

Пусть каждая единица содержит одно и то же число m элементов и $p_i = a_i/m$ есть доля элементов i -й единицы, относящихся к классу C. Доля единиц выборки, относящихся к классу C, равна:

$$p = \frac{\sum a_i}{nm} = \frac{1}{n} \sum p_i.$$

т. е. оценка p представляет собой невзвешенное среднее величин p_i . Следовательно, истинную дисперсию p и ее оценку можно получить, применяя непосредственно формулы гл. 2, если в них вместо y_i подставить p_i .

$$V(p) = \frac{1-f}{n} \frac{\sum (p_i - p)^2}{N-1}. \quad (3.23)$$

Несмещенной выборочной оценкой этой дисперсии будет

$$v(p) = \frac{1-f}{n} \frac{\sum (p_i - p)^2}{n-1}. \quad (3.24)$$

Пример 1. Группа из 61 больного проказой получала некоторый препарат в течение 48 недель. Для того чтобы определить действие препарата на бациллы проказы, бактериологически определялось наличие бацилл на шести участках тела каждого больного. Из 366 участков на 153, или на 41,8%, результат был отрицательным. Какова стандартная ошибка этого процента?

Хотя это пример скорее контролируемого эксперимента, чем обследования, он показывает, насколько ошибочной может быть биномиальная схема. Для биномиальной схемы $n = 366$ и стандартная ошибка $(p) = \sqrt{pq/(n-1)} = \sqrt{41,8 \cdot 58,2/365} = 2,58\%$.

В данном случае каждый пациент представляет собой гнездовую единицу с $m = 6$ элементами (участками тела). Для того чтобы найти стандартную ошибку по правильной формуле, нужно знать распределение частот 61 значения p_i . Удобнее вести вычисления по распределению y_i , числа участков тела с отрицательным результатом на одного пациента. Если p_i выразить в процентах, то $p_i = 100y_i/6$. Согласно распределению из табл. 3.4 находим $\sum y_i^2 = 669$ и

$$\begin{aligned} \text{стандартная ошибка } (\bar{y}) &= \sqrt{\frac{\sum y_i (y_i - \bar{y})^2}{n(n-1)}} = \\ &= \sqrt{\frac{669 - (153)^2/61}{61 \cdot 60}} = 0,279. \end{aligned}$$

Следовательно,

$$\text{стандартная ошибка } (p) = \frac{100}{6} \text{ стандартная ошибка } (\bar{y}) = 4,65\%.$$

Это значение приблизительно в 1,8 раза больше величины, полученной по биномиальной формуле. Биномиальная схема требует предположения о том, что результаты на различных участках тела одного и того же пациента независимы, хотя в действительности они имеют сильную положительную корреляцию. В последней строке табл. 3.4 указаны ожидаемые значения числа пациентов с 0, 1, 2, ... участками тела, давшими отрицательный результат, подсчитанные по разложению би-

нома $(0,58 + 0,42)^6$. Обратите внимание на заметный избыток наблюдаемых частот пациентов, f , без отрицательных результатов и с пятью и шестью отрицательными результатами.

Таблица 3.4
Число участков с отрицательным результатом,
приходящихся на одного пациента

$n_i = 6p_i/100$	0	1	2	3	4	5	6	Итого
f	17	11	4	4	7	14	4	61
f/n_i	0	11	8	12	28	70	24	153
f_{exp}	2,3	10,1	18,3	17,6	9,6	2,8	0,3	61,0

Для случая, когда объем гнезда — переменная величина, обозначим через m_i число элементов в i -й гнездовой единице и положим $p_i = a_i/m_i$. Доля единиц выборки, принадлежащих классу C , равна

$$p = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i}.$$

По своей структуре это типичная оценка-отношение, или оценка по отношению, которая была рассмотрена в параграфе 2.9 и будет рассматриваться далее в гл. 6. Она несколько смещена, хотя это смещение, по-видимому, редко имеет практическое значение.

Если мы подставим в (2.29) a_i вместо y_i и m_i вместо x_i , то получим приближенное выражение для дисперсии p :

$$V(p) \approx \frac{1-f}{n\bar{M}^2} \frac{\sum_{i=1}^N (a_i - Pm_i)^2}{N-1},$$

где P — доля элементов класса C в совокупности и $\bar{M} = \sum_{i=1}^N m_i/N$ — среднее число элементов на одно гнездо. Это выражение можно записать и в другой форме:

$$V(p) \approx \frac{1-f}{n} \sum_{i=1}^N \left(\frac{m_i}{\bar{M}} \right)^2 \frac{(p_i - P)^2}{N-1}. \quad (3.25)$$

Отсюда видно, что приближенное значение дисперсии содержит взвешенную сумму квадратов отклонений p_i от значения для совокупности, P .

В качестве оценки дисперсии имеем

$$v(p) = \frac{1-f}{n\bar{m}^2} \frac{\sum a_i^2 - 2p \sum a_i m_i + p^2 \sum m_i^2}{n-1}, \quad (3.26)$$

где $\bar{m} = \sum m_i/n$ есть среднее число элементов на одно гнездо в выборке.

Пример 2. Простая случайная выборка объемом в 30 домохозяйств была извлечена из материалов переписи, проведенной в 1947 г. в 6 и 7 кварталах Восточного округа здравоохранения г. Балтимора. Совокупность содержит приблизительно 15 000 домохозяйств. В табл. 3.5 члены каждого домохозяйства классифицированы (а) согласно тому, посетили ли они врача в течение предшествующих 12 месяцев, (б) по полу.

Таблица 3.5
ДАННЫЕ ПРОСТОЙ СЛУЧАЙНОЙ ВЫБОРКИ ОБЪЕМОМ В 30 ДОМОХОЗЯЙСТВ

Домохозяйства	Число членов домохозяйства m_i	Число		Посетили ли врача в предшествующем году	
		мужчин a_i	женщин	да a_i	нет
1	5	1	4	5	0
2	6	3	3	0	6
3	3	1	2	2	1
4	3	1	2	3	0
5	2	1	1	0	2
6	3	1	2	0	3
7	3	1	2	0	3
8	3	1	2	0	3
9	4	2	2	0	4
10	4	3	1	0	4
11	3	2	1	0	3
12	2	1	1	0	2
13	7	3	4	0	7
14	4	3	1	4	0
15	3	2	1	1	2
16	5	3	2	2	3
17	4	3	1	0	4
18	4	3	1	0	4
19	3	2	1	1	2
20	3	1	2	3	0
21	4	1	3	2	2
22	3	2	1	0	3
23	3	2	1	0	3
24	1	0	1	0	1
25	2	1	1	2	0
26	4	3	1	2	2
27	3	1	2	0	3
28	4	2	2	2	2
29	2	1	1	0	2
30	4	2	2	1	3
Итого	104	53	51	30	74

Мы хотим сопоставить формулу дисперсии для оценки по отношению с неприемлемой здесь биномиальной формулой. Рассмотрим сначала долю людей, посетивших врача. Для биномиальной формулы мы приняли бы

$$n = 104; \quad p = \frac{30}{104} = 0,2885.$$

Следовательно (bin — от английского «binomial» — биномиальный),

$$v_{bin}(p) = \frac{pq}{n} = \frac{0,2885 \cdot 0,7115}{104} = 0,00197.$$

В случае формулы для отношения мы замечаем, что имеется 30 гнезд, и полагаем

$$\begin{aligned} n &= 30; \\ m_i &\text{— общее число лиц в } i\text{-м домохозяйстве;} \\ a_i &\text{— число лиц в } i\text{-м домохозяйстве, посетивших врача;} \\ p &= 0,2885, \text{ как и ранее;} \\ \bar{m} &= \frac{104}{30} = 3,4667; \end{aligned}$$

$$\Sigma a_i^2 = 86; \quad \Sigma m_i^2 = 404; \quad \Sigma a_i m_i = 113.$$

Пкс можно не учитывать. Следовательно, согласно (3.26)

$$v(p) = \frac{86 - 2 \cdot 0,2885 \cdot 113 + 0,2885^2 \cdot 404}{30 \cdot 29 \cdot 3,4667^2} = 0,00520.$$

Дисперсия, полученная по методу отношения (0,00520), значительно больше той, которую дает биномиальная формула (0,00197). По различным причинам семьи отличаются по частоте, с которой члены семьи обращаются к врачу. Для выборки в целом доля лиц, посетивших врача, немногим больше одной четверти, хотя есть несколько семей, где врача посетили все члены семьи. Аналогичные результаты получились бы для любого признака, относительно которого члены одной и той же семьи проявляют тенденцию действовать сходным образом.

При оценивании доли мужчин в совокупности результаты будут иными. Произведя аналогичные вычисления, получим:

биномиальная формула $v(p) = 0,00240$;

формула отношения $v(p) = 0,00114$.

Здесь биномиальная формула *преувеличивает* дисперсию. Интересна причина этого. Большинство домохозяйств сложилось в результате брака и поэтому включает, по крайней мере, одного мужчину и одну женщину. Следовательно, доля мужчин на одну семью отличается от 1/2 меньше, чем следовало бы из биномиальной формулы. Ни одна из 30 семей, за исключением одной, состоящей из одного человека, не содержит только мужчин или только женщин. Если бы распределение членов домохозяйств по полу было биномиальным с истинным P , составляющим приблизительно 1/2, то домохозяйства, состоящие из людей одного пола, насчитывали бы четверть всех домохозяйств из 3 человек и одну восьмую домохозяйств из 4 человек. Это свойство соотношения численностей полов в семьях было рассмотрено Хансеном и Хервицем (Hansen and Hurwitz, 1942). Другие примеры ошибок, вызванных неправомерным применением биномиальной формулы в социологических исследованиях, приводит Киш (Kish, 1957).

Упражнения

3.1. Для совокупности с $N = 6$, $A = 4$, $A' = 2$ найдите значение a для всех возможных простых случайных выборок объема 3. Проверьте справедливость теорем, дающих значения среднего и дисперсии $p = a/n$. Убедитесь, что

$$\frac{N-p}{(n-1)N} pq$$

представляет собой несмещенную оценку дисперсии p .

3.2. В простой случайной выборке 200 колледжей из совокупности в 2000 колледжей 120 одобрили некоторое предложение, 57 были против и 23 колледжа не выразили своего мнения. Укажите 95%-ные доверительные границы для числа колледжей в совокупности, одобряющих это предложение.

3.3. Обеспечивают ли результаты предыдущей выборки твердую уверенность в том, что большинство колледжей в изучаемой совокупности одобряют предложение?

3.4. Совокупность объема $N = 7$ состоит из элементов $B_1, C_1, C_2, C_3, D_1, D_2$ и D_3 . Для того чтобы оценить долю C_i -х по отношению к $(C_i + D_i)$ -м извлекается простая случайная выборка объема 4. Найдите условное распределение этой доли p и проверьте формулу для ее условной дисперсии.

3.5. В предыдущем упражнении какова вероятность того, что выборка объема 4 содержит B_2 ? Найдите среднее значение дисперсии p по всем простым случайным выборкам объема 4 и убедитесь, что оно составляет 11/280.

3.6. Простая случайная выборка объемом в 290 домохозяйств извлечена из городского района, содержащего 14 828 домохозяйств. Каждой семье задавались вопросы: владеет ли она домом или нанимает квартиру и пользуется ли она отдельным туалетом. Результаты имеют вид:

	Домовладельцы		Нанимающие квартиру		Все вместе
	Да	Нет	Да	Нет	
Пользование отдельным туалетом					
Число семей	141	6	109	34	290

(а) Для семей, нанимающих квартиру, оцените процент пользующихся отдельным туалетом и укажите стандартную ошибку вашей оценки; (б) оцените общее число семей в районе, нанимающих квартиру, которые не пользуются отдельным туалетом, и вычислите стандартную ошибку этой оценки.

3.7. Предположим, что для упражнения 3.6 общее число семей в городском районе, нанимающих квартиру, равно 7526. Получите новую оценку общего числа семей, нанимающих квартиру, которые не пользуются отдельным туалетом, и вычислите стандартную ошибку этой оценки.

3.8. При оценивании общего числа единиц класса C в области 1 (параграф 3.10) рекомендовалось пользоваться оценкой $\hat{A}_1 = N_1 p_1$, если N_1 известно и оценкой $\hat{A}_1' = N_2 / n$, если N_1 неизвестно. Покажите, что для больших выборок, если пренебречь пкс, отношение дисперсии \hat{A}_1 к дисперсии \hat{A}_1' составляет приблизительно $Q_1/(Q_1 + P_1 n)$, где n — доля единиц совокупности, не принадлежащих области 1, а P_1 — как и в параграфе 3.10 — доля единиц класса C в области 1. Укажите условия, при которых знание N_1 обеспечивает значительное уменьшение дисперсии.

3.9. В простой случайной выборке объема 5 из совокупности объемом в 30 единиц ни одна из единиц не принадлежит классу C . С помощью гипергеометрического распределения найдите верхнюю границу A , числа единиц класса C в совокупности, соответствующую 95%-ной односторонней доверительной вероятности.

Найдите также приближенное значение A_U , вычислив 95%-ную верхнюю биномиальную границу P_U и сократив интервал, как описано в параграфе 3.6. Попробуйте также применить метод, изложенный в примере 3 параграфа 3.6.

3.10. Студенческой службе здоровья известны общее число студентов, N , и общее число посещений врача студентами в течение года, Y . Некоторые студенты не посещали врача. Необходимо получить оценку среднего числа посещений Y/N_1 для N_1 студентов, посетивших врача хотя бы один раз, однако значение N_1 неизвестно. Извлекается простая случайная выборка объемом n из студентов. Из них n_1 студентов посетили врача хотя бы по разу, а общее число посещений для них равно y . Пкс можно пренебречь. (а) Покажите, что y/n_1 представляет собой несмещенную оценку Y/N_1 и что ее условная дисперсия равна S^2/n_1 , где S^2 — дисперсия числа посещений студентами, побывавшими у врача хотя бы раз. (б) Второй метод оценивания Y/N_1 заключается в том, чтобы оценить N_1 с помощью $\hat{N}_1 = Nn_1/n$ и, следовательно, Y/N_1 — с помощью Yn/Nn_1 . Покажите, что эта оценка смещена и что отношение смещения к истинному значению Y/N_1 составляет приблизительно $(N - N_1)/N_1$. Найдите приближенное выражение для дисперсии оценки Yn/Nn_1 и покажите, что оценка из (а) имеет большую дисперсию при

$$S^2 > \frac{(N - N_1)n_1}{N_1 n} \left(\frac{Y}{N_1} \right)^2.$$

Указание. Если p представляет собой биномиальную оценку P , основанную на n испытаниях, тогда приблизительно

$$E\left(\frac{1}{p}\right) = \frac{1}{P} + \frac{Q}{nP^2}; \quad V\left(\frac{1}{p}\right) = \frac{Q}{nP^3}.$$

3.11. Какая из двух предыдущих оценок будет более точной при следующих условиях? $N = 2004$, $Y = 3011$. Выборка при $n = 100$ показала, что 73 студента посетили врача хотя бы один раз. Общее число их посещений равно 152, а оценка дисперсии, s^2 , равна 1,55.

3.12. Простая случайная выборка n гнездовых единиц, каждая из которых содержит m элементов, извлекается из совокупности с долей элементов класса C , равной P . Если внутривнездовая корреляция не меняется, то какими будут наибольшее и наименьшее значения истинной дисперсии p (выборочной оценки P) и как они соотносятся с дисперсией при биномиальном распределении? Пкс можно пренебречь.

3.13. Для выборки объемом в 30 домохозяйства из табл. 3.5 на стр. 85 приведены данные о числе посещений зубного врача в течение прошедшего года. Оцените дисперсию доли лиц, посетивших зубного врача, и сравните эту оценку с оценкой соответствующей дисперсии при биномиальном распределении.

3.14. Один из возможных способов отбора для исследования редкого признака состоит в том, чтобы продолжать извлечение простой случайной выборки до тех пор, пока в выборку не попадет m единиц, обладающих этим признаком, причем m определяется заранее (Haldane, 1945). Докажите, пренебрегая пкс, что вероятность того, что полный объем необходимой для этого выборки составит n единиц, равна

$$\frac{(n-1)!}{(m-1)!(n-m)!} p^m Q^{n-m} \quad (n > m),$$

где P — частота редкого признака. Найдите средний объем полной выборки и покажите, что $p = (m-1)/(n-1)$ представляет собой несмещенную оценку P . (Подробнее эти вопросы изложены в статьях Финни (Finney, 1949) и Санделиуса (Sandelius, 1951); они рассматривают схему, при которой отбор продолжается до тех пор, пока либо будет получено m единиц с редким признаком, либо общий объем выборки достигнет заранее установленного числа единиц n_0).

Число членов семьи	Посетили ли зубного врача		Число членов семьи	Посетили ли зубного врача	
	да	нет		да	нет
5	1	4	5	1	4
6	0	6	4	4	0
3	1	2	4	1	3
3	2	1	3	1	2
2	0	2	3	0	3
3	0	3	4	1	3
3	1	2	3	0	3
3	1	2	3	1	2
4	1	3	1	0	1
4	0	4	2	0	2
3	1	2	4	0	4
2	0	2	3	1	2
7	2	5	4	1	3
4	1	3	2	0	2
3	0	3	4	0	4

ЛИТЕРАТУРА

- Chung J. H. and DeLury D. B. (1950). *Confidence limits for the hypergeometric distribution*. University of Toronto Press.
- Finney D. J. (1949). On a method of estimating frequencies. *Biometrika*, 36, 233—234.
- Fisher R. A. (1958). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, thirteenth edition, 21. Есть русский перевод: Фишер Р. А. Статистические методы для исследователей. М., Госстатиздат, 1958.
- Fisher R. A. and Yates F. (1957). *Statistical tables for biological, agricultural and medical research*. Oliver and Boyd, Edinburgh, fifth edition.
- Haldane J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, 33, 222—225.
- Hansen M. H. and Hurwitz W. N. (1942). Relative efficiencies of various sampling units in population inquiries. *Jour. Amer. Stat. Ass.*, 37, 89—94.
- Harvard Computation Laboratory (1955). *Tables of the cumulative binomial probability distribution*. Harvard University Press, Cambridge, Mass.
- Kish L. (1957). Confidence limits for clustered samples. *Amer. Soc. Rev.*, 22, 154—165.
- Lieberman G. J. and Owen D. B. (1961). *Tables of the hypergeometric probability distribution*. Stanford University Press.
- National Bureau of Standards (1950). *Tables of the binomial probability distribution*. U. S. Government Printing Office, Washington D. C.
- Romig H. G. (1952). *50—100 binomial tables*. John Wiley and Sons, New York.
- Sandelius M. (1951). Truncated inverse binomial sampling. *Skandinavisk Aktuarietidskrift*, 34, 41—44.

ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ

4.1. ГИПОТЕТИЧЕСКИЙ ПРИМЕР

При планировании выборочного обследования всегда наступает момент, когда нужно решить, каким должен быть объем выборки. Это решение играет важную роль. Слишком большая выборка требует излишних затрат, слишком малая уменьшает полезность результатов. Принятое решение не всегда бывает удовлетворительным, поскольку часто из-за недостатка информации мы не можем быть уверены в том, что определили наилучший объем выборки. Теория выборочного метода дает возможность научного обоснования такого решения.

Осветить этапы нахождения этого решения поможет нам гипотетический пример. Антрополог собирается изучать обитателей некоторого острова. Кроме прочего, он намерен оценить процент населения с группой крови 0. Работа организована таким образом, что можно получить простую случайную выборку. Каким должен быть объем этой выборки?

На этот вопрос нельзя ответить, не получив сначала ответа на другой вопрос. Насколько достоверно антрополог хочет знать процент людей с группой крови 0? В ответ он заявляет, что будет удовлетворен, если этот процент окажется правильным в пределах $\pm 5\%$ в том смысле, что если по данным выборки группу крови 0 будут иметь 43% людей, то для всех жителей острова этот процент будет наверняка находиться между 38 и 48%.

Для того чтобы избежать недоразумения, мы должны объяснить антропологу, что для отдельного обследования нельзя дать абсолютной гарантии достоверности результатов в пределах 5%, если только не проводить сплошного обследования. Как бы велико ни было n , всегда есть шанс получить столь неудачную выборку, что ее ошибка превысит желаемые 5%. Антрополог холодно замечает, что он знает об этом, что он готов рискнуть, если шанс получить неудачную выборку составит 1 к 20, и что вообще он просит сообщить ему значение n , а не читать лекцию по статистике.

Теперь мы можем получить грубую оценку значения n . Для упрощения дела пкс не принимается во внимание и предполагается, что выборочный процент, p , распределен нормально. Допустимы ли эти

предположения, можно будет проверить после нахождения для n первого приближения.

Математически p должно находиться в интервале $(P \pm 5)$ с вероятностью 19/20. Поскольку p предполагается нормально распределенным с математическим ожиданием P , с вероятностью 19/20 оно будет находиться в интервале $(P \pm 2\sigma_p)$. Далее,

$$\sigma_p \approx \sqrt{PQ/n}.$$

Следовательно, мы можем положить

$$2 \sqrt{PQ/n} = 5 \quad \text{или} \quad n = \frac{4PQ}{25}.$$

Здесь возникает трудность, присущая всем задачам, связанным с определением объема выборки. Формула для n получена, однако n зависит от некоторой характеристики подвергаемой выборочному изучению совокупности. В нашем примере такой характеристикой служит величина P , которую мы еще только хотим оценить. Поэтому мы спрашиваем антрополога, не представляет ли он, каким может быть возможное значение P . Он отвечает, что, судя по данным предшествующих обследований других этнических групп и по его предположениям о расовом происхождении островитян, P вряд ли выходит за пределы интервала от 30 до 60%.

Этих сведений достаточно для получения удовлетворительного ответа. Для любого значения P между 30 и 60 произведение PQ заключено между 2100 и максимальным значением 2500 при $P = 50$. Соответствующее значение n находится между 336 и 400. Для полной гарантии в качестве первого приближения возьмем значение $n = 400$.

Теперь можно проанализировать предположения, сделанные в ходе наших рассуждений. При $n = 400$ и P между 30 и 60 распределение p будет близко к нормальному. Требуется ли пкс, зависит от числа людей на острове. Если численность совокупности превышает 8000, то доля отбора будет меньше 5% и исправлять n с учетом пкс не потребуется. Метод исправления полученного n , если это понадобится, рассматривается в параграфе 4.4.

4.2. АНАЛИЗ ПРОБЛЕМЫ

Основные этапы определения объема выборки таковы:

1. Следует сформулировать некоторое утверждение, описывающее требования к выборке. Оно может заключаться в указании желательных пределов ошибки выборки, как в предыдущем примере, или же в указании на некоторое решение или действие, которые должны быть основаны на результатах выборки. Ответственность за формулирование такого утверждения лежит прежде всего на лицах, желающих воспользоваться результатами обследования, хотя они часто нуждаются в помощи для изложения своих пожеланий в числовом виде.

2. Следует найти уравнение, связывающее n с желательным уровнем точности выборки. Это уравнение будет меняться в зависимости

от содержания требования к точности и от предполагаемого способа отбора. Одно из преимуществ вероятностного отбора заключается в том, что он дает возможность построить такое уравнение.

3. Это уравнение будет содержать в качестве параметров некоторые неизвестные характеристики совокупности. Их следует оценить, чтобы иметь возможность указать конкретные значения.

4. Часто оказывается, что необходимо публиковать данные по некоторым крупным подразделениям совокупности и что желательные пределы ошибок должны быть заданы для каждого подразделения. В этом случае значения n вычисляются отдельно для каждого подразделения, а общее значение n находят суммируя их.

5. Обычно при выборочном обследовании измеряется более одного признака или характеристики; иногда число признаков довольно велико. Если желательная степень точности указана для каждого признака отдельно, то вычисления приводят к набору значений n для разных признаков, противоречащих одно другому. Следует найти некоторый способ согласования этих значений.

6. В заключение необходимо проверить, совместимо ли полученное значение n с ресурсами, выделенными для выборочного обследования. Для этого нужно оценить затраты денежных средств, труда, времени и материалов, требующиеся для получения выборки предполагаемого объема. Иногда выясняется, что n должно быть значительно уменьшено. В этом случае мы сталкиваемся со сложным вопросом — обследовать ли выборку гораздо меньшего объема, снизив тем самым точность результатов, или отказаться от исследования до получения больших ресурсов.

В следующих параграфах некоторые из этих вопросов будут рассмотрены более подробно.

4.3. ЗАДАНИЕ УРОВНЯ ТОЧНОСТИ

Уровень желательной точности можно задать, указав величину ошибки выборочных оценок, с которой мы готовы примириться. Лучшие всего определять эту величину, исходя из целей применения выборочных данных. Иногда бывает трудно решить, сколь большую ошибку можно было бы допустить, особенно если результаты должны применяться для различных целей. Предположим, что мы спрашиваем антрополога, почему он хочет, чтобы значение процента людей с группой крови 0 было правильным в пределах 5%, а не, скажем, 4 или 6%. Он может ответить, что данные по группам крови будут применены прежде всего для классификации по расам. У него есть сильное подозрение, что островитяне принадлежат или к расовому типу с P , составляющим приблизительно 35%, или к типу с P , составляющим приблизительно 50%. Предел ошибки оценки, составляющий 5%, кажется ему достаточно малым для того, чтобы отнести островитян к одному из этих типов. У него нет, однако, особых возражений против пределов ошибки в 4 или в 6%.

Таким образом, выбор антропологом 5%-ного предела ошибки был до некоторой степени произвольным. Этот пример типичен в том отношении, что часто значение ошибки выбирается именно так. В сущности наш антрополог знает чего он хочет, пожалуй, более твердо, чем многие другие ученые и административные работники. Когда впервые возникает вопрос о желательной степени точности, такие люди сознаются, что никогда не думали об этом и не знают, что ответить. Однако, как показывает мой опыт, после обсуждения они часто могут, по крайней мере, приблизительно указать кажущееся для них приемлемым значение ошибки.

Во многих практических ситуациях мы не можем продвинуться дальше этого в определении желательной степени точности. Трудность состоит отчасти в том, что обычно недостаточно хорошо известно, к каким последствиям в отношении практических решений, принимаемых по результатам обследований, могут привести ошибки того или иного размера. При этом, даже если такие последствия известны, результатами многих важных обследований пользуются разные люди для разных целей, причем некоторые из таких целей нельзя предугадать в то время, когда планируется обследование. Следовательно, элемент догадки будет, вероятно, и в дальнейшем играть значительную роль при задании уровня точности.

Если выборка извлекается для очень конкретной цели, например для принятия решения, которое состоит лишь в ответе «да» или «нет», или для того, чтобы решить, сколько средств израсходовать с некоторым риском, то необходимая точность обычно может быть задана более определенно через указание последствий ошибочного решения. Общий подход к такого рода проблемам излагается в параграфе 4.9, который, хотя и в сжатом виде, дает логические основы для их разрешения.

4.4. ФОРМУЛА ДЛЯ n ПРИ ОТБОРЕ ДЛЯ ОЦЕНИВАНИЯ ДОЛЕЙ

Единицы разделены на два класса: C и C' . Мы условливаемся относительно некоторого предельного значения ошибки d оценки p , доли единиц класса C , и хотим, чтобы лишь с небольшим риском α фактическая ошибка была больше d , т. е. чтобы

$$\Pr(|p - P| \geq d) = \alpha.$$

Предполагается, что производят простой случайный отбор и p имеет нормальное распределение. По теореме 3.2 из параграфа 3.2

$$\sigma_p = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{PQ}{n}}.$$

Следовательно, формула, связывающая n с желательной степенью точности, имеет вид

$$d = t \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{PQ}{n}},$$

где t — значение абсциссы для кривой нормального распределения, отсекающее на «хвостах» площадь α . Разрешая равенство относительно n , получаем

$$n = \frac{\frac{t^2 pq}{d^2}}{1 + \frac{1}{N} \left(\frac{t^2 pq}{d^2} - 1 \right)} \quad (4.1)$$

Для практического применения вместо P в эту формулу подставляется некоторая предварительная оценка, p . Если N велико, то в качестве первого приближения берется

$$n_0 = \frac{t^2 pq}{d^2} = \frac{pq}{V} \quad (4.2)$$

где $V = \frac{d^2}{t^2}$ — желательная дисперсия выборочной доли.

Практически сначала вычисляется n_0 . Если n_0/N пренебрежимо мало, то n_0 будет удовлетворительным приближением для n из (4.1). В противном случае из сравнения (4.1) и (4.2) следует, что n можно получить по формуле

$$n = \frac{n_0}{1 + (n_0 - 1)/N} \approx \frac{n_0}{1 + (n_0/N)} \quad (4.3)$$

Пример. В гипотетическом примере с группами крови мы имели

$$d = 0,05; p = 0,5; \alpha = 0,05; t = 2.$$

Таким образом,

$$n_0 = \frac{4 \cdot 0,5 \cdot 0,5}{0,0025} = 400.$$

Предположим, что на острове живет только 3200 человек. Необходимо учесть пкс, и мы получаем

$$n = \frac{n_0}{1 + (n_0 - 1)/N} = \frac{400}{1 + \frac{399}{3200}} = 356.$$

Формула для n_0 справедлива также, если величины d , p и q все выражены не в долях единицы, а в процентах. Поскольку произведение pq растет, когда p стремится к $1/2$, или к 50%, более надежная оценка n получится, если в интервале, где, как предполагается, находится p , принять в качестве p ближайшее к $1/2$ число. Например, если предполагается, что p находится между 5 и 9%, то для определения n мы принимаем значение 9%.

4.5. ФОРМУЛА ДЛЯ n В СЛУЧАЕ НЕПРЕРЫВНЫХ ПЕРЕМЕННЫХ

Пусть \bar{y} — среднее значение наблюдений при простой случайной выборке, и мы хотим, чтобы

$$\Pr(|\bar{y} - \bar{Y}| \geq d) = \alpha,$$

где d — выбранное предельное значение ошибки и α — некоторая малая вероятность. Предположим, что \bar{y} распределено нормально; согласно следствию 1 из теоремы 2.2 его стандартная ошибка равна:

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N}} \frac{S}{\sqrt{n}}.$$

Следовательно,

$$d = t \sqrt{\frac{N-n}{N}} \frac{S}{\sqrt{n}} \quad (4.4)$$

Это дает

$$n = \frac{\left(\frac{tS}{d}\right)^2}{1 + \frac{1}{N} \left(\frac{tS}{d}\right)^2}.$$

Как и в предыдущем параграфе, в качестве первого приближения принимаем

$$n_0 = \left(\frac{tS}{d}\right)^2 = \frac{S^2}{V} \quad (4.5)$$

Оно пригодно, если только n_0/N мало; в противном случае мы вычисляем n как

$$n = \frac{n_0}{1 + n_0/N} \quad (4.6)$$

Если оценивается суммарное значение для совокупности, Y , с предельной ошибкой d , то в качестве первого приближения вместо (4.5) следует взять

$$n_0 = \left(\frac{NtS}{d}\right)^2 = \frac{(NS)^2}{V}.$$

Равенство 4.6 остается без изменений.

Пример. В питомниках, выращивающих на продажу саженцы деревьев, в конце зимы или в начале весны целесообразно оценивать число имеющихся здоровых саженцев, поскольку от него зависит удовлетворение заявок и прием заказов. Выборочные методы оценивания общего числа саженцев были изучены Джонсоном (Johnson, 1943). Приведенные далее данные были получены при обследовании гряды саженцев серебристого клена, ширина которой 1 фут и длина 430 фу-

тов. Единицей отбора служил 1 фут длины гряды, так что $N = 430$. Путем сплошного обследования гряды было найдено, что истинные значения для совокупности равны: $\bar{Y} = 19$, $S^2 = 85,6$.

Сколько нужно выбрать единиц при простом случайном отборе для того, чтобы оценить \bar{Y} с ошибкой в пределах 10% при вероятности 19/20, что она не выйдет за эти пределы?

Из (4.5) получаем

$$n_0 = \frac{t^2 S^2}{d^2} = \frac{4 \cdot 85,6}{1,9^2} = 95.$$

Поскольку n_0/N пренебречь нельзя, принимаем

$$n = \frac{95}{1 + \frac{95}{430}} = 78.$$

Для того чтобы получить желательную точность, необходимо сосчитать саженцы почти на 20% гряды.

Формулы для n , приведенные здесь, применимы только при простом случайном отборе, когда в качестве оценки \bar{Y} принимается выборочное среднее. Соответствующие формулы для других способов отбора и оценивания указываются при рассмотрении этих методов.

4.6. ПРЕДВАРИТЕЛЬНЫЕ ОЦЕНКИ ДИСПЕРСИИ ДЛЯ СОВОКУПНОСТИ

Пример с питомником нетипичен в том отношении, что в нем была известна дисперсия для совокупности, S^2 . На практике оценки дисперсий совокупности, нужные для определения объема выборки, можно получить четырьмя способами: (1) производить отбор в два этапа, извлекая на первом простую случайную выборку объема n_1 и по ее данным получая значения S^2 или P и требуемое n ; (2) по данным пробного обследования; (3) по предыдущим обследованиям этой или подобных ей совокупностей; (4) на основании предположений о структуре совокупности, дополненных некоторыми математическими расчетами.

Первый способ дает наиболее надежные оценки S^2 или P , но применяется редко, поскольку он увеличивает сроки всего обследования. Для тех случаев, когда он осуществим, Кохс (Cox, 1952), следуя работе Штейна (Stein, 1945), показал, как вычислять n по A_1^2 или p_1 так, чтобы окончательная оценка \bar{y} или p имела заданную дисперсию V , заданный предел ошибки d или заданный коэффициент вариации. Первая выборка предполагается достаточно большой, чтобы можно было пренебречь членами порядка $1/n_1^2$. Приведем некоторые из этих результатов.

Оценивание \bar{Y} при заданной дисперсии V

Если s_1^2 — значение дисперсии, полученное из первой выборки, то нужно отобрать столько дополнительных единиц, чтобы сделать

общий объем выборки равным:

$$n = \frac{s_1^2}{V} \left(1 + \frac{2}{n_1} \right). \quad (4.7)$$

Распределение y предполагается приблизительно нормальным. Если бы S было известно точно, то требующийся объем выборки составил бы S^2/V . Незнание S увеличивает объем выборки в среднем в $(1 + 2/n_1)$ раз.

Оценивание P при заданной дисперсии V

Пусть оценка P , полученная из первой выборки, равна p_1 . Тогда совместный объем первой и второй выборок должен быть

$$n = \frac{p_1 q_1}{V} + \frac{3 - 8p_1 q_1}{p_1 q_1} + \frac{1 - 3p_1 q_1}{V n_1}. \quad (4.8)$$

Первый член в правой части равенства соответствует объему, который потребовался бы, если бы было известно, что P равно p_1 . Для этого способа обычная биномиальная оценка p , вычисленная по полной выборке объема n , окажется несколько смещенной. Для того чтобы внести поправку на смещение, следует принять

$$\hat{P} = p + \frac{V(1-2p)}{pq}.$$

Оценивание P при заданном коэффициенте вариации, равном \sqrt{C}

Нужно взять

$$n = \frac{q_1}{C p_1} + \frac{3}{p_1 q_1} + \frac{1}{C p_1 n_1}. \quad (4.9)$$

Оценкой служит $\hat{P} = p - C p/q$. Во всех трех приведенных результатах пкс во внимание не принимается.

Пример. Обследователь хочет оценить P с коэффициентом вариации равным 0,1 (или 10%). Он предполагает, что P заключено между 5 и 20%. Такой интервал слишком велик, чтобы можно было сразу получить хорошее приближение требуемого n . Поскольку коэффициент вариации P равен $\sqrt{Q/nP}$, легко проверить, что $n = 400$ подходит для $P = 20\%$, но если P равно лишь 5%, то необходимо значение $n = 1900$. Поэтому сначала он рассматривает выборку с $n_1 = 400$ и находит $p_1 = 0,105$. Поскольку $\sqrt{C} = 0,1$; $C = 0,01$. Из равенства (4.9) следует, что

$$n = \frac{0,895}{0,01 \cdot 0,105} + \frac{3}{0,0940} + \frac{1}{0,01 \cdot 42} = 925.$$

Совместная выборка дает $np = 88$; $p = 88/925 = 0,0951$. Поправка на смещение, Cp/q , равна 0,0011, что дает окончательную оценку, равную 0,0940, или 9,4%.

Второй способ — небольшое пробное обследование — удобен для многих целей, особенно если осуществимость основного обследования находится под сомнением. Если пробное обследование проводится по простой случайной выборке, то применимы описанные только что методы. Однако часто пробное обследование ограничивается лишь той частью совокупности, которая более удобна для изучения или помогает выявить значение определенных факторов. Этот избирательный характер пробного обследования следует учитывать, применяя его результаты для оценки S^2 или P . Например, обычно предварительное исследование ограничивается несколькими гнездами единиц. Вычисляемая при этом s^2 будет отражать, главным образом, вариацию внутри этих гнезд и может привести к преуменьшению соответствующего S^2 . Связь между внутри- и междугнездовой вариацией рассматривается в гл. 9. Та же проблема возникает при гнездовом отборе для оценивания долей, когда формула pq/n может преуменьшать эффект вариации между гнездами. Определение объема выборки при гнездовом отборе для оценивания долей хорошо проиллюстрировал Корнфилд (Cornfield, 1951).

Третий способ — применение результатов предыдущих обследований — предполагает, что имеются те или иные данные о средних квадратичных отклонениях в предыдущих обследованиях или, по крайней мере, существует возможность эти данные получить. К сожалению, стоимость вычисления средних квадратичных отклонений в сложных обследованиях довольно высока даже при применении электронных вычислительных машин, и обычно вычисляются и хранятся только те значения средних квадратичных отклонений, которые нужны, чтобы получить общее представление о точности основных оценок. Если есть подходящие данные за прошлые годы, то к величине S^2 может понадобиться поправка на ее изменение во времени. Для периодически получаемых данных, когда во времени меняется \bar{Y} , часто оказывается, что S^2 меняется со скоростью, заключенной между $k\bar{Y}$ и $k\bar{Y}^2$, где k — некоторое постоянное число. Так, если предполагается, что за время, прошедшее после предыдущего обследования, \bar{Y} увеличилось на 10%, то мы должны увеличить нашу первоначальную оценку S^2 на 10—20%.

Наконец, иногда можно вычислить вполне пригодную оценку S^2 , располагая лишь сравнительно ограниченными сведениями о характере совокупности. В ранних исследованиях числа червей-проволочников в почве в качестве единицы отбора принимался образец грунта размером $9 \times 9 \times 5$ дюймов из верхнего слоя почвы. Для определения n исследователю нужно было знать среднее квадратичное отклонение числа червей-проволочников, найденных при взятии одного образца. Если бы черви были распределены в верхнем слое случайным образом, то число их в небольшом объеме грунта распределялось бы по закону Пуассона, для которого $S^2 = \bar{Y}$. Поскольку, возможно, существует тенденция к расположению червей группами, было решено положить $S^2 = 1,2\bar{Y}$, где множитель 1,2 взят произвольно для увеличения надежности. Хотя \bar{Y} не было известно точно, можно было ука-

зать величины \bar{Y} , имеющие экономическое значение в отношении потерь урожая. Эти две отправных точки дали информацию, которая позволила удовлетворительно определить объем выборки.

Деминг (Deming, 1960) показал, как можно применить для оценивания S^2 некоторые простые вероятностные распределения, если известен размах распределения и имеется общее представление о его форме. Если имеется распределение биномиального типа с долей наблюдений на одном конце интервала значений, равной p , и долей q на другом, то $S^2 = pqh^2$, где h — длина интервала значений (размах). Если $p = q = 1/2$, то максимальное значение S^2 при данном h равно $0,25h^2$. Другие полезные соотношения: $S^2 = 0,083h^2$ для равномерного распределения, $S^2 = 0,056h^2$ для распределения, имеющего форму прямоугольного треугольника, и $S^2 = 0,042h^2$ для распределения, имеющего форму равнобедренного треугольника.

Эти формулы мало помогают, если h велико или известно неточно. Однако при больших h в практике выборочного исследования успешно применяется расслоение совокупности (см. гл. 5) таким образом, чтобы размах внутри каждого слоя стал значительно меньше. Обычно внутри слоя и вид распределения становится проще (ближе к равномерному). Таким образом, эти формулы выгодны при определении S^2 , а следовательно, и n внутри отдельных слоев.

4.7. ОБЪЕМ ВЫБОРКИ ПРИ ИЗУЧЕНИИ НЕСКОЛЬКИХ ПРИЗНАКОВ

В большинстве обследований сведения собирают не по одному, а по нескольким признакам. Один из методов определения объема выборки в этом случае состоит в том, чтобы указать пределы ошибок для признаков, считающихся в обследовании наиболее важными. Необходимый объем выборки определяется сначала отдельно по каждому из этих важнейших признаков.

После того, как значения n для отдельных признаков вычислены, можно рассмотреть ситуацию в целом. Может оказаться, что все полученные значения n достаточно близки одно к другому. Если наибольшее значение n приемлемо с точки зрения бюджета обследования, то оно и выбирается. Чаше, однако, значения n довольно сильно отличаются одно от другого, причем выбор наибольшего из них нежелателен или по финансовым соображениям или потому, что общий уровень точности окажется значительно выше намечавшегося первоначально. В этом случае, чтобы оправдать применение меньшего значения n , желательный уровень точности для некоторых признаков можно снизить.

В некоторых случаях значения n , требующиеся для различных признаков, настолько расходятся, что от изучения некоторых признаков нужно вообще отказаться, поскольку при имеющихся ресурсах ожидаемая для них точность будет совершенно недостаточной. Возникающие трудности не обязательно связаны с объемом выборки. Для разных признаков могут понадобиться разные способы отбора. Для совокупностей, которые обследуются повторно, полезно накапливать информацию относительно тех признаков, которые экономически вы-

годно объединить в обследовании общего типа, и тех, которые требуют особых методов. В качестве примера в табл. 4.1 приведена классификация признаков по четырем типам, основанная на опыте региональных сельскохозяйственных обследований. В этой классификации под обследованием общего типа понимается такое обследование, при котором единицы отбора распределены довольно равномерно по некоторому региону, как, например, при простом случайном отборе.

Таблица 4.1

ПРИМЕР РАЗЛИЧНЫХ ТИПОВ ПРИЗНАКОВ
ПРИ РЕГИОНАЛЬНЫХ ОБСЛЕДОВАНИЯХ

Типы	Характеристики признаков	Необходимый тип отбора
1	Широко распространен по всему региону, достаточно часто встречается во всех его частях	Обследование общего типа с низкой долей отбора
2	Широко распространен по всему региону, но встречается не часто	Обследование общего типа, но с более высокой долей отбора
3	Встречается достаточно часто в большинстве частей региона, но распределен более спорадически: отсутствует в одних частях региона и сильно сконцентрирован в других	Для получения хороших результатов нужна расслоенная выборка с разной интенсивностью в разных частях региона (см. гл. 5). Иногда может быть изучен в обследовании общего типа при дополнительном отборе
4	Распределен очень спорадически или сосредоточен в небольшой части региона	Не пригоден для обследования общего типа, требуется отбор, учитывающий характер распределения

4.8. ОБЪЕМ ВЫБОРКИ ПРИ НЕОБХОДИМОСТИ ПОЛУЧИТЬ ОЦЕНКИ ДЛЯ ПОДРАЗДЕЛЕНИЯ СОВОКУПНОСТИ

Довольно часто планируется получение оценок не только для совокупности в целом, но и для отдельных ее подразделений. Если эти подразделения можно выделить заранее, как в случае нескольких географических районов, то n вычисляется отдельно для каждого подразделения. Предположим, что среднее по каждому подразделению нужно оценить с определенной дисперсией V . Для i -го подразделения имеем $n_i = S_i^2/V$, так что общий объем выборки $n = \sum S_i^2/V$. Отдельные S_i^2 будут, в среднем, меньше S^2 , дисперсии для совокупности, но часто это различие весьма незначительно. Так, если имеется k подразделений, то $n \approx kS^2/V$, в то время как для получения такой же оценки только для совокупности в целом нужно было бы взять $n = S^2/V$.

Таким образом, если желательно получить оценки с дисперсией V для каждого из k подразделений совокупности, то объем выборки должен быть приблизительно в k раз больше, чем для получения оценки той же точности для всей совокупности. Лица, не имеющие опыта

проведения обследований, часто упускают из вида это обстоятельство при вычислении объема выборки.

Если подразделения совокупности соответствуют классификации по таким переменным, как возраст, пол, доход, число лет обучения в школе, то к какому подразделению относится то или иное лицо становится известно только тогда, когда выборка уже получена. В этом случае предварительный объем выборки еще можно определить, если известны доли π_i единиц, принадлежащих различным подразделениям. Если извлекается простая случайная выборка объема n , то ожидаемое значение объема выборки из i -го подразделения будет $n\pi_i$. Если $n\pi_i$ велико, то дисперсия среднего значения для выборки из этого подразделения будет в среднем

$$V(\bar{y}_i) = E\left(\frac{S_i^2}{n_i}\right) \approx \frac{S_i^2}{n\pi_i}.$$

Следовательно, для того чтобы получить $V(\bar{y}_i) = V$, нужно взять $n \approx S_i^2/\pi_i V$. Если это справедливо для каждого подразделения, то

$$V \approx \max\left(\frac{S_i^2}{\pi_i V}\right) \approx \frac{S^2}{V} \max\left(\frac{1}{\pi_i}\right).$$

Если подразделения совокупности приблизительно равны по величине, то $\pi_i \approx 1/k$, но если некоторые подразделения малочисленны, то множитель $\max(1/\pi_i)$ может быть значительно больше k . В этом случае мы должны или увеличить значение V для таких подразделений или найти некоторый способ заранее находить принадлежащие к ним единицы, так чтобы долю отбора для них сделать выше. Для этой цели иногда полезен метод двойного отбора (см. гл. 12).

Еще выше требования к объему выборки в аналитических обследованиях, для которых ограничения имеют вид

$$V(\bar{y}_i - \bar{y}_j) \leq V$$

для каждой пары подразделений совокупности (областей изучения). В этом случае

$$n \approx \max_{i,j} \frac{1}{V} \left(\frac{S_i^2}{\pi_i} + \frac{S_j^2}{\pi_j} \right).$$

Если S_i^2 не очень отличаются от S^2 , то n будет равно $2kS^2/V$ при k областях изучения, имеющих равные объемы, и еще больше, когда их объемы неодинаковы. Влияние пкс, которую мы здесь не учитывали, сказывается в некотором уменьшении требующихся n .

4.9. ОБЪЕМ ВЫБОРКИ С ТОЧКИ ЗРЕНИЯ ТЕОРИИ РЕШЕНИЙ

Более логичный подход к определению объема выборки может быть иногда предложен в тех случаях, когда по результатам выборки должно приниматься практическое решение. Такое решение будет, по-видимому, более обоснованным, если выборочная оценка имеет малую ошибку, чем если она имеет большую ошибку. Мы можем вычислить

в денежном выражении потери $l(z)$, которые принесет решение, принимаемое при равной z ошибке оценки. Хотя действительное значение z заранее предсказать нельзя, теория выборочного метода дает нам возможность найти распределение частот $f(z, n)$ величины z , которое при определенном способе отбора будет зависеть от объема выборки n . Следовательно, при данном объеме выборки *ожидаемые потери* будут

$$L(n) = \int l(z) f(z, n) dz.$$

Мы стремимся извлечь выборку так, чтобы уменьшить эти потери. Если $C(n)$ — издержки на получение выборки объема n , то целесообразно выбрать n так, чтобы минимизировать

$$C(n) + L(n),$$

поскольку это выражение представляет собой общие расходы, связанные с получением выборки и принятием решения по ее результатам. Выбор n определяет как оптимальный объем выборки, так и наиболее выгодный уровень точности.

К решению той же задачи можно подойти и иначе, не через потери, связанные с ошибками выборочных оценок, а через денежный *выигрыш*, получаемый от сведений приносимых выборкой. Если вводится понятие денежного выигрыша, то мы рассматриваем при извлечении выборки объема n ожидаемый выигрыш $G(n)$, который равен нулю, если выборка не извлекается. Теперь мы *максимизируем*

$$G(n) - C(n).$$

В такой форме наш подход эквивалентен правилу максимизации прибыли в классической экономике.

Наиболее простой вид функции потерь $l(z) = \lambda z^2$, где λ — постоянная величина. Отсюда

$$L(n) = \lambda E(z^2).$$

Например, если \hat{Y} — выборочная оценка \bar{Y} и $z = \hat{Y} - \bar{Y}$, то при простом случайном отборе

$$L(n) = \lambda V(\hat{Y}) = \frac{\lambda S^2}{n} - \frac{\lambda S^2}{N}.$$

Наиболее простой вид функции издержек на получение выборки

$$C(n) = c_0 + c_1 n,$$

где c_0 — накладные расходы. Дифференцируя, получаем значение n , минимизирующее издержки плюс потери

$$n = \sqrt{\lambda S^2 / c_1}.$$

В более общем виде этот результат приводит Йейтс (Yates, 1960). Те же рассуждения применимы к любому способу отбора и оценивания, при котором дисперсия оценки обратно пропорциональна n и стоимость — линейная функция n .

Блис (Blythe, 1945) описал применение рассмотренного только что подхода к оцениванию объема древесины на участке леса с целью

ее продажи (см. упражнение 4.11). Нордин (Nordin, 1944) изучал оптимальный объем выборки для определения потенциального объема продаж при сбыте новых изделий. Если такой объем продаж можно предвидеть точно, то появляется возможность распределить имеющееся оборудование и объем производства во времени так, чтобы максимизировать ожидаемую прибыль производителя. Гранди и другие (Grundy et al., 1954, 1956) рассматривают оптимальный объем второй выборки, если уже известны результаты первой.

Этот подход был далее существенно развит в работах по статистической теории решений. В них рассматриваются также обобщения, как применение в качестве меры стоимости и потерь полезности, а не денежной оценки, применение субъективной априорной информации о неизвестных параметрах, представленной в виде «априорных» вероятностных распределений этих неизвестных параметров, исследование различных видов функций издержек и потерь и различных видов как количественных, так и качественных данных. Подробное изложение этого метода можно найти в книге Райфа и Шлайфера (Raiffa and Schlaifer, 1961). Еще не совсем ясно, в какой степени различные задачи будут поддаваться полному разрешению с помощью описанного здесь подхода. Однако его ценность заключается уже в том, что он стимулирует ясное понимание роли различных факторов в получении хорошего решения. Одной из областей, где такой подход может найти применение, служит при массовом производстве выборочное изучение изделий с тем, чтобы на основе оценки их качества решить, принять или отклонить данную партию изделий. Ситтиг (Sittig, 1951) рассматривал экономические аспекты определения объема выборки, учитывая расходы на проверку и издержки, связанные с наличием дефектных изделий в принятой партии и доброкачественных в отвергнутой.

Упражнения

4.1. Для некоторого района, насчитывающего 4000 домов, нужно оценить процент домовладельцев со стандартной ошибкой, не превышающей 2%, и процент домохозяйств, имеющих два автомобиля, со стандартной ошибкой, не превышающей 1%. (Величины 1 и 2% — это абсолютные значения, а не коэффициенты вариации.) Предполагается, что истинный процент домовладельцев заключен между 45 и 65%, а процент домохозяйств, имеющих два автомобиля — между 5 и 10%. Каким должен быть объем выборки, чтобы получить оценки обеих характеристик с указанной точностью?

4.2. Какого объема должна быть выборка из совокупности, содержащей 676 листов подписей под петицией (табл. 2.1, с. 42), чтобы оценить общее число подписей с ошибкой не более чем в 1000 подписей и с риском ошибиться в одном случае из 20? Считайте, что значение s^2 , приведенное на с. 42, это S^2 для совокупности.

4.3. Необходимо провести обследование, чтобы установить распространенность общих заболеваний в некоторой большой совокупности. Для каждого заболевания, поражающего не менее 1% лиц этой совокупности, желательно оценить общее число случаев заболевания с коэффициентом вариации, не превышающим 20%. (а) Каким должен быть объем простой случайной выборки при условии, что наличие заболевания устанавливается безошибочно? (б) Каким должен быть объем выборки, если необходимо оценить с одинаковой точностью общее число случаев заболевания отдельно у мужчин и женщин?

4.4. При исследовании червей-проволочников необходимо на каждом участке, где плотность их численности в верхнем 5-дюймовом слое почвы превышает

200 000 на один акр, оценить число проволоочников на один акр с ошибкой в пределах 30% на 95%-ном доверительном уровне. Единичей отбора служит проба грунта размером $9 \times 9 \times 5$ дюймов. Считая, что распределение числа червей в отдельной пробе имеет лишь незначительно большую изменчивость, чем распределение Пуассона, мы полагаем $S^2 = 1,2\bar{Y}$. Каким должен быть объем простой случайной выборки? (1 акр = 43560 квадратным футам, 1 квадратный фут = 144 квадратным дюймам).

4.5. При сельскохозяйственном обследовании в штате Айова, где единицей отбора служил участок земли площадью в 1 квадратную милю, были получены следующие значения коэффициентов вариации в расчете на одну единицу (данные Джессена (R. J. Jessen)):

Признак	Оценки коэффициента вариации, %
Площадь под строениями	38
Площадь под пшеницей	39
Площадь под овсом	44
Число работников — членов семьи	100
Число наемных работников	110
Число безработных	317

Планируется обследование, для того чтобы оценить значения площадей с коэффициентом вариации 2,5% и значения чисел работников (кроме безработных) с коэффициентом вариации 5%. Сколько нужно взять единиц при простом случайном отборе? Какой точности можно ожидать от этой выборки для оценки числа безработных?

4.6. Нужно оценить путем опыта среднее значение некоторой случайной переменной с дисперсией $V = 0,0005$. Значения этой случайной переменной для первых 20 наблюдений приведены в таблице. Сколько еще наблюдений нужно сделать? (Воспользуйтесь формулой 4.7.)

Номер наблюдений	Значение случайной величины	Номер наблюдений	Значение случайной величины
1	0,0725	11	0,0712
2	0,0755	12	0,0748
3	0,0759	13	0,0878
4	0,0739	14	0,0710
5	0,0732	15	0,0754
6	0,0843	16	0,0712
7	0,0727	17	0,0757
8	0,0769	18	0,0737
9	0,0730	19	0,0704
10	0,0727	20	0,0723

4.7. Планируется обследование домохозяйств для того, чтобы оценить долю семей, обладающих определенными признаками. Предполагается, что для основных интересующих нас признаков значение P заключено между 30 и 70%. Какими должны быть значения n при простом случайном отборе, необходимые для того, чтобы оценить со стандартной ошибкой, не превосходящей 3%, следующие средние: (а) общее среднее значение P ; (б) отдельные средние P_i для групп семей с доходом: до 5000 долл., от 5000 до 10000 долл., свыше 10000 долл. ($i = 1, 2, 3$); (в) разности средних ($P_i - P_j$) для каждой пары групп из (б)? Дайте ответ от-

дельно для (а), (б) и (в). Статистика доходов указывает, что доли семей в трех группах по доходу составляют 50, 38 и 12%.

4.8. Четырехгодичные колледжи в США были разделены на четыре группы по размеру в соответствии с числом студентов в 1952—1953 гг. Средние квадратичные отклонения в каждой группе следующие:

	Группа			
	1	2	3	4
Число студентов S	до 1 000 236	1 000—3 000 625	3 000—10 000 2 008	свыше 10 000 10 023

Если вам известны границы групп, но не известны значения S , насколько правильно вы сможете их установить, пользуясь простыми математическими средствами (см. параграф 4.6)? Ни в одном из колледжей не обучается менее 200 студентов, а в наибольшем из них приблизительно 50 000 студентов.

4.9. При квадратичной функции потерь и линейной функции издержек, приведенных в параграфе 4.9, удалось, применяя некоторую улучшенную схему отбора, добиться того, что значение S^2 оказалось возможным заменить меньшим значением S'^2 , а c_0 , c_1 и λ остались без изменения. Пусть n' , V' обозначают новый оптимальный объем выборки и соответствующую $V(\bar{Y})$. Покажите, что $n' < n$ и что $V' < V$ при условии, что $\lambda < N/2$.

4.10. Покажите, что если функция потерь, связанных с ошибкой при оценивании \bar{y} имеет вид $\lambda |\bar{y} - \bar{Y}|$ и издержки $C = c_0 + c_1 n$, то при простом случайном отборе, пренебрегая пкс, наиболее экономичным значением n будет

$$\left(\frac{\lambda S}{c_1 \sqrt{2\pi}} \right)^{2/3}.$$

4.11. [Из работы (Blythe, 1945)]. Продажная цена строевого леса на некотором участке равна UW , где U — цена единицы объема древесины и W — объем древесины на участке. На участке подсчитывается общее число стволов N и по простой случайной выборке n стволов оценивается средний объем древесины в расчете на один ствол. В соответствии с этой оценкой определяется стоимость леса и эту сумму покупатель предварительно уплачивает продавцу. Позднее покупатель определяет точный объем купленной древесины, и если он оплатил больший объем, чем получил, то продавец возмещает ему разницу. Если же покупатель оплатил объем древесины меньший полученного, то он продавцу ничего не возвращает.

Постройте функцию потерь для продавца. Предполагая, что издержки на наблюдение n стволов равны cn , найдите оптимальное значение n . Среднее квадратичное отклонение объема древесины в расчете на один ствол можно считать равным S , а пкс пренебречь.

ЛИТЕРАТУРА

- Blythe R. H. (1945). The economics of sample size applied to the scaling of saw-logs. *Biom. Bull.*, 1, 67—70.
 Cornfield J. (1951). The determination of sample size. *Amer. Jour. Pub. Health*, 41, 654—661.
 Cox D. R. (1952). Estimation by double sampling. *Biometrika*, 39, 217—227.
 Deming W. E. (1960). *Sample design in business research*. John Wiley and Sons. New York.

- Grundy P. M., Healy M. J. R. and Rees D. H. (1954). Decision between two alternatives — how many experiments? *Biometrics*, 10, 317—323.
- Grundy P. M., Healy M. J. R. and Rees D. H. (1956). Economic choice of the amount of experimentation. *Jour. Roy. Stat. Soc.*, B. 18, 32—55.
- Johnson F. A. (1943). A statistical study of sampling methods for tree nursery inventories. *Jour. Forestry*, 41, 674—679.
- Nordin J. A. (1944). Determining sample size. *Jour. Amer. Stat. Assoc.*, 39, 497—506.
- Raiffa H. and Schlaifer R. (1961). *Applied statistical decision theory*. Harvard Business School, Boston.
- Sittig J. (1951). The economic choice of sampling system in acceptance sampling. *Bull. Int. Stat. Inst.*, 33, V, 51—84.
- Stein C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Stat.*, 16, 243—258.
- Yates F. (1960). *Sampling methods for censuses and surveys*. Charles Griffin and Co., London, Third Edition. Есть русский перевод: Йейтс Ф. Выборочный метод в переписях и обследованиях. М. «Статистика», 1965.

РАССЛОЕННЫЙ СЛУЧАЙНЫЙ ОТБОР

5.1. ОПИСАНИЕ

При расслоенном отборе* совокупность, содержащая N единиц, сначала подразделяется на подсовокупности, состоящие соответственно из N_1, N_2, \dots, N_L единиц. Эти подсовокупности не содержат общих единиц и вместе исчерпывают всю совокупность, так что

$$N_1 + N_2 + \dots + N_L = N.$$

Такие подсовокупности называются *слоями*. Для того чтобы можно было полностью воспользоваться выгодами от расслоения, значения N_L должны быть известны. Когда слои определены, выборка извлекается из каждого слоя, причем отбор в разных слоях производится независимо. Объемы выборок внутри слоев обозначаются соответственно через n_1, n_2, \dots, n_L .

Если в каждом слое берут простую случайную выборку, то способ отбора в целом называется *расслоенным случайным отбором*.

Расслоение — довольно распространенный прием. Это обусловлено многими причинами; перечислим основные из них.

1. Если желательно получить с определенной точностью данные о некоторых подразделениях совокупности, то каждое такое подразделение рекомендуется рассматривать на правах самостоятельной «совокупности».

2. Применение расслоения может быть продиктовано организационными соображениями, например агентство, проводящее обследование, может иметь районные отделения, каждое из которых обеспечивает проведение обследования какой-либо части совокупности.

3. Проблемы, связанные с отбором в разных частях совокупности, могут сильно различаться. При выборочных обследованиях населения людей, находящихся в таких заведениях, как гостиницы, больницы,

* В советской статистической литературе этот вид отбора называется чаще всего типическим или районированным, поскольку подразумевается, что деление изучаемой совокупности на типические группы преследует цель не только сократить вариацию признака, но и выделить типы явлений. Однако автор, следуя принятому в западной статистической литературе, ограничивается при подразделении совокупности лишь требованием однородности каждой из ее частей. Поэтому английский термин «stratified» был переведен как «расслоенный», что больше отвечает содержанию описываемого приема. — *Примеч. ред.*

туры, часто выделяют в отдельный слой в отличие от людей, живущих в обычных домах, поскольку к отбору в этих двух случаях требуется разный подход. При обследовании, предпринятом с целью изучения деловой активности, мы можем составить список крупных фирм, выделив их в отдельный слой. Для более мелких фирм можно применить один из видов территориального отбора.

4. Расслоение может дать выигрыш в точности при оценивании характеристик всей совокупности. Иногда неоднородную совокупность удается подразделить на подсовокупности, каждая из которых внутренне однородна. Это и подразумевается под названием *слой* по аналогии с разделением на слои в геологии. Если каждый слой однороден в том смысле, что результаты измерений в нем очень мало изменяются от единицы к единице, то можно получить точную оценку среднего значения для любого слоя по небольшой выборке в этом слое. Затем эти оценки можно объединить в одну точную оценку для всей совокупности.

В теории расслоенного отбора рассматриваются свойства оценок, полученных по расслоенным выборкам, и условия определения наилучших объемов выборки по слоям, n_h , для получения максимальной точности. Здесь предполагается, что сами слои уже образованы. Как образовать слои и определить, сколько их должно быть, рассматривается далее (параграф 5А.6).

5.2. ОБОЗНАЧЕНИЯ

Индексы h и i обозначают соответственно номер слоя и номер единицы внутри слоя. Это естественное обобщение обозначений, введенных ранее. Далее приведены символы, относящиеся к слою h

N_h	— общее число единиц;
n_h	— число единиц в выборке;
y_{hi}	— значение, полученное для i -й единицы;
$W_h = \frac{N_h}{N}$	— вес слоя;
$f_h = \frac{n_h}{N_h}$	— доля отбора в слое;
$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$	— истинное среднее;
$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$	— выборочное среднее;
$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1}$	— истинная дисперсия.

Заметим, что в знаменателе дисперсии стоит $(N_h - 1)$.

5.3. СВОЙСТВА ОЦЕНОК

В качестве оценки среднего значения на единицу для совокупности при расслоенном отборе применяется \bar{y}_{st} [st — от английского «stratified» — расслоенный]*:

$$\bar{y}_{st} = \frac{\sum_{h=1}^L N_h \bar{y}_h}{N}, \quad (5.1)$$

где $N = N_1 + N_2 + \dots + N_L$.

Оценка \bar{y}_{st} , вообще говоря, не совпадает с выборочным средним. Выборочное среднее, \bar{y} , можно записать в виде

$$\bar{y} = \frac{\sum_{h=1}^L n_h \bar{y}_h}{n}. \quad (5.2)$$

Различие состоит в том, что в \bar{y}_{st} оценкам, полученным по отдельным слоям, придаются их правильные веса N_h/N . Очевидно, что \bar{y} совпадает с \bar{y}_{st} при условии, что для каждого слоя

$$\frac{n_h}{n} = \frac{N_h}{N} \quad \text{или} \quad \frac{n_h}{N_h} = \frac{n}{N} \quad \text{или} \quad f_h = f.$$

Это значит, что доля отбора одинакова для всех слоев. Такое расслоение называется расслоением с *пропорциональным* размещением n_h . Оно обеспечивает *равновзвешенную* выборку. Если нужно вычислять многочисленные оценки, то равновзвешенная выборка дает экономию времени.

Основные свойства оценки \bar{y}_{st} изложены в следующих далее теоремах. Первые две теоремы относятся к расслоенному отбору вообще, а не только к расслоенному случайному отбору; иными словами, выборка из каждого слоя не обязательно должна быть простой случайной выборкой.

Теорема 5.1. Если для каждого слоя выборочная оценка \bar{y}_h есть несмещенная оценка, то \bar{y}_{st} служит несмещенной оценкой среднего для совокупности, \bar{Y} .

Доказательство.

$$E(\bar{y}_{st}) = E \frac{\sum_{h=1}^L N_h \bar{y}_h}{N} = \frac{\sum_{h=1}^L N_h \bar{Y}_h}{N},$$

* При переводе была сохранена транскрипция индексов, указывающих на способ отбора или оценивания, с помощью которых получена данная величина. Разъяснение индекса, независимо от того, дает ли его автор, а также перевод этого разъяснения на русский язык приводятся в квадратных скобках. — *Примеч. ред.*

поскольку оценки по отдельным слоям несмещенные. Но среднее значение для совокупности, \bar{Y} , можно записать в виде

$$\bar{Y} = \frac{\sum_{h=1}^L \sum_{i=1}^{N_h} y_{hi}}{N} = \frac{\sum_{h=1}^L N_h \bar{y}_h}{N}.$$

Теорема доказана.

Следствие. Поскольку при простом случайном отборе внутри слоев \bar{y}_h есть несмещенные оценки \bar{Y}_h , при расслоенном случайном отборе \bar{y}_{st} будет несмещенной оценкой \bar{Y} .

Теорема 5.2. При расслоенном отборе дисперсия \bar{y}_{st} , оценки среднего для совокупности, \bar{Y} , имеет вид

$$V(\bar{y}_{st}) = \frac{\sum_{h=1}^L N_h^2 V(\bar{y}_h)}{N^2} = \sum_{h=1}^L W_h^2 V(\bar{y}_h), \quad (5.3)$$

где

$$V(\bar{y}_h) = E(\bar{y}_h - \bar{Y}_h)^2.$$

Теорема доказывается при двух ограничениях: (а) \bar{y}_h должны быть несмещенными оценками \bar{Y}_h и (б) выборки в разных слоях должны извлекаться независимо.

Доказательство.

$$\bar{y}_{st} - \bar{Y} = \frac{\sum N_h \bar{y}_h}{N} - \frac{\sum N_h \bar{Y}_h}{N} = \frac{\sum N_h (\bar{y}_h - \bar{Y}_h)}{N}, \quad (5.4)$$

причем суммирование распространяется на все слои. Заметим, что ошибка оценки, $(\bar{y}_{st} - \bar{Y})$, выражена теперь как взвешенное среднее ошибок, сделанных при оценивании по отдельным слоям. Следовательно,

$$(\bar{y}_{st} - \bar{Y})^2 = \frac{\sum N_h^2 (\bar{y}_h - \bar{Y}_h)^2}{N^2} + \frac{2 \sum N_h N_j (\bar{y}_h - \bar{Y}_h) (\bar{y}_j - \bar{Y}_j)}{N^2},$$

причем в последнем члене справа суммирование распространяется на все пары слоев.

Теперь возьмем среднее по всем возможным выборкам. Начнем с того, что для каждого члена с несовпадающими индексами будем считать неизменной выборку в слое h и возьмем среднее по всем выборкам в слое j . Поскольку отбор по этим двум слоям происходит независимо, то какую бы выборку в слое h мы не извлекли, возможные выборки в слое j и соответствующие вероятности будут теми же самыми. Но так как \bar{y}_j предполагается несмещенной, усреднение $(\bar{y}_j - \bar{Y}_j)$ даст нуль. Следовательно, все члены с несовпадающими индексами исчезают.

Квадратичные члены дают

$$V(\bar{y}_{st}) = \frac{\sum N_h^2 E(\bar{y}_h - \bar{Y}_h)^2}{N^2} = \frac{\sum N_h^2 V(\bar{y}_h)}{N^2}.$$

Важная особенность этого результата заключается в том, что дисперсия \bar{y}_{st} зависит только от дисперсий оценок средних \bar{Y}_h для отдельных слоев. Если бы совокупность с большой вариацией значений можно было разделить на слои так, чтобы внутри каждого слоя все единицы имели одинаковые значения признака, то мы могли бы оценить \bar{Y} без всякой ошибки. Как показывает равенство (5.4), к такому результату приводит применение правильных весов N_h/N при построении оценки \bar{y}_{st} .

Теорема 5.3. При расслоенном случайном отборе дисперсия оценки \bar{y}_{st} имеет вид

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{S_h^2}{n_h} = \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h} (1 - f_h). \quad (5.5)$$

Доказательство. Поскольку \bar{y}_h представляет собой несмещенную оценку \bar{Y}_h , можно применить теорему 5.2. Кроме того, по теореме 2.2, справедливой для отдельного слоя,

$$V(\bar{y}_h) = \frac{S_h^2}{n_h} \frac{N_h - n_h}{N_h}.$$

Подставляя эти выражения в (5.3), получаем

$$\begin{aligned} V(\bar{y}_{st}) &= \frac{1}{N^2} \sum_{h=1}^L N_h^2 V(\bar{y}_h) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{S_h^2}{n_h} = \\ &= \sum W_h^2 \frac{S_h^2}{n_h} (1 - f_h). \end{aligned}$$

Некоторые частные случаи этой формулы приведены в следствиях 1, 2, 3.

Следствие 1. Если доли отбора n_h/N_h пренебрежимо малы во всех слоях, то

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum \frac{N_h^2 S_h^2}{n_h} = \sum \frac{W_h^2 S_h^2}{n_h}. \quad (5.6)$$

Эта формула применима в тех случаях, когда можно пренебречь поправками на конечность совокупности.

Следствие 2. В случае пропорционального размещения подставим в (5.5)

$$n_h = \frac{n N_h}{N}.$$

Дисперсия принимает вид

$$V(\bar{y}_{st}) = \sum \frac{N_h}{N} \frac{S_h^2}{n} \left(\frac{N-n}{N} \right) = \frac{1-f}{n} \sum W_h S_h^2. \quad (5.7)$$

Следствие 3. Если применяется пропорциональное размещение и дисперсии во всех слоях имеют одно и то же значение, S_w^2 , то получаем простой результат

$$V(\bar{y}_{st}) = \frac{S_w^2}{n} \left(\frac{N-n}{N} \right). \quad (5.8)$$

Теорема 5.4. Если $\hat{Y}_{st} = N\bar{y}_{st}$ есть оценка суммарного значения для совокупности, Y , то

$$V(\hat{Y}_{st}) = \sum N_h (N_h - n_h) \frac{S_h^2}{n_h}. \quad (5.9)$$

Это следует непосредственно из теоремы 5.3.

Таблица 5.1
ВЕЛИЧИНЫ 64 ГОРОДОВ (в тыс. жителей) в 1920 и 1930 гг.

Величина города в 1920 г. (x_{hi})				Величина города в 1930 г. (y_{hi})			
Слой (h)				Слой (h)			
1	2	3	4	1	2	3	4
797	314	172	121	900	364	209	113
773	298	172	120	822	317	183	115
748	296	163	119	781	328	163	123
734	258	162	118	805	302	253	154
588	256	161	118	670	288	232	140
577	243	159	116	1238	291	260	119
507	238	153	116	573	253	201	130
507	237	144	113	634	291	147	127
457	235	138	113	578	308	292	100
438	235	138	110	487	272	164	107
415	216	138	110	442	284	143	114
401	208	138	108	451	255	169	111
387	201	136	106	459	270	139	163
381	192	132	104	464	214	170	116
324	180	130	101	400	195	150	122
315	179	126	100	366	260	143	134

Замечание. Города расположены на обе даты в одном и том же порядке.

СУММАРНЫЕ ЗНАЧЕНИЯ И СУММЫ КВАДРАТОВ

	1920 г.		1930 г.	
Слой	$\Sigma(x_{hi})$	$\Sigma(x_{hi}^2)$	$\Sigma(y_{hi})$	$\Sigma(y_{hi}^2)$
1	8 349	4 756 619	10 070	7 145 450
2	7 941	1 474 871	9 498	2 141 729

Пример. В табл. 5.1 приведены данные 1920 и 1930 гг. о числе жителей 64 больших городов в США (в тысячах). Эти города в общем списке городов США, упорядоченном по числу жителей в 1920 г., занимали места с пятого по шестьдесят восьмое. Города разделены на два слоя, первый из которых содержит 16 наиболее крупных городов, а второй — остальные 48 городов.

Суммарное число жителей всех 64 городов в 1930 г. нужно оценить по выборке, состоящей из 24 городов. Найдем стандартную ошибку оценки суммарного числа для: (1) простой случайной выборки, \hat{Y}_{ran} , (2) расслоенной случайной выборки с пропорциональным размещением, \hat{Y}_{prop} , (3) расслоенной случайной выборки, содержащей по 12 единиц из каждого слоя, \hat{Y}_{equal} .

Эта совокупность походит на совокупности разного рода предприятий и учреждений в том отношении, что значения признака у некоторых единиц — в нашем примере числа жителей в больших городах — составляют значительную часть суммарного значения и отличаются гораздо большей колеблемостью, чем у остальных.

Суммарные значения для слоев и суммы квадратов приведены вслед за табл. 5.1. В этом примере мы пользуемся только данными 1930 г., данные 1920 г. найдут применение в другом примере.

Для всей совокупности в 1930 г. находим

$$Y = 19\,568; \quad S^2 = 52\,448.$$

1. Для простого случайного отбора согласно следствию 2 теоремы 2.2

$$V(\hat{Y}_{ran}) = \frac{N^2 S^2}{n} \frac{N-n}{N} = \frac{64^2 \cdot 52\,448}{24} \left(\frac{40}{64} \right) = 5\,594\,453.$$

Стандартная ошибка равна:

$$\sigma(\hat{Y}_{ran}) = 2365.$$

2. Дисперсии для отдельных слоев равны:

$$S_1^2 = 53\,843; \quad S_2^2 = 5581.$$

Заметим, что дисперсия для слоя, состоящего из крупных городов, приблизительно в 10 раз больше дисперсии для другого слоя.

При пропорциональном размещении имеем: $n_1 = 6$; $n_2 = 18$. Из формулы (5.7), умножая обе части равенства на N^2 , получаем

$$V(\hat{Y}_{prop}) = \frac{N-n}{n} \sum N_h S_h^2 = \frac{40}{24} [16 \cdot 53\,843 + 48 \cdot 5581] = 1\,882\,293;$$

$$\sigma(\hat{Y}_{prop}) = 1372.$$

3. При $n_1 = n_2 = 12$ воспользуемся общей формулой (5.9):

$$V(\hat{Y}_{equal}) = \sum N_h (N_h - n_h) \frac{S_h^2}{n_h} = \frac{16 \cdot 4 \cdot 53\,843}{12} + \frac{48 \cdot 36 \cdot 5581}{12} = 1\,090\,827.$$

$$\sigma(\hat{Y}_{equal}) = 1044.$$

В нашем примере оценка по выборке, содержащей одинаковое число единиц из каждого слоя, оказалась более точной, чем оценка по выборке с пропорциональным размещением. Обе они значительно точнее оценки при простом случайном отборе.

5.4. ОЦЕНКА ДИСПЕРСИИ И ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ

Если в каждом слое извлекается простая случайная выборка, то несмещенной оценкой S_h^2 (по теореме 2.4) служит

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2. \quad (5.10)$$

Таким образом, мы получаем следующую теорему.

Теорема 5.5. При расслоенном случайном отборе несмещенной оценкой дисперсии \bar{y}_{st} служит:

$$v(\bar{y}_{st}) = s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{s_h^2}{n_h}. \quad (5.11)$$

Другой, удобный для вычислений, вид этой формулы

$$s^2(\bar{y}_{st}) = \sum_{h=1}^L \frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^L \frac{W_h s_h^2}{N}. \quad (5.12)$$

Второй член в правой части равенства характеризует уменьшение дисперсии, обусловленное пкс.

Для того чтобы вычислить эту оценку, из каждого слоя должны быть извлечены, по крайней мере, две единицы. Оценивание дисперсии для случая, когда расслоение таково, что в каждом слое отбирается только одна единица, рассматривается в параграфе 5А.11.

Следствие. В некоторых случаях есть основания предполагать, что S_h^2 имеет одно и то же значение во всех слоях. Из дисперсионного анализа выборки следует, что объединенной (т. е. получаемой на основе наблюдений во всех слоях. — *Примеч. пер.*) оценкой этого общего значения дисперсии будет

$$s_w^2 = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n - L}.$$

Поскольку в таких случаях обычно применяют пропорциональный отбор, оценка дисперсии \bar{y}_{st} (согласно следствию 3 теоремы 5.3) принимает простой вид

$$v(\bar{y}_{st}) = \frac{s_w^2}{n} \frac{N - n}{N}$$

с $n - L$ степенями свободы.

Формулы для доверительных границ имеют вид:

среднее значение для совокупности:

$$\bar{y}_{st} \pm ts(\bar{y}_{st}); \quad (5.13)$$

суммарное значение для совокупности:

$$N\bar{y}_{st} \pm tNs(\bar{y}_{st}). \quad (5.14)$$

В этих формулах предполагается, что \bar{y}_{st} распределено нормально и что $s(\bar{y}_{st})$ определена без погрешностей, так что множитель t можно найти по таблицам нормального распределения.

Если каждый слой обеспечивает только небольшое число степеней свободы, то ошибки выборки, связанные с величинами типа $s(\bar{y}_{st})$, обычно учитывают, получая значения t не из таблиц нормального распределения, а из таблиц t -распределения Стьюдента. Вообще распределение $s(\bar{y}_{st})$ слишком сложно, чтобы можно было прямо применить этот прием. Приближенный метод придания $s(\bar{y}_{st})$ эффективного числа степеней свободы состоит в следующем (Satterthwaite, 1946).

Мы можем записать

$$s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L g_h s_h^2, \text{ где } g_h = \frac{N_h (N_h - n_h)}{n_h}.$$

Эффективное число степеней свободы n_e равно:

$$n_e = \frac{(\sum g_h s_h^2)^2}{\sum \frac{g_h^2 s_h^4}{n_h - 1}}. \quad (5.15)$$

Величина n_e всегда заключена между наименьшей из величин $(n_h - 1)$ и их суммой. Такое приближение учитывает тот факт, что S_h^2 может меняться от слоя к слою. Его применение требует предположения о том, что y_{hi} распределены нормально, поскольку дисперсия s_h^2 принята равной $2\sigma_h^4/(n_h - 1)$. Как следует из формулы (2.51), с. 60, если распределение y_{hi} имеет положительный эксцесс, то дисперсия s_h^2 будет больше этой величины. В этом случае формула (5.15) преувеличивает эффективное число степеней свободы.

5.5. ОПТИМАЛЬНОЕ РАЗМЕЩЕНИЕ

При расслоенном отборе величины объемов выборок n_h в соответствующих слоях определяет обследователь. Их можно выбрать так, чтобы минимизировать $V(\bar{y}_{st})$ при определенных издержках на получение выборки или минимизировать издержки при определенной величине $V(\bar{y}_{st})$.

Простейшая функция издержек имеет вид

$$\text{издержки} = C = c_0 + \sum c_h n_h. \quad (5.16)$$

Для каждого слоя издержки пропорциональны объему выборки, но издержки в расчете на одну единицу, c_h , могут меняться от слоя к слою. Член c_0 соответствует накладным расходам. Такая функция оправдана, если основную часть затрат составляют расходы на измерение каждой единицы. Если же существенную часть затрат составляют расходы на передвижение от одной единицы к другой, то, как показали практические и теоретические исследования, этим путевым расходам больше соответствует выражение $\sum t_h / \sqrt{n_h}$, где t_h — путевые расходы в расчете на единицу (Beardwood et al., 1959). Здесь рассматривается только линейная функция издержек вида (5.16).

Теорема 5.6. При расслоенном случайном отборе с функцией издержек вида (5.16) дисперсия \bar{y}_{st} , оценки среднего, минимальна, когда n_h пропорциональны $N_h S_h / \sqrt{c_h}$.

Доказательство. Задача состоит в минимизации выражения

$$V(\bar{y}_{st}) = \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} (1 - f_h) = \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^L \frac{W_h^2 S_h^2}{N_h}$$

при условии

$$c_1 n_1 + c_2 n_2 + \dots + c_L n_L = C - c_0.$$

Применяя метод множителей Лагранжа, выберем n_h и множитель λ , минимизирующие

$$V(\bar{y}_{st}) + \lambda (\sum c_h n_h - C + c_0) = \sum \frac{W_h^2 S_h^2}{n_h} - \sum \frac{W_h^2 S_h^2}{N_h} + \lambda (c_1 n_1 + c_2 n_2 + \dots + c_L n_L - C + c_0).$$

Дифференцирование по n_h приводит к уравнениям

$$-\frac{W_h^2 S_h^2}{n_h^2} + \lambda c_h = 0 \quad (h = 1, 2, \dots, L)$$

или

$$n_h \sqrt{\lambda} = \frac{W_h S_h}{\sqrt{c_h}}. \quad (5.17)$$

Суммируя по всем слоям, получаем

$$n \sqrt{\lambda} = \sum \frac{W_h S_h}{\sqrt{c_h}}. \quad (5.18)$$

Окончательно, разделив (5.17) на (5.18), получаем

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum (W_h S_h / \sqrt{c_h})} = \frac{N_h S_h / \sqrt{c_h}}{\sum (N_h S_h / \sqrt{c_h})}. \quad (5.19)$$

Эта теорема приводит к следующим правилам отбора. В данном слое берите выборку большего объема, если:

- 1) слой больше;
- 2) в слое больше вариация признака;

3) отбор в слое обходится дешевле.

Для того чтобы завершить размещение, осталось сделать еще один шаг. Уравнение (5.19) указывает n_h в долях n , однако мы еще не знаем, какое значение n взять. Решение последнего вопроса зависит от того, должна ли выборка обеспечить определенные общие издержки C или определенную дисперсию V для \bar{y}_{st} . Если неизменны издержки, то нужно подставить оптимальные значения n_h в функцию издержек (5.16) и разрешить получившееся уравнение относительно n . Это дает

$$n = \frac{(C - c_0) \sum (N_h S_h / \sqrt{c_h})}{\sum (N_h S_h \sqrt{c_h})}.$$

Если неизменна V , то нужно подставить оптимальные значения n в формулу для $V(\bar{y}_{st})$. Получаем

$$n = \frac{(\sum W_h S_h \sqrt{c_h})^2}{V + (1/N) \sum W_h^2 S_h^2},$$

где $W_h = N_h/N$.

Важный частный случай возникает при $c_h = c$, т. е. когда издержки в расчете на единицу во всех слоях одинаковы. Общие издержки принимают вид $C = c_0 + cn$, и оптимальное размещение при неизменных издержках сводится к оптимальному размещению при неизменном объеме выборки. В этом частном случае теорема 5.6 принимает следующий вид.

Теорема 5.7. При расслоенном случайном отборе с неизменным общим объемом выборки n дисперсия \bar{y}_{st} минимальна, если

$$n_h = n \frac{W_h S_h}{\sum W_h S_h} = n \frac{N_h S_h}{\sum N_h S_h}. \quad (5.20)$$

Такое размещение иногда называют *неймановым размещением*, поскольку оно приобрело известность после работы Неймана (Neuman, 1934). Как позднее обнаружилось, этот результат был получен ранее Чупровым (Tschuprow, 1923).

Минимальное значение дисперсии при неизменном n получается, если подставить значение n_h из (5.20) в общую формулу для $V(\bar{y}_{st})$. В результате имеем

$$V_{min}(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h^2 S_h^2}{N}. \quad (5.21)$$

Второй член в правой части равенства соответствует пкс.

Другое доказательство утверждения об оптимальном размещении (Stuart, 1954) опирается на неравенство Коши — Шварца*. Мини-

* В отечественной литературе это неравенство называется неравенством Коши — Буняковского. — *Примеч. пер.*

мизация V при неизменном C или минимизация C при неизменном V эквивалентны минимизации произведения

$$V' C' = \left(\sum \frac{W_h^2 S_h^2}{n_h} \right) (\sum c_h n_h),$$

поскольку V' и C' представляют собой те части V и C , которые зависят от n_h . Согласно неравенству, если a_h и b_h — два набора положительных чисел, то

$$(\sum a_h^2) (\sum b_h^2) \geq (\sum a_h b_h)^2, \quad (5.22)$$

причем равенство имеет место только тогда, когда отношение b_h/a_h постоянно для всех h . Положим

$$a_h = \frac{W_h S_h}{\sqrt{n_h}}; \quad b_h = \sqrt{c_h n_h}.$$

Из неравенства (5.22) следует

$$V' C' = \left(\sum \frac{W_h^2 S_h^2}{n_h} \right) (\sum c_h n_h) = (\sum a_h^2) (\sum b_h^2) \geq (\sum W_h S_h \sqrt{c_h n_h})^2.$$

Минимальное значение $V' C'$ достигается при

$$\frac{b_h}{a_h} = \frac{n_h \sqrt{c_h}}{W_h S_h} = \text{постоянному числу,}$$

что согласуется с теоремой 5.6.

5.6. СРАВНИТЕЛЬНАЯ ТОЧНОСТЬ РАССЛОЕННОГО СЛУЧАЙНОГО ОТБОРА И ПРОСТОГО СЛУЧАЙНОГО ОТБОРА

При рациональном применении расслоение почти всегда приводит к уменьшению дисперсии оценок средних или суммарных значений по сравнению с простой случайной выборкой того же объема. Однако неверно считать, что всякая расслоенная случайная выборка дает меньшую дисперсию, чем простая случайная выборка. Если значения n_h далеки от оптимальных, то расслоенный отбор может дать более высокую дисперсию. В действительности при неизменном общем объеме выборки более высокую дисперсию может дать даже расслоение с оптимальным размещением, хотя такой результат имеет скорее академический интерес, чем какое-либо практическое значение.

В этом параграфе мы сравним простой случайный отбор с расслоенным случайным отбором при пропорциональном и при оптимальном размещении¹. Это сравнение покажет, из чего складывается выигрыш, получаемый от расслоения. Пкс не учитывается.

Дисперсии оценок средних значений обозначаются соответственно через $V_{\text{ран}}$ — для простого случайного отбора, $V_{\text{проп}}$ — для расслоенного отбора с пропорциональным размещением и $V_{\text{опт}}$ — для расслоенного отбора с оптимальным размещением.

¹ Интересный анализ этой проблемы приведен в работах Армитаджа (Armitage, 1947) и Эванса (Evans, 1951).

Теорема 5.8. Если пренебречь величинами порядка n_h/N_h , то справедливы неравенства

$$V_{\text{опт}} \leq V_{\text{проп}} \leq V_{\text{ран}}, \quad (5.23)$$

где оптимальное размещение рассматривается при неизменном n , т. е. при $n_h \propto N_h S_h$.

Доказательство. Если пренебречь пкс, то

$$V_{\text{ран}} = \frac{S^2}{n}, \quad (5.24)$$

$$V_{\text{проп}} = \frac{\sum N_h S_h^2}{nN} \quad [\text{согласно равенству (5.7), параграф 5.3}]; \quad (5.25)$$

$$V_{\text{опт}} = \frac{(\sum N_h S_h)^2}{nN^2} \quad [\text{согласно равенству (5.21), параграф 5.5}]. \quad (5.26)$$

На основании стандартного тождества дисперсионного анализа для расслоенной совокупности имеем

$$(N-1) S^2 = \sum_h \sum_i (y_{hi} - \bar{Y})^2 = \sum_h \sum_i (y_{hi} - \bar{Y}_h)^2 + \sum_h N_h (\bar{Y}_h - \bar{Y})^2 = \\ = \sum_h (N_h - 1) S_h^2 + \sum_h N_h (\bar{Y}_h - \bar{Y})^2. \quad (5.27)$$

Поскольку члены порядка $1/N_h$ пренебрежимо малы, это равенство можно записать в виде

$$NS^2 = \sum_h N_h S_h^2 + \sum_h N_h (\bar{Y}_h - \bar{Y})^2.$$

Следовательно,

$$V_{\text{ран}} = \frac{S^2}{n} = \frac{\sum_h N_h S_h^2}{nN} + \frac{\sum_h N_h (\bar{Y}_h - \bar{Y})^2}{nN} = V_{\text{проп}} + \frac{\sum_h N_h (\bar{Y}_h - \bar{Y})^2}{nN}. \quad (5.28)$$

Согласно определению $V_{\text{опт}}$ должно выполняться неравенство $V_{\text{проп}} \geq V_{\text{опт}}$. Разность этих дисперсий равна:

$$V_{\text{проп}} - V_{\text{опт}} = \frac{1}{nN} \left[\sum_h N_h S_h^2 - \frac{(\sum_h N_h S_h)^2}{N} \right] = \\ = \frac{1}{nN} \sum_h N_h (S_h - \bar{S})^2, \quad (5.29)$$

где $\bar{S} = \sum_h N_h S_h / N$. Из (5.29) и (5.28) имеем

$$V_{\text{ран}} = V_{\text{опт}} + \frac{\sum_h N_h (S_h - \bar{S})^2}{nN} + \frac{\sum_h N_h (\bar{Y}_h - \bar{Y})^2}{nN}. \quad (5.30)$$

Из этого равенства следует, что когда мы от простого случайного отбора переходим к расслоенному с оптимальным размещением, уменьшение дисперсии складывается из двух частей. Первая из них (крайний правый член равенства) имеет своим источником исключение различий между средними значениями для слоев; вторая (средний член справа) — исключение эффекта различий между средними квадратичными откло-

нениями для слоев. Этот второй компонент отражает различие между дисперсиями при пропорциональном и оптимальном размещении.

Если пкс пренебречь нельзя, то аналогичные выкладки приводят к равенству

$$V_{\text{ran}} = V_{\text{prop}} + \frac{N-n}{nN(N-1)} \left[\sum N_h (\bar{Y}_h - \bar{Y})^2 - \frac{1}{N} \sum (N - N_h) S_h^2 \right]. \quad (5.31)$$

Отсюда следует, что дисперсия в случае расслоенного отбора с пропорциональным размещением будет больше, чем дисперсия при простом случайном отборе, если

$$\sum N_h (\bar{Y}_h - \bar{Y})^2 < \frac{1}{N} \sum (N - N_h) S_h^2. \quad (5.32)$$

Теоретически это может иметь место. Предположим, что все S_h^2 равны S_w^2 , так что пропорциональное размещение оптимально в смысле Неймана. Тогда (5.32) примет вид

$$\sum N_h (\bar{Y}_h - \bar{Y})^2 < (L-1) S_w^2$$

или

$$\frac{\sum N_h (\bar{Y}_h - \bar{Y})^2}{L-1} < S_w^2.$$

Те, кто знаком с дисперсионным анализом, узнают это неравенство, требующее, чтобы средний квадрат между слоями был меньше, чем средний квадрат внутри слоев, т. е. чтобы F -отношение было меньше единицы.

5.7. ПРИ КАКИХ УСЛОВИЯХ РАССЛОЕНИЕ ОБЕСПЕЧИВАЕТ БОЛЬШОЙ ВЫИГРЫШ В ТОЧНОСТИ?

Идеальной переменной для расслоения служит сама переменная y — признак, наблюдаемый при обследовании. Если бы мы могли производить расслоение по значениям y , то слои не пересекались бы и дисперсия внутри слоев была бы много меньше, чем вся дисперсия совокупности, особенно при большом числе слоев. Это положение иллюстрирует пример, приведенный в параграфе 5.3, с. 109. Совокупность состояла из 64 городов, была известна их величина (число жителей в 1930 г.), причем расслоение производилось по величине города. Даже при разделении совокупности только на два слоя пропорциональное размещение уменьшило стандартную ошибку \hat{Y} с 2365 до 1372. Расслоение с $n_1 = n_2 = 12$, соответствующее оптимальному нейманову размещению, привело к дальнейшему уменьшению ошибки до 1044.

Конечно, на практике мы не можем производить расслоение по самим значениям y . Но во многих важных случаях к этому можно подойти довольно близко и, следовательно, получить большой выигрыш в точности, если соблюдаются следующие три условия:

1. Совокупность состоит из объектов, величина которых варьирует в широких пределах.

2. Значения основных переменных, подлежащих наблюдению, тесно связаны с величиной объекта.

3. Обследователь располагает хорошей мерой величины объектов, по которой можно произвести расслоение.

Примерами могут служить предприятия и учреждения определенного типа, например бакалейные магазины (при обследованиях, связанных с объемом товарооборота или числом служащих), школы (при обследованиях, касающихся числа учеников), больницы (при изучении их вместимости и контингента больных), налоговые декларации (при изучении признаков, тесно связанных с величиной дохода, облагаемого налогом). В США сильно различаются по величине, если измерять ее общей площадью или валовым доходом, также фермы. Однако обычные сельскохозяйственные характеристики, такие, как объем производства тех или иных культур или того или иного вида продукции животноводства, часто обнаруживают лишь умеренную корреляцию с величиной фермы, так что выигрыш от расслоения по величине ферм невелик.

Если величина объекта остается постоянной во времени, хотя бы в течение коротких периодов, то практически наиболее удобной ее мерой обычно служит величина того же объекта в ближайший по времени момент, когда была проведена перепись. Случай, когда для расслоения имеются хорошие прежние данные, иллюстрирует пример из параграфа 5.3. В табл. 5.2 приведены S_h и оптимальные $n_h \propto N_h S_h$, полученные при размещении выборки городов по данным переписей 1920 и 1930 гг.

Таблица 5.2

ПОЛУЧЕНИЕ ОПТИМАЛЬНОГО РАЗМЕЩЕНИЯ

Слой	N_h	По данным 1920 г.			По данным 1930 г.		
		S_h	$N_h S_h$	n_h	S_h	$N_h S_h$	n_h
1	16	163,30	2612,80	11,56	232,04	3712,64	12,21
2	48	58,55	2810,40	12,44	74,71	3586,08	11,79
Итого	64		5423,20	24,00		7298,72	24,00

По данным 1920 г. n_1 равно 11,56 вместо «истинного» оптимального значения 12,21 по данным 1930 г. При округлении до целых чисел оба ряда данных приводят к одинаковому размещению: выборка состоит по 12 единиц из каждого слоя.

Отметим, что оптимальная доля отбора составляет 75% в слое 1 и только 25% в слое 2. Часто оказывается, что из-за большой колеблемости в слое, состоящем из наиболее крупных объектов, формула требует 100%-ного отбора из этого слоя. Вообще размещение может требовать даже более чем 100%-ного отбора (см. параграф 5.8). Заметим также, что S_h в 1920 г. были меньше, чем в 1930 г. Данные 1920 г. дают слишком оптимистическое представление о точности, которой можно было достигнуть при отборе городов в 1930 г. Как указывалось

в параграфе 4.6, если применяются данные прошлых лет, то всегда следует учитывать, что значения S_h могли с тех пор измениться, даже если поправка на такое изменение может быть сделана лишь ориентировочно.

Часто ради организационных удобств или необходимости получить отдельные данные по каждому слою применяется географическое расслоение, при котором слоями служат компактные территории, такие, как графства (county) или городские районы. Обычно такое расслоение сопровождается некоторым увеличением точности, поскольку существует много факторов, в силу которых люди из одного района или растения с одного участка обнаруживают сходство своих основных характеристик. Однако выигрыш от географического расслоения бывает обычно умеренным. Например, в табл. 5.3 приведены данные об эффективности географического расслоения для ряда типичных сельскохозяйственных экономических характеристик, опубликованные Джессеном (Jessen, 1942), Джессеном и Хауземаном (Jessen and Houseman, 1944).

Таблица 5.3
ОТНОСИТЕЛЬНАЯ ТОЧНОСТЬ ВЫБОРКИ (в %) ПРИ РАЗЛИЧНЫХ ВАРИАНТАХ
ГЕОГРАФИЧЕСКОГО РАССЛОЕНИЯ

Штат	Число признаков	В качестве слоя принято			
		тауншип	графство	область распространения	штат в целом
Айова, 1938 г.	18	115	100	96	91
Айова, 1939 г.	19	121	100	97	91
Флорида, 1942 г.					
Область цитрусового садоводства	14	144	100	...	
Область полеводства	15	111	100	...	
Калифорния, 1942 г.	17	113	100	97	

Таблица характеризует четыре значения слоев по величине — тауншип (township), графство, область распространения «типа сельского хозяйства» и штат. Чтобы дать некоторое представление о сравнительной величине слоев, укажем, например, что в штате Айова существовало около 1600 тауншипов, 100 графств и 5 областей распространения «типа сельского хозяйства».

Точность способа расслоения выражена в таблице величиной, обратно пропорциональной величине $V(\bar{y}_{st})$ для этого способа. Таким образом, точность способа 1 по сравнению со способом 2 измеряется отношением $V_2(\bar{y}_{st})/V_1(\bar{y}_{st})$, выраженным в процентах. Приведенные данные — это средние данные по ряду признаков, число которых указано во втором столбце. В каждом случае за основу сравнения принято расслоение по графствам. Как уже было отмечено, выигрыш в точности оказался умеренным. Для Айовы разделение совокупности на 1600 слоев (тауншипов) по сравнению с отсутствием расслоения

(штат в целом) увеличило точность приблизительно на 30%, т. е. уменьшило дисперсию приблизительно на 25%.

Что касается сравнения пропорционального расслоения с оптимальным, то существуют два случая, когда оптимальное расслоение дает значительный выигрыш. В первом случае, уже рассмотренном, совокупность состоит из больших и малых объектов и расслоение производится в соответствии с некоторой мерой их величины. Дисперсия S_h^2 обычно значительно больше для крупных объектов, чем для мелких, поэтому пропорциональное расслоение оказывается неэффективным. Второй случай относится к обследованию, в которых отбор из одних слоев обходится гораздо дороже, чем из других. Влияние множителя $\sqrt{c_h}$ может привести к тому, что пропорциональное размещение будет непригодным.

При планировании размещения выборки, для которого оценки оптимальных n_h незначительно отличаются от пропорциональных, стоит оценить, насколько увеличится $V(\bar{y}_{st})$ или $V(\bar{Y}_{st})$, если применить пропорциональное размещение. Оптимум в задаче размещения выражен довольно слабо (см. параграф 5A.2) и увеличение дисперсии может оказаться удивительно небольшим. Более того, выигрыш от применения оптимального размещения, вычисленный по оценкам величин S_h , всегда преувеличен из-за ошибок при их оценивании. Простота пропорционального размещения и его свойство равнозначности, по-видимому, оправдывают 10—20%-ное увеличение дисперсии.

5.8. РАЗМЕЩЕНИЕ, ТРЕБУЮЩЕЕ БОЛЕЕ ЧЕМ 100%-НОГО ОТБОРА

Как уже упоминалось в параграфе 5.7, формула для вычисления объемов выборок по слоям при оптимальном размещении может давать для некоторых слоев значения n_h большие, чем соответствующие N_h . Вернемся к примеру с городами, рассмотренному в параграфе 5.3. Выборка объемом в 24 города оптимально размещалась в двух слоях так, что в первом слое нужно было отобрать 12 из 16 городов, во втором — 12 из 48. Если бы объем выборки составлял 48, то для оптимального размещения потребовалось бы взять в первом слое 24 города из 16. Лучшее, что можно в таком случае сделать, это включить в выборку все города из первого слоя, а остальные 32 города взять из второго, вместо 24 требующихся по формуле. Такое положение возникает только тогда, когда общая доля отбора значительна и вариация в одном из слоев гораздо больше, чем в других. Подобные случаи встречались на практике.

Находя ожидаемую дисперсию при таком размещении или сравнивая это размещение с другими, следует соблюдать осторожность и применять надлежащие формулы. Формула (5.5) из параграфа 5.3, если в нее подставить значения n_h , соответствующие пересмотренному оптимальному размещению, остается справедливой. Формула (5.21), указывающая минимальное значение дисперсии при неизменном n ,

$$V_{\min}(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}$$

перестает быть верной! Если слой 1 — это единственный слой, в котором доля отбора должна превысить 100%, то правильная формула для V_{min} имеет вид

$$V_{min}(\bar{y}_{st}) = \frac{1}{N^2} \frac{(\sum' N_h S_h)^2}{n - N_1} - \frac{1}{N^2} \sum' N_h S_h^2,$$

где \sum' означает суммирование по всем слоям за исключением слоя 1.

5.9. ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ В СЛУЧАЕ НЕПРЕРЫВНЫХ ПЕРЕМЕННЫХ

В параграфе 5.5 были приведены формулы для определения n при предположительно оптимальном размещении выборки. В этом параграфе предлагаются формулы для любого вида размещения и выделяются некоторые особые случаи. Предполагается, что оценка имеет определенную дисперсию V . Если вместо нее определен предел ошибки d (см. параграф 4.4), то можно воспользоваться равенством $V = (d/t)^2$, где t — квантиль нормального распределения, соответствующий допустимой вероятности того, что ошибка превысит желательные пределы.

Оценивание среднего значения для совокупности, \bar{Y}

Пусть s_h — оценка S_h и пусть $n_h = w_h n$, где w_h уже выбраны. В этих обозначениях ожидаемое значение $V(\bar{y}_{st})$ по теореме 5.3 (параграф 5.3) равно:

$$V = \frac{1}{n} \sum \frac{W_h^2 s_h^2}{w_h} - \frac{1}{N} \sum W_h s_h^2 \quad (5.33)$$

где $W_h = N_h/N$. Это дает общую формулу для n :

$$n = \frac{\sum \frac{W_h^2 s_h^2}{w_h}}{V + \frac{1}{N} \sum W_h s_h^2} \quad (5.34)$$

Если пренебречь пкс, то в качестве первого приближения имеем

$$n_0 = \frac{1}{V} \sum \frac{W_h^2 s_h^2}{w_h} \quad (5.35)$$

Если n_0/N пренебречь нельзя, то можно вычислить n по формуле

$$n = \frac{n_0}{1 + \frac{1}{NV} \sum W_h s_h^2} \quad (5.36)$$

Для частных случаев формулы можно записать в различном, более удобном для вычислений, виде. Приведем некоторые из них.

Предположительно оптимальное размещение (при неизменном n):
 $w_h \propto W_h s_h$.

$$n = \frac{(\sum W_h s_h)^2}{V + \frac{1}{N} \sum W_h s_h^2} \quad (5.37)$$

Пропорциональное размещение: $w_h = W_h = N_h/N$.

$$n_0 = \frac{\sum W_h s_h^2}{V}; \quad n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (5.38)$$

Оценивание суммарного значения для совокупности

Если V — желательное значение $V(\bar{y}_{st})$, то основные формулы принимают следующий вид:

Общий случай:

$$n = \frac{\sum \frac{N_h^2 s_h^2}{w_h}}{V + \sum N_h s_h^2} \quad (5.39)$$

Предположительно оптимальное размещение (при неизменном n):

$$n = \frac{(\sum N_h s_h)^2}{V + \sum N_h s_h^2} \quad (5.40)$$

Пропорциональное размещение:

$$n_0 = \frac{N}{V} \sum N_h s_h^2; \quad n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (5.41)$$

Пример. Этот пример взят из работы Корнелла (Cornell, 1947), в которой описывается выборка колледжей и университетов США, взятая в 1946 г. Федеральным управлением просвещения (U. S. Office of Education) для того, чтобы оценить число студентов в 1946—1947 учебном году. Для иллюстрации рассматривается совокупность, включающая 196 педагогических колледжей и училищ. Они были распределены по семи слоям, из которых один небольшой слой рассматриваться не будет. Первые пять слоев образованы в соответствии с величиной колледжей, шестой состоит только из женских колледжей. Оценки s_h величины S_h были вычислены по данным 1943—1944 учебного года. Было применено «оптимальное» размещение, основанное на этих s_h .

Требовалось оценить суммарное число студентов с коэффициентом вариации 5%. В 1943 г. общее число студентов для этой группы колледжей составляло 56 472. Таким образом, желательная стандартная ошибка равна:

$$0,05 \cdot 56472 = 2824,$$

так что желательная дисперсия будет

$$V = (2824)^2 = 7\,974\,976.$$

Можно возразить, что число студентов в 1946 г. должно было быть больше, чем в 1943 г., и что на это увеличение необходимо сделать поправку. В действительности при вычислениях предполагается только, что коэффициент вариации в расчете на колледж одинаков в 1943 и в 1946 гг. — такое предположение вполне допустимо.

В табл. 5.4 приведены значения N_h , s_h и $N_h s_h$, которые были известны до определения n .

Таблица 5.4

ДАННЫЕ ДЛЯ ОПРЕДЕЛЕНИЯ ОБЪЕМА ВЫБОРКИ

Слой	N_h	s_h	$N_h s_h$	n_h
1	13	325	4 225	9
2	18	190	3 420	7
3	26	189	4 914	10
4	42	82	3 444	7
5	73	86	6 278	13
6	24	190	4 560	10
Итого	196		26 841	56

Подходящей формулой для определения n служит (5.40), которая применима при оценивании суммарного значения в случае «оптимального» размещения. Когда совокупность состоит только из 196 единиц, пкс не может быть пренебрежимо малой. Однако в целях иллюстрации, найдем первое приближение, не учитывающее пкс. Это будет

$$n_0 = \frac{(\sum N_h s_h)^2}{V} = \frac{(26\,841)^2}{7\,974\,976} = 90,34.$$

Очевидно, необходима поправка. В качестве правильного n на основании (5.40) получаем

$$n = \frac{n_0}{1 + \frac{1}{V} \sum N_h s_h^2} = \frac{90,34}{1 + \frac{4\,640\,387}{7\,974\,976}} = 57,1.$$

Была взята выборка объемом в 56 единиц¹. Значения n_h для отдельных слоев указаны в крайнем правом столбце табл. 5.4.

5.10. РАССЛОЕННЫЙ ОТБОР ДЛЯ ОЦЕНИВАНИЯ ДОЛЕЙ

Если мы хотим оценить долю единиц совокупности, относящихся к некоторому определенному классу C , то идеальное расслоение мы

¹ Числовые данные несколько отличаются от приведенных Корнеллом (Cornell, 1947).

получили бы, если бы смогли включить в первый слой все единицы из класса C , а во второй — все остальные. Не имея такой возможности, мы пытаемся образовать слои таким образом, чтобы доля единиц из класса C варьировала от слоя к слою как можно больше.

Пусть

$$P_h = \frac{A_h}{N_h}; \quad p_h = \frac{a_h}{n_h}$$

— доли единиц из класса C соответственно в слое h и в выборке из этого слоя. Естественной оценкой доли единиц для всей совокупности при расслоенном случайном отборе будет

$$p_{st} = \sum \frac{N_h p_h}{N}. \quad (5.42)$$

Теорема 5.9. При расслоенном случайном отборе дисперсия p_{st} равна:

$$V(p_{st}) = \frac{1}{N^2} \sum \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}. \quad (5.43)$$

Доказательство. Теорема представляет собой частный случай общей теоремы о дисперсии оценки среднего значения. По теореме 5.3

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{S_h^2}{n_h}.$$

Пусть y_{hi} — переменная, которая принимает значение 1, если единица принадлежит классу C , и значение 0 в противном случае. В параграфе 3.2 (равенство 3.4) было показано, что для такой переменной

$$S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h.$$

Отсюда следует утверждение теоремы.

Замечание. Почти во всех приложениях, даже если нельзя пренебречь пкс, можно отбросить члены порядка $1/N_h$ и поэтому применять несколько более простую формулу

$$V(p_{st}) = \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{P_h Q_h}{n_h} = \sum \frac{W_h^2 P_h Q_h}{n_h} (1 - f_h). \quad (5.44)$$

Следствие 1. Если пкс можно пренебречь, то

$$V(p_{st}) = \sum W_h^2 \frac{P_h Q_h}{n_h}. \quad (5.45)$$

Следствие 2. При пропорциональном размещении

$$V(p_{st}) = \frac{N - n}{N} \frac{1}{nN} \sum \frac{N_h^2 P_h Q_h}{N_h - 1} \approx \quad (5.46)$$

$$\approx \frac{1 - f}{n} \sum W_h P_h Q_h. \quad (5.47)$$

Для получения выборочной оценки дисперсии нужно в любую из приведенных формул подставить $p_h q_h / (n_h - 1)$ вместо неизвестных $P_h Q_h / n_h$.

Правило выбора n_h , минимизирующих $V(p_{st})$, следует из общей теории, изложенной в параграфе 5.5.

Минимальная дисперсия при неизменном общем объеме выборки.

$$n_h \approx N_h \sqrt{N_h / (N_h - 1)} \sqrt{P_h Q_h} \approx N_h \sqrt{P_h Q_h}.$$

Отсюда

$$n_h \approx n \frac{N_h \sqrt{P_h Q_h}}{\sum N_h \sqrt{P_h Q_h}}. \quad (5.48)$$

Минимальная дисперсия при неизменных издержках, где издержки равны $c_0 + \sum c_h n_h$.

$$n_h \approx n \frac{N_h \sqrt{P_h Q_h / c_h}}{\sum N_h \sqrt{P_h Q_h / c_h}}. \quad (5.49)$$

Значение n определяется так же, как и в параграфе 5.5.

5.11. ВЫИГРЫШ В ТОЧНОСТИ ПРИ РАССЛОЕННОМ ОТБОРЕ ДЛЯ ОЦЕНИВАНИЯ ДОЛЕЙ

Если издержки в расчете на единицу одинаковы во всех слоях, то можно указать два удобных рабочих правила: (а) расслоенный отбор дает небольшой выигрыш в точности по сравнению с простым случайным отбором, если только P_h не меняются сильно от слоя к слою; (б) оптимальное размещение при заданном n дает небольшой выигрыш по сравнению с пропорциональным, если все P_h заключены между 0,1 и 0,9.

Для иллюстрации первого утверждения в табл. 5.5 сравниваются расслоенный случайный отбор (пропорциональное размещение) и простой случайный отбор при трех слоях одинаковой величины ($W_h = 1/3$). Рассматриваются четыре случая, в первом — значения P_h для трех слоев равны 0,4; 0,5 и 0,6 и в последнем (самый крайний случай) $P_h =$

Таблица 5.5
ОТНОСИТЕЛЬНАЯ ТОЧНОСТЬ РАССЛОЕННОГО И ПРОСТОГО СЛУЧАЙНОГО ОТБОРА

P_h	Простой	Расслоенный	Относительная точность, %
	$nV(p)/(1-f) = PQ$	$nV(p_{st})/(1-f) = \frac{1}{3} \sum P_h Q_h$	
0,4; 0,5; 0,6	2 500	2 433	103
0,3; 0,5; 0,7	2 500	2 233	112
0,2; 0,5; 0,8	2 500	1 900	132
0,1; 0,5; 0,9	2 500	1 433	174

$= 0,1; 0,5$ и $0,9$. В двух средних столбцах приведены значения дисперсий оценок долей, умноженные на $n/(1-f)$, и в последнем столбце указаны значения относительной точности расслоенного случайного отбора по сравнению с простым. Выигрыш в точности велик только в двух последних случаях.

Для того чтобы сравнить при неизменном n пропорциональное размещение с оптимальным, укажем, что если пренебречь пкс, то

$$V_{opt} = \frac{(\sum W_h \sqrt{P_h Q_h})^2}{n}; \quad V_{prop} = \frac{\sum W_h P_h Q_h}{n}.$$

Следовательно, относительная точность при пропорциональном размещении по сравнению с оптимальным будет

$$\frac{V_{opt}}{V_{prop}} = \frac{(\sum W_h \sqrt{P_h Q_h})^2}{\sum W_h P_h Q_h}.$$

Если все P_h заключены между двумя значениями P_0 и $(1 - P_0)$, то нас может интересовать наименьшее возможное значение относительной точности. Для простоты рассмотрим случай двух слоев одинаковой величины ($W_1 = W_2$). Минимальная относительная точность достигается при $P_1 = \frac{1}{2}$ и $P_2 = P_0$. При этом ее значение становится равным:

$$\frac{V_{opt}}{V_{prop}} = \frac{(0,5 + \sqrt{P_0 Q_0})^2}{2(0,25 + P_0 Q_0)}.$$

Некоторые значения этой функции приведены в табл. 5.6. Даже при P_0 , равном 0,1 или 0,9, относительная точность составляет 94%. В большинстве случаев простота пропорционального размещения и его свойство равновзвешенности более чем компенсируют такую небольшую потерю в точности.

Таблица 5.6
ОТНОСИТЕЛЬНАЯ ТОЧНОСТЬ ПРИ ПРОПОРЦИОНАЛЬНОМ РАЗМЕЩЕНИИ ПО СРАВНЕНИЮ С ОПТИМАЛЬНЫМ

P_0	0,4 или 0,6	0,3 или 0,7	0,2 или 0,8	0,1 или 0,9	0,05 или 0,95
Относительная точность, %	100,0	99,8	98,8	94,1	86,6

Необходимо указать на ограничения, принятые в примере. В нем не учитывались различия в издержках на отбор в разных слоях. В некоторых обследованиях P_h очень малы, но меняются по слоям, скажем, от 0,001 до 0,05. Здесь более значительный выигрыш может дать оптимальное расслоение.

5.12. ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ ПРИ ОЦЕНИВАНИИ ДОЛЕЙ

Нужные формулы можно вывести из более общих формул, приведенных в параграфе 5.9. Пусть V — желательное значение дисперсии при оценивании доли P для всей совокупности. Формулы для двух основных типов размещения имеют следующий вид.

Пропорциональное размещение:

$$n_0 = \frac{\sum W_h P_h q_h}{V}; \quad n = \frac{n_0}{1 + \frac{n_0}{N}}. \quad (5.50)$$

Предположительно оптимальное размещение:

$$n_0 = \frac{(\sum W_h \sqrt{P_h q_h})^2}{V}; \quad n = \frac{n_0}{1 + \frac{1}{NV} \sum W_h P_h q_h}. \quad (5.51)$$

где n_0 — первое приближение, когда пкс не учитывается, и n — уточненное значение, учитывающее пкс. При выводе этих формул множители $N_h/(N_h - 1)$ считались равными единице.

Полученные результаты относятся к случаю оценивания доли. Если предпочтительно иметь дело с процентным соотношением, то в случае, когда P_h , Q_h , V и т. д. выражены в процентах, применимы те же формулы. При оценивании суммарного числа единиц совокупности, принадлежащих классу C , т. е. NP , все дисперсии умножаются на N^2 .

Упражнения

5.1. В совокупности с $N = 6$ и $L = 2$ значения y_{hi} составляют 0, 1, 2 для слоя 1 и 4, 6, 11 для слоя 2. Должна быть извлечена выборка объемом $n = 4$. (а) Покажите, что оптимальные n_h , соответствующие нейманову размещению, округленные до целых чисел, будут $n_1 = 1$ в слое 1 и $n_2 = 3$ в слое 2. (б) Вычислите оценки \bar{y}_{st} для всех выборок, которые можно получить при оптимальном размещении и при пропорциональном размещении. Проверьте, что эти оценки будут несмещенными. Пользуясь этим, найдите $V_{opt}(\bar{y}_{st})$ и $V_{prop}(\bar{y}_{st})$ непосредственно. (в) Проверьте, что полученное значение $V_{opt}(\bar{y}_{st})$ совпадает со значением, получаемым по формуле (5.5), а значение $V_{prop}(\bar{y}_{st})$ — со значением, получаемым по формуле (5.7), с. 108. (г) Применение формулы (5.21), с. 113 для вычисления $V_{opt}(\bar{y}_{st})$ дает не совсем верный ответ, поскольку при этом не учитывается то обстоятельство, что n_h были округлены до целых чисел. Насколько получаемое по этой формуле значение отличается от правильного?

5.2. Нужно провести выборочное обследование домохозяйств в городе с целью оценить среднюю стоимость движимого имущества на одно домохозяйство. Домохозяйства разделены на два слоя: с высоким и низким уровнем квартирной платы. Предполагается, что стоимость имущества на одно домохозяйство в слое с высоким уровнем квартирной платы приблизительно в девять раз больше, чем в слое с низким уровнем, и что S_h могут быть пропорциональны квадратному корню из среднего значения для слоя.

Слой с высоким уровнем квартирной платы насчитывает 4000 домохозяйств, а слой с низким — 20 000. (а) Как бы вы распределили выборку объемом в 1000 домохозяйств между двумя слоями? (б) Как нужно было бы распределить выбор-

ку, если бы мы преследовали цель оценить разность стоимостей имущества на одно домохозяйство в двух слоях?

5.3. В таблице приводятся данные о расслоении всех ферм некоторого графства по размеру фермы и о среднем числе акров под маисом на одну ферму в каждом слое.

Размер фермы, акров	Число ферм N_h	Среднее число акров под маисом \bar{Y}_h	Среднее квадратичное отклонение S_h
0—40	394	5,4	8,3
41—80	461	16,3	13,3
81—120	391	24,3	15,1
121—160	334	34,5	19,8
161—200	169	42,1	24,5
201—240	113	50,1	26,0
241—	148	63,8	35,2
Суммарное или среднее значение	2 010	26,3	

Для выборки объемом в 100 ферм вычислите объемы выборок в каждом слое при (а) пропорциональном размещении, (б) оптимальном размещении. Сравните точность этих методов с точностью при простом случайном отборе.

5.4. Докажите справедливость формулы (5.31) из параграфа 5.6:

$$V_{ran} = V_{prop} + \frac{(N-n)}{nN(N-1)} \left[\sum N_h (\bar{Y}_h - \bar{Y})^2 - \frac{1}{N} \sum (N - N_h) S_h^2 \right].$$

5.5. Некоторая совокупность разделена на два слоя с относительными объемами: W_1 , W_2 . Обследователь полагает, что S_1 , S_2 можно считать равными, но что c_2 заключено между $2c_1$ и $4c_1$. Он предпочел бы пропорциональное размещение, но не хочет, чтобы дисперсия была много больше, чем при оптимальном размещении. Покажите, что при заданных издержках $C = c_1 n_1 + c_2 n_2$, если пренебречь пкс, то

$$\frac{V_{prop}(\bar{y}_{st})}{V_{opt}(\bar{y}_{st})} = \frac{W_1 c_1 + W_2 c_2}{(W_1 \sqrt{c_1} + W_2 \sqrt{c_2})^2}.$$

Для случая $W_1 = W_2$ вычислите относительное увеличение дисперсии, вызванное применением пропорционального размещения при отношении c_2/c_1 , равном 2 и 4.

5.6. Обследователь предполагает произвести расслоенный случайный отбор. Он ожидает, что издержки на собственно обследование будут $\sum c_h n_h$. Его предварительные оценки соответствующих величин для двух слоев следующие:

Слой	W_h	S_h	C_h (в долл.)
1	0,4	10	4
2	0,6	20	9

(а) Найдите значения n_1/n и n_2/n , минимизирующие общие издержки на собственно обследование при заданном значении $V(\bar{y}_{st})$. (б) Найдите объем вы-

борки, необходимый для того, чтобы при таком оптимальном размещении $V(\bar{y}_{st}) = 1$. Пкс не учитывайте. (в) Какими будут общие издержки собственно обследования?

5.7. После того как была получена выборка, о которой говорится в упражнении 5.6, исследователь установил, что издержки на собственно обследование в расчете на одну единицу составили в действительности 2 долл. в слое 1 и 12 долл. в слое 2. (а) Насколько увеличились расходы по сравнению с предполагававшимися? (б) Если бы исследователю были известны правильные значения издержек заранее, мог ли бы он рассчитывать на $V(\bar{y}_{st}) = 1$ при такой же сумме расходов, как в упражнении 5.6?

Указание. Ответ на этот вопрос можно получить с помощью неравенства Коши — Шварца при $V = 1$ (см. с. 114), не находя нового размещения.

5.8. Некоторая совокупность разделена на два слоя. Значения W_h и S_h для этих слоев равны:

Слой	W_h	S_h
1	0,8	2
2	0,2	4

Вычислите объемы выборок n_1, n_2 для двух слоев, необходимые, чтобы удовлетворить следующим условиям. Каждый случай требует отдельных вычислений. (Пкс не учитывайте.) (а) Нужно минимизировать общий объем выборки $n = n_1 + n_2$ при условии, что стандартная ошибка оценки среднего значения для совокупности, \bar{y}_{st} , должна быть равна 0,1. (б) Стандартная ошибка оценки среднего значения в каждом слое должна быть равна 0,1. (в) Стандартная ошибка разности двух оценок средних значений для слоев должна равняться 0,1, причем, как и ранее, минимизируется общий объем выборки.

5.9. Для случая двух слоев исследователь по практическим соображениям вместо значений, задаваемых неймановым размещением, хотел бы иметь $n_1 = n_2$. Обозначим через $V(\bar{y}_{st})$ и $V_{opt}(\bar{y}_{st})$ дисперсии соответственно для случая $n_1 = n_2$ и для нейманова размещения. Покажите, что относительный прирост дисперсии равен:

$$\frac{V(\bar{y}_{st}) - V_{opt}(\bar{y}_{st})}{V_{opt}(\bar{y}_{st})} = \left(\frac{r-1}{r+1} \right)^2,$$

где $r = n_1/n_2$ вычисляется по значениям для нейманова размещения. Каким будет относительный прирост дисперсии, если вместо значений, оптимальных для слоев из упражнения 5.8, случай (а), взять $n_1 = n_2$.

5.10. Покажите, что если функция издержек имеет вид $C = c_0 + \sum t_h V n_h$, где c_0 и t_h — известные числа, то для того чтобы минимизировать $V(\bar{y}_{st})$ при неизменных общих издержках, n_h должны быть пропорциональны

$$\left(\frac{W_h^2 S_h^2}{t_h} \right)^{2/3}.$$

Найдите n_h для выборки объемом в 1000 единиц при следующих данных:

Слой	W_h	S_h	t_h
1	0,4	4	1
2	0,3	5	2
3	0,2	6	4

5.11. Пусть $V_{prop}(\bar{y}_{st})$ — дисперсия оценки среднего для расслоенной случайной выборки объема n при пропорциональном размещении и $V(\bar{y})$ — дисперсия оценки среднего для простой случайной выборки объема n . Покажите, что отношение

$$\frac{V_{prop}(\bar{y}_{st})}{V(\bar{y})}$$

не зависит от объема выборки, но отношение

$$\frac{V_{min}(\bar{y}_{st})}{V_{prop}(\bar{y}_{st})}$$

при увеличении n уменьшается. Это означает, что оптимальное при неизменном n размещение становится более эффективным по сравнению с пропорциональным размещением при увеличении n . [Воспользуйтесь формулами (5.7) и (5.21)].

5.12. Сравните значения, получаемые для $V(\bar{y}_{st})$, при пропорциональном размещении и при оптимальном размещении для выборки неизменного объема в следующих двух совокупностях. Величина всех слоев одинакова. Пкс можно пренебречь.

Совокупность 1		Совокупность 2	
Слой	P_h	Слой	P_h
1	0,1	1	0,01
2	0,5	2	0,05
3	0,9	3	0,10

Какую общую закономерность можно увидеть на примере этих двух совокупностей?

5.13. Покажите, что для случая оценивания долей утверждения, соответствующие теореме 5.8, должны быть выражены следующим образом:

$$V_{ran} = V_{prop} + \frac{\sum W_h (P_h - P)^2}{n};$$

$$V_{prop} = V_{opt} + \frac{\sum W_h (\sqrt{P_h Q_h} - \sqrt{P_h Q_h})^2}{n},$$

где

$$\sqrt{P_h Q_h} = \sum W_h \sqrt{P_h Q_h}.$$

5.14. В некоторой фирме 62% всех работающих составляют квалифицированные или неквалифицированные работники-мужчины, 31% — конторские работники-женщины и 7% — руководящие работники. Фирма хотела бы по выборке объемом в 400 работников оценить долю тех, кто пользуется во время отдыха определенным инвентарем. Согласно грубым прикидкам этим видом инвентаря пользуются от 40 до 50% мужчин, от 20 до 30% женщин и от 5 до 10% руководящих работников. (а) Как бы вы разместили выборку среди этих трех групп? (б) Если бы истинные доли составляли соответственно 48, 21 и 4%, то какой была бы стандартная ошибка оценки доли p_{st} ? (в) Какой была бы стандартная ошибка p для простой случайной выборки объемом $n = 400$?

- Armitage P. (1947). A comparison of stratified with unrestricted random sampling from a finite population. *Biometrika*, 34, 273—280.
- Beardwood J., Halton J. H. and Hammersley J. M. (1959). The shortest path through many points. *Proc. Cambridge Phil. Soc.*, 55, 299—327.
- Cornell F. G. (1947). A stratified random sample of a small finite population. *Jour. Amer. Stat. Assoc.*, 42, 523—532.
- Evans W. D. (1951). On stratification and optimum allocations. *Jour. Amer. Stat. Assoc.*, 46, 95—104.
- Jessen R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agr. Exp. Sta. Res. Bull.* 304.
- Jessen R. J. and Houseman E. E. (1944). Statistical investigations of farm sample surveys taken in Iowa, Florida and California. *Iowa Agr. Exp. Sta. Res. Bull.* 329.
- Neyman J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Jour. Roy. Stat. Soc.*, 97, 558—606.
- Stuart A. (1954). A simple presentation of optimum sampling results. *Jour. Roy. Stat. Soc. B*, 16, 239—241.
- Satterthwaite F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110—114.
- Tschuprow A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2, 461—493, 646—683.

ДРУГИЕ ПРОБЛЕМЫ РАССЛОЕННОГО ОТБОРА

5А.1. ЭФФЕКТ ОТКЛОНЕНИЙ ОТ ОПТИМАЛЬНОГО РАЗМЕЩЕНИЯ

В следующих параграфах рассматривается ряд частных вопросов практического применения расслоенного отбора. Параграфы с 5А.1 по 5А.7 посвящены проблемам, возникающим при планировании выборки, а параграфы с 5А.8 по 5А.12 — методам анализа полученных данных, включая упрощенные методы вычисления стандартных ошибок. В заключение дается первоначальное представление о некоторых полезных результатах для тех случаев, когда обследование проводится в аналитических целях (параграф 5А.13). В настоящем параграфе рассматриваются потери в точности, вызываемые отклонением фактического размещения выборки от оптимального.

Предположим, что мы собираемся воспользоваться оптимальным размещением при заданном n . Объем выборки n'_h из слоя h должен быть равен:

$$n'_h = \frac{n(W_h S_h)}{\sum W_h S_h}. \quad (5А.1)$$

Согласно равенству (5.21, с. 113), соответствующая минимальная дисперсия равна:

$$V_{\min}(\bar{y}_{st}) = \frac{1}{n} (\sum W_h S_h)^2 - \frac{1}{N} \sum W_h S_h^2. \quad (5А.2)$$

На практике, поскольку S_h не известны, мы можем воспользоваться размещением лишь приближенным к оптимальному. Если \hat{n}_h — объем выборки, извлекаемый из слоя h , то фактически получаемая дисперсия согласно равенству (5.5, с. 107) равна:

$$V(\bar{y}_{st}) = \sum \frac{W_h^2 S_h^2}{\hat{n}_h} - \frac{1}{N} \sum W_h S_h^2. \quad (5А.3)$$

Увеличение дисперсии, вызванное неоптимальностью размещения, составит

$$V(\bar{y}_{st}) - V_{\min}(\bar{y}_{st}) = \sum \frac{W_h^2 S_h^2}{\hat{n}_h} - \frac{1}{n} (\sum W_h S_h)^2.$$

Таблица 5А.1
ЭФФЕКТ ОТКЛОНЕНИЙ ОТ ОПТИМАЛЬНОГО РАЗМЕЩЕНИЯ

Слой	n'_h (опт.)	\hat{n}_h (факт.)	$\frac{ \hat{n}_h - n'_h }{\hat{n}_h}$	$\frac{(\hat{n}_h - n'_h)^2}{\hat{n}_h}$
1	200	150	0,33	16,7
2	100	120	0,17	3,3
3	40	70	0,43	12,9
Итого	340	340	—	32,9

Подставим в первый член правой части этого равенства значения $W_h S_h$, выраженные через n'_h , из (5А.1). Это даст интересный результат:

$$V(\bar{y}_{st}) - V_{min}(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n^2} \left(\sum \frac{n'^2_h}{\hat{n}_h} - n \right) = \frac{(\sum W_h S_h)^2}{n^2} \sum \frac{(\hat{n}_h - n'_h)^2}{\hat{n}_h}. \quad (5А.4)$$

Возвращаясь к равенству 5А.2, заметим, что если пкс (последним членом в правой части равенства) можно пренебречь, то

$$\frac{V_{min}(\bar{y}_{st})}{n} = \frac{(\sum W_h S_h)^2}{n^2}.$$

Следовательно, относительное увеличение дисперсии, вызванное отклонением от оптимального размещения, будет

$$\frac{V(\bar{y}_{st}) - V_{min}(\bar{y}_{st})}{V_{min}(\bar{y}_{st})} = \frac{1}{n} \sum_{h=1}^L \frac{(\hat{n}_h - n'_h)^2}{\hat{n}_h}, \quad (5А.5)$$

где \hat{n}_h — фактический, а n'_h — оптимальный объем выборки из слоя h . Если пкс пренебречь нельзя, то знак $=$ в (5А.5) заменяется знаком \geq .

Практическую значимость этого результата трудно представить без числовых примеров. Полезно одно общее, хотя и несколько ограничивающее, следствие из последнего равенства. Пусть g — наибольшая из величин $|\hat{n}_h - n'_h|/\hat{n}_h$ по всем слоям. Тогда из (5А.5) следует, что

$$\frac{V - V_{min}}{V_{min}} \leq \frac{1}{n} \sum \frac{\hat{n}_h^2 g^2}{\hat{n}_h} = g^2.$$

Например, если максимальное отклонение $|\hat{n}_h - n'_h|$, выраженное в долях \hat{n}_h , равно 0,2, или 20%, то относительное увеличение дисперсии не может превысить $(0,2)^2 = 0,04$, или 4%. Если $g = 30\%$, то это

относительное увеличение не превысит 9%. В этом смысле можно сказать, что оптимум выражен слабо.

Это правило довольно грубо и обычно дает увеличение дисперсии значительно больше действительного. В табл. 5А.1 приведен пример с тремя слоями при $n = 340$. Оптимальное размещение требует объемов выборок 200, 100 и 40, тогда как фактические объемы выборок равны 150, 120 и 70.

Поскольку величина g равна 0,43 (слой 3), согласно нашему общему правилу относительное увеличение дисперсии составляет 18%. Из данных крайнего правого столбца следует, что в действительности дисперсия увеличивается на $32,9/340 = 9,7\%$.

Эванс (Evans, 1951) изучал тот же вопрос с точки зрения эффекта ошибок при оценивании S_h и разработал приближенное правило, которое показывает, можно ли считать, что предположительно оптимальное размещение будет более точным, чем пропорциональное размещение. Он предположил, что коэффициент вариации оценки S_h одинаков для всех слоев. Это предположение оправдано, если значения S_h уже оценены на основании некоторой предварительной выборки одинакового объема во всех слоях. Эванс показал, как вычислить объем такой предварительной выборки, необходимый для того, чтобы «оптимальное» размещение было в среднем лучше пропорционального. Ранее Сукхатм (Sukhatme, 1935) показал, что небольшая предварительная выборка обычно дает большую вероятность того, что «оптимальное» размещение будет предпочтительнее простого случайного отбора.

5А.2. ЭФФЕКТ ОШИБОК В ЗНАЧЕНИЯХ ВЕЛИЧИНЫ СЛОЕВ

Объемы слоев, N_h , необходимые для желательного вида расслоения, могут быть известны неточно, если они были получены по уже устаревшим данным переписи. Вместо истинных весов слоев, W_h , мы располагаем их оценками w_h . Выборочной оценкой \bar{Y} служит $\sum w_h \bar{y}_h$.

Укажем (в общих чертах) последствия применения ошибочных весов.

1. Выборочная оценка оказывается смещенной. По этой причине мы измеряем точность оценки с помощью среднего квадрата ошибки относительно \bar{Y} , а не с помощью дисперсии оценки относительно ее собственного среднего (см. параграф 1.8).

2. При увеличении объема выборки смещение остается неизменным. Следовательно, всегда существует некоторый объем выборки, для которого оценка становится менее точной, чем при простом случайном отборе, и весь выигрыш в точности от расслоения теряется.

3. Обычная оценка, $s(\bar{y}_{st})$, преуменьшает истинную ошибку \bar{y}_{st} , поскольку она не учитывает слагаемого общей ошибки, возникающего из-за смещения.

Для того чтобы обосновать эти утверждения, заметим, что при многократном отборе среднее значение оценки равно $\sum w_h \bar{Y}_h$. Следовательно, смещение составляет

$$\sum (w_h - W_h) \bar{Y}_h.$$

СРАВНЕНИЕ ЗНАЧЕНИЙ $V(\bar{y})$

n	Простой случайный отбор	Расслоенный случайный отбор	
		(а)	(б)
50	0,0200	0,0186	0,0074
100	0,0100	0,0095	0,0055
200	0,0050	0,0049	0,0045
300	0,0033	0,0034	0,0042
400	0,0025	0,0027	0,0041
1000	0,0010	0,0013	0,0038

Оно не зависит от объема выборки. Нетрудно убедиться, что при нахождении среднего квадрата ошибки (СКО) оценки член, соответствующий дисперсии, можно получить по обычной формуле, подставляя w_h вместо W_h . Следовательно,

$$\text{СКО}(\bar{y}_{st}) = \sum \frac{w_h^2 S_h^2}{n_h} (1 - f_h) + [\sum (w_h - W_h) \bar{Y}_h]^2. \quad (5A.6)$$

Это выражение было найдено Стиваном (Stephan, 1941). Остается добавить, что обычная формула для $s^2(\bar{y}_{st})$ очевидно, представляет собой несмещенную оценку первого члена в правой части (5A.6), но не учитывает второй член.

Пример. Этот пример иллюстрирует потерю в точности вследствие неправильных весов в случаях, когда расслоение (а) малоэффективно, (б) высокоэффективно. Рассмотрим большую совокупность с $S^2 = 1$, разделенную на два слоя так, что $W_1 = 0,9$ и $W_2 = 0,1$. Предположим, что $S_1 = S_2 = S_h$. Тогда, пренебрегая членами порядка $1/N_h$, имеем

$$S^2 \approx \sum W_h S_h^2 + \sum W_h (\bar{Y}_h - \bar{Y})^2 = S_h^2 + W_1 W_2 (\bar{Y}_1 - \bar{Y}_2)^2. \quad (5A.7)$$

т. е.

$$1 = S_h^2 + 0,09 (\bar{Y}_1 - \bar{Y}_2)^2.$$

Для случая (а) положим $\bar{Y}_1 - \bar{Y}_2 = 1$. Тогда $S_h^2 = 0,91$ и пропорциональное расслоение по сравнению с простым случайным отбором уменьшает дисперсию на 9%.

Для случая (б) положим $\bar{Y}_1 - \bar{Y}_2 = 3$, откуда $S_h^2 = 0,19$ и уменьшение дисперсии составляет более 80%.

Для случая двух слоев смещение можно записать в виде

$$(w_1 - W_1) (\bar{Y}_1 - \bar{Y}_2),$$

так как $(w_1 - W_1) = -(w_2 - W_2)$. Предположим, что оценки весов равны $w_1 = 0,92$ и $w_2 = 0,08$. Смещение составляет $0,02 \cdot 1 = 0,02$ для (а) и $0,06$ для (б). Таким образом, при выборке объема n мы имеем для сравнения следующие дисперсии.

$$\text{Простой случайный отбор: } V(\bar{y}) = \frac{1}{n}.$$

Расслоенный случайный отбор:

$$(а) \quad V(\bar{y}_{st}) = \frac{0,91}{n} + 0,0004;$$

$$(б) \quad V(\bar{y}_{st}) = \frac{0,19}{n} + 0,0036.$$

В случае (а) простой случайный отбор оказывается предпочтительнее начиная с $n = 300$. Однако до $n = 1000$ между этими двумя способами нет особого различия.

В случае (б), когда способы различаются более существенно, до $n = 200$ расслоение предпочтительнее, хотя уже при выборке такого объема большая часть потенциального выигрыша теряется. Для n , больших 300, расслоенный отбор становится заметно хуже простого случайного отбора. Таким образом, достоверно оценить W_h особенно важно, когда расслоение высокоэффективно или когда объем выборки велик.

В некоторых обследованиях для того, чтобы оценить W_h , можно взять предварительно большую выборку объема n' . Такой метод, известный под названием *двойного отбора* или *двухфазного отбора*, имеет широкое применение, он рассматривается в гл. 12. В ней будет показано, что при применении двойного отбора средний квадрат ошибки \bar{y}_{st} приближенно равен

$$\frac{\sum W_h S_h^2}{n} + \frac{\sum W_h (\bar{Y}_h - \bar{Y})^2}{n'}.$$

Сравнивая этот СКО с выражением для S^2/n , которое дает равенство (5A.7), мы видим, что большая часть выигрыша от расслоения сохраняется при условии, что n' гораздо больше n . В более общем виде можно утверждать, что набор задаваемых весов сохраняет большую часть потенциального выигрыша от расслоения, если эти веса оценены гораздо более достоверно, чем если бы они были оценены по простой случайной выборке объема n .

5А.3. ПРОБЛЕМА РАЗМЕЩЕНИЯ ПРИ ИЗУЧЕНИИ НЕСКОЛЬКИХ ПРИЗНАКОВ

Поскольку наилучшее размещение для одного признака не будет, вообще говоря, наилучшим для другого, при обследовании совокупности по нескольким признакам нужно принимать некоторое компромиссное решение. Первый шаг в этом направлении состоит в том, чтобы сократить число признаков, учитываемых при размещении, до сравнительно небольшого, оставив только наиболее важные. Если имеются надежные данные прошлых обследований, то мы можем после этого вычислить оптимальные размещения для каждого признака отдельно и посмотреть, до какой степени они расходятся. В специализированных

обследованиях связь между признаками может быть очень тесной и размещения могут различаться сравнительно мало.

Пример. Данные, приведенные Джессеном (Jessen, 1942), относятся к такого рода сельскохозяйственному обследованию. Штат Айова был разделен на пять географических районов, обозначенных в соответствии с основным направлением сельского хозяйства в каждом из них. Предположим, что эти районы должны служить слоями при обследовании молочного животноводства. Наибольший интерес представляют три признака: среднеедневное число дойных коров, среднее число галлонов молока за день и общая прибыль, полученная за год от сбыта молочной продукции. Оценки средних квадратичных отклонений внутри слоев, s_h , сделанные по данным обследования 1938 г., приведены в табл. 5А.3.

Таблица 5А.3

СРЕДНИЕ КВАДРАТИЧНЫЕ ОТКЛОНЕНИЯ ВНУТРИ СЛОЕВ				
Слой	$W_h = \frac{N_h}{N}$	\bar{s}_h для числа коров	s_h для числа галлонов молока	s_h для прибыли от сбыта молочной продукции (в долл.)
Северо-восточный молочный	0,197	4,6	11,7	332
Зерновой	0,191	3,4	9,8	357
Западный животноводческий	0,219	3,3	7,0	246
Южный пастбищный	0,184	2,8	6,5	173
Восточный животноводческий	0,208	3,7	9,8	279

Таблица 5А.4

ОБЪЕМЫ ВЫБОРОК ВНУТРИ СЛОЕВ (n=1000)					
Слой	Размещение				среднее m _h
	пропор- циональное	оптимальное для			
		числа коров	числа галлонов	прибыли	
Северо-восточный молочный	197	254	258	236	250
Зерновой	191	182	209	246	212
Западный животноводческий	219	203	171	194	189
Южный пастбищный	184	145	134	115	131
Восточный животноводчес- кий	208	216	228	209	218

В табл. 5А.4 указаны основанные на этих s_h оптимальные наименьшие размещения по отдельным признакам для выборки объемом в 1000 ферм.

Оптимальные размещения для отдельных признаков отличаются одно от другого лишь умеренно. За единственным исключением, все

Таблица 5А.5

ОЖИДАЕМЫЕ ДИСПЕРСИИ ОЦЕНКИ СРЕДНЕГО

Вид размещения	Для числа коров	Для числа галлонов	Для прибыли
Оптимальное	0,0127	0,0800	76,9
Компромиссное	0,0128	0,0802	77,6
Пропорциональное	0,0131	0,0837	80,9

три отклоняются от пропорционального размещения в одну и ту же сторону. Так, для первого слоя пропорциональное размещение требует 197 ферм, а размещения для отдельных признаков — от 236 до 258 ферм. Средние из оптимальных объемов выборок для трех признаков, приведенные в последнем столбце табл. 5А.4, обеспечивают удовлетворительное компромиссное размещение.

В табл. 5А.5 указаны ожидаемые выборочные дисперсии \bar{y}_{st} , вычисленные по отдельным признакам для оптимального (opt), компромиссного (comp) и пропорционального (prop) размещений. Соответствующие формулы имеют вид:

$$v_{opt} = \frac{(\sum W_h s_h)^2}{n}, \quad v_{comp} = \sum \frac{(W_h s_h)^2}{m_h},$$

$$v_{prop} = \frac{\sum W_h s_h^2}{n}.$$

Компромиссное размещение дает почти столь же точные результаты, как и отдельные оптимальные размещения для каждого признака, если бы они были возможны. Еще более примечательно, что пропорциональное размещение лишь незначительно менее точно, чем компромиссное или оптимальные для отдельных признаков. Кроме того, в табл. 5А.5 точность оптимальных и компромиссного размещений преувеличена, поскольку эти размещения получены на основе выборочных оценок дисперсий. Все это еще раз подтверждает, что оптимум в задаче размещения слабо выражен, о чем упоминалось в параграфе 5А.1.

5А.4. ДРУГИЕ СПОСОБЫ РАЗМЕЩЕНИЯ ПРИ ИЗУЧЕНИИ НЕСКОЛЬКИХ ПРИЗНАКОВ

В некоторых обследованиях оптимальные размещения для отдельных признаков различаются столь сильно, что очевидного компромиссного решения не существует. Все же необходимо иметь некоторый принцип, который позволил бы задавать единое размещение, хотя понятно, что никакой принцип не будет, по-видимому, универсальным. Далее излагаются два полезных подхода к этой проблеме, предложенных Йейтсом (Yates, 1960).

Первый относится к обследованиям специализированного характера, для которых потери, связанные с определенным значением ошибки оценки, можно представить в денежном выражении или через полезность. Такой подход уже рассматривался в параграфе 4.9. При v

переменных и квадратичной функции потерь целесообразно выразить суммарные ожидаемые потери в виде линейной функции

$$L(n_h) = a_1 V_1 + a_2 V_2 + \dots + a_v V_v, \quad (5A.8)$$

где a_j — некоторые заранее известные числа и $V_j = V(\bar{y}_{st})$ для j -й переменной. При линейной функции издержек на отбор имеем

$$C = c_0 + \sum c_h n_h. \quad (5A.9)$$

Необходимо определить n_h , при которых $(C + L)$ минимально. С помощью обычных методов математического анализа находим

$$n_h = \frac{W_h}{V c_h} \sqrt{\sum_{j=1}^v a_j S_{jh}^2}, \quad (5A.10)$$

где S_{jh}^2 — дисперсия j -й переменной в слое h .

При втором подходе мы задаем желательную величину стандартной ошибки или дисперсии, V_j ($j=1, 2, \dots, v$), для каждой переменной. Если мы должны оценивать средние значения для совокупности, это означает, что

$$\sum_{h=1}^L \frac{W_h^2 S_{jh}^2}{n_h} - \sum_{h=1}^L \frac{W_h S_{jh}^2}{N} \leq V_j \quad (j=1, 2, \dots, v). \quad (5A.11)$$

Мы пользуемся знаком неравенства, потому что наиболее экономичное размещение может для некоторых признаков дать меньшую дисперсию, чем желаемая.

При этом подходе минимизируются издержки (равенство 5A.9) при заданных V_j .

Первый шаг состоит в том, чтобы для каждой отдельной переменной определить оптимальное размещение и найти издержки при соответствующей дисперсии этой переменной. Далее, возьмем переменную, скажем y_1 , для которой издержки C_1 наибольшие, и проверим, не будет ли оптимальное размещение для y_1 удовлетворять всем остальным ($v-1$) ограничениям. Если это так, то мы принимаем это размещение и задача решена, поскольку никакое другое размещение не будет удовлетворять ограничению на V_1 для y_1 при издержках, меньших чем C_1 .

Если же это размещение не удовлетворяет некоторым ограничениям, то задача осложняется. Далениус (Dalenius, 1957) предложил остроумный графический метод решения для случая, когда имеется только два слоя. Более общий математический подход был предложен Йейтсом (Yates, 1960). Эти методы иллюстрируются следующими примерами.

Пример 1. (Два слоя, три переменных.) Значения W_h и S_{jh} приведены в столбцах 1—4 табл. 5A.6. Предполагается, что пкс можно пренебречь и что c_h — постоянная. Дополнительные вычисления, требующиеся в том случае, когда эти предположения не выполняются, крайне незначительны. При оптимальном размещении для j -й переменной

$$V(\bar{y}_{st}) = \frac{(\sum W_h S_{jh})^2}{n}.$$

В столбцах 5—7 приведены данные для вычисления отдельных оптимальных дисперсий.

Таблица 5A.6
УСЛОВНАЯ СОВОКУПНОСТЬ С ДВУМЯ СЛОЯМИ И ТРЕМЯ ПЕРЕМЕННЫМИ

Номер столбца	1	2	3	4	5	6	7	8	9
Слой	W_h	S_{1h}	S_{2h}	S_{3h}	$W_h S_{1h}$	$W_h S_{2h}$	$W_h S_{3h}$	$(стб.6)^2 / (стб.5)$	$(стб.7)^2 / (стб.5)$
1	0,8	4	2	1	3,2	1,6	0,8	0,8	0,2
2	0,2	4	6	8	0,8	1,2	1,6	1,8	3,2
Итого					4,0	2,8	2,4	2,6	3,4

Случай 1. Он служит примером ситуации, когда существует простое решение. Предположим, что для каждой оценки желательная стандартная ошибка равна 0,1, так что каждое $V_j = 0,01$. По итогам столбцов 5—7 видно, что наибольшая выборка требуется для первой переменной: $(4,0)^2/0,01 = 1600$. Из данных столбца 5 следует, что соответствующим оптимальным размещением будет ${}_1n_1 = 1280$, ${}_1n_2 = 320$, где левый индекс означает, что размещение оптимально для y_1 .

Посмотрим теперь, будет ли это решение удовлетворять двум другим ограничениям. Для этого воспользуемся следующим общим соотношением: если принимается оптимальное размещение для j -й переменной, то дисперсия для k -й переменной равна:

$${}_jV(\bar{y}_{st}) = \sum_{h=1}^L \frac{W_h^2 S_{kh}^2}{f_h^2} = \frac{\sum W_h S_{kh}}{n} \left(\sum \frac{W_h^2 S_{jh}^2}{W_h S_{jh}} \right), \quad (5A.12)$$

так как $f_h^2 = n W_h S_{jh} / \sum W_h S_{jh}$. Мы применим это соотношение при $j=1, k=2, 3$. В столбцах 8 и 9 по данным столбцов 5, 6 и 7 вычислены нужные величины. Отсюда

$${}_1V(\bar{y}_{st}) = \frac{4,0 \cdot 2,6}{1600} = 0,0065; \quad {}_1V(\bar{y}_{st}) = \frac{4,0 \cdot 3,4}{1600} = 0,0085.$$

Рассматриваемое размещение легко удовлетворяет обоим ограничениям.

Случай 2. Для тех же данных положим желательные стандартные ошибки равными 0,1 для y_1 и y_2 и лишь 0,08 для y_3 , так что $V_3 = 0,0064$. Полученное ранее решение здесь не годится, так как для него $V_3 = 0,0085$. Применим графический метод Далениуса. Из неравенства (5A.11) следует, что для любой конкретной переменной значения n_1, n_2 , удовлетворяющие ограничению со знаком равенства, лежат на гиперболе относительно n_1 и n_2 . На рис. 5A.1 показаны три гиперболы для нашей задачи. Например, для y_1 уравнение гиперболы имеет вид

$$\frac{(W_1 S_{11})^2}{n_1} + \frac{(W_2 S_{12})^2}{n_2} = \frac{10,24}{n_1} + \frac{0,64}{n_2} = 0,01.$$

Область, в которой выполняются все три ограничения, расположена выше и правее пунктирных линий AB , BC . В этой области мы ищем точку, для которой $n_1 + n_2$ минимально. Ею, очевидно, будет точка B . Следовательно, наш график дает решение $n_1 = 1200$, $n_2 = 430$, $n = 1630$. Это решение можно найти чисто арифметически, поскольку для него ограничения для переменных y_1 и y_2 выполняются со знаком равенства. Читатель может проверить, что арифметическое решение дает значения $n_1 = 1200$, $n_2 = 437$.

В случае двух слоев, графический метод применим при любом числе переменных. Более сложную ситуацию, когда число слоев больше двух, иллюстрирует пример 2.

Пример 2. (Четыре слоя, две переменных.) Исходные данные приведены в столбцах 1—3 табл. 5А.7. Необходимо найти наименьший объем выборки, при котором

$$V_1 \leq 0,04; V_2 \leq 0,01.$$

Таблица 5А.7
Условная совокупность с четырьмя слоями и двумя переменными

Номер столбца	1	2	3	4	5	6	7
Слой	W_h	S_{1h}	S_{2h}	$W_h S_{1h}$	$W_h S_{2h}$	(стб. 5) ² : (стб. 4)	(стб. 4) ² : (стб. 5)
1	0,4	5	1	2,0	0,4	0,08	10,00
2	0,3	5	2	1,5	0,6	0,24	3,75
3	0,2	5	4	1,0	0,8	0,64	1,25
4	0,1	5	8	0,5	0,8	1,28	0,31
Итого				5,0	2,6	2,24	15,31

Как и ранее, сначала находим оптимальные размещения и соответствующие объемы выборок для каждой переменной. По данным столбцов 4 и 5 имеем

$${}_1V(\bar{y}_{1st}) = \frac{25}{n}; \quad {}_1n = \frac{25}{0,04} = 625;$$

$${}_2V(\bar{y}_{2st}) = \frac{6,76}{n}; \quad {}_2n = \frac{6,76}{0,01} = 676.$$

Из равенства (5А.12) и данных столбца 7 следует, что если принять размещение 2 (т. е. размещение, удовлетворяющее ограничению по переменной 2) с $n = 676$, то дисперсия y_1 будет равна:

$${}_2V(\bar{y}_{1st}) = \frac{2,6 \cdot 15,31}{676} = \frac{39,81}{676} = 0,0589.$$

Это больше, чем значение 0,04, заданное для V_1 .

Значит, следует искать компромиссное размещение, в точности удовлетворяющее обоим ограничениям. Вводя множители Лагранжа λ_1 и λ_2 , находим значения n_h , минимизирующие

$$\sum_{h=1}^L c_h n_h + \lambda_1 \sum_{h=1}^L \frac{W_h^2 S_{1h}^2}{n_h} + \lambda_2 \sum_{h=1}^L \frac{W_h^2 S_{2h}^2}{n_h}.$$

В нашем примере $c_h = 1$, но далее рассматривается общий случай. Дифференцирование по n_h приводит к

$$n_h = \frac{n \sqrt{\lambda_1 \frac{W_h^2 S_{1h}^2}{c_h} + \lambda_2 \frac{W_h^2 S_{2h}^2}{c_h}}}{\sum \sqrt{\lambda_1 \frac{W_h^2 S_{1h}^2}{c_h} + \lambda_2 \frac{W_h^2 S_{2h}^2}{c_h}}}. \quad (5А.13)$$

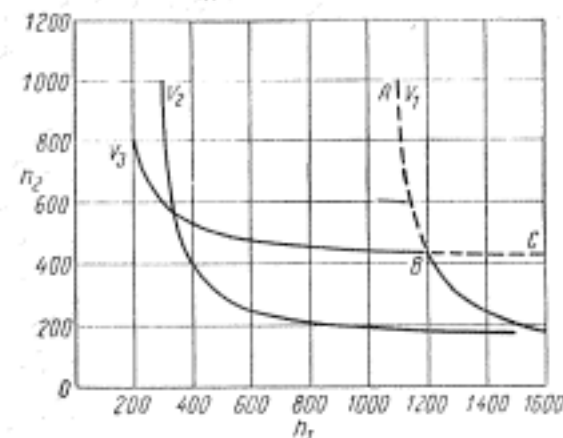


Рис. 5А.1. Графическое решение задачи размещения выборки (три переменных, два слоя)

Для определения n_h нужно найти величины λ_1 , λ_2 и n , удовлетворяющие ограничениям на обе дисперсии и условию минимальных издержек. Поскольку простого и точного решения не существует, необходимо принять тот или иной метод последовательного приближения. Мы применим подход, при котором сначала определяются n_h .

Как следует из (5А.13), оптимальные n_h представляют собой некоторую комбинацию S_{1h}^2 и S_{2h}^2 , взвешенных тем или иным образом. Понятно, что λ_1 и λ_2 входят в (5А.13) только в виде отношения λ_1/λ_2 . Так как S_{1h}^2 и S_{2h}^2 могут сильно различаться по величине, найти хорошее первое приближение для λ_1/λ_2 довольно трудно. Чтобы облегчить эту задачу, сделаем некоторую замену переменных.

Если λ_1/λ_2 устремить к бесконечности, то n_h в (5А.13) станет равным n_h , оптимальному для размещения 1. Аналогично, если $\lambda_1/\lambda_2 \rightarrow 0$,

то $n_h = {}_2n_h$. Отсюда следует, что для значения $\lambda = \lambda_1/(\lambda_1 + \lambda_2)$, соответствующего искомым λ_1, λ_2 , (5A.13) эквивалентно

$$n_h = \frac{n \sqrt{\lambda ({}_1n_h)^2 + (1-\lambda) ({}_2n_h)^2}}{\sum \sqrt{\lambda ({}_1n_h)^2 + (1-\lambda) ({}_2n_h)^2}}. \quad (5A.14)$$

Для произвольного значения λ запишем:

$$V(\bar{y}_{1st}) = \frac{\phi_1(\lambda)}{n}; \quad V(\bar{y}_{2st}) = \frac{\phi_2(\lambda)}{n}.$$

Мы хотим найти λ и n такие, что

$$\frac{\phi_1(\lambda)}{n} = V_1 = 0,04; \quad \frac{\phi_2(\lambda)}{n} = V_2 = 0,01. \quad (5A.15)$$

По прежним вычислениям мы знаем, что при $\lambda = 1$ $\phi_1(\lambda) = 25$ представляет собой минимальное значение функции $\phi_1(\lambda)$ и что при $\lambda = 0$ $\phi_1(\lambda) = 39,81$. В качестве приближения примем, что $\phi_1(\lambda)$ — парабола по λ , с вершиной при $\lambda = 1$. Это дает

$$\frac{\phi_1(\lambda)}{n} \approx \frac{25 + 14,81(1-\lambda)^2}{n} = 0,04. \quad (5A.16)$$

Для $\phi_2(\lambda)$ минимум достигается при $\lambda = 0$ и равен 6,76. Значение $\phi_2(\lambda)$ при $\lambda = 1$ можно найти, вычислив $V(\bar{y}_{2st})$ для размещения 1. По итогам столбцов 4 и 6 таблицы 5A.7 получаем $\phi_2(1) = 5,0 \cdot 2,24 = 11,20$. Приближение параболой дает

$$\frac{\phi_2(\lambda)}{n} = \frac{6,76 + 4,44\lambda^2}{n} = 0,01. \quad (5A.17)$$

Уравнения 5A.16 и 5A.17 легко решаются и дают в качестве первого приближения $\lambda = 0,41$ и $n = 751$.

Следующий шаг состоит в вычислении n_h (обозначим их ${}_h n_h$) для размещения, соответствующего полученным n и λ . После чего мы находим объемы выборок, необходимые для того, чтобы удовлетворить ограничениям на дисперсии по каждой переменной. Если найденное размещение близко к оптимальному, то эти два объема будут почти одинаковыми (и, как мы ожидаем, близкими к $n = 751$).

Величины ${}_h n_h$ получаем из уравнения (5A.14). Если

$$r_h = \sqrt{\lambda ({}_1n_h/n)^2 + (1-\lambda) ({}_2n_h/n)^2},$$

то (5A.14) можно записать в виде

$$\frac{n_h}{n} = \frac{r_h}{\sum r_h}.$$

Такая запись удобна, поскольку по правилу размещения Неймана

$$\frac{{}_1n_h}{n} = \frac{W_h S_{1h}}{\sum W_h S_{1h}}; \quad \frac{{}_2n_h}{n} = \frac{W_h S_{2h}}{\sum W_h S_{2h}}$$

и величины $W_h S_{1h}, W_h S_{2h}$ уже вычислены и записаны в столбцах 4 и 5 табл. 5A.7.

Таблица 5A.8

ПРОВЕРКА ПЕРВОГО ПРИБЛИЖЕНИЯ К ОПТИМАЛЬНОМУ РАЗМЕЩЕНИЮ

Номер столбца	1	2	3	4	5	6	7
Слой	$\left(\frac{{}_1n_h}{n}\right)^2$	$\left(\frac{{}_2n_h}{n}\right)^2$	r_h	$\frac{r_h}{\sum r_h} = \frac{n_h}{n}$	$\frac{W_h^2 S_{1h}^2}{n_h/n}$	$\frac{W_h^2 S_{2h}^2}{n_h/n}$	n_h
1	0,16	0,0225	0,2808	0,2652	15,08	0,60	194
2	0,09	0,0529	0,2610	0,2465	9,13	1,46	180
3	0,04	0,0961	0,2704	0,2554	3,92	2,51	187
4	0,01	0,0961	0,2466	0,2329	1,07	2,75	171
Итого			1,0588	1,0000	29,20	7,32	732

В табл. 5A.8 приведены результаты оставшихся вычислений, в столбце 4 указаны n_h/n . Для отыскания соответствующих дисперсий для выборки объема n воспользуемся равенством

$$\lambda V(\bar{y}_{1st}) = \sum \frac{W_h^2 S_{1h}^2}{n_h} = \frac{1}{n} \sum \frac{W_h^2 S_{1h}^2}{n_h/n}.$$

Величины $W_h^2 S_{1h}^2/n_h/n$ приведены в столбцах 5 и 6. Из итогов по столбцам имеем

$$\lambda V(\bar{y}_{1st}) = \frac{29,20}{n} = 0,04; \quad n = 730;$$

$$\lambda V(\bar{y}_{2st}) = \frac{7,32}{n} = 0,01; \quad n = 732.$$

Два полученных значения n столь близки, что мы принимаем это размещение и полагаем $n = 732$. Значения n_h , приведенные в столбце 7, получены умножением данных столбца 4 на 732.

Если два значения n , полученные при первом приближении, сильно различаются, то необходимо вычислить второе приближение для λ и n либо графически, либо по параболическим функциям, пользуясь уже вычисленными значениями $\phi_1(\lambda)$ и $\phi_2(\lambda)$. В случае двух переменных этот метод применим при любом числе слоев.

Для случая, когда как число переменных, так и число слоев больше двух, хорошего метода вычислений, в сущности, нет. Здесь могут оказаться полезными некоторые приемы математического программирования. Возникают также и более сложные проблемы; например, может оказаться желательным задать как общие дисперсии для совокупности, так и верхние границы дисперсий для некоторых ее подразделений.

5A.5. РАССЛОЕНИЕ ПО ДВУМ ПРИЗНАКАМ ДЛЯ НЕБОЛЬШИХ ВЫБОРОК

Предположим, что имеются два критерия расслоения, скажем, по R строкам и по C столбцам, в результате чего получается RC клеток. Если $n \geq RC$, то в выборке может быть представлена каждая клетка. Осложнение возникает, когда $n < RC$, а мы хотели бы, чтобы каждый критерий расслоения получил в выборке пропорциональное представительство. Брайант, Хартли и Джессен (Bryant, Hartley and Jessen,

Таблица 5A.9
ЧИСЛО И ДОЛЯ ШКОЛ В КАЖДОЙ КЛЕТКЕ

Величина города	Расходы на одного ученика					Итого	$n_{i.}$	
	A	B	C	D				
I	m_{1j}	15	21	17	9	$m_{1.}$	62	4
	P_{1j}	0,091	0,127	0,103	0,055	$P_{1.}$		
II	m_{2j}	10	8	13	7	$m_{2.}$	38	2
	P_{2j}	0,061	0,049	0,079	0,042	$P_{2.}$		
III	m_{3j}	6	9	5	8	$m_{3.}$	28	2
	P_{3j}	0,036	0,055	0,030	0,049	$P_{3.}$		
IV	m_{4j}	4	3	6	6	$m_{4.}$	19	1
	P_{4j}	0,024	0,018	0,036	0,036	$P_{4.}$		
V	m_{5j}	3	2	5	8	$m_{5.}$	18	1
	P_{5j}	0,018	0,012	0,030	0,049	$P_{5.}$		
Итого	$m_{.j}$	38	43	46	38	165		
	$P_{.j}$	0,230	0,261	0,278	0,231	1,000		
$n_{.j}$		2	3	3	2			

Таблица 5A.10
КВАДРАТ 10×10 ДЛЯ ИЗВЛЕЧЕНИЯ ВЫБОРКИ

Строка		Столбец							
		1 A	2	3 B	4	5	6 C	7	8 D
1	I	×							
2				×					
3			×						
4							×		
5	II					×			
6							×		
7	III			×					
8								×	
9	IV				×				
10	V								×

1960) разработали простой метод, единственное требование которого состоит в том, чтобы n превосходило наибольшее из чисел R и C .

Для иллюстрации этого метода рассмотрим небольшую совокупность из 165 школ, подразделенную на пять групп по величине города и на четыре группы по средней величине расходов на одного ученика. Числа школ m_{ij} и доли их общего числа $P_{ij} = m_{ij}/165$ в каждой из 20 клеток приведены в табл. 5A.9.

Задача состоит в том, чтобы каждая школа имела приблизительно одинаковые шансы быть отобранной и в то же время каждая из групп была представлена в выборке пропорционально своей численности. В нашем примере $n = 10$. Вычислим значения $n_{i.} = nP_{i.}$ и $n_{.j} = nP_{.j}$ и округлим эти произведения до ближайших целых чисел (с небольшой дополнительной поправкой, если нужно, так, чтобы суммы как $n_{i.}$, так и $n_{.j}$ были равны n). Эти числа приведены в табл. 5A.9.

Следующий шаг состоит в отборе $n = 10$ клеток с вероятностью быть отобранной для ij -й клетки, равной $n_{i.}n_{.j}/n^2$ *. Для этого строится квадрат $n \times n$ (табл. 5A.10). В строке 1 случайным образом отбирается один столбец. В строке 2 случайным образом отбирается один из оставшихся столбцов и т. д. **. (Быстрее всего это можно сделать с помощью случайной перестановки чисел от 1 до 10.) Результаты одного конкретного извлечения отмечены в табл. 5A.10 крестиками.

Заметим, что слою A, соответствующему группе A, приписаны столбцы 1 и 2, поскольку $n_{.1} = 2$. Аналогично слою 1, соответствующему группе I, приписаны строки с 1 по 4, так как $n_{1.} = 4$ и т. д. Тем самым выборка размещается по 20 клеткам. В более компактной форме это размещение записано в табл. 5A.11. После этого в клетке IA случайным образом отбираются две школы из 15 и т. д. Вероятность быть отобранной для некоторой школы из i -й строки и j -го столбца (табл. 5A.9) пропорциональна $n_{i.}n_{.j}/P_{ij}$ ***. Таким образом, вероятности быть отобранными для всех школ не равны, хотя они будут приблизительно одинаковыми, если $P_{ij} \approx n_{i.}n_{.j}/n^2$.

Несмещенной оценкой среднего значения на одну школу служит

$$\bar{y}_0 = \frac{1}{n} \sum \frac{n_{i.}n_{.j}}{P_{ij}} y_{ij},$$

где y_{ij} — суммарное значение для выборки в ij -й клетке. Если, однако, $P_{ij} \approx n_{i.}n_{.j}/n^2$, то выборочное среднее \bar{y} будет вероятно предпо-

*В действительности, в дальнейшем идет речь о таком отборе $n = 10$ клеток из $RC = 20$, при котором каждая клетка может попадать в выборку несколько раз и среднее число таких попаданий для ij -й клетки равно $n_{i.}n_{.j}/n$. — Примеч. пер.

** В конце концов мы получим n пар строк и столбцов $(i_1, i_2), (i_2, i_3), \dots, (i_n, i_1)$. В этих парах участвуют все строки и столбцы, каждая строка и каждый столбец только один раз. Эти пары определяют номера извлекаемых клеток. — Примеч. пер.

*** Если ij -я клетка при описанной процедуре попадает в выборку объемом в 10 клеток k раз, то для некоторой школы из этой клетки вероятность быть отобранной пропорциональна k/P_{ij} и, следовательно, безусловная вероятность быть отобранной пропорциональна среднему числу попаданий, деленному на P_{ij} , т. е. $n_{i.}n_{.j}/P_{ij}$. — Примеч. пер.

Таблица 5А.11

РАЗМЕЩЕНИЕ ВЫБОРКИ ПО 20 КЛЕТКАМ

	A	B	C	D	Итого
I	2	1	1	0	4
II	0	0	2	0	2
III	0	1	0	1	2
IV	0	1	0	0	1
V	0	0	0	1	1
Итого	2	3	3	2	10

читательнее, поскольку для него в этом случае смещение будет пренебрежимо мало. Выборочную оценку дисперсии можно применять как для несмещенной, так и для смещенной оценки, при условии что n , по крайней мере, вдвое больше R и C и что в каждой строке и в каждом столбце отбирается не меньше двух единиц.

Если в некоторых клетках P_{ij} заметно отличается от $n_{i.}n_{.j}/n^2$, то нужно принять дополнительные меры к тому, чтобы сделать вероятность извлечения каждой школы ближе к постоянной. После вычисления $n_{i.}$ и $n_{.j}$, нужно рассмотреть величины $D_{ij} = nP_{ij} - n_{i.}n_{.j}/n$, округлив их до целых чисел. Если для какой-нибудь клетки D_{ij} — положительное целое число, то этой клетке автоматически приписывается D_{ij} единиц для отбора. После этого величины $n_{i.}$ и $n_{.j}$ уменьшаются в соответствии с этим фиксированным размещением, а затем применяется описанная процедура размещения.

Ранее совокупность приемов для решения описанной в этом параграфе задачи, включая случай, когда имеется значительное число пустых клеток, было названо Гудменом и Кишем (Goodman and Kish, 1950) *контролируемым отбором*. В их трактовке этого метода строки отражают расслоение по основному признаку и из каждой строки извлекается по одной единице. Гудмен и Киш показали, как найти ограниченное число допустимых размещений (каждое с соответствующей вероятностью), таких, чтобы клетки отбирались с вероятностями P_{ij} .

5А.6. ФОРМИРОВАНИЕ СЛОЕВ

В связи с этой темой возникает несколько вопросов. По какому признаку лучше всего производить расслоение? Как нужно определять границы между слоями? Сколько должно быть слоев?

Для отдельного признака или отдельной переменной y расслоение, очевидно, лучше всего производить по распределению частот самой величины y . Следующая по удобству характеристика — это, по-видимому, распределение частот некоторой другой величины, сильно коррелированной с y . Уравнения для определения наилучших границ слоев при пропорциональном и при неймановом размещении, когда число слоев задано, были получены Далениусом (Dalenius, 1957), а ря-

дом исследователей были предложены быстросходящиеся приближенные методы их решения. Мы рассмотрим нейманово размещение, потому что в тех совокупностях, где выигрыш от расслоения наиболее значителен, это размещение обычно предпочтительнее пропорционального. Сначала предполагается, что слои образуются по значениям самой переменной y .

Пусть y_0, y_L — наименьшее и наибольшее значения y в совокупности. Задача состоит в отыскании таких промежуточных границ слоев y_1, y_2, \dots, y_{L-1} , при которых

$$V(\bar{y}_{st}) = \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \quad (5A.18)$$

минимальна. Если пкс можно пренебречь, то достаточно минимизировать $\sum W_h S_h$. Поскольку y_h встречается в этой сумме только в членах $W_h S_h$ и $W_{h+1} S_{h+1}$, имеем

$$\frac{\partial}{\partial y_h} (\sum W_h S_h) = \frac{\partial}{\partial y_h} (W_h S_h) + \frac{\partial}{\partial y_h} (W_{h+1} S_{h+1}).$$

Если $f(y)$ — функция распределения частот y , то

$$W_h = \int_{y_{h-1}}^{y_h} f(t) dt; \quad \frac{\partial W_h}{\partial y_h} = f(y_h). \quad (5A.19)$$

Далее,

$$W_h S_h^2 = \int_{y_{h-1}}^{y_h} t^2 f(t) dt - \frac{\left[\int_{y_{h-1}}^{y_h} t f(t) dt \right]^2}{\int_{y_{h-1}}^{y_h} f(t) dt}. \quad (5A.20)$$

Дифференцируя (5A.20), получаем

$$S_h^2 \frac{\partial W_h}{\partial y_h} + 2W_h S_h \frac{\partial S_h}{\partial y_h} = y_h^2 f(y_h) - 2y_h \mu_h f(y_h) + \mu_h^2 f(y_h),$$

где μ_h — среднее значение y в слое h . Добавим $S_h^2 \partial W_h / \partial y_h$ к левой части равенства и равную ей величину $S_h^2 f(y_h)$ — к правой. Это дает после деления на $2S_h$

$$\frac{\partial (W_h S_h)}{\partial y_h} = S_h \frac{\partial W_h}{\partial y_h} + W_h \frac{\partial S_h}{\partial y_h} = \frac{1}{2} f(y_h) \frac{(y_h - \mu_h)^2 + S_h^2}{S_h}.$$

Аналогично находим

$$\frac{\partial (W_{h+1} S_{h+1})}{\partial y_h} = -\frac{1}{2} f(y_h) \frac{(y_h - \mu_{h+1})^2 + S_{h+1}^2}{S_{h+1}}.$$

Следовательно, уравнения для вычисления y_h имеют вид

$$\frac{(y_h - \mu_h)^2 + S_h^2}{S_h} = \frac{(y_h - \mu_{h+1})^2 + S_{h+1}^2}{S_{h+1}} \quad (h = 1, 2, \dots, L-1). \quad (5A.21)$$

Таблица 5А.12
ВЫЧИСЛЕНИЕ ГРАНИЦ СЛОЕВ С ПОМОЩЬЮ ПРАВИЛА НАКОПЛЕННЫХ

ЗНАЧЕНИЙ $\sqrt{f(y)}$					
Доля про- мышленных ссуд в общей сумме ссуд (в %)	$f(y)$	Накопленные значения $\sqrt{f(y)}$	Доля про- мышленных ссуд в общей сумме ссуд (в %)	$f(y)$	Накопленные значения $\sqrt{f(y)}$
0—5	3 464	58,9	50—55	126	340,3
5—10	2 516	109,1	55—60	107	350,6
10—15	2 157	155,5	60—65	82	359,7
15—20	1 581	195,3	65—70	50	366,8
20—25	1 142	229,1	70—75	39	373,0
25—30	746	256,4	75—80	25	378,0
30—35	512	279,0	80—85	16	382,0
35—40	376	298,4	85—90	19	386,4
40—45	265	314,7	90—95	2	387,8
45—50	207	329,1	95—100	3	389,5

К сожалению, эти уравнения плохо приспособлены для практического счета, поскольку как μ_h , так и S_h зависят от y_h . Мы изложим простой приближенный метод Далениуса и Ходжеса (Dalenius and Hodges, 1959). Пусть

$$Z(y) = \int_{y_0}^y \sqrt{f(t)} dt.$$

Если слоев много и границы их достаточно узки, то $f(y)$ внутри данного слоя можно приближенно считать постоянной величиной. Следова-

$$W_h = \int_{y_{h-1}}^{y_h} f(t) dt \approx f_h(y_h - y_{h-1});$$

$$S_h \approx \frac{1}{\sqrt{12}} (y_h - y_{h-1});$$

$$Z_h - Z_{h-1} = \int_{y_{h-1}}^{y_h} \sqrt{f(t)} dt \approx \sqrt{f_h} (y_h - y_{h-1}),$$

где f_h — «постоянное» значение $f(y)$ в слое h . Пользуясь этими приближенными равенствами, находим

$$\sqrt{12} \sum_{h=1}^L W_h S_h \approx \sum_{h=1}^L f_h (y_h - y_{h-1})^2 \approx \sum_{h=1}^L (Z_h - Z_{h-1})^2. \quad (5A.22)$$

Так как $(Z_L - Z_0)$ неизменно, то легко проверить, что сумма в правой части равенства минимальна, если все $(Z_h - Z_{h-1})$ равны между собой.

Таким образом, при заданном $f(y)$ можно построить ряд накопленных значений $\sqrt{f(y)}$ и выбрать y_h так, чтобы они образовали на шкале накопленных значений равные интервалы. Табл. 5А.12 иллюстрирует применение этого правила.

Пример. Имеются данные о распределении совокупности 13 435 банков США по величине процента промышленных ссуд в общей сумме ссуд (McEvoy, 1956). Распределение асимметрично и его мода расположена на левом конце интервала значений. В столбце накопленных значений \sqrt{f} , $58,9 = \sqrt{3464}$, $109,1 = \sqrt{3464} + \sqrt{2516}$ и т. д.

Предположим, что мы хотим сформировать пять слоев. Поскольку максимальное накопленное значение \sqrt{f} равно 389,5, значениями границ на этой шкале должны быть 77,9; 155,8; 233,7 и 311,6. Ближайшие подходящие значения указаны ниже:

	Слой				
	1	2	3	4	5
Границы слоя	0—5%	5—15%	15—25%	25—45%	45—100%
Интервал накопленных значений \sqrt{f}	58,9	96,6	73,6	85,6	74,8

Первые два интервала, 58,9 и 96,6, довольно сильно различаются, но их нельзя улучшить без дальнейшего подразделения исходных групп.

Если интервалы групп исходного распределения y разной длины, то методика разбиения шкалы немного меняется. Если длина интервала меняется от d к ud , то значение \sqrt{f} для второго интервала при построении ряда накопленных значений \sqrt{f} умножается на \sqrt{u} .

Хотя описанное правило получено при довольно грубых с математической точки зрения предположениях, оно хорошо зарекомендовало себя в приложении как к теоретическим, так и к реальным распределениям (Cochran, 1961). При другом оправдавшем себя методе (Ekman, 1959) границы слоев образуются так, чтобы $W_h (y_h - y_{h-1})$ были постоянными.

Наше приближенное правило имеет интересное следствие. Из уравнения (5А.22) следует, что это правило эквивалентно тому, чтобы делать $W_h S_h$ приблизительно одинаковым, как предложили Далениус и Герни (Dalenius and Gurney, 1951). Но при постоянном $W_h S_h$ равномерное размещение дает постоянный объем выборки во всех слоях $n_h = n/L$. Поскольку оптимум мало чувствителен к вариации n_h (см. параграф 5А.1), применение правила накопленных значений \sqrt{f} и извлечение в образованных таким путем слоях выборок одинакового объема дает большой эффект.

До сих пор мы исходили из нереалистичного предположения о том, что расслоение можно производить по значениям самой переменной y . На практике для этого пользуются некоторой другой переменной x (например, величиной y по данным последней переписи). Далениус

(Dalenius, 1957) исследовал уравнения для определения границ x , минимизирующих $\sum W_h S_{y|h}$, если известна регрессия y по x . Если эта регрессия нелинейна, то такие границы могут значительно отличаться от оптимальных для случая, когда наблюдению подлежит сама переменная x . Однако из уравнений следует, что если регрессия y по x линейна и между y и x внутри всех слоев существует тесная корреляционная связь, то границы слоев в обоих случаях будут приблизительно одинаковы. Пусть

$$y = \alpha + \beta x + e,$$

где $E(e) = 0$ для всех x , а e и x некоррелированы. Дисперсия e внутри слоя h равна S_{eh}^2 . Тогда границы по x , при которых $V(\bar{y}_{st})$ минимальна, удовлетворяют уравнениям (Dalenius, 1957):

$$\frac{\beta^2 [(x_h - \mu_{xh})^2 + S_{xh}^2] + 2S_{eh}^2}{\beta S_{xh} \sqrt{1 + S_{eh}^2/\beta^2 S_{xh}^2}} = \frac{\beta^2 [(x_{h+1} - \mu_{x, h+1})^2 + S_{x, h+1}^2] + 2S_{e, h+1}^2}{\beta S_{x, h+1} \sqrt{1 + S_{e, h+1}^2/\beta^2 S_{x, h+1}^2}}.$$

Если $S_{eh}^2/\beta^2 S_{xh}^2$ малы для всех h , то эти уравнения сводятся к уравнению (5A.21), из которого определяются оптимальные границы для x . В общем случае $S_{eh}^2/\beta^2 S_{xh}^2 = (1 - \rho_h^2)/\rho_h^2$, где ρ_h — коэффициент корреляции между y и x в слое h .

Несмотря на необходимость дальнейшего исследования из полученного результата можно сделать вывод о том, что правило накопленных значений \sqrt{f} , примененное к x , должно обеспечивать эффективное расслоение для другой переменной y , которая имеет линейную регрессию по x при сильной корреляции. Некоторые численные результаты Кокрена (Cochran, 1961) подтверждают это предположение. Более того, если ρ_h принимают лишь умеренные значения, как бывает при большом числе слоев, то отклонение от оптимальных границ для x будет иметь менее заметное влияние на точность оценивания по y .

Предыдущие рассуждения, конечно, относились в основном к выборочному исследованию объектов, расслоенных по некоторой мере их величины. Полученные результаты применимы также и в тех случаях, когда главный интерес в обследовании представляет изучение нескольких переменных при условии, что все они в той или иной степени связаны с одной и той же мерой величины. Предположим, например, что некоторые переменные приблизительно пропорциональны какой-либо мере величины, другие пропорциональны квадратному корню из нее, а остальные от нее почти не зависят. Границы, определенные по правилу накопленных значений \sqrt{f} , будут приблизительно оптимальными для первой группы переменных и довольно хорошими для второй (для которой, впрочем, выигрыш от расслоения в любом случае невелик). Для третьей группы переменных точность может несколько снизиться из-за применения неравных долей отбора.

Иное положение возникает в том случае, когда одна группа переменных тесно связана с определенной мерой величины, а вторая группа переменных тесно связана с некоторой другой мерой, причем распределения частот в обоих случаях заметно различаются. В этом случае применим общий подход, изложенный в параграфе 5A.4, однако хорошие

численные методы для получения границ, обеспечивающих желательные уровни дисперсий, еще не разработаны.

При географическом расслоении проблема еще менее поддается точному математическому анализу, поскольку формировать слои можно самыми различными способами. Обычный прием состоит в том, что выбирается небольшое число переменных, сильно коррелированных с основными изучаемыми при обследовании признаками, после чего путем логического анализа в сочетании с методом проб и ошибок определяются границы слоев, наиболее подходящие для этих основных переменных. Поскольку выигрыш от расслоения в этом случае будет, вероятно, невелик, не имеет особого смысла затрачивать большие усилия на улучшение этих границ. Принципы расслоения при изучении экономических характеристик рассматривали Стивен (Stephan, 1941) и Хегуд и Бернерт (Hagood and Bernert, 1945), а при изучении сельскохозяйственных — Кинг и Маккарти (King and McCarty, 1941).

5A.7. ЧИСЛО СЛОЕВ

Решение о том, каким должно быть число слоев L , зависит от ответов на два вопроса: (а) В какой мере уменьшается дисперсия при увеличении L ? (б) Как влияет увеличение L на издержки на проведение обследования?

Что касается (а), то предположим сначала, что слои образуются по значениям y . Рассмотрим простейший случай, когда y распределено равномерно в интервале $(a, a + d)$. Тогда S_y^2 до расслоения равна $d^2/12$, так что для простой случайной выборки объема n $V(\bar{y}) = d^2/12n$. Если образованы L слоев одинаковой величины, то дисперсия внутри каждого слоя равна $S_{yh}^2 = d^2/12L^2$. Следовательно, для расслоенной выборки при $W_h = 1/L$ и $n_h = n/L$

$$V(\bar{y}_{st}) = \frac{1}{n} \left(\sum_{h=1}^L W_h S_{yh} \right)^2 = \frac{1}{n} \left(\sum_{h=1}^L \frac{1}{L} \cdot \frac{d}{\sqrt{12L}} \right)^2 = \frac{d^2}{12nL^2} = \frac{V(\bar{y})}{L^2}.$$

Таким образом, для равномерного распределения дисперсия уменьшается обратно пропорционально квадрату числа слоев. Весьма примечательно, что такая зависимость сохраняется приблизительно и для асимметричных распределений с конечным размахом вариации, когда при их расслоении выбираются оптимальные границы для нейманова размещения. Кокрен (Cochran, 1961) нашел, что для восьми распределений переменных того вида, который, по-видимому, часто встречается на практике, средние значения $V(\bar{y}_{st})/V(\bar{y})$ при $L = 2, 3, 4$ были соответственно 0,232; 0,098 и 0,053 по сравнению с 0,250; 0,111 и 0,062 для равномерного распределения.

Тот же анализ, из которого следует выгода увеличения числа слоев, дает обманчивую картину в том случае, когда для формирования слоев применяется некоторая другая переменная x . Если $\phi(x) = E(y|x)$ есть регрессия y по x , то мы можем записать

$$y = \phi(x) + e,$$

Таблица 5А.13
 $V(\bar{y}_{st})/V(\bar{y})$ КАК ФУНКЦИЯ L ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИОННОЙ
 МОДЕЛИ И НЕКОТОРЫХ РЕАЛЬНЫХ ДАННЫХ

L	Линейная регрессионная модель				Реальные данные, ряд		
	$\rho =$				1	2	3
	0,99	0,95	0,90	0,85			
2	0,265	0,323	0,392	0,458	0,197	0,295	0,500
3	0,129	0,198	0,280	0,358	0,108	0,178	0,375
4	0,081	0,154	0,241	0,323	0,075	0,142	0,244
5	0,059	0,134	0,222	0,306	0,065	0,106	0,241
6	0,047	0,123	0,212	0,298	0,050	0,104	0,212
∞	0,020	0,098	0,190	0,277	—	—	—

ВИД ДАННЫХ

Ряд	Данные	x	y	Источник
1	Число студентов в колледжах	1952	1958	Cochran (1961)
2	Величина городов	1940	1950	Cochran (1961)
3	Доходы семей	1929	1933	Dalenius and Gurney (1951)

где ϕ и ϵ некоррелированы. Следовательно,

$$S_y^2 = S_\phi^2 + S_\epsilon^2.$$

Из предыдущих рассуждений следует, что создание L оптимальных слоев по x может уменьшить S_ϕ^2 до S_ϕ^2/L^2 , если $\phi(x)$ линейна, или в несколько меньшей степени, если $\phi(x)$ нелинейна. Но расслоение по x не уменьшает члена S_ϵ^2 . При увеличении L рано или поздно наступит момент, когда член S_ϵ^2 будет преобладать. Дальнейшее увеличение L приведет лишь к незначительному относительному уменьшению $V(\bar{y}_{st})$.

Насколько быстро достигается точка, после которой увеличение числа слоев дает все меньший выигрыш, зависит от нескольких факторов, в частности, от соотношения между S_ϵ^2 и S_ϕ^2 и от вида $\phi(x)$. В литературе приводится всего лишь несколько примеров, основанных на реальных данных. Дополним их, применив сравнительно простой теоретический подход. Предположим, что выбор оптимальных границ слоев на основании переменной x при выборках равного объема n/L в каждом слое уменьшает $V(\bar{x}_{st})$ пропорционально $1/L^2$. Таким образом,

$$V(\bar{x}_{st}) = \frac{L}{n} \sum_{h=1}^L W_h^2 S_{xh}^2 = \frac{S_x^2}{nL^2}. \quad (5A.23)$$

Предположим также, что регрессия y по x линейна, т. е.

$$y = \alpha + \beta x + \epsilon,$$

причем S_ϵ^2 постоянная. Тогда

$$V(\bar{y}_{st}) = \frac{L}{n} \sum_{h=1}^L W_h^2 S_{yh}^2 = \frac{L\beta^2}{n} \sum_{h=1}^L W_h^2 S_{xh}^2 + \frac{LS_\epsilon^2}{n} \sum_{h=1}^L W_h^2.$$

Для любого набора L слоев $\sum W_h^2 \geq \frac{1}{L}$. Пользуясь (5A.23), получаем

$$V(\bar{y}_{st}) \geq \frac{1}{n} \left(\frac{\beta^2 S_x^2}{L^2} + S_\epsilon^2 \right) = \frac{S_y^2}{n} \left[\frac{\rho^2}{L^2} + (1 - \rho^2) \right], \quad (5A.24)$$

где ρ — коэффициент корреляции между y и x в нерасслоенной совокупности.

Для этой модели в табл. 5А.13 указаны $V(\bar{y}_{st})/V(\bar{y})$ при $\rho = 0,99; 0,95; 0,90$ и $0,85$ и при L от 2 до 6 в предположении, что в (5A.24) имеет место равенство. В трех столбцах правой части таблицы приведены $V(\bar{y}_{st})/V(\bar{y})$ для трех рядов реальных данных, указанных под таблицей. Переменной x для них служат значения y на некоторый предшествующий момент.

Результаты для регрессионной модели показывают, что при ρ , не превосходящих 0,95, для $L > 6$ можно ожидать лишь небольшого уменьшения дисперсии. Ряды реальных данных 2 и 3 подтверждают этот вывод, хотя для данных о числе студентов в колледжах (ряд 1) может оказаться выгодным некоторое дальнейшее увеличение L .

Для того чтобы закончить наш анализ, необходимо ввести функцию, показывающую, как зависят от L издержки на проведение обследования. Далениус (Dalenius, 1957) предложил соотношение $C = LC_s + nC_n$. Отношение издержек C_s/C_n будет меняться в зависимости от вида обследования. Увеличение числа слоев требует дополнительных усилий при планировании обследования и при извлечении выборки и увеличивает число весов, применяемых при вычислении оценок, если только оценки не равновзвешенные. В некоторых обследованиях в организации собственно наблюдения не требуется почти никаких изменений, зато в других — работу в каждом слое нужно организовать особо. Независимо от вида функции издержек, из результатов табл. 5А.13 следует, что если увеличение числа слоев сверх 6 вызовет необходимость существенно уменьшить n , чтобы сохранить издержки на обследование неизменными, то такое увеличение вряд ли принесет выгоду.

Наши рассуждения в этом параграфе ограничивались обследованиями, в которых оценки требуются только для совокупности в целом. Если мы хотим получить оценки также для географических подразделений совокупности, то доводы за увеличение числа слоев становятся более весомыми.

5А.8. РАССЛОЕНИЕ ПОСЛЕ ИЗВЛЕЧЕНИЯ ВЫБОРКИ

В отношении некоторых, удобных для расслоения, характеристик слоев, к которому принадлежит та или иная единица, нельзя указать заранее, до того как данные собраны. Типичными примерами могут слу-

жить такие характеристики человека, как возраст, пол, раса и уровень образования. Объемы слоев N_h могут быть довольно точно известны по данным официальной статистики, но распределить по слоям сами единицы можно только после получения выборки.

Можно, например, извлечь простую случайную выборку объема n и затем распределить отобранные единицы по слоям. Вместо выборочного среднего \bar{y} , мы пользуемся в этом случае оценкой $\bar{y}_w = \sum W_h \bar{y}_h$, где \bar{y}_h — среднее значение по единицам выборки, попавшим в слой h , и $W_h = N_h/N$. Этот способ обладает почти такой же точностью, как и пропорциональный расслоенный отбор при условии, что (а) выборка достаточно велика, скажем, в каждый слой попадает более 20 единиц, и (б) ошибками в значениях весов W_h можно пренебречь (см. параграф 5А.2).

Чтобы показать это, введем m_h — число единиц выборки, попавших в слой h , причем m_h , вообще говоря, меняется от выборки к выборке. Для выборок, в которых m_h неизменно,

$$V(\bar{y}_w) = \sum \frac{W_h^2 S_h^2}{m_h} - \frac{1}{N} \sum W_h S_h^2.$$

Теперь нужно вычислить среднее значение этой величины для многократных выборок объема n . При этом необходима некоторая осторожность, поскольку одна или несколько из величин m_h могут быть равны нулю. Если это произошло, то прежде чем строить оценку, следовало бы объединить два или более слоев, причем оценка стала бы немного менее точной. Однако при увеличении n вероятность того, что какая-либо из m_h равна нулю, становится столь малой, что влиянием этого обстоятельства на дисперсию можно пренебречь.

Для случая, когда обращение m_h в нуль не принимается во внимание, Стивен (Stephan, 1945) показал, что с точностью до членов порядка n^{-2}

$$E\left(\frac{1}{m_h}\right) = \frac{1}{nW_h} + \frac{1-W_h}{n^2 W_h^2}.$$

Следовательно,

$$E[V(\bar{y}_w)] = \frac{1-L}{n} \sum W_h S_h^2 + \frac{1}{n^2} \sum (1-W_h) S_h^2.$$

Первый член равен значению $V(\bar{y}_{st})$ при пропорциональном расслоении. Второй соответствует увеличению дисперсии, вызванному тем, что сами m_h не распределены пропорционально. Однако

$$\frac{1}{n^2} \sum (1-W_h) S_h^2 = \frac{1}{n} \left(\frac{L}{n} \right) \bar{S}_h^2 - \frac{1}{n^2} \sum W_h S_h^2 = \frac{1}{nL} \bar{S}_h^2 - \frac{1}{n^2} \sum W_h S_h^2,$$

где \bar{S}_h^2 — среднее значение S_h^2 и $\bar{n}_h = n/L$ есть среднее число единиц на один слой. Таким образом, если S_h^2 различаются не очень сильно,

то член, характеризующий увеличение дисперсии, приблизительно в $1/\bar{n}_h$ раз меньше дисперсии для пропорционального расслоения. Поэтому увеличение дисперсии будет небольшим, если \bar{n}_h достаточно велико.

Описанный метод можно применить также и к выборке, уже расслоенной по другой переменной, например на пять географических районов, при условии, что W_h известны отдельно внутри каждого района.

5А.9. ОТБОР КВОТАМИ

При исследовании общественного мнения и изучении рынка широко применяется и другой способ, при котором n_h для каждого слоя определяются заранее, так что расслоение получается заведомо пропорциональным. Обследователю предписывается продолжать отбор до тех пор, пока в каждом слое не будет заполнена необходимая «квота». Наиболее распространенными признаками для такого расслоения служат географический район, возраст, пол, раса и некоторые характеристики имущественного положения. Если бы обследователи отбирали опрашиваемых внутри географических районов случайным образом и после этого относили каждого к соответствующему слою, то такой способ отбора совпадал бы с расслоенным случайным отбором. Для того чтобы заполнить при этом все квоты, необходимо проделать большую работу, поскольку на более поздних этапах обследования многие из опрашиваемых попадали бы в уже заполненные квоты.

Для того чтобы облегчить заполнение квот, обследователю предоставляется некоторая свобода в отношении того, включать ли данное лицо или домохозяйство в выборку. Степень такой свободы различна в зависимости от того, какая организация проводит обследование, но вообще отбор квотами можно охарактеризовать как расслоенный отбор с более или менее преднамеренным отбором единиц внутри слоев. По этой причине к результатам отбора квотами нельзя с уверенностью применять формулы вычисления ошибок выборки. Подробное сравнение результатов вероятностного отбора и отбора квотами проведено Стивеном и Маккарти (Stephan and McCarthy, 1958), которые дали также замечательный анализ достоинств и недостатков обоих методов. По-видимому, для таких характеристик, как доход, образование и занятия, метод квот дает смещенные выборки, хотя при изучении мнений и психологических установок он часто приносит результаты, которые хорошо согласуются с результатами вероятностного отбора.

5А.10. ОЦЕНИВАНИЕ ПО ВЫБОРКЕ ВЫИГРЫША, ПОЛУЧАЕМОГО ОТ РАССЛОЕНИЯ

Если был применен расслоенный случайный отбор, то в качестве ориентира для будущих обследований может представить интерес оценка достигнутого выигрыша в точности по сравнению с простым случайным отбором.

По выборке мы получаем в этом случае значения N_h , n_h , \bar{y}_h и s_h^2 . Из параграфа 5.4 следует, что оценкой дисперсии взвешенного сред-

него по данным расслоенной выборки будет

$$v(\bar{y}_{st}) = \sum \frac{W_h^2 s_h^2}{n_h} - \sum \frac{W_h s_h^2}{N}.$$

Задача состоит в сравнении этой величины с оценкой дисперсии среднего, которая была бы получена при простом случайном отборе. Один из применяемых иногда способов предполагает вычисление знакомого нам среднего квадрата отклонений от выборочного среднего

$$s^2 = \frac{\sum (y_{st} - \bar{y})^2}{n-1},$$

которое не учитывает расслоения. Это выражение рассматривается как оценка дисперсии на единицу для простой случайной выборки. Такой способ вполне удовлетворителен, если размещение пропорционально, поскольку простая случайная выборка сама распределена по слоям приблизительно пропорционально. Но если было принято размещение, сильно отличающееся от пропорционального, то полученная выборка уже не будет сходна с простой случайной выборкой и s^2 может оказаться плохой оценкой. Изложим общий подход к получению указанной оценки.

Истинная дисперсия среднего для простой случайной выборки, согласно алгебраическому тождеству для S^2 , равна:

$$V_{ran} = \frac{(N-n)}{nN} S^2 = \frac{(N-n)}{nN} \left[\frac{\sum (N_h - 1) S_h^2 + \sum N_h (\bar{Y}_h - \bar{Y})^2}{(N-1)} \right]. \quad (5A.25)$$

В первом члене внутри скобок нужно только подставить s_h^2 вместо S_h^2 . Второй член необходимо исследовать. В качестве оценки $\sum N_h (\bar{Y}_h - \bar{Y})^2$ естественно попробовать $\sum N_h (\bar{y}_h - \bar{y}_{st})^2$. Оказывается, что это выражение дает преувеличенную оценку и, следовательно, нуждается в поправке. Поскольку соответствующее утверждение будет полезно и в дальнейшем, сформулируем его в виде теоремы.

Теорема 5A.1. Для расслоенного случайного отбора

$$\begin{aligned} E[\sum N_h (\bar{y}_h - \bar{y}_{st})^2] &= \sum N_h (\bar{Y}_h - \bar{Y})^2 + \\ &+ \sum \frac{S_h^2 (N_h - n_h)}{n_h} \left(1 - \frac{N_h}{N}\right) = \sum N_h (\bar{Y}_h - \bar{Y})^2 + \\ &+ \sum \frac{N_h S_h^2}{n_h} (1 - f_h) (1 - W_h). \end{aligned}$$

Доказательство. Мы можем записать

$$\sum N_h (\bar{y}_h - \bar{y}_{st})^2 = \sum N_h [(\bar{Y}_h - \bar{Y}) + (\bar{y}_h - \bar{Y}_h) - (\bar{y}_{st} - \bar{Y})]^2.$$

Раскроем теперь скобки и возьмем среднее по всем возможным выборкам. Можно проверить, что среднее значение каждого из двух удвоенных произведений, содержащих $(\bar{Y}_h - \bar{Y})$, равно нулю. Тогда

$$E \sum N_h (\bar{y}_h - \bar{y}_{st})^2 = \sum N_h (\bar{Y}_h - \bar{Y})^2 + E \sum N_h (\bar{y}_h - \bar{Y}_h)^2 +$$

$$+ E \sum N_h (\bar{y}_{st} - \bar{Y})^2 - 2E \sum N_h (\bar{y}_h - \bar{Y}_h) (\bar{y}_{st} - \bar{Y}). \quad (5A.26)$$

Но

$$\sum N_h (\bar{y}_h - \bar{Y}_h) (\bar{y}_{st} - \bar{Y}) = N (\bar{y}_{st} - \bar{Y})^2$$

по определению \bar{y}_{st} и \bar{Y} . Следовательно, два последних члена в (5A.26) совместно дают

$$-EN (\bar{y}_{st} - \bar{Y})^2 = -\sum \frac{N_h (N_h - n_h)}{N} \frac{S_h^2}{n_h},$$

так как выражение слева в N раз больше дисперсии \bar{y}_{st} . Для второго члена в правой части (5A.26) справедливо

$$E \sum N_h (\bar{y}_h - \bar{Y}_h)^2 = \sum \frac{N_h (N_h - n_h)}{N_h} \frac{S_h^2}{n_h} = \sum (N_h - n_h) \frac{S_h^2}{n_h},$$

потому что внутри каждого слоя \bar{y}_h представляет собой среднее для простой случайной выборки. Следовательно,

$$\begin{aligned} E \sum N_h (\bar{y}_h - \bar{y}_{st})^2 &= \sum N_h (\bar{Y}_h - \bar{Y})^2 + \\ &+ \sum (N_h - n_h) \frac{S_h^2}{n_h} - \sum \frac{N_h (N_h - n_h)}{N} \frac{S_h^2}{n_h} = \sum N_h (\bar{Y}_h - \bar{Y})^2 + \\ &+ \sum \frac{S_h^2 (N_h - n_h)}{n_h} \left(1 - \frac{N_h}{N}\right) = \sum N_h (\bar{Y}_h - \bar{Y})^2 + \\ &+ \sum \frac{N_h S_h^2}{n_h} (1 - f_h) (1 - W_h). \end{aligned}$$

Следствие. Несмещенной оценкой для $\sum N_h (\bar{Y}_h - \bar{Y})^2$ служит

$$\sum N_h (\bar{y}_h - \bar{y}_{st})^2 - \sum \frac{N_h S_h^2}{n_h} (1 - f_h) (1 - W_h).$$

Если это выражение подставить в (5A.25), мы получим, что несмещенной оценкой для V_{ran} будет

$$\begin{aligned} v_{ran} &= \frac{N-n}{n(N-1)} \left[\sum W_h s_h^2 - \sum \frac{W_h s_h^2}{n_h} + \sum \frac{W_h^2 s_h^2}{n_h} - \frac{\sum W_h s_h^2}{N} + \right. \\ &\quad \left. + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2 \right]. \end{aligned}$$

Это выражение неудобно для вычислений. Но на практике его почти всегда можно несколько упростить. Приведем два случая.

$N > 50$. Это выполняется почти для всех совокупностей. Четвертый член внутри скобок можно опустить, поскольку он равен первому, деленному на N . Имеем

$$v_{ran} = \frac{N-n}{nN} \left[\sum W_h s_h^2 - \sum \frac{W_h s_h^2}{n_h} + \sum \frac{W_h^2 s_h^2}{n_h} + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2 \right]. \quad (5A.27)$$

Все $n_h > 50$. Второй и третий члены внутри скобок можно отбросить, что дает

$$v_{\text{ран}} = \frac{N-n}{nN} [\sum W_h s_h^2 + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2]. \quad (5A.28)$$

Пример. Вычисления иллюстрируются данными для первых трех слоев из выборки педагогических колледжей (параграф 5.9). Данные выборки 1946 г. приведены в табл. 5A.14. Средние значения представляют собой числа студентов на один колледж (в тысячах).

Таблица 5A.14
ОСНОВНЫЕ ДАННЫЕ, ПОЛУЧЕННЫЕ ПО РАССЛОЕННОЙ ВЫБОРКЕ

Слой	N_h	n_h	\bar{y}_h	s_h^2
1	13	9	2,200	1,615
2	18	7	1,638	0,063
3	26	10	0,992	0,077
Итоги	57	26		

Выборка столь мала, что для $v_{\text{ран}}$ нужно применить выражение (5A.27). Дополнительные вычисления приведены в табл. 5A.15.

Таблица 5A.15
ПРОВЕДЕНИЕ ВЫЧИСЛЕНИЙ

Слой	W_h	$W_h s_h^2$	$W_h \bar{y}_h^2 / n_h$	$W_h^2 s_h^2 / n_h$	$W_h \bar{y}_h$
1	0,228	0,36822	0,04091	0,00933	0,50160
2	0,316	0,01991	0,00284	0,00090	0,51761
3	0,456	0,03511	0,00351	0,00160	0,45235
Итоги	1,000	0,42324	0,04726	0,01183	1,47156

Вычисления производятся следующим образом:

$$v_{st} = \sum \frac{W_h^2 s_h^2}{n_h} - \sum \frac{W_h s_h^2}{N} = 0,01183 - 0,00743 = 0,0044;$$

$$v_{\text{ран}} = \frac{31}{57 \cdot 26} [0,4232 - 0,0473 + 0,0118 + 2,4000 - 2,1655] = 0,0130.$$

Таким образом, оказывается, что расслоение уменьшает дисперсию приблизительно до одной трети ее значения для простой случайной выборки.

Пропорциональное размещение. В этом случае обычно подходит оценка $v_{\text{ран}}$, получаемая на основании суммы квадратов отклонений наблюдений от выборочного среднего. Действительно, пользуясь обычным тождеством дисперсионного анализа, имеем

$$s^2 = \frac{\sum (y_{hi} - \bar{y})^2}{n-1} = \frac{1}{n-1} \left[\sum (n_h - 1) s_h^2 + \sum n_h \bar{y}_h^2 - \frac{(\sum n_h \bar{y}_h)^2}{n} \right].$$

Если членами порядка $1/n_h$ можно пренебречь, то это выражение эквивалентно

$$\sum W_h s_h^2 + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2,$$

так как $W_h = n_h/n$. Это выражение, в свою очередь, равно величине, заключенной в квадратные скобки в (5A.28). Таким образом, если размещение пропорционально n членами порядка $1/n_h$ можно пренебречь, то удовлетворительной оценкой будет

$$v_{\text{ран}} = \frac{(N-n)}{N} \frac{s^2}{n}.$$

5A.11. ОЦЕНИВАНИЕ ДИСПЕРСИИ ПРИ ОДНОЙ ЕДИНИЦЕ НА СЛОЙ

Если совокупность отличается большой вариацией и существует много эффективных критериев расслоения, то совокупность может быть расслоена до такой степени, что выборка будет содержать только по одной единице в каждом слое. В этом случае выведенной ранее формулой для оценки $V(\bar{y}_{st})$ воспользоваться нельзя. Можно попытаться получить оценку, сгруппировав слои попарно. Будем предполагать, что слои, образующие пару, имеют одинаковые объемы N_h . Небольшие расхождения в этих объемах на методе не сказываются. Средние значения для слоев, \bar{Y}_h , для двух членов пары не должны сильно различаться, однако объединение в пары должно быть произведено до получения результатов выборки по причинам, которые будут ясны позднее. Число слоев должно быть не меньше 20, чтобы выборочная дисперсия имела, по крайней мере, 10 степеней свободы.

Обозначим наблюдения в отдельной паре через y_{j1}, y_{j2} , где j принимает значения от 1 до $L/2$. Далее, возьмем среднее по всем выборкам из этой пары:

$$E(y_{j1} - y_{j2})^2 = (\bar{Y}_{j1} - \bar{Y}_{j2})^2 + \frac{N_j - 1}{N_j} (S_{j1}^2 + S_{j2}^2), \quad (5A.29)$$

где $N_j = N_h$ есть объем каждого слоя в этой паре. Рассмотрим оценку

$$v(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^{L/2} N_j^2 (y_{j1} - y_{j2})^2. \quad (5A.30)$$

Согласно (5A.29) математическое ожидание этой величины

$$Ev(\bar{y}_{st}) = \frac{1}{N^2} \left[\sum_{h=1}^L N_h (N_h - 1) S_h^2 + \sum_{j=1}^{L/2} N_j^2 (\bar{Y}_{j1} - \bar{Y}_{j2})^2 \right]. \quad (5A.31)$$

Первый член в правой части представляет собой правильную дисперсию (по теореме 5.3 при $n_h = 1$); второй соответствует положительному смещению. Величина этого смещения зависит от того, насколько удачно объединены в пары те слои, истинные средние которых мало различаются. Вид оценки (5A.30) показывает, что при образовании пар не следует стремиться к тому, чтобы как можно меньше отличались выборочные значения, так как это очень преуменьшило бы дисперсию. Описанный метод иногда называют способом «совмещенных слоев».

При другом способе отбора каждая пара слоев рассматривается как отдельный слой и отбирается случайным образом по две единицы в каждом из таких $L/2$ слоев. Несмещенную оценку $V(\bar{y}_{st})$ для такого рода отбора можно получить из обычной формулы. Читатель может проверить, что

$$V(\bar{y}_{st}) = \frac{1}{N^2} \left[\sum_{h=1}^L N_h(N_h-1) \frac{2N_h-2}{2N_h-1} S_h^2 + \sum_{j=1}^{L/2} N_j^2 \frac{N_j-1}{2N_j-1} (\bar{Y}_{j1} - \bar{Y}_{j2})^2 \right]. \quad (5A.32)$$

Из сравнения с (5A.31) становится ясным, что формула (5A.30) преувеличивает не только истинную дисперсию при одной единице в каждом слое, но и дисперсию, возникающую при вдвое более крупных слоях.

Предпочтительны ли слои меньшего объема, в свете этого утверждения остается спорным. К сожалению, хотя при одной единице в каждом слое выигрыш в точности больше, чем при двух единицах в паре слоев, в этом случае и оценка в большей степени преувеличивает дисперсию.

5A.12. УПРОЩЕННОЕ ВЫЧИСЛЕНИЕ СТАНДАРТНЫХ ОШИБОК

Одно из достоинств вероятностного отбора состоит в том, что для любой оценки, полученной по выборке, можно вычислить стандартную ошибку. К сожалению, вычисление стандартных ошибок более трудоемко, чем вычисление самих оценок. В сложных, охватывающих всю страну обследованиях, где учитываются сотни и тысячи признаков, вычисление стандартных ошибок и указание их при результатах обследования составляют главную проблему. Для уменьшения затрат труда и средств был выдвинут целый ряд рекомендаций. Насколько они приемлемы, зависит, разумеется, от того, какие должны применяться при этом вычислительные машины и методы.

Напомним, что формулы дисперсий оценок среднего и суммарного значений для совокупности имеют вид:

$$V(\bar{y}_{st}) = \sum_{h=1}^L \frac{w_h^2 S_h^2}{n_h} (1-f_h); \quad (5A.33)$$

$$V(\hat{Y}_{st}) = \sum_{h=1}^L \frac{N_h^2 S_h^2}{n_h} (1-f_h). \quad (5A.34)$$

Рассмотрим сначала обследование, в котором число слоев небольшое, скажем, ≤ 10 , а выборка сравнительно велика. В таких обследованиях S_h^2 и f_h могут сильно меняться от слоя к слою. Один из приемов состоит в том, чтобы вычислять S_h^2 в слое h по подвыборке объема n_h . В большинстве случаев для этого будет достаточной подвыборка объемом в 20 единиц. Подвыборка должна извлекаться случайным образом. Быстрее можно получить *систематическую* подвыборку (содержащую, например, каждую шестую единицу в списке отобранных единиц, если n_h составляет около 120), но она, вероятно, даст преувеличенную оценку S_h^2 , если отобранные единицы регистрировались в каком-либо определенном порядке (см. параграф 8.3).

Если отобранные единицы располагаются в основном в случайном порядке, то другая возможность сокращения объема вычислительной работы состоит в том, чтобы по мере сложения единиц для подсчета выборочного суммарного значения для слоя фиксировать некоторые промежуточные суммарные значения. Например, если $n_h = 123$, мы можем подводить промежуточные итоги через каждые 10 единиц, получив, таким образом, 12 промежуточных суммарных значений T_1, T_2, \dots, T_{12} . Последние три единицы в них не учитываются, хотя они включены в суммарное значение для выборки. Тогда величина $\sum (T_i - \bar{T})^2/110$ даст оценку S_h^2 с 11 степенями свободы. Если распределение y отличается от нормального, то такая оценка будет, вероятно, более точной, чем оценка, вычисленная по 12 случайным образом выбранным единицам, потому что применение промежуточных суммарных значений уменьшает влияние эксцесса (параграф 2.14) на $V(\bar{y}_{st})$.

Предположим теперь, что число слоев велико и объемы выборок внутри слоев малы, как, например, в случае географического расслоения, охватывающего большую территорию. Для большой группы слоев допустимо предположить, что S_h^2 меняются от слоя к слою незначительно. Тогда мы можем извлечь случайным образом подвыборку слоев (например, отобрать 8 слоев из 40—50), вычислить s_h^2 в каждом из восьми слоев и образовать объединенное s_h^2 для применения ко всем 40—50 слоям. Такой метод сопряжен с большим риском, чем оценивание S_h^2 в каждом слое с помощью случайной подвыборки объемом в две или три единицы, и предпочтителен лишь тогда, когда его применение дает выигрыш во времени.

Когда n_h одинаковы во всех слоях или в большой группе слоев, применимы другие методы. Предположим, что $n_h = 2$ и N_h одинаковы. Пусть y_{h1}, y_{h2} — значения признака у двух единиц и пусть $d_h = y_{h1} - y_{h2}$. Тогда

$$E(d_h^2) = 2S_h^2.$$

Следовательно, несмещенной оценкой $V(\bar{Y}_{st})$ по этой группе слоев будет

$$v(\bar{Y}_{st}) = N_h^2 (1-f_h) \sum_{h=1}^L \frac{d_h^2}{4}.$$

Этот результат можно проверить, сравнив его с (5A.34).

Эта оценка имеет L степеней свободы — более чем необходимо, если L велико. Слои можно объединить в k групп, которые могут содержать разное число слоев. Пусть $D_j = \sum d_{hj}$, взятых по слоям в j -й группе. Нетрудно показать, что несмещенной оценкой $V(\hat{Y}_{st})$ с k степенями свободы будет

$$v(\hat{Y}_{st}) = N_k^2 (1 - f_k) \sum_{j=1}^k \frac{D_j^2}{4}.$$

Подобные же методы применимы, когда n_h одинаковы и больше 2. Предположим, что $n_h = 4$ и что мы хотим получить оценку $v(\hat{Y}_{st})$ с 18 степенями свободы. Объединим слои в шесть групп. В каждом слое четырем отобранным единицам присвоим номера с первого по четвертый каким-либо подходящим для нас способом. Пусть $T_{j1}, T_{j2}, T_{j3}, T_{j4}$ будут суммарными значениями в группе j , взятыми по первым, вторым, третьим и четвертым единицам, а \bar{T}_j — их средним значением. Тогда

$$v(\hat{Y}_{st}) = N_k^2 (1 - f_k) \sum_{j=1}^6 \sum_{u=1}^4 \frac{(T_{ju} - \bar{T}_j)^2}{12}.$$

В общем случае в знаменателе вместо 12 будет стоять $n_h(n_h - 1)$. Если n_h одинаковы, а N_h различаются, то группировка становится более трудоемкой. Вместо суммарных значений T_{ju} мы должны образовывать суммарные значения

$$\hat{Y}_{ju} = \sum N_h y_{hju} \quad (u = 1, 2, \dots, n_h),$$

где y_{hju} обозначает единицу, имеющую номер « ju » в слое, и сумма берется по j -й группе слоев. При k группах слоев можно показать, что

$$E \left[\sum_{j=1}^k \sum_{u=1}^{n_h} \frac{(\hat{Y}_{ju} - \hat{Y}_j)^2}{n_h(n_h - 1)} \right] = \sum_{h=1}^L \frac{N_h^2 S_h^2}{n_h}$$

с $k(n_h - 1)$ степенями свободы. Величина в квадратных скобках увеличивает $V(\hat{Y}_{st})$, так как ее математическое ожидание не содержит нужной поправки, $1 - f_k$. При необходимости, в качестве грубой поправки оценку можно умножить на $1 - n/N$.

Если $k = 1$, то все n_h величины \hat{Y}_{ju} будут несмещенными оценками суммарного значения для совокупности, \bar{Y} , и описанный способ дает оценку $V(\hat{Y}_{st})$ с $n_h - 1$ степенями свободы.

В обследованиях, учитывающих большое число признаков, зависимость между $s(\hat{Y})$ и \hat{Y} может быть в основном сходной для широкого класса признаков. Это можно проверить, нанося $s(\hat{Y})$ и \hat{Y} для различных признаков на график и подбирая шкалы, при которых простая зависимость дает хорошее согласие. В этом случае такого рода «график ошибок», названный так Йейтсом (Yates, 1960), оказывается весьма полезным. Если существует хороший график ошибок, то вычисление

стандартных ошибок при небольшом числе степеней свободы становится более надежным. Для менее важных признаков стандартные ошибки можно вообще не вычислять, а находить их оценки по графику. Представление стандартных ошибок с помощью графика или полученной на его основе таблицы позволяет иногда избежать печатания сотен отдельных значений. Примеры графиков ошибок можно найти у Йейтса (Yates, 1960), Хансена и др. (Hansen et al., 1953).

При повторных исследованиях зависимость между $s(\hat{Y})$ и \hat{Y} может остаться неизменной или измениться со временем очень мало. Это дает возможность для дальнейшей экономии. Если график ошибок уже построен, то в будущем вычислять стандартные ошибки нужно через определенные интервалы времени, причем для того, чтобы обнаружить, не изменился ли существенно вид графика ошибок, достаточно лишь небольшого числа данных.

Обзор упрощенных методов вычислений приводит Жаркович (Žarković, 1960), у Кейфитца (Keyfitz, 1957) изложены некоторые остроумные методы для случая $n_h = 2$.

5A.13. СЛОИ КАК ОБЛАСТИ ИЗУЧЕНИЯ

В этом параграфе рассматриваются обследования, главная цель которых состоит в том, чтобы сравнить между собой различные слои, в предположении, что сами слои определены заранее. В этом случае правила размещения объема выборки по слоям отличаются от правил, применяемых с целью получить оценки для всей совокупности. Если имеется только два слоя, мы можем выбрать n_1, n_2 так, чтобы дисперсия разности оценок средних для слоев, $(\bar{y}_1 - \bar{y}_2)$, была минимальной. Опуская по причинам, указанным в параграфе 2.12, имеем

$$V(\bar{y}_1 - \bar{y}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}.$$

Если функция издержек линейна и имеет вид

$$C = c_0 + c_1 n_1 + c_2 n_2,$$

V будет минимальной, когда

$$n_1 = \frac{\frac{nS_1}{\sqrt{c_1}}}{S_1/\sqrt{c_1} + S_2/\sqrt{c_2}}; \quad n_2 = \frac{\frac{nS_2}{\sqrt{c_2}}}{S_1/\sqrt{c_1} + S_2/\sqrt{c_2}}. \quad (5A.35)$$

Если имеется L слоев и $L > 2$, то оптимальное размещение зависит от желательных уровней точности при сравнении различных слоев. Например, можно минимизировать издержки при наличии $L(L - 1)/2$ ограничений вида $V(\bar{y}_h - \bar{y}_i) \leq V_{hi}$, где значения V_{hi} выбраны в соответствии с точностью, которая считается необходимой для удовлетворительного сравнения слоев h и i .

Часто применим более простой метод размещения, особенно если S_h и c_h различаются не очень сильно. Один из подходов состоит в том,

чтобы минимизировать среднее значение дисперсий разности оценок для всех $L(L-1)/2$ пар слоев, т. е. минимизировать

$$\bar{V} = \frac{2}{L} \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \dots + \frac{S_L^2}{n_L} \right).$$

При неизменном C \bar{V} минимизируется согласно правилу (5А.35), когда

$$n_h \propto \frac{S_h}{\sqrt{c_h}}.$$

Этот прием может привести к тому, что для некоторых пар слоев сравнение оценок будет произведено с меньшей, чем хотелось бы, точностью, а для остальных — с большей. Другой подход состоит в выборе n_h таким образом, чтобы стандартные ошибки разности оценок для каждой пары слоев были бы одинаковыми и равнялись, скажем, \sqrt{V} . Это будет выполняться, если положить $S_h^2/n_h = V/2$ для каждого слоя. При неизменных издержках этот метод дает меньшую общую точность, чем первый метод. Читатель может проверить, что два оптимальных размещения дадут

$$\bar{V} = \frac{2(\sum S_h \sqrt{c_h})^2}{L(C-c_0)}; \quad V = \frac{2(\sum S_h^2 c_h)}{(C-c_0)}.$$

Из неравенства Коши—Шварца следует, что V всегда больше, чем \bar{V} , если только $S_h \sqrt{c_h}$ не постоянная. Если V существенно больше \bar{V} , то иногда, после нескольких попыток, удастся найти такое компромиссное размещение, при котором средняя дисперсия близка к \bar{V} и одновременно $V(\bar{y}_h - \bar{y}_l)$ достаточно постоянна.

Иногда нужно получить как оценки для каждого слоя, так и общие оценки для всей совокупности. При планировании обследования мы могли бы задать следующие ограничения:

$$V(\bar{y}_h) = \frac{S_h^2}{n_h} (1-f_h) \leq V_h; \quad V(\bar{y}_{st}) = \sum \frac{N_h^2 S_h^2}{n_h} (1-f_h) \leq V.$$

Здесь пкс учитывается, поскольку цель состоит в том, чтобы указать точность, с которой должны быть оценены средние значения в конечной совокупности. Условия на $V(\bar{y}_{st})$ определяют нижние границы значений n_h . Если оказывается, что эти нижние границы обеспечивают выполнение условия, накладываемого на $V(\bar{y}_{st})$, то задача размещения решена. Для случая, когда условие на $V(\bar{y}_{st})$ не выполняется, Далениус (Dalenius, 1957) предложил графический метод решения.

5А.14. ОЦЕНИВАНИЕ СУММАРНЫХ И СРЕДНИХ ЗНАЧЕНИЙ ДЛЯ ПОДСОВОКУПНОСТЕЙ

Часто подсовокупности, или области изучения, оказываются представленными во всех слоях. Например, при географическом расслоении могут понадобиться отдельные оценки для всей совокупности, для

мужчин и женщин, для разных возрастных групп, для пользующихся или не пользующихся какой-либо зубной пастой и т. д. Решение этой задачи вызывает некоторые трудности. Основные формулы были выведены Йейтсом (Yates, 1953), дальнейшее обсуждение и доказательства содержатся у Дербина (Dugbin, 1958) и Хартли (Hartley, 1959). Методы, применимые к отдельным слоям, были рассмотрены в параграфах 2.10 и 2.11.

Следующие обозначения относятся к единицам из слоя h , принадлежащим области j .

Обозначения

Число единиц: N_{hj} ; $\sum_j N_{hj} = N_h$.

Число единиц в выборке: n_{hj} ; $\sum_j n_{hj} = n_h$.

Результат наблюдения для отдельной единицы: y_{hij} .

Выборочное среднее: $\bar{y}_{hj} = \sum_{i=1}^{n_{hj}} \frac{y_{hij}}{n_{hj}}$.

Среднее по области изучения: $\bar{Y}_{hj} = \sum_{i=1}^{N_{hj}} \frac{y_{hij}}{N_{hj}}$.

Суммарное и среднее значение по всем слоям для области j , соответственно

$$Y_j = \sum_h N_{hj} \bar{Y}_{hj}; \quad \bar{Y}_j = \frac{Y_j}{N_j},$$

где

$$N_j = \sum_h N_{hj}.$$

Осложнения возникают в связи с тем, что n_{hj} — случайные переменные. Если известны N_{hj} , то задача упрощается. В качестве оценок Y_j и \bar{Y}_j можно воспользоваться

$$\hat{Y}_j = \sum_h N_{hj} \bar{y}_{hj}; \quad \hat{\bar{Y}}_j = \frac{\hat{Y}_j}{N_j}.$$

Как показано в параграфе 2.10, для $V(\bar{y}_{hj})$ остается справедливой обычная формула. Следовательно,

$$V(\hat{Y}_j) = \sum_h \frac{N_{hj}^2 S_{hj}^2}{n_{hj}} \left(1 - \frac{n_{hj}}{N_{hj}} \right),$$

где S_{hj}^2 — дисперсия единиц, принадлежащих области j в слое h . На практике, однако, N_{hj} бывают известны редко.

Оценивание суммарных значений для областей изучения

Если N_{hj} неизвестны, то суммарное значение по определенной области изучения в каждом слое оценивается как в параграфе 2.11.

Для того чтобы получить оценку суммарного значения по этой области, нужно сложить суммарные значения по слоям, т. е.

$$\hat{Y}_j = \sum_h \frac{N_h}{n_h} \sum_i^{n_{hj}} y_{hij}.$$

Истинная дисперсия \hat{Y}_j и ее оценка получаются с помощью того же приема, что и в параграфе 2.11. Вводится переменная y'_{hi} , которая равна y_{hij} для всех единиц в области j и равна нулю для всех остальных единиц совокупности. Как показано в параграфе 2.11, оценка дисперсии в этом случае принимает вид

$$v(\hat{Y}_j) = \sum_h \frac{N_h^2}{n_h(n_h-1)} (1-f_h) \left[\sum_i^{n_{hj}} y_{hij}^2 - \frac{(\sum y_{hij})^2}{n_h} \right].$$

Оценивание средних значений для областей изучения

Для того чтобы оценить среднее значение для области, Y_j/N_j , необходима выборочная оценка N_j . Несмещенной его оценкой будет

$$\hat{N}_j = \sum_h \frac{N_h}{n_h} n_{hj}.$$

Следовательно, полагаем

$$\hat{\bar{Y}}_j = \frac{\hat{Y}_j}{\hat{N}_j} = \frac{\sum_h \frac{N_h}{n_h} \sum_i y_{hij}}{\sum_h \frac{N_h}{n_h} n_{hj}}.$$

При пропорциональном расслоении $\hat{\bar{Y}}_j$ сводится к обычному выборочному среднему по единицам, принадлежащим области j . В общем случае эта оценка называется *совместной оценкой по отношению*, она рассматривается дальше в параграфе 6.11. Для того чтобы показать справедливость этого утверждения, введем новую фиктивную переменную x'_{hi} , которая равна 1 для всех единиц из области j и 0 для всех остальных единиц, причем i принимает теперь значения от 1 до N_h . Очевидно,

$$\begin{aligned} \bar{x}'_h &= \frac{\sum_i^{n_h} x'_{hi}}{n_h} = \frac{n_{hj}}{n_h}; \\ \bar{y}'_h &= \frac{\sum_i^{n_{hj}} y_{hij}}{n_h} = \frac{\sum_i^{n_h} y_{hi}}{n_h} = \frac{n_{hj}}{n_h} \bar{y}_{hj}, \end{aligned} \quad (5A.36)$$

так что оценку среднего для области можно записать в виде

$$\hat{\bar{Y}}_j = \frac{\sum_h \frac{N_h}{n_h} \sum_i y_{hij}}{\sum_h \frac{N_h}{n_h} n_{hj}} = \frac{\sum_h N_h \bar{y}'_h}{\sum_h N_h \bar{x}'_h} = \frac{\bar{y}'_{st}}{\bar{x}'_{st}}.$$

Мы получили формулу совместной оценки по отношению для двух переменных y'_{hi} и x'_{hi} . Как следует из параграфа 6.11, оценку дисперсии можно приближенно выразить в виде

$$v(\hat{\bar{Y}}_j) \approx \frac{1}{\hat{N}_j^2} \sum_h \frac{N_h^2 (1-f_h)}{n_h(n_h-1)} \sum_i^{n_h} [y'_{hi} - \hat{\bar{Y}}_j x'_{hi} - (\bar{y}'_h - \hat{\bar{Y}}_j \bar{x}'_h)]^2. \quad (5A.37)$$

Второе суммирование на основании (5A.36) можно записать в виде

$$\begin{aligned} \sum_i^{n_h} (y'_{hi} - \hat{\bar{Y}}_j x'_{hi})^2 &= n_h (\bar{y}'_h - \hat{\bar{Y}}_j \bar{x}'_h)^2 = \\ &= \sum_i^{n_{hj}} (y_{hij} - \hat{\bar{Y}}_j)^2 - \frac{n_{hj}^2}{n_h} (\bar{y}_{hj} - \hat{\bar{Y}}_j)^2. \end{aligned} \quad (5A.38)$$

Далее, первый член в (5A.38) можно иначе записать как

$$\sum_i^{n_{hj}} (y_{hij} - \bar{y}_{hj})^2 + n_{hj} (\bar{y}_{hj} - \hat{\bar{Y}}_j)^2.$$

Подставляя эти выражения в (5A.37), получаем окончательно оценку дисперсии в виде

$$\begin{aligned} v(\hat{\bar{Y}}_j) &\approx \frac{1}{\hat{N}_j^2} \sum_h \frac{N_h^2 (1-f_h)}{n_h(n_h-1)} \left[\sum_i (y_{hij} - \bar{y}_{hj})^2 + \right. \\ &\quad \left. + n_{hj} \left(1 - \frac{n_{hj}}{n_h}\right) (\bar{y}_{hj} - \hat{\bar{Y}}_j)^2 \right]. \end{aligned} \quad (5A.39)$$

Второй член выражения в квадратных скобках представляет часть дисперсии, обусловленную вариацией между слоями. Различия между средними для слоев не исключаются полностью из дисперсии оценки среднего любой подсовокупности. Часть дисперсии, обусловленная вариацией между слоями, мала, если малы члены $1 - n_{hj}/n_h$, т. е. если подсовокупность почти так же велика, как и вся совокупность.

Как указал Дербин (Durbin, 1958), формула (5A.39) справедлива также для оценок средних для всей совокупности, если выборка неполна по тем или иным причинам, например, вследствие неполучения ответа при условии, конечно, что в качестве оценки принимается $\hat{\bar{Y}}_j$. В этом случае $\hat{\bar{Y}}_j$ интерпретируется как оценка среднего для той части совокупности, по которой при применявшихся методах сбора данных ответы будут получены. Здесь, однако, возникает дополнительное осложнение, связанное с тем, что среднее значение по «неответившей» части совокупности часто отличается от среднего по «ответившей». Тогда $\hat{\bar{Y}}_j$ будет смещенной оценкой среднего для всей совокупности, а это смещение в (5A.39) не учитывается.

5А.1. При планировании обследования продаж в магазинах определенного типа, при $n = 550$, имеются хорошие оценки S_h по результатам предыдущего обследования для двух слоев из трех. Третий слой состоит из новых магазинов и магазинов, не имевших продаж по данным предыдущего обследования, так что значение S_h можно лишь предполагать. Если S_h в действительности равно 10, вычислите $V(\bar{y}_{st})$, которая была бы получена при неймановом размещении в случае, когда S_h предполагается равной (а) 5, (б) 20. Покажите, что в обоих случаях относительное увеличение дисперсии по сравнению с истинным оптимальным ее значением немногим превышает 2%.

Слой	W_h	Истинное S_h	Оценки S_h	
			(а)	(б)
1	0,3	30	30	30
2	0,6	20	20	20
3	0,1	10	5	20

5А.2. Покажите, что если все S_h , кроме S_L , оценены правильно, а в качестве оценки S_L принято $\hat{S}_L = S_L(1 + \lambda)$, то относительное увеличение $V_{opt}(\bar{y}_{st})$ при неймановом размещении, когда вместо истинного S_L применяется \hat{S}_L , равно

$$\frac{\lambda^2 n_L' (n - n_L')}{(1 + \lambda) n^2},$$

где n_L' — объем выборки в слое L при истинном неймановом размещении. Проверьте, что результаты упражнения 5А.1 с этой формулой согласуются. (Совпадение будет не совсем точным из-за округления n_h до целых чисел.) Выведите отсюда, что 50%-ное преуменьшение S_L имеет тот же эффект, что и ее 100%-ное преувеличение.

5А.3. Покажите, что если имеется два слоя и если ϕ есть отношение фактического n_1/n_2 к нейманову оптимальному n_1/n_2 , то при любых значениях N_1 , N_2 , S_1 и S_2 отношение $V_{min}(\bar{y}_{st})/V(\bar{y}_{st})$ всегда больше $4\phi/(1 + \phi)^2$.

5А.4. Результаты простой случайной выборки объема $n = 1000$ можно распределить по трем «слоям» с $\bar{y}_h = 10,2$; $12,6$ и $17,1$; $s_h^2 = 10,82$ (одна и та же в каждом слое) и $s^2 = 17,66$. Оценками весов для слоев служат значения $w_h = 0,5$; $0,3$; $0,2$ соответственно. Известно, что эти веса неточны, но предполагается, что неточность не превышает 5%, так что наиболее неблагоприятными будут случаи либо $W_h = 0,525$; $0,285$ и $0,190$, либо $W_h = 0,475$; $0,315$ и $0,210$. Основываясь на результатах из параграфа 5А.2, рекомендуете ли вы производить расчленение? (При необходимости полагайте $\bar{y}_h = \bar{Y}_h$ и $s_h^2 = S_h^2$.)

5А.5. Планируется обследование с тремя слоями для оценки процента семей, имеющих сбережения в банке, и средней величины вклада на семью. Предварительные оценки процентов P_h и значений S_h внутри слоев для величины вклада следующие:

Слой	W_h	$P_h(\%)$	S_h (в долл.)
1	0,6	20	90
2	0,3	40	180
3	0,1	70	520

Вычислите наименьший объем выборки n и соответствующие n_h , удовлетворяющие следующим условиям: (а) Процент семей должен оцениваться со стандартной ошибкой, равной 2, и средняя величина вклада — со стандартной ошиб-

кой, равной 5 долл. (б) Процент семей должен оцениваться со стандартной ошибкой, равной 1,5, и средняя величина вклада — со стандартной ошибкой 5 долл.

5А.6. В таблице приведено распределение 911 городов по величине города. Совокупность включает города от 10 000 до 60 000 жителей по группам

Величина города (в тыс. жителей)	f	y'	\sqrt{f}	Накопленные f	Накопленные \sqrt{f}	$f y'$
10—11	205	0	14,3	205	14,3	0
12—13	135	1	11,6	340	25,9	135
14—15	106	2	10,3	446	36,2	212
16—17	82	3	9,1	528	45,3	246
18—19	61	4	7,8	589	53,1	244
20—21	42	5	6,5	631	59,6	210
22—23	32	6	5,7	663	65,3	192
24—25	30	7	5,5	693	70,8	210
26—27	27	8	5,2	720	76,0	216
28—29	18	9	4,2	738	80,2	162
30—31	22	10	4,7	760	84,9	220
32—33	21	11	4,6	781	89,5	231
34—35	19	12	4,4	800	93,9	228
36—37	16	13	4,0	816	97,9	208
38—39	14	14	3,7	830	101,6	196
40—41	17	15	4,1	847	105,7	255
42—43	9	16	3,0	856	108,7	144
44—45	8	17	2,8	864	111,5	136
46—47	11	18	3,3	875	114,8	198
48—49	9	19	3,0	884	117,8	171
50—51	7	20	2,6	891	120,4	140
52—53	4	21	2,0	895	122,4	84
54—55	5	22	2,2	900	124,6	110
56—57	5	23	2,2	905	126,8	115
58—59	6	24	2,4	911	129,2	144
Итого	911		129,2			4407

$$\Sigma f y'^2 = 50\ 395$$

с интервалом в 2000. Для упрощения вычислений приведены условные значения y' и значения \sqrt{f} , накопленные f , накопленные $f y'$ и $\Sigma f y'^2$. Применяя способ Даланнуса — Ходжеса, сформируйте два слоя для оптимального размещения в смысле Неймана. Найдите значения W_h и S_h для каждого из этих слоев. Проверьте, (а) что оптимальные объемы выборки в двух слоях почти одинаковы, и найдя S^2 для всей совокупности, (б) что

$$\frac{V(\bar{y})}{V_{opt}(\bar{y}_{st})} \approx 4,8.$$

5А.7. Распределение, плотность которого имеет форму равнобедренного треугольника, $f(y) = 2(1 - y)$, $0 < y < 1$, разделено на два слоя с границей в точке a . (а) Покажите, что

$$W_1 = a(2 - a); \quad W_2 = (1 - a)^2;$$

$$S_1^2 = \frac{a^2(6 - 6a + a^2)}{18(2 - a)^2}; \quad S_2^2 = \frac{(1 - a)^2}{18}.$$

(б) Покажите, что согласно правилу накопленных значений V/\bar{y} наилучшим значением a будет $1 - 1/4 = 0,37$ и что для этой границы оптимальное n_1/n_2 равно приблизительно 27/25, а $V(\bar{y}_{st})$ составляет приблизительно 27% величины, получаемой при простом случайном отборе.

5А.8. На получение расслоенной выборки отпущено 5000 долл. Предполагают, что функция издержек в обозначениях параграфа 5А.7 приблизительно имеет вид $C = 200L + 10n$ и

$$V(\bar{y}_{st}) \approx \frac{S^2}{n} \left[\frac{\rho^2}{L^2} + (1 - \rho^2) \right],$$

где ρ — коэффициент корреляции между переменной, по значениям которой формируются слои, и переменной, наблюдаемой в обследовании. Вычислите оптимальное L при $\rho = 0,95; 0,9$ и $0,8$. Найдите хорошее компромиссное число слоев, пригодное для всех трех значений ρ .

5А.9. Приведенные дальше данные получены по расслоенной выборке торговцев автомобильными шинами, отобранной в марте 1945 г. (Deming and Simmons, 1946). Торговцы были распределены по слоям в соответствии с числом новых шин, имевшихся у них во время предыдущей переписи. Выборочные средние \bar{y}_h представляют собой средние числа новых шин на одного торговца. (а) Оцените выигрыш в точности, получаемой от такого расслоения. (б) Сравните этот выигрыш с выигрышем, который был бы получен при пропорциональном размещении.

Границы слоев	N_h	W_h	\bar{y}_h	s_h^2	n_h
1—9	19 850	0,8032	4,1	34,8	3 000
10—19	3 250	0,1315	13,0	92,2	600
20—29	1 007	0,0407	25,0	174,2	340
30—39	606	0,0245	38,2	320,4	230
Итого	24 713	0,9999			4 170

5А.10. Для совокупности с $N = 6$, $L = 2$ значения y_{hi} составляют 0, 1, 3 для первого слоя и 5, 6, 9 — для второго слоя. Вычислите (а) $V(\bar{y})$ для простой случайной выборки объема $n = 2$, (б) $V(\bar{y}_{st})$ для расслоенной случайной выборки, содержащей по одной единице в каждом слое, (в) среднее значение оценки $\bar{v}(\bar{y}_{st})$, получаемой с помощью метода совмещенных слоев. Проверьте, что $\bar{v}(\bar{y}_{st}) > V(\bar{y})$.

ЛИТЕРАТУРА

- Bryant E. C., Hartley H. O. and Jessen R. J. (1960). Design and estimation in two-way stratification. *Jour. Amer. Stat. Assoc.*, 55, 105—124.
 Cochran W. G. (1961). Comparison of methods for determining stratum boundaries. *Bull. Int. Stat. Inst.*, 38, 2, 345—358.
 Dalenius T. (1957). *Sampling in Sweden*. Contributions to the methods and theories of sample survey practice. Almqvist and Wiksell, Stockholm.
 Dalenius T. and Gurney M. (1951). The problem of optimum stratification. II. *Skand. Akt.*, 34, 133—148.
 Dalenius T. and Hodges J. L. Jr. (1959). Minimum variance stratification. *Jour. Amer. Stat. Assoc.*, 54, 88—101.
 Deming W. E. and Simmons W. R. (1946). On the design of a sample for dealer inventories. *Jour. Amer. Stat. Assoc.*, 41, 16—33.

- Durbin J. (1958). Sampling theory for estimates based on fewer individuals than the number selected. *Bull. Int. Stat. Inst.*, 36, 3, 113—119.
 Ekman G. (1959). An approximation useful in univariate stratification. *Ann. Math. Stat.*, 30, 219—229.
 Evans W. D. (1951). On stratification and optimum allocations. *Jour. Amer. Stat. Assoc.*, 46, 95—104.
 Goodman R. and Kish L. (1950). Controlled selection—a technique in probability sampling. *Jour. Amer. Stat. Assoc.*, 45, 350—372.
 Hagood M. J. and Bernert E. H. (1945). Component indexes as a basis for stratification. *Jour. Amer. Stat. Assoc.*, 40, 330—341.
 Hartley H. O. (1959). *Analytic studies of survey data*. Istituto di Statistica, Rome, volume in onora di Corrado Gini.
 Jessen R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agr. Exp. Sta. Res. Bull.* 304.
 Keyfitz N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Jour. Amer. Stat. Assoc.*, 52, 503—510.
 King A. J. and McCarty D. E. (1941). Application of sampling to agricultural statistics with emphasis on stratified sampling. *Jour. Marketing*, 5, 462—474.
 McEvoy R. H. (1956). Variation in bank asset portfolios. *Jour. Finance*, 11, 463—473.
 Stephan F. F. (1941). Stratification in representative sampling. *Jour. Marketing*, 6, 38—46.
 Stephan F. F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate. *Ann. Math. Stat.*, 16, 50—61.
 Stephan F. F. and McCarthy P. J. (1958). *Sampling opinions*. John Wiley and Sons, New York.
 Sukhatme P. V. (1935). Contribution to the theory of the representative method. *Supp. Jour. Roy. Stat. Soc.*, 2, 253—268.
 Yates F. (1953). *Sampling methods for censuses and surveys*. Hafner, New York, Second edition; (1960) third edition.
 Zarković, S. S. (1960). Computation of errors for sample estimates. *Monthly Bull. Agr. Econ. and Stat.* F. A. O. Rome, 9, No. 4, 1—9.

ОЦЕНКИ ПО ОТНОШЕНИЮ

6.1. МЕТОДЫ ОЦЕНИВАНИЯ

Одним из признаков развития теоретической статистики служит возникновение большого теоретического раздела, посвященного тому, как получать по имеющимся данным хорошие оценки. Однако в развитии теории выборочных обследований достижения этого раздела применялись мало. На мой взгляд, это объясняется двумя основными причинами. Во-первых, в обычных обследованиях, учитывающих большое число признаков, значительным преимуществом обладают методы оценивания, которые не требуют приемов более сложных, чем простое сложение, тогда как тонкие методы оценивания в статистической теории, скажем, такие, как метод максимального правдоподобия, могут требовать для получения оценок ряда последовательных приближений. Во-вторых, между этими двумя направлениями исследования существует некоторое различие в подходе к их методам. В теоретической статистике большинство методов оценивания предполагает, что заранее известна функциональная форма распределения частот, которому подчиняются данные выборки, и метод оценивания тщательно подбирается к этому виду распределения. В теории выборочных обследований обычно делаются лишь общие предположения о характере этого распределения частот (что оно сильно скошено или же весьма симметрично), а его конкретная функциональная форма не рассматривается. Такой подход вполне оправдан при обработке материалов обследований, в которых вид распределения может меняться от одного признака к другому, причем мы не хотим, прерывая работу, исследовать его, чтобы решить, как получать каждую оценку.

Таким образом, в настоящее время количество методов оценивания, применяемых при обработке материалов выборочных обследований, ограничено. Мы рассмотрим два метода: метод отношения в этой главе и метод линейной регрессии в гл. 7. Применение более сложных методов, возможно, будет расширяться, по крайней мере, в небольших специализированных обследованиях, потому что выигрыш в точности от применения более эффективных методов оценивания часто может быть получен недорогой ценой, поскольку вычисления при этом усложняются только на заключительном этапе.

6.2. ОЦЕНКА ПО ОТНОШЕНИЮ

При способе оценивания по отношению для каждой единицы выборки наблюдается некоторая вспомогательная переменная x_i , коррелированная с y_i . Суммарное значение X переменной x_i для совокупности должно быть известным. На практике x_i часто бывает значением y_i на тот предшествующий момент времени, когда производилась сплошная перепись. Цель этого способа заключается в том, чтобы повысить точность, пользуясь тем, что между y_i и x_i существует корреляция. Здесь мы предполагаем, что производится простой случайный отбор.

Оценка по отношению величины Y , суммарного значения y_i для совокупности, имеет вид

$$\hat{Y}_R = \frac{y}{x} X = \frac{\bar{y}}{\bar{x}} X, \quad (6.1)$$

где y , x — суммарные выборочные значения соответственно y_i и x_i . [Индекс R — от английского «ratio» — отношение.]

Если x_i — это значение y_i на некоторый предшествующий момент времени, то при способе отношения выборка применяется для того, чтобы оценить относительное изменение, Y/X , происшедшее с этого момента. Для получения оценки текущего суммарного значения для совокупности оценка относительного изменения, y/x , умножается на суммарное значение для совокупности, X , известное из предшествующего обследования. Если отношение y_i/x_i приблизительно одинаково для всех единиц, то значения y/x мало меняются от выборки к выборке и оценка по отношению обладает высокой точностью. В другом случае за x_i может быть принята, скажем, общая посевная площадь на ферме, а за y_i — площадь, занятая некоторой культурой. В этом случае оценка по отношению даст хорошие результаты, если все фермеры отводят под эту культуру приблизительно один и тот же процент общей площади.

Если оценивается величина \bar{Y} — среднее значение y_i для совокупности, то оценкой по отношению будет

$$\hat{\bar{Y}}_R = \frac{y}{x} \bar{X}.$$

Часто бывает, что мы хотим оценить не суммарное или среднее значение, а некоторое отношение, например отношение площади под пшеницей к площади под зерновыми культурами, отношение затрат на рабочую силу к общим затратам или отношение срочных вкладов к общей сумме вкладов. Выборочной оценкой служит $\hat{R} = y/x$. В этом случае X знать не нужно. Применение оценок по отношению для этих целей уже рассматривалось в параграфе 2.9 и в параграфе 3.12 (при описании гнездового отбора для оценки долей).

Пример. В табл. 6.1 указано число жителей (в тысячах) каждого из 49 городов, попавших в простую случайную выборку из совокупности, насчитывающей 196 больших городов (см. пример в параграфе 2.13). Нужно оценить общее число жителей 196 городов в 1930 г. Пред-

Таблица 6.1

ВЕЛИЧИНА 49 КРУПНЫХ ГОРОДОВ США (по числу жителей, в тыс.)
в 1920 (x_i) и 1930 гг. (y_i)

x_i	y_i	x_i	y_i	x_i	y_i
76	80	2	50	243	291
138	143	507	634	87	105
67	67	179	260	30	111
29	50	121	113	71	79
381	464	50	64	256	288
23	48	44	58	43	61
37	63	77	89	25	57
120	115	64	63	94	85
61	69	64	77	43	50
387	459	56	142	298	317
93	104	40	60	36	46
172	183	40	64	161	232
78	106	38	52	74	93
66	86	136	139	45	53
60	57	116	130	36	54
46	65	46	53	50	58
				48	75

полагается известным истинное суммарное значение для 1920 г., X . Оно равно 22 919.

В этом примере можно применить оценку по отношению. В большинстве городов в выборке число жителей увеличилось с 1920 по 1930 г. примерно на 20%. По данным выборки имеем

$$y = \sum y_i = 6262, \quad x = \sum x_i = 5054.$$

Следовательно, оценка по отношению числа жителей всех 196 городов в 1930 г. будет

$$\hat{Y}_R = \frac{y}{x} X = \frac{6262}{5054} \cdot 22\,919 = 28\,397.$$

Соответствующая оценка, основанная на выборочном среднем в расчете на один город, будет

$$\hat{Y} = N\bar{y} = \frac{196 \cdot 6262}{49} = 25\,048.$$

Правильное суммарное значение для 1930 г. составляет 29 351.

На рис. 6.1 изображены распределения значений оценки по отношению и оценки, основанной на выборочном среднем числе жителей на один город (оценка простым распространением), полученных по 200 простым случайным выборкам из этой совокупности объемом в 49 единиц каждая. Существенное увеличение точности при применении способа отношения очевидно.

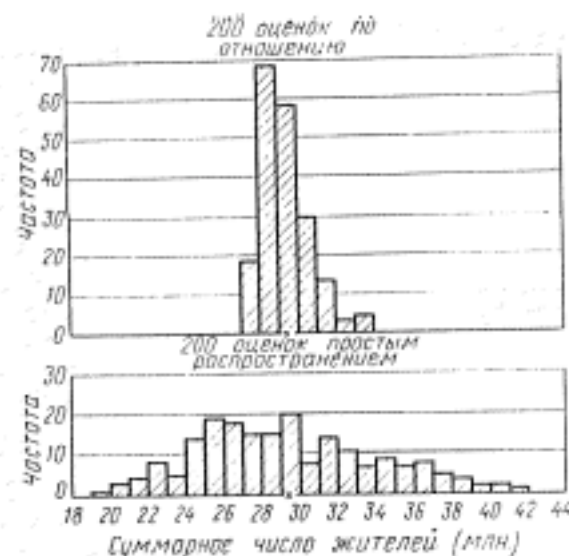


Рис. 6.1. Экспериментальное сравнение оценки по отношению с оценкой, основанной на выборочном среднем. X обозначает действительное суммарное число жителей

6.3. ПРИБЛИЖЕННАЯ ДИСПЕРСИЯ ОЦЕНКИ ПО ОТНОШЕНИЮ

Распределение оценки по отношению оказывается крайне неудобным для исследования из-за того, что и y и x меняются от выборки к выборке. По сравнению с тем, что хотелось бы знать для практического применения этого способа, теоретические результаты, известные нам до сих пор, недостаточны. Сформулируем основные положения сначала без доказательства.

Оценка по отношению есть, очевидно, состоятельная оценка. За исключением некоторых особых видов совокупностей, эта оценка будет смещенной, хотя при больших выборках смещением можно пренебречь. Распределение оценки по отношению при возрастании n стремится к нормальному при некоторых не очень жестких ограничениях относительно вида совокупности, подвергающейся выборочному исследованию. Для совокупностей, к которым чаще всего применяется этот способ, при выборках умеренного объема распределение обнаруживает тенденцию к положительной асимметрии. Точных формул для смещения и для выборочной дисперсии оценки у нас нет, имеются только приближенные формулы, справедливые для больших выборок.

Из приведенных положений вытекает, что не возникает никаких затруднений, если выборка достаточно велика для того, чтобы (а) отношение было распределено приблизительно нормально и (б) для его дисперсии была справедлива формула для больших выборок. В качестве рабочего правила можно принять, что результаты, относящиеся к большим выборкам, можно применять, если объем выборки

превышает 30 и достаточно велик, чтобы коэффициенты вариации \bar{x} и \bar{y} оба были меньше 10%.

Теорема 6.1. Оценки по отношению суммарного значения для совокупности, \bar{Y} , среднего для совокупности, \bar{Y} , и отношения Y/X есть соответственно

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}} \bar{X}; \quad \hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X}; \quad \hat{R} = \frac{\bar{y}}{\bar{x}}.$$

В простой случайной выборке объема n (n велико)

$$V(\hat{Y}_R) \approx \frac{N^2(1-f)}{n} \left[\frac{\sum_{i=1}^N (y_i - R x_i)^2}{N-1} \right]; \quad (6.2)$$

$$V(\hat{\bar{Y}}_R) \approx \frac{1-f}{n} \left[\frac{\sum_{i=1}^N (y_i - R x_i)^2}{N-1} \right]; \quad (6.3)$$

$$V(\hat{R}) \approx \frac{1-f}{n \bar{X}^2} \left[\frac{\sum_{i=1}^N (y_i - R x_i)^2}{N-1} \right], \quad (6.4)$$

где $f = n/N$ есть доля отбора.

Приближенная формула (6.4) была обоснована при доказательстве теоремы 2.5. Поскольку $\hat{Y}_R = \bar{X} \hat{R}$, $\hat{\bar{Y}}_R = N \bar{X} \hat{R}$, она сразу же приводит к формулам (6.2) и (6.3).

Следствие 1. Полученным результатам можно придать различную форму. Поскольку $\bar{Y} = R \bar{X}$, можно записать

$$\begin{aligned} V(\hat{Y}_R) &= \frac{N^2(1-f)}{n(N-1)} \sum_{i=1}^N [(y_i - \bar{Y}) - R(x_i - \bar{X})]^2 = \\ &= \frac{N^2(1-f)}{n(N-1)} [\sum (y_i - \bar{Y})^2 + R^2 \sum (x_i - \bar{X})^2 - \\ &\quad - 2R \sum (y_i - \bar{Y})(x_i - \bar{X})]. \end{aligned}$$

Коэффициент корреляции ρ между y_i и x_i для конечной совокупности определяется равенством

$$\rho = \frac{E(y_i - \bar{Y})(x_i - \bar{X})}{\sqrt{E(y_i - \bar{Y})^2 E(x_i - \bar{X})^2}} = \frac{\sum (y_i - \bar{Y})(x_i - \bar{X})}{(N-1) S_y S_x}.$$

Это приводит к

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x). \quad (6.5)$$

Эквивалентной записью будет

$$V(\hat{Y}_R) = (1-f) \frac{Y^2}{n} \left(\frac{S_y^2}{\bar{Y}^2} + \frac{S_x^2}{\bar{X}^2} - \frac{2S_{yx}}{\bar{Y}\bar{X}} \right), \quad (6.6)$$

где $S_{yx} = \rho S_y S_x$ есть ковариация между y_i и x_i . Это соотношение можно также записать в виде

$$V(\hat{Y}_R) = (1-f) \frac{Y^2}{n} (C_{yy} + C_{xx} - 2C_{yx}), \quad (6.7)$$

где C_{yy} , C_{xx} — квадраты коэффициентов вариации [в дальнейшем обозначаемых cv — от английского «coefficient of variation»] соответственно, y_i и x_i , а C_{yx} — относительная ковариация.

Следствие 2. Поскольку \hat{Y}_R , $\hat{\bar{Y}}_R$ и \hat{R} отличаются только известными множителями, коэффициент вариации (т. е. стандартная ошибка, деленная на оцениваемую величину) одинаков для всех трех оценок. Из (6.7) следует, что квадрат этого коэффициента вариации равен:

$$(cv)^2 = \frac{V(\hat{Y}_R)}{Y^2} = \frac{1-f}{n} (C_{yy} + C_{xx} - 2C_{yx}). \quad (6.8)$$

Величина $(cv)^2$ была названа Хансеном и др. (Hansen et al, 1953) *относительной дисперсией*. Она дает возможность не повторять формулы дисперсий для таких связанных одна с другой величин, как оценки среднего и суммарного значений для совокупности.

6.4. ДОСТОВЕРНОСТЬ ПРИБЛИЖЕННОГО ЗНАЧЕНИЯ ДИСПЕРСИИ

Сукхатм (Sukhatme, 1954) исследовал ошибку, связанную с применением приближенной формулы для $V(\hat{R})$. Напомним (см. теорему 2.5), что приближенная формула была получена на основе равенства

$$\hat{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}}$$

с заменой затем в знаменателе \bar{x} величиной \bar{X} . Вместо этого запишем:

$$\frac{1}{\bar{x}} = \frac{1}{\bar{X} + (\bar{x} - \bar{X})} = \frac{1}{\bar{X}} \left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}} \right)^{-1}$$

и разложим член, стоящий в скобках в правой части равенства, в ряд Тейлора. Это даст

$$\hat{R} - R = \frac{\bar{y} - R\bar{x}}{\bar{X}} \left[1 - \frac{\bar{x} - \bar{X}}{\bar{X}} + \frac{(\bar{x} - \bar{X})^2}{\bar{X}^2} - \dots \right]. \quad (6.9)$$

Возводя это равенство в квадрат, мы можем выразить $E(\hat{R} - R)^2$ через моменты совместного распределения y и x . К сожалению, полученное выражение будет слишком сложным, чтобы его можно было применить в практических целях, и здесь мы его не приводим.

Если y и x подчиняются двумерному нормальному распределению, то результат значительно упрощается. Сукхатм показал (Sukhatme, 1954), что с точностью до членов порядка $1/n^2$

$$E\left(\frac{\hat{R} - R}{R}\right)^2 \approx V_1 \left(1 + \frac{3C_{xx}}{n} \right) + \frac{6C_{xx}(\rho^2 C_{yy} + C_{xx} - 2C_{yx})}{n^2}, \quad (6.10)$$

где

$$V_1 = \frac{1}{n} (C_{yy} + C_{xx} - 2C_{yx})$$

служит первым приближением к относительной дисперсии \hat{R} , задаваемой (6.8), если пренебречь пкс. Вынося в (6.10) V_1 как общий множитель, имеем

$$E \left(\frac{\hat{R} - R}{R} \right)^2 \approx V_1 \left(1 + \frac{3C_{xx}}{n} + \frac{6C_{xx}}{n} \cdot \frac{\rho^2 C_{yy} + C_{xx} - 2C_{yx}}{C_{yy} + C_{xx} - 2C_{yx}} \right). \quad (6.11)$$

Поскольку последний член в скобках в правой части равенства меньше, чем $6C_{xx}/n$, это даст

$$E \left(\frac{\hat{R} - R}{R} \right)^2 < V_1 \left(1 + \frac{9C_{xx}}{n} \right) \quad (6.12)$$

с точностью до членов порядка $1/n^2$. Здесь C_{xx}/n — квадрат коэффициента вариации \bar{x} . Таким образом, если n достаточно велико для того, чтобы коэффициент вариации \bar{x} был меньше 0,1, то применение V_1 не приведет к преуменьшению оценки более чем на 9%. Практически множитель 9 в (6.12), по-видимому, слишком велик по сравнению с тем, что даст (6.11). Например, если $C_{xx} = C_{yy}$, то (6.11) сводится к

$$E \left(\frac{\hat{R} - R}{R} \right)^2 \approx V_1 \left[1 + \frac{C_{xx}}{n} (6 - 3\rho) \right].$$

Поскольку при применении способа отношения ρ почти всегда положительно, более оправданной была бы величина множителя, заключенная между 3 и 6. С другой стороны, отклонение от нормальности в распределениях y и x влияет также и на член порядка $1/n^2$.

Выражение $E(\hat{R} - R)^2$ представляет собой средний квадрат ошибки \hat{R} относительно истинного значения отношения R , а не дисперсию \hat{R} . Поскольку \hat{R} , вообще говоря, смещенная оценка, средний квадрат ошибки будет более правильной мерой ее достоверности, чем дисперсия.

6.5. СМЕЩЕНИЕ ОЦЕНКИ ПО ОТНОШЕНИЮ

Оценка по отношению имеет смещение порядка $1/n$. Поскольку стандартная ошибка оценки имеет порядок $1/\sqrt{n}$, отношение смещения к стандартной ошибке также имеет порядок $1/\sqrt{n}$ и становится пренебрежимо малым при достаточно больших n . На практике смещение обычно оказывается несущественным даже для выборок умеренного объема. Приведем три полезных результата, относящихся к смещению.

Первый из них указывает главный член смещения при его разложении в ряд Тейлора. Из (6.9), оставляя два первых члена, получаем

$$\hat{R} - R \approx \frac{\bar{y} - R\bar{x}}{\bar{x}} \left(1 - \frac{\bar{x} - \bar{X}}{\bar{x}} \right).$$

Так как

$$E(\bar{y} - R\bar{x}) = \bar{Y} - R\bar{X} = 0,$$

то главный член смещения определяется вторым членом внутри скобок. Далее, по теореме 2.3 (с. 39) и определению ρ

$$E\bar{y}(\bar{x} - \bar{X}) = E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) = \frac{1-f}{n} \rho S_y S_x.$$

Кроме того,

$$E\bar{x}(\bar{x} - \bar{X}) = E(\bar{x} - \bar{X})^2 = \frac{1-f}{n} S_x^2.$$

Следовательно, главный член смещения имеет вид

$$E(\hat{R} - R) \approx \frac{1-f}{n \bar{X}^2} (R S_x^2 - \rho S_y S_x). \quad (6.13)$$

Относительное смещение (т. е. смещение/ R), которое одинаково для \hat{R} , \hat{Y}_R и $\hat{\bar{Y}}_R$, будет

$$\frac{E(\hat{R} - R)}{R} \approx \frac{1-f}{n \bar{X} \bar{Y}} (R S_x^2 - \rho S_y S_x) \approx \frac{1-f}{n} (C_{xx} - C_{yx}). \quad (6.14)$$

Поскольку выборочные оценки R , S_x , S_y , ρ , \bar{X} и \bar{Y} можно вычислить, формулы (6.13) и (6.14) иногда применяются для грубого определения величины смещения в конкретной выборке.

Второй результат состоит в том, что оценка по отношению есть несмещенная оценка, если регрессия y по x выражается прямой, проходящей через начало координат. Это означает, что $E(y|x) = \beta x$. Для конечных совокупностей это равенство предполагает, что (а) если несколько единиц имеют одно и то же значение x , то среднее из их значений y равно βx , и (б) если некоторое значение x встречается только у одной единицы из совокупности, то значение y для этой единицы равно βx . Такие условия в конечной совокупности, вероятно, не будут выполняться в точности. Однако часто они будут выполняться приближенно, поскольку, как указывалось в параграфе 6.2, оценка по отношению, вероятно, будет применяться тогда, когда есть основания полагать, что y/x приблизительно постоянно.

Теорема 6.2. Если $E(y|x) = \beta x$ для всех значений x в конечной совокупности, то оценки \hat{Y}_R , $\hat{\bar{Y}}_R$ и \hat{R} в простых случайных выборках объема n будут несмещенными.

Доказательство. Запишем

$$y = \beta x + e.$$

Тогда

$$E(e|x) = 0 \quad (6.15)$$

для любого значения x . Беря среднее по всей совокупности, имеем

$$\bar{Y} = \beta \bar{X},$$

так что $R = \beta$. Далее, беря среднее по выборке, получаем

$$\frac{\bar{y}}{\bar{x}} = \beta + \frac{\bar{e}}{\bar{x}},$$

т. е.

$$\hat{R} = R + \frac{\bar{e}}{\bar{x}}.$$

Возьмем среднее \hat{R} по всем выборкам объема n , содержащим один и тот же набор значений x , так что \bar{x} при этом усреднении не изменится. Предположим, что какое-то определенное значение x , скажем x' , встречается у m' единиц выборки и у M' единиц совокупности. Тогда при таком усреднении каждая из M' единиц будет встречаться в выборках одинаково часто. Но согласно (6.15)

$$E(e | x') = 0.$$

Следовательно, для рассматриваемого множества выборок $E(\bar{e}) = 0$. Отсюда вытекает, что $E(\hat{R}) = R$ по этому множеству выборок и что $E(\hat{R}) = R$ по всем простым случайным выборкам объема n .

Третий результат, полученный Хартли и Россом (Hartley and Ross, 1954), указывает верхнюю границу отношения смещения к стандартной ошибке. Рассмотрим, в простых случайных выборках объема n , ковариацию величин \hat{R} и \bar{x} . Имеем

$$\text{cov}(\hat{R}, \bar{x}) = E\left(\frac{\bar{y}}{\bar{x}} \cdot \bar{x}\right) - E(\hat{R})E(\bar{x}) = \bar{Y} - \bar{X}E(\hat{R}).$$

Следовательно,

$$E(\hat{R}) = \frac{\bar{Y}}{\bar{X}} - \frac{1}{\bar{X}} \text{cov}(\hat{R}, \bar{x}) = R - \frac{1}{\bar{X}} \text{cov}(\hat{R}, \bar{x}). \quad (6.16)$$

Таким образом, смещение \hat{R} равно $-\text{cov}(\hat{R}, \bar{x})/\bar{X}$. В отличие от приближенного выражения смещения в (6.13) при разложении в ряд Тейлора здесь мы получили его точное выражение.

Далее,

$$|\text{смещение } \hat{R}| = \frac{|\rho_{\hat{R}, \bar{x}} \sigma_{\hat{R}} \sigma_{\bar{x}}|}{\bar{X}} \leq \frac{\sigma_{\hat{R}} \sigma_{\bar{x}}}{\bar{X}},$$

поскольку коэффициент корреляции между \hat{R} и \bar{x} не может быть больше 1. Следовательно,

$$\frac{|\text{смещение } \hat{R}|}{\sigma_{\hat{R}}} \leq \frac{\sigma_{\bar{x}}}{\bar{X}} = \text{коэффициент вариации } \bar{x}. \quad (6.17)$$

Та же самая граница применима, конечно, и для смещений оценок \hat{Y}_R и $\hat{\bar{Y}}_R$. Таким образом, если коэффициент вариации \bar{x} меньше 0,1, то смещение вполне можно считать пренебрежимо малым по сравнению со стандартной ошибкой.

6.6. ОЦЕНИВАНИЕ ДИСПЕРСИИ ПО ВЫБОРКЕ

Согласно (6.2)

$$V(\hat{Y}_R) \approx \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^N (y_i - R x_i)^2}{N-1}.$$

Как уже указывалось в параграфе 2.9, в качестве выборочной оценки дисперсии для совокупности мы принимаем

$$\frac{\sum_{i=1}^n (y_i - \hat{R} x_i)^2}{(n-1)}.$$

Эта оценка имеет смещение порядка $1/n$.

Для оценки дисперсии, $v(\hat{Y}_R)$, получаем

$$\begin{aligned} v(\hat{Y}_R) &= \frac{N(N-n)}{n(n-1)} \sum_{i=1}^n (y_i - \hat{R} x_i)^2 = \\ &= \frac{N(N-n)}{n(n-1)} (\sum y_i^2 + \hat{R}^2 \sum x_i^2 - 2\hat{R} \sum y_i x_i). \end{aligned} \quad (6.18)$$

Этот вид наиболее удобен для вычислений.

Пример. Этот пример иллюстрирует вычисление стандартной ошибки оценки по отношению суммарного значения для совокупности. Данные взяты из табл. 6.1 (с. 174). Сначала вычислим $y = \sum y_i = 6262$; $x = \sum x_i = 5054$; $\hat{R} = \frac{y}{x} = 1,239019$. Из (6.18)

$$v(\hat{Y}_R) = \frac{N(N-n)}{n(n-1)} (\sum y_i^2 + \hat{R}^2 \sum x_i^2 - 2\hat{R} \sum y_i x_i).$$

Для вычисления выражения в скобках суммы квадратов и произведения расположены в тех же строках, что и соответствующие множители

	Множитель
$\sum y_i^2 = 1\,527\,882$	1
$\sum x_i^2 = 1\,044\,504$	$1,535168 = \hat{R}^2$
$\sum y_i x_i = 1\,251\,630$	$2,478038 = 2\hat{R}$

Следовательно,

$$v(\hat{Y}_R) = \frac{196 \cdot 147}{49 \cdot 48} \cdot (29\,784) = 364\,854;$$

$$s(\hat{Y}_R) = 604.$$

6.7. ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ

Если выборка достаточно велика для того, чтобы можно было применить аппроксимацию нормальным распределением, то доверительные границы для Y и R можно получить по формулам:

$$Y: \hat{Y}_R \pm t \sqrt{v(\hat{Y}_R)}, \quad (6.19)$$

$$R: \hat{R} \pm t \sqrt{v(\hat{R})}, \quad (6.20)$$

где t — квантиль нормального распределения, соответствующий заданной доверительной вероятности.

В параграфе 6.3 указывалось, что аппроксимация нормальным распределением вполне удовлетворительна, если объем выборки не менее 30 и достаточно велик для того, чтобы коэффициенты вариации \bar{y} и \bar{x} оба были меньше 0,1. Если эти условия не выполняются, то формула для $v(\hat{R})$ будет давать (чаще всего) слишком маленькие значения, а положительная асимметрия распределения \hat{R} может оказаться заметно выраженной.

Другой метод вычисления доверительных границ применялся в биологических исследованиях (Fieller, 1932; Paulson, 1942). При этом подходе требуются менее ограничительные предположения, чем аппроксимация нормальным распределением, и в определенной степени учитывается асимметрия распределения \hat{R} .

Этот метод предполагает, что \bar{y} и \bar{x} подчиняются двумерному нормальному распределению, так что $(\bar{y} - R\bar{x})$ имеет нормальное распределение. Из этого следует, что величина

$$\frac{\bar{y} - R\bar{x}}{\sqrt{[(N-n)/Nn] \left(\frac{s_y^2}{N} + R^2 \frac{s_x^2}{N} - 2R s_y s_x \right)}} \quad (6.21)$$

распределена приблизительно нормально со средним значением, равным нулю, и дисперсией, равной единице. (Мы подставляем выборочные оценки s_y^2 и т. д. вместо соответствующих дисперсий и ковариаций для совокупности и предполагаем, что выборка достаточно велика для того, чтобы ошибкой, вносимой этой заменой, можно было пренебречь. В биологических исследованиях, в которых выборки могут быть весьма малыми, введенная только что величина должна считаться подчиняющейся t -распределению Стьюдента.)

Значение R неизвестно, однако на основании данных выборки можно считать отвергнутым какое-либо предположительное значение R , при котором соответствующий квантиль оказывается слишком большим. Следовательно, доверительные границы для R можно найти, приравняв (6.21) к $\pm t$ и решая полученное таким образом квадратное уравнение относительно R . Эти доверительные границы будут приближенными, поскольку, если мы попытаемся проверить их, производя многократный отбор из какой-либо неизменной совокупности с известным R , то некоторые значения \bar{y} и \bar{x} будут такими, что оба корня квад-

ратного уравнения окажутся мнимыми. Таких значений будет меньше, если коэффициенты вариации \bar{y} и \bar{x} меньше 0,3.

После некоторых преобразований оба корня можно записать в виде:

$$R = \hat{R} \frac{(1 - t^2 c_{yx}) \pm t \sqrt{(c_{yy} + c_{xx} - 2c_{yx}) - t^2 (c_{yy} c_{xx} - c_{yx}^2)}}{1 - t^2 c_{xx}}, \quad (6.22)$$

где

$$c_{yy} = \frac{N-n}{Nn} \cdot \frac{s_y^2}{\bar{y}^2}$$

есть квадрат оценки коэффициента вариации \bar{y} . Аналогично определяются c_{xx} и c_{yx} . Если каждая из величин $t^2 c_{yy}$, $t^2 c_{xx}$ и $t^2 c_{yx}$ мала по сравнению с 1, то значения границ сводятся к

$$R = \hat{R} \pm t \hat{R} \sqrt{c_{yy} + c_{xx} - 2c_{yx}}.$$

Мы получаем то же выражение, что и при аппроксимации нормальным распределением (6.20).

Того же рода квадратичные границы для Y получают, подставляя в (6.22) \hat{Y}_R вместо \hat{R} .

Хотя эти квадратичные границы должны быть, вообще говоря, более точными, чем границы при нормальной аппроксимации в (6.20), поскольку они получены при менее ограничительных предположениях, Гаек (Hájek, 1958) показал, что если линия регрессии y по x проходит через начало координат, то при больших выборках нормальные границы содержат R с большей частотой, чем квадратичные.

6.8. СРАВНЕНИЕ ОЦЕНКИ ПО ОТНОШЕНИЮ С ОЦЕНКОЙ ПО СРЕДНЕМУ НА ЕДИНИЦУ

В предшествующих главах исследовалась оценка Y вида $N\bar{y}$, где \bar{y} — среднее значение на единицу в выборке (при простом случайном отборе) или взвешенное среднее на единицу (при расслоенном случайном отборе). Оценки такого типа называются оценками по среднему на единицу или оценками, полученными простым распространением.

Теорема 6.3. Для больших выборок при простом случайном отборе оценка по отношению \hat{Y}_R имеет меньшую дисперсию, чем оценка $\hat{Y} = N\bar{y}$, получаемая простым распространением, если

$$\rho > \frac{1}{2} \left(\frac{S_x}{\bar{x}} \right) \left(\frac{S_y}{\bar{y}} \right) = \frac{\text{коэффициент вариации } x_i}{2 (\text{коэффициент вариации } y_i)}.$$

Доказательство. Для \hat{Y} имеем

$$V(\hat{Y}) = \frac{N^2(1-f)}{n} S_y^2.$$

Для оценки по отношению имеем из (6.5)

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x).$$

Следовательно, оценка по отношению имеет меньшую дисперсию, если

$$S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x < S_y^2,$$

т. е. если

$$\rho > \frac{RS_x}{2S_y} = \frac{1}{2} \left(\frac{S_x}{\bar{X}} \right) / \left(\frac{S_y}{\bar{Y}} \right).$$

Эта теорема показывает, что оценка по отношению может быть как более, так и менее точной, чем оценка, полученная простым распространением, в зависимости от величины коэффициента корреляции между y_i и x_i и от коэффициентов вариации этих двух переменных. Изменчивость вспомогательной переменной x_i играет важную роль: если ее коэффициент вариации более чем в два раза превышает коэффициент вариации y_i , то оценка по отношению всегда будет менее точной, поскольку ρ не может превзойти 1. Если величиной x_i служит значение y_i на некоторый предшествующий момент времени, то оба коэффициента вариации могут быть приблизительно одинаковыми. В этом случае оценка по отношению будет лучшей, если ρ больше 0,5.

Теорема 6.3 применима только к достаточно большим выборкам, когда справедлива приближенная формула для $V(\bar{Y}_R)$. Для выборок меньшего объема метод отношения, по-видимому, дает меньший выигрыш, чем это вытекает из теоремы, потому что приближенная формула обычно преуменьшает значение дисперсии.

6.3 УСЛОВИЯ, ПРИ КОТОРЫХ ОЦЕНКА ПО ОТНОШЕНИЮ ОПТИМАЛЬНА

В регрессионном анализе хорошо известна теорема, указывающая вид совокупности, для которой оценка по отношению есть наилучшая в широком классе оценок. Теорема относится к бесконечным совокупностям.

Теорема 6.4. При простом случайном отборе из бесконечной совокупности оценка по отношению есть «наилучшая линейная несмещенная оценка» для \bar{Y} , если выполнены два условия:

1) зависимость между y_i и x_i выражается прямой, проходящей через начало координат;

2) дисперсия y_i относительно этой прямой пропорциональна x_i .

«Наилучшая линейная несмещенная оценка» определяется следующим образом. Рассмотрим все оценки, представляющие собой линейные функции выборочных значений y_i , т. е. имеющие вид

$$l_1 y_1 + l_2 y_2 + \dots + l_n y_n,$$

где величины l_i не зависят от y_i , хотя могут быть функциями x_i .

При этом рассматриваются лишь те l_i , которые дают несмещенные оценки \bar{Y} . Оценка с наименьшей при этих условиях дисперсией называется «наилучшей линейной несмещенной оценкой».

Доказательство. Математическая модель имеет вид

$$y_i = Bx_i + e_i,$$

где e_i не зависят от x_i . При условии, что x_i неизменно, e_i имеет среднее значение нуль и дисперсию λx_i . Следовательно,

$$\bar{Y} = B\bar{X}.$$

Еще Гаусс показал, что наилучшей линейной несмещенной оценкой $B\bar{X}$ служит $b\bar{X}$, где b — оценка B , полученная методом наименьших квадратов [см., например, работу Дейвида и Неймана (David and Neuman, 1938)]. Такая оценка имеет вид

$$b = \frac{\sum w_i y_i x_i}{\sum w_i x_i^2}, \quad \text{где} \quad w_i = \frac{1}{\sigma_{e_i}^2} = \frac{1}{\lambda x_i}.$$

Это дает

$$b = \frac{\sum y_i}{\sum x_i} = \frac{\bar{y}}{\bar{x}}.$$

Следовательно, оптимальной оценкой \bar{Y} будет оценка по отношению $(\bar{y}/\bar{x}) \bar{X}$.

Практическая ценность этого результата состоит в том, что он указывает условия, при которых оценка по отношению не только превосходит оценку по среднему на единицу, но и есть наилучшая в целом классе оценок. Для того чтобы решить какого рода оценку применить, полезно построить график, на котором значения x_i поставлены в соответствие значения y_i . Если точки на этом графике хорошо приближаются прямой, проходящей через начало координат, и дисперсия точек y_i относительно этой прямой увеличивается приблизительно пропорционально x_i , то мы имеем веский довод в пользу применения оценки по отношению.

Иногда зависимость между y_i и x_i выражается прямой, проходящей через начало координат, но дисперсия y_i при условии, что x_i неизменно, не пропорциональна x_i . При выборочном обследовании населения в Греции Джессен и др. (Jessen et al., 1947) установили, что дисперсия увеличивается приблизительно как x_i^2 . Это означает, что регрессия в данном случае взвешенная, причем веса $w_i \propto 1/x_i^2$. Для оценки по методу наименьших квадратов, b , это дает

$$b = \frac{\sum w_i y_i x_i}{\sum w_i x_i^2} = \frac{1}{n} \sum \left(\frac{y_i}{x_i} \right).$$

В этой ситуации наилучшей оценкой \bar{Y} будет $b\bar{X}$, где b — среднее значение отношений y_i/x_i для отдельных единиц отбора.

6.10. ОЦЕНКА ПО ОТНОШЕНИЮ ПРИ РАССЛОЕННОМ СЛУЧАЙНОМ ОТБОРЕ

Оценка по отношению суммарного значения для совокупности, Y , может быть получена двумя способами. Один из них состоит в том, чтобы получить *раздельные* оценки по отношению суммарного значения для каждого слоя и сложить их. Если y_h, x_h — суммарные выборочные значения в h -м слое и X_h — суммарное значение x_{hi} для этого слоя, такая раздельная оценка, \hat{Y}_{Rs} [с — от английского «separate» — раздельный], имеет вид

$$\hat{Y}_{Rs} = \sum_h \frac{y_h}{x_h} X_h = \sum_h \frac{\bar{y}_h}{\bar{x}_h} X_h. \quad (6.23)$$

Предположения о том, что истинные отношения неизменны от слоя к слою, не делается. Однако отдельные суммарные значения, X_h , должны быть известны.

Теорема 6.5. Если объемы выборки, n_h , во всех слоях велики, то

$$V(\hat{Y}_{Rs}) = \sum_h \frac{N_h^2(1-f_h)}{n_h} (S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h \rho_h S_{y_h} S_{x_h}), \quad (6.24)$$

где $R_h = Y_h/X_h$ есть истинное значение отношения в слое h , а ρ_h определяется для каждого слоя, как и ранее.

Доказательство. Запишем

$$\hat{Y}_{Rh} = \frac{y_h}{x_h} X_h.$$

Тогда

$$\hat{Y}_{Rs} - Y = \sum_h (\hat{Y}_{Rh} - Y_h).$$

Следовательно,

$$V(\hat{Y}_{Rs}) = E(\hat{Y}_{Rs} - Y)^2 = \sum_h E(\hat{Y}_{Rh} - Y_h)^2 + 2 \sum_h \sum_{j > h} E(\hat{Y}_{Rh} - Y_h)(\hat{Y}_{Rj} - Y_j).$$

Поскольку \hat{Y}_{Rh} есть оценка по отношению, полученная для простой случайной выборки в слое h , мы можем воспользоваться формулой (6.5) для приближенного выражения дисперсии \hat{Y}_{Rh} , а именно

$$V(\hat{Y}_{Rh}) = \frac{N_h^2(1-f_h)}{n_h} (S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h \rho_h S_{y_h} S_{x_h}).$$

Члены двойной суммы обращаются в нуль, потому что отбор по различным слоям происходит независимо и с точностью до порядка приближения, принятого в формуле дисперсии, \hat{Y}_{Rh} будет несмещенной оценкой Y_h . Отсюда следует формула (6.24).

Эта формула справедлива только для случая, когда выборка во всех слоях достаточно велика, чтобы для каждого из них можно было применить приближенную формулу дисперсии. В практической работе нужно помнить об этом условии.

Более того, как следует из приводимых далее общих соображений, когда n_h малы, а число слоев L велико, смещение \hat{Y}_{Rs} по сравнению со стандартной ошибкой этой оценки может оказаться таким, что пренебречь им будет невозможно.

Для отдельного слоя было показано (параграф 6.5), что

$$\frac{|\text{смещение } \hat{Y}_{Rh}|}{\sigma(\hat{Y}_{Rh})} \leq \text{коэффициент вариации } \bar{x}_h.$$

Если бы смещение во всех слоях имело один и тот же знак, что вполне возможно, то смещение \hat{Y}_{Rs} было бы приблизительно в L раз больше, чем смещение \hat{Y}_{Rh} . Но стандартная ошибка \hat{Y}_{Rs} в этом случае была бы больше стандартной ошибки \hat{Y}_{Rh} только примерно в \sqrt{L} раз. Следовательно, отношение

$$\frac{|\text{смещение } \hat{Y}_{Rs}|}{\sigma(\hat{Y}_{Rs})}$$

имеет величину порядка

$$\sqrt{L} \times (\text{коэффициент вариации } \bar{x}_h).$$

Например, если имеется 50 слоев и коэффициент вариации \bar{x}_h в каждом слое равен приблизительно 0,1, то смещение оценки \hat{Y}_{Rs} может составлять 0,7 величины ее стандартной ошибки. В этом случае смещение составило бы около одной трети среднего квадрата ошибки \hat{Y}_{Rs} .

Хотя на практике смещение бывает обычно гораздо меньше его возможной верхней границы, в случае, когда оценка по отношению вычисляется отдельно для каждого слоя и $\sqrt{L} \times$ (коэффициент вариации \bar{x}_h) больше, скажем, 0,3, опасность смещения необходимо иметь в виду.

6.11. СОВМЕСТНАЯ ОЦЕНКА ПО ОТНОШЕНИЮ

Другая оценка выводится из единственного *совместного* отношения (Hansen, Hurwitz and Gurney, 1946). По данным выборки мы вычисляем

$$\hat{Y}_{st} = \sum_h N_h \bar{y}_h; \quad \hat{X}_{st} = \sum_h N_h \bar{x}_h.$$

Это обычные оценки суммарных значений для совокупности соответственно Y и X , сделанные по расслоенной выборке. Совместная оценка по отношению, \hat{Y}_{Rc} [с — от английского «combined» — совместный], имеет вид

$$\hat{Y}_{Rc} = \frac{\hat{Y}_{st}}{\hat{X}_{st}} X = \frac{\bar{y}_{st}}{\bar{x}_{st}} X,$$

где $\bar{y}_{st} = \hat{Y}_{st}/N$, $\bar{x}_{st} = \hat{X}_{st}/N$ есть оценки средних для совокупности, полученные по расслоенной выборке.

Оценка \hat{Y}_{Rc} не требует знания X_h , но должно быть известно X . Совместная оценка гораздо менее подвержена опасности смещения, чем раздельная оценка. Пользуясь уже упомянутым в параграфе 6.5 приемом, предложенным Хартли и Россом, и считая $\hat{R}_c = \bar{y}_{st}/\bar{x}_{st}$, имеем

$$\text{cov}(\hat{R}_c, \bar{x}_{st}) = E\left(\frac{\bar{y}_{st}}{\bar{x}_{st}} \cdot \bar{x}_{st}\right) - E(\hat{R}_c) E(\bar{x}_{st}) = \bar{Y} - \bar{X} E(\hat{R}_c).$$

Следовательно,

$$E(\hat{R}_c) = R - \frac{1}{\bar{X}} \text{cov}(\hat{R}_c, \bar{x}_{st})$$

и

$$\frac{|\text{смещение } \hat{R}_c|}{\sigma_{\hat{R}_c}} = \frac{|\rho_{\hat{R}_c, \bar{x}_{st}} \cdot \sigma_{\bar{x}_{st}}|}{\bar{X}} \leq \text{cv}\bar{x}_{st}.$$

Таким образом, смещения оценок \hat{R}_c и \hat{Y}_{Rc} пренебрежимо малы по отношению к их стандартным ошибкам при единственном условии, что коэффициент вариации \bar{x}_{st} меньше 0,1.

Теорема 6.6. Если общий объем выборки n велик, то

$$V(\hat{Y}_{Rc}) = \sum_h \frac{N_h^2(1-f_h)}{n_h} (S_{yh}^2 + R^2 S_{xh}^2 - 2R\rho_h S_{yh} S_{xh}). \quad (6.25)$$

Доказательство. Следуем тем же рассуждениям, что и в теореме 2.5. В данном случае основное равенство имеет вид

$$(\hat{Y}_{Rc} - Y) = \frac{N\bar{X}}{\bar{x}_{st}} (\bar{y}_{st} - R\bar{x}_{st}) \approx N(\bar{y}_{st} - R\bar{x}_{st}). \quad (6.26)$$

Рассмотрим теперь переменную $u_{hi} = y_{hi} - Rx_{hi}$. Правая часть (6.26) равна $N\bar{u}_{st}$, где \bar{u}_{st} — взвешенное среднее значение переменной u_{hi} для расслоенной выборки. Далее, среднее значение u_{hi} для совокупности, \bar{U} , равно нулю, так как $R = \bar{Y}/\bar{X}$.

Следовательно, мы можем применить к \bar{u}_{st} теорему 5.3 о дисперсии оценки среднего для расслоенной случайной выборки. Это дает

$$V(\hat{Y}_{Rc}) = N^2 V(\bar{u}_{st}) = \sum_h \frac{N_h(N_h - n_h)}{n_h} S_{uh}^2,$$

где

$$S_u^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (u_{hi} - \bar{U}_h)^2 =$$

$$= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} [(y_{hi} - \bar{Y}_h) - R(x_{hi} - \bar{X}_h)]^2.$$

После раскрытия квадратных скобок получаем равенство (6.25).

Интересно отметить, что в уравнениях (6.24) и (6.25) приближенные формулы дисперсий \hat{Y}_{Rc} и \hat{Y}_{Rd} имеют почти один и тот же вид; различие состоит лишь в том, что все отношения для отдельных слоев R_h в (6.24) заменяются отношением R в (6.25).

6.12. СРАВНЕНИЕ СОВМЕСТНОЙ И РАЗДЕЛЬНОЙ ОЦЕНОК

Мы можем записать

$$\begin{aligned} V(\hat{Y}_{Rc}) - V(\hat{Y}_{Rd}) &= \\ &= \sum_h \frac{N_h^2(1-f_h)}{n_h} [(R^2 - R_h^2) S_{xh}^2 - 2(R - R_h) \rho_h S_{yh} S_{xh}] = \\ &= \sum_h \frac{N_h^2(1-f_h)}{n_h} [(R - R_h)^2 S_{xh}^2 + 2(R_h - R) (\rho_h S_{yh} S_{xh} - R_h S_{xh}^2)]. \end{aligned}$$

В случаях, когда применима оценка по отношению, последний член в правой части равенства обычно невелик. (Если внутри каждого слоя зависимость между y_{hi} и x_{hi} выражается прямой, проходящей через начало координат, то этот член исчезает.) Таким образом, если только R_h во всех слоях не одинаковы, раздельная оценка по отношению, вероятно, окажется более точной. При этом предполагается, однако, что выборка в каждом слое достаточно велика, чтобы оставалась справедливой приближенная формула для дисперсии $V(\hat{Y}_{Rd})$. Если же в каждом слое выборка содержит лишь небольшое число единиц, то рекомендуется совместная оценка, коль скоро против нее нет убедительных практических соображений.

Для получения оценок этих дисперсий по данным выборки мы подставляем на соответствующие места выборочные оценки R_h и R . Вместо соответствующих дисперсий подставляются выборочные средние квадраты s_{yh}^2 и s_{xh}^2 , а вместо члена $\rho_h S_{yh} S_{xh}$ — выборочная ковариация. Выборочные средние квадраты и ковариация должны быть вычислены отдельно для каждого слоя.

Пример. Данные получены при переписи всех ферм графства Джефферсон, штат Айова. В этом примере y_{hi} обозначает число акров под пшеницей и x_{hi} — общее число акров на ферме. Совокупность разделена на два слоя, первый слой содержит фермы размером до 160 акров. Предполагается, что выборка содержит 100 ферм. В том случае, когда применяется расслоенный отбор, будем считать, что из слоя 1 отобрано 70 ферм, а из слоя 2 — 30. Это примерно соответствует оптимальному размещению. Данные приведены в табл. 6.2. Последние три величины в ней Q_h , V_h' и V_h'' как вспомогательные понадобятся при вычислениях, причем две последние будут определены позднее.

Таблица 6.2
ДАННЫЕ ПО ГРАФСТВУ ДЖЕФФЕРСОН, ШТАТ АЙОВА

Слой	Размер ферм (в акрах)	N_h	S_{yh}^2	S_{ykh}	S_{xh}^2	R_h
1	0—160	1580	312	494	2055	0,2350
2	Более 160	430	922	858	7357	0,2109
Для всей совокупности		2010	620	1453	7619	0,2242
Слой	\bar{Y}_h	\bar{X}_h	n_h	$Q_h = W_h^2/n_h$	V_h'	V_h''
1	19,40	82,56	70	0,008828	193	194
2	51,63	244,85	30	0,001525	887	907
Для всей совокупности	26,30	117,28	100			

Мы рассмотрим пять способов оценивания средней площади под пшеницей на одну ферму для всех ферм графства. Пкс не учитывается.

1. Простая случайная выборка. Оценка по среднему на одну ферму:

$$V_1 = \frac{S_y^2}{n} = \frac{620}{100} = 6,20.$$

2. Простая случайная выборка. Оценка по отношению:

$$V_2 = \frac{1}{n} (S_y^2 + R^2 S_x^2 - 2RS_{yx}) = \\ = \frac{1}{100} [620 + (0,2242)^2 \cdot 7619 - 2 \cdot 0,2242 \cdot 1453] = 3,51.$$

3. Расслоенная случайная выборка. Оценка по среднему на одну ферму:

$$V_3 = \sum \frac{W_h^2}{n_h} S_{yh}^2 = \sum Q_h S_{yh}^2 = 4,16.$$

4. Расслоенная случайная выборка. Оценка по отношению, с применением в каждом слое отдельного отношения:

$$V_4 = \sum Q_h (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{ykh}) = \sum Q_h V_h' = 3,06.$$

5. Расслоенная случайная выборка. Оценка по отношению, с применением совместного отношения:

$$V_5 = \sum Q_h (S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{ykh}) = \sum Q_h V_h'' = 3,10.$$

Данные сравнительной точности разных способов отбора и оценивания сведены в следующую таблицу:

Метод отбора	Способ оценивания	Относительная точность
1. Простой случайный	Среднее на ферму	100
2. Простой случайный	Отношение	177
3. Расслоенный случайный	Среднее на ферму	149
4. Расслоенный случайный	Раздельное отношение	203
5. Расслоенный случайный	Совместное отношение	200

Эти результаты обнаруживают интересное обстоятельство, имеющее практическое значение. Расслоение по величине фермы преследует ту же общую цель, что и оценка по отношению, в знаменателе которой стоит величина фермы. Оба приема уменьшают влияние вариации величины фермы на ошибку выборки, которой подвержена оценка средней площади под пшеницей на одну ферму. Выигрыш в точности от применения оценки по отношению составляет, например, 77% при простом случайном отборе и только 36% (203 по сравнению со 149) — при расслоенном случайном отборе.

При планировании обследования мы можем оказаться перед выбором: учитывать ли некоторый фактор путем расслоения или же применяя тот или иной способ оценивания. Наилучшее решение зависит от различных обстоятельств. В частности, нужно принять во внимание следующие соображения: (а) некоторые факторы, например географическое расположение, легче учесть при расслоении, чем в способе оценивания; (б) решение зависит от характера зависимости между y_i и x_i . Все простые способы оценивания дают больший эффект, если эта зависимость линейна. Если зависимость имеет сложный или разрывный характер, то более эффективным может оказаться расслоение, поскольку при достаточно большом числе слоев оно будет исключать влияние почти любого вида зависимости между y_i и x_i ; (в) если некоторые важные переменные приблизительно пропорциональны x_i , а другие приблизительно пропорциональны другой переменной z_i , то лучше принять x_i и z_i в качестве знаменателей для оценок по отношению, чем производить расслоение по одной из этих переменных.

6.13. УПРОЩЕННОЕ ВЫЧИСЛЕНИЕ ДИСПЕРСИИ

Для случая $n_h = 2$ для всех слоев Кейфитц (Keyfitz, 1957) предложил упрощенные методы вычисления относительной дисперсии \hat{Y}_{Re} или \hat{R}_e . На основании (6.25), подставляя оценки, полученные по выборке, имеем

$$\frac{v(\hat{Y}_{Re})}{\hat{Y}_{Re}^2} = \sum_h \frac{N_h^2 (1-f_h)}{2} \left(\frac{s_{yh}^2}{\bar{Y}_{st}^2} + \frac{s_{xh}^2}{\bar{X}_{st}^2} - \frac{2r_{yh} s_{yh} s_{xh}}{\bar{Y}_{st} \bar{X}_{st}} \right).$$

Пусть

$$y'_{h1} = \frac{N_h y_{h1}}{2}; \quad y'_{h2} = \frac{N_h y_{h2}}{2}; \quad dy'_h = y'_{h1} - y'_{h2}.$$

Аналогичные определения введем для x . Тогда при $n_h = 2$ легко показать как алгебраическое тождество, что

$$\frac{N_h^2 s_{y'h}^2}{2\bar{y}_{st}^2} = \frac{1}{\bar{y}_{st}^2} \left(\frac{N_h (y_{h1} - y_{h2})}{2} \right)^2 = \left(\frac{dy'_h}{\bar{y}_{st}} \right)^2.$$

Аналогичные выражения можно получить для членов, содержащих s_{xh}^2 и $r_h s_{yxh}$. Следовательно, относительную дисперсию можно вычислять как

$$\sum_h \left(\frac{dy'_h}{\bar{y}_{st}} - \frac{dx'_h}{\bar{x}_{st}} \right)^2.$$

В этом выражении члены, учитывающие пкс, опущены. Если f_h приблизительно постоянны, то можно применить множитель $1 - \bar{f}_h$, где $\bar{f}_h = \Sigma f_h / L$.

6.14. ОПТИМАЛЬНОЕ РАЗМЕЩЕНИЕ ДЛЯ ОЦЕНКИ ПО ОТНОШЕНИЮ

Оптимальное размещение n_h для оценки по отношению может быть иным, чем для оценки по среднему на единицу. Рассмотрим сначала случайную переменную \hat{Y}_{Rs} . По теореме 6.5 ее дисперсия будет

$$V(\hat{Y}_{Rs}) = \sum_h \frac{N_h(N_h - n_h)}{n_h} (S_{y'h}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_h S_{yxh} S_{xh}) = \\ = \sum_h \frac{N_h(N_h - n_h)}{n_h} S_{dh}^2, \quad \text{где} \quad S_{dh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} d_{hi}^2 \quad (6.27)$$

и где $d_{hi} = y_{hi} - R_h x_{hi}$ представляет собой отклонение y_{hi} от $R_h x_{hi}$. В соответствии с методами нахождения оптимального размещения, приведенными в гл. 5, выражение (6.27) минимально для случая, когда суммарные издержки имеют вид $\Sigma c_h n_h$ при

$$n_h \propto \frac{N_h S_{dh}}{\sqrt{c_h}}.$$

Напомним, что для оценки по среднему на единицу n_h , минимизирующие дисперсию, выбирались пропорциональными $N_h S_{yh} / \sqrt{c_h}$.

При планировании выборки размещение для оценки по отношению может вызвать некоторое осложнение, потому что делать предположения о вероятных значениях S_{dh} довольно трудно. Полезны два правила. Для совокупности, в которой оценка по отношению есть наилучшая линейная несмещенная оценка, S_{dh} будут приблизительно пропорциональны $\sqrt{\bar{X}_h}$ (по теореме 6.4). В этом случае n_h должны быть пропорциональны $N_h \sqrt{\bar{X}_h} / \sqrt{c_h}$. Иногда дисперсия d_{hi} будет ближе к величине, пропорциональной \bar{X}_h^2 . Это приводит к размещению с n_h , пропорциональными $N_h \bar{X}_h / \sqrt{c_h}$, т. е. суммарным значениям x_{hi} для слоев,

деленным на квадратный корень из величины издержек в расчете на единицу. Такого рода пример был рассмотрен Хансеном, Хервицем и Гурни (Hansen, Hurwitz and Gurney, 1946) для случая выборки, предназначенной для оценки объема продаж в розничной торговле.

Те же общие доводы сохраняют силу и в случае, когда применяется оценка \hat{Y}_{Re} .

Пример. Различные способы размещения можно сравнить, пользуясь данными сплошного учета 257 коммерческих персиковых садов штата Северная Каролина в июне 1946 г. (Finkner, 1950). Цель обследования состояла в том, чтобы найти наиболее эффективный метод отбора, который позволил бы оценить объем производства персиков в этом районе. Были собраны сведения о количестве персиковых деревьев и оценке величины урожая персиков в каждом саду. Сильная корреляция между этими двумя переменными диктовала применение оценки по отношению. Один очень большой сад был исключен из рассмотрения.

Для нашего примера район считается разделенным в географическом отношении на три слоя. Число персиковых деревьев в саду обозначается x_{hi} , оценка урожая персиков в бушелях y_{hi} . Мы рассмотрим только первую оценку \hat{Y}_{Rs} (основанную на вычислении отдельного отношения для каждого слоя), поскольку правила размещения в принципе одинаковы для обоих типов оценок по отношению при расслоенном отборе.

Сравниваются четыре способа размещения: (а) n_h пропорциональны N_h ; (б) n_h пропорциональны $N_h S_{yh}$; (в) n_h пропорциональны $N_h \sqrt{\bar{X}_h}$ и (г) n_h пропорциональны $N_h \bar{X}_h = X_h$. Объем выборки равен 100. Данные для сравнения приведены в табл. 6.3.

Таблица 6.3
ДАННЫЕ ОБСЛЕДОВАНИЯ ПЕРСИКОВЫХ САДОВ ШТАТА СЕВЕРНАЯ КАРОЛИНА

Слой	S_{xh}^2	S_{yxh}	$S_{y'h}^2$	S_{xh}	S_{yh}	\bar{X}_h	\bar{Y}_h	R_h	S_{dh}^2
1	5 186	6 462	8 699	72,01	93,27	53,80	69,48	1,29133	658
2	2 367	3 100	4 614	48,65	67,93	31,07	43,64	1,40475	573
3	4 877	4 817	7 311	69,83	85,51	56,97	66,39	1,16547	2 706
Для совокупности	3 898	4 434	6 409	62,43	80,06	44,45	56,47	1,27053	1 433
Слой	N_h	(а)	$N_h S_{yh}$	(б)	$\sqrt{\bar{X}_h}$	$N_h \sqrt{\bar{X}_h}$	(в)	$N_h \bar{X}_h$	(г)
1	47	18	4 384	22	7,33	344,5	20	2 529	22
2	118	46	8 016	40	5,57	657,3	39	3 666	32
3	91	36	7 781	38	7,55	687,1	41	5 184	46
Для совокупности	256	100	20 181	100	20,45	1588,9	100	11 379	100

В верхней части таблицы указаны основные данные. Для вычисления четырех необходимых для сравнения дисперсий сначала были найдены n_h для каждого типа размещения. Их значения указаны в столбцах (а), (б), (в), (г) в нижней части таблицы. Так, для варианта размещения (а) $n_h = nN_h/N$, так что для первого слоя

$$n_1 = \frac{(100)(47)}{256} = 18.$$

После того как были получены n_h , соответствующие $V(\hat{Y}_{Rd})$ отыскивались путем подстановки n_h в формулу

$$V(\hat{Y}_{Rd}) = \sum_h \frac{N_h(N_h - n_h)}{n_h} S_{dh}^2,$$

где

$$S_{dh}^2 = S_{ph}^2 + R_h^2 S_{zh}^2 - 2R_h S_{yzh}.$$

Величины S_{dh}^2 одинаковы для всех четырех вариантов размещения; они приведены в крайнем правом столбце верхней части таблицы.

Значения дисперсий и относительной точности указаны в табл. 6.4.

Таблица 6.4
СРАВНЕНИЕ ЧЕТЫРЕХ СПОСОБОВ РАЗМЕЩЕНИЯ

Способ размещения: n_h пропорциональным	Дисперсия				Относительная точность
	слой			общая	
	1	2	3		
1. N_h	49824	105833	376215	531872	100
2. $N_h S_{ph}^2$	35144	131847	343446	510437	104
3. $N_h \sqrt{\bar{X}_h}$	41750	136964	300312	479026	111
4. $N_h \bar{X}_h$	35144	181710	240888	457742	116

Полученные размещения не дают особых различий, как можно было бы ожидать, поскольку n_h для всех четырех способов различаются мало. Способ 4, при котором объем выборки из слоя пропорционален общему количеству деревьев в слое, по-видимому, чуть-чуть лучше остальных.

6.15. НЕСМЕЩЕННЫЕ ОЦЕНКИ ТИПА ОЦЕНОК ПО ОТНОШЕНИЮ

В последние годы был замечен значительный интерес к исследованию оценок типа оценок по отношению, но не смещенных или имеющих меньшее смещение, чем обычные оценки по отношению. Такие оценки могут быть полезны в обследованиях, где выделено много слоев, а объемы выборки в каждом слое невелики, если кажется приемлемой раздельная оценка по отношению.

Одну из оценок, предложенную Хартли и Россом (Hartley and Ross, 1954), можно получить, исходя из среднего \bar{r} отношений y_i/x_i , введением поправки на смещение.

$$\bar{r} = \frac{1}{n} \sum r_i = \frac{1}{n} \sum \frac{y_i}{x_i}.$$

Далее,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N r_i (x_i - \bar{X}) &= \frac{1}{N} \sum_{i=1}^N \frac{y_i}{x_i} \cdot x_i - \left(\frac{1}{N} \sum_{i=1}^N r_i \right) \bar{X} = \\ &= \bar{Y} - \bar{X} \bar{r} = \bar{X} (R - \bar{r}). \end{aligned}$$

Но при простом случайном отборе $E(\bar{r}) = E(r_i)$. Следовательно,

$$\text{смещение } \bar{r} = E(\bar{r}) - R = -\frac{1}{\bar{X}N} \sum_{i=1}^N r_i (x_i - \bar{X}). \quad (6.28)$$

Согласно теореме 2.3, несмещенной выборочной оценкой

$$\frac{1}{N-1} \sum_{i=1}^N r_i (x_i - \bar{X})$$

служит

$$\frac{1}{n-1} \sum_{i=1}^n r_i (x_i - \bar{x}) = \frac{n}{n-1} (\bar{y} - \bar{r}\bar{x}).$$

Вычитая из оценки \bar{r} оценку ее смещения (6.28), получаем новую оценку

$$\tilde{r} = \bar{r} + \frac{n(N-1)}{(n-1)N\bar{X}} (\bar{y} - \bar{r}\bar{x}). \quad (6.29)$$

Соответствующей несмещенной оценкой суммарного значения для совокупности \tilde{Y} служит

$$\tilde{r}X = \bar{r}X + \frac{n(N-1)}{n-1} (\bar{y} - \bar{r}\bar{x}). \quad (6.30)$$

Пример. Вычислим оценку суммарного значения для слоя по представленной далее простой случайной выборке объемом в 8 единиц из слоя с $N = 16$, $X = 106$. Из табл. 6.5 имеем

$$\begin{aligned} \tilde{r}X &= 2,389 \cdot 106 + \frac{8 \cdot 15}{7} [11,000 - 2,389 \cdot 5,5] = \\ &= 253,2 - 36,7 = 216,5. \end{aligned}$$

Таблица 6.5

ВЫЧИСЛЕНИЕ ОЦЕНКИ ХАРТЛИ—РОССА

	Номер единицы								Среднее
	1	2	3	4	5	6	7	8	
y_i	8	15	5	7	5	13	11	24	11,000
x_i	8	6	1	4	3	5	4	13	5,500
r_i	1,000	2,500	5,000	1,750	1,667	2,600	2,750	1,846	2,389

Существует точная формула $V(\bar{r}'X)$ при любом объеме выборки. Если пкс можно пренебречь, то эта формула имеет вид

$$V(\bar{r}'X) = \frac{N^2}{n} (S_y^2 + \bar{r}_p^2 S_x^2 - 2\bar{r}_p S_{yx}) + \frac{N^2}{n(n-1)} (S_r^2 S_x^2 + S_{rx}^2), \quad (6.31)$$

где \bar{r}_p и S_r^2 — среднее значение и дисперсия r_i для совокупности и S_{rx} — ковариация r_i и x_i для совокупности. Эта формула была предложена Гудменом и Хартли (Goodman and Hartley, 1958), которые получили также несмещенную выборочную оценку $V(\bar{r}'X)$. Формула истинной дисперсии, когда пкс пренебречь нельзя, приведена в работе Робсона (Robson, 1957).

Общее сравнение точности $\bar{r}'X$ и обычной оценки по отношению \hat{Y}_R можно произвести только при достаточно больших выборках, когда справедлива приближенная формула для $V(\hat{Y}_R)$. При больших n второй член в (6.31) можно опустить. Первый член можно переписать в виде

$$V(\bar{r}'X) \approx \frac{N^2}{n} \sum_{i=1}^N \frac{[y_i - \bar{Y} - \bar{r}_p(x_i - \bar{X})]^2}{N-1}.$$

Соответствующее выражение для \hat{Y}_R имеет вид

$$V(\hat{Y}_R) \approx \frac{N^2}{n} \sum_{i=1}^N \frac{(y_i - R x_i)^2}{N-1}.$$

Таким образом, как указывают Гудмен и Хартли, при больших выборках $\bar{r}'X$ будет более точной, если прямая $\bar{Y} + \bar{r}_p(x_i - \bar{X})$ лучше приближает точки графика значений y_i , чем прямая $R x_i$. Хотя тщательного сравнения не проводилось, кажется вероятным, что в большинстве случаев, когда применимы оценки по отношению, для больших выборок $V(\hat{Y}_R)$ будет меньше. Было бы интересно сравнить эти оценки для малых выборок, потому что именно для них отсутствие смещения может играть важную роль.

Исследуя другой подход, Лахири (Lahiri, 1951) показал, что обычная оценка по отношению будет несмещенной, если выборка извлекается с вероятностью, пропорциональной Σx_i . Существует два способа получения такой выборки. Один, указанный Лахири, состоит в следую-

щем. Производится обычный отбор без возвращения. Если T есть сумма n наибольших значений x_i в совокупности, то выбирается случайное число между 1 и T , скажем, v . Если для выборки выполняется $\Sigma x_i \geq v$, то она принимается. В противном случае вся она возвращается в совокупность, вместо нее берется другая и испытание повторяется с извлечением нового случайного числа для каждой испытываемой выборки до тех пор, пока не будет получена выборка, которая может быть принята. Ясно, что вероятность, с которой выборка принимается, пропорциональна Σx_i . При втором способе, предложенном Мидзуно (Midzuno, 1951), нужно извлечь первую единицу выборки с вероятностью, пропорциональной x_i , а остальные $(n-1)$ единиц — с одинаковыми вероятностями.

Покажем теперь, что \hat{R} — несмещенная оценка.

Если сложить Σx_i по всем простым случайным выборкам объема n , то общая сумма будет равна $C \frac{n-1}{N-1} X$, поскольку каждая единица участвует в $C \frac{n-1}{N-1}$ выборках. Следовательно, вероятность того, что будет извлечена выборка с некоторым конкретным значением Σx_i , равна:

$$P = \frac{\Sigma x_i}{C \frac{n-1}{N-1} X}.$$

Для рассматриваемого метода извлечения выборки при $\hat{R}_L = \Sigma y_i / \Sigma x_i$

$$E(\hat{R}_L) = \sum_{SrS} \left(P \frac{\Sigma y_i}{\Sigma x_i} \right),$$

где \sum_{SrS} означает суммирование по всем простым случайным выборкам.

Подставляя значение P , получаем

$$E(\hat{R}_L) = \sum_{SrS} \left[\frac{\Sigma x_i}{C \frac{n-1}{N-1} X} \right] \cdot \frac{\Sigma y_i}{\Sigma x_i} = \frac{C \frac{n-1}{N-1} Y}{C \frac{n-1}{N-1} X} = R.$$

Точное выражение для $V(\hat{R}_L)$ не получено. Легко показать, что для больших выборок \hat{R}_L имеет ту же приближенную дисперсию, что и \hat{R} .

Пример. В этом примере рассматривается искусственная совокупность, для которой оценка Лахири оказывается наиболее удачной. Совокупность содержит три слоя с $N_h = 4$, $n_h = 2$ для каждого слоя. Совокупность специально построена таким образом, что (а) R_h заметно меняются от слоя к слою и потому оказывается предпочтительнее раздельная оценка по отношению и (б) оценка по отношению

внутри каждого слоя сильно смещена. Сравниваются пять методов оценивания \bar{Y} , суммарного значения для совокупности:

Простое распространение: $\sum N_h \bar{y}_h$.

Совместная оценка по отношению: $(\bar{y}_{st}/\bar{x}_{st}) X$.

Раздельная оценка по отношению: $\sum_h (\bar{y}_h/\bar{x}_h) X_h$.

Раздельная оценка Хартли — Росса: $\sum_h (\bar{r}_h' X_h)$.

Раздельная оценка Лахири: $\sum_h (\bar{y}_h/\bar{x}_h) X_h$.

Всего существует $6^3 = 216$ возможных выборок. Поскольку значения оценок получены для каждой выборки, приводятся точные значения смещений и дисперсий.

Таблица 6.6
НЕБОЛЬШАЯ ИСКУССТВЕННАЯ СОВОКУПНОСТЬ

	Слой					
	I		II		III	
	y	x	y	x	y	x
	2	2	2	1	3	1
	3	4	5	4	7	3
	4	6	9	8	9	4
	11	20	24	23	25	12
Итоги	20	32	40	36	44	20
R_h	0,625		1,111		2,200	

Таблица 6.7
СРАВНЕНИЕ РАЗЛИЧНЫХ ОЦЕНОК \bar{Y}

Способ получения оценки	Дисперсия	(Смещение) ²	Средний квадрат ошибки
Простое распространение	820,3	0,0	820,3
Совместная по отношению	262,8	4,5	269,3
Раздельная по отношению	35,9	24,1	60,0
Раздельная Хартли — Росса	153,6	0,0	153,6
Раздельная Лахири	19,6	0,0	19,6

Полученные результаты обнаруживают несколько интересных особенностей. Для совместной оценки по отношению квадрат смещения составляет лишь незначительную часть среднего квадрата ошибки, хотя применительно к этой оценке совокупность наиболее «неудобна». Из-за сильной вариации R_h раздельная оценка по отношению оказывается гораздо более достоверной, чем совместная оценка, но она сильно смещена. Раздельная оценка Хартли — Росса лучше совместной оценки по отношению, но хуже (если судить по среднему квадрату ошибки) раздельной оценки по отношению. Раздельная оценка Лахири значительно превосходит все остальные оценки. Метод Лахири оказался

подходящим для этой совокупности, потому что четвертые единицы в каждом слое имеют большую вероятность быть отобранными, а выборки, содержащие эти единицы, дают хорошие оценки R_h .

Из этого примера нельзя сделать общих выводов. Практическое применение метода Лахири ограничено тем, что обследователь вряд ли пожелает извлекать выборки с вероятностями, пропорциональными Σx_i , если только он не собирается применять оценку такого типа для всех основных признаков, изучаемых в обследовании.

Кенуй (Kenouille, 1956) разработал метод, позволяющий уменьшить порядок смещения с $1/n$ до $1/n^2$ и применимый к широкому классу оценок. Ценность этого метода применительно к оценкам по отношению была отмечена Дербином (Durbine, 1959). Смещение оценок типа \hat{R} , \hat{Y}_R может быть разложено в ряд Тейлора вида

$$E(\hat{R}) = R + \frac{b_1}{n} + \frac{b_2}{n^2} + \dots \quad (6.32)$$

Пусть выборка подразделяется случайным образом на g групп, величиной m каждая, где $n = gm$. Из (6.32) следует, что

$$E(g\hat{R}) = gR + \frac{b_1}{m} + \frac{b_2}{gm^2} + \dots \quad (6.33)$$

Пусть теперь \hat{R}_j — обычное отношение $\Sigma y/\Sigma x$, вычисленное по выборке после того, как из нее исключена j -я группа. Так как \hat{R}_j получено по простой случайной выборке объема m ($g-1$), мы имеем

$$E(\hat{R}_j) = R + \frac{b_1}{(g-1)m} + \frac{b_2}{(g-1)^2 m^2} + \dots$$

Следовательно,

$$E[(g-1)\hat{R}_j] = (g-1)R + \frac{b_1}{m} + \frac{b_2}{(g-1)m^2} + \dots$$

Вычитая последнее выражение из (6.33), получаем с точностью до членов порядка n^{-2}

$$E[g\hat{R} - (g-1)\hat{R}_j] = R - \frac{b_2}{g(g-1)m^2} = R - \frac{b_2}{n^2} \frac{g}{(g-1)}$$

Таким образом, смещение теперь представляет собой величину порядка $1/n^2$. Мы можем построить g оценок такого типа, по одной для каждой группы. Кенуй показал, что если взять среднее этих оценок, т. е.

$$\hat{R}_Q = g\hat{R} - (g-1) \frac{\hat{R}_1 + \hat{R}_2 + \dots + \hat{R}_g}{g},$$

то его дисперсия отличается от дисперсии \hat{R} членами порядка $1/n^2$. Следовательно, возможное увеличение дисперсии, связанное с такой поправкой на смещение, будет пренебрежимо малым для выборок не слишком большого объема.

Простейшую оценку такого вида можно получить, положив $g = 2$. Оценками \hat{R}_1 и \hat{R}_2 служат оценки, получаемые по двум половинам выборки, и

$$\hat{R}_0 = 2\hat{R} - \frac{\hat{R}_1 + \hat{R}_2}{2}.$$

Другой крайностью будет случай, когда $g = n$. Преимущества выбора одного значения g по сравнению с другим не изучались.

Рассмотренный метод вряд ли сможет нам помочь, если объемы выборки по слоям невелики, как в случае искусственной совокупности с $n_h = 2$ в этом параграфе (см. с. 198). В выборках умеренного объема из совокупностей с большой вариацией x его стоило бы применять из осторожности.

6.16. СРАВНЕНИЕ ДВУХ ОТНОШЕНИЙ

В аналитических обследованиях часто оказывается необходимым определять разность $\hat{R} - \hat{R}'$ оценок двух отношений и вычислять стандартную ошибку $\hat{R} - \hat{R}'$. Далее приводятся формулы для выборочных оценок дисперсии $\hat{R} - \hat{R}'$, поскольку чаще всего бывают нужны именно они. По причинам, изложенным в параграфе 2.12, пкс опускаются.

Сначала рассматривается простой случайный отбор. Необходимо различать три случая.

Два отношения независимы

Это случается, когда единицы распределены на два различных класса и мы хотим сравнить отношения, оцениваемые отдельно для каждого из этих классов. Например, при изучении бюджетов домохозяйств простая случайная выборка домохозяйств может быть подразделена на дома, находящиеся в личном владении и арендуемые, для сравнения долей дохода, затрачиваемых в обоих классах на содержание дома. Если обозначить оценки отношений через $\hat{R} = \bar{y}/\bar{x}$, $\hat{R}' = \bar{y}'/\bar{x}'$, то

$$v(\hat{R} - \hat{R}') = v(\hat{R}) + v(\hat{R}').$$

Два отношения имеют общий знаменатель

Когда единицей отбора служит группа семей, нам может понадобиться, например, сравнить долю мужчин, пользующихся электрическими бритвами, с долей мужчин, пользующихся безопасными бритвами. Пусть для любой единицы y — число мужчин, пользующихся электрическими бритвами, y' — число мужчин, пользующихся безопасными бритвами, и x — общее число мужчин.

$$\hat{R} - \hat{R}' = \frac{\bar{y} - \bar{y}'}{\bar{x}}.$$

Если $d_i = y_i - y'_i$, то оценку дисперсии разности $\hat{R} - \hat{R}'$ можно вычислять по формуле

$$v(\hat{R} - \hat{R}') \approx \frac{1}{n(n-1)\bar{x}^2} \sum_{i=1}^n [d_i - (\hat{R} - \hat{R}')\bar{x}]^2.$$

Два отношения имеют разные знаменатели, но могут быть коррелированы

Примером может служить сравнение доли курящих мужчин с долей курящих женщин в обследовании, в котором единицей отбора служит группа домов. С математической точки зрения это наиболее общий случай:

$$v(\hat{R} - \hat{R}') = v(\hat{R}) + v(\hat{R}') - 2\text{cov}(\hat{R}\hat{R}').$$

Новый здесь только член $\text{cov}(\hat{R}\hat{R}')$. Записывая как обычно

$$\hat{R} - R \approx \frac{\bar{y} - R\bar{x}}{\bar{x}}; \quad \hat{R}' - R' \approx \frac{\bar{y}' - R'\bar{x}'}{\bar{x}'},$$

получаем

$$\text{cov}(\hat{R}\hat{R}') \approx \frac{1}{n\bar{x}\bar{x}'} \text{cov}(y_i - Rx_i)(y'_i - R'x'_i).$$

Выборочную оценку можно вычислять по формуле

$$\text{cov}(\hat{R}\hat{R}') \approx \frac{1}{n(n-1)\bar{x}\bar{x}'} \sum (y_i y'_i - \hat{R} y'_i x_i - \hat{R}' y_i x'_i + \hat{R}\hat{R}' x_i x'_i).$$

Пример. В 1954 г. проводились испытания вакцины Солка против полиомиелита среди учащихся первых трех классов во всех школах нескольких графств. Графства отбирались неслучайным образом: предпочтение отдавалось тем, для которых имелись сведения о предшествующих вспышках полиомиелита. Однако для нашего примера будем считать, что они составляют случайную выборку из некоторой совокупности.

Дети, чьи родители не дали согласия на участие в испытаниях, составили группу, названную группой «непривитых», и, конечно, не получили уколов. Половина детей, для которых такое согласие было дано, получила по три укола нейтральной жидкости и они составили группу, названную группой «плацебо». По данным табл. 6.8 нужно сравнить частоты \hat{R} и \hat{R}' случаев поражения параличом в группе «непривитых» и группе «плацебо». Для сокращения объема данных сравнение ограничено 34 графствами, в каждом из которых обе группы вместе включают более 4000 детей.

При этих данных любое изменение в уровне заболеваемости полиомиелитом от графства к графству будет соответствовать некоторой положительной корреляции между \hat{R} и \hat{R}' .

По итогам табл. 6.8 получены следующие величины:

$$\text{Плацебо: } \hat{R} = \frac{88}{167,4} = 0,525687; \quad \bar{x} = \frac{167,4}{34} = 4,9235.$$

$$\text{Непривитые: } \hat{R}' = \frac{99}{284,6} = 0,347857; \quad \bar{x}' = \frac{284,6}{34} = 8,3706.$$

Таблица 6.8

ЧИСЛО ДЕТЕЙ (x, x')* И ЧИСЛО СЛУЧАЕВ ПОРАЖЕНИЯ
ПАРАЛИЧОМ (y, y')** В СРЕДНЕМ НА ОДНО ГРАФСТВО

x	x'	y	y'	x	x'	y	y'
4,1	2,4	0	0	13,8	25,6	3	3
3,5	8,0	1	6	10,5	8,1	2	0
4,1	6,1	7	2	21,6	25,9	10	7
2,6	4,6	2	1	3,5	6,7	2	2
2,4	1,5	2	1	6,8	7,3	3	8
2,2	1,9	0	0	2,3	3,7	0	1
1,1	4,0	1	1	2,6	2,9	2	0
1,6	4,0	1	2	6,0	11,1	3	1
5,7	7,8	1	4	11,0	14,8	7	11
3,3	11,0	3	7	19,4	42,5	11	14
1,0	3,8	0	1	6,8	13,7	6	2
2,0	5,2	1	0	1,2	4,0	3	1
8,3	19,0	4	4	5,4	9,3	11	6
1,0	3,7	1	5	1,7	2,6	0	2
1,1	4,2	0	1	2,1	2,3	0	0
2,3	6,8	1	2	1,5	2,6	0	0
1,9	3,5	0	2	3,0	4,0	0	2
Итого				167,4	284,6	88	99

* x, x' — числа детей в группе плацебо и группе непривитых (в тысячах).

** y, y' — числа случаев поражения параличом в группе плацебо и в группе непривитых.

Для получения $v(\hat{R})$, $v(\hat{R}')$ и $\text{cov}(\hat{R}\hat{R}')$ нужно знать все суммы квадратов без поправки и суммы произведений для четырех переменных.

$$v(\hat{R}) = \frac{1}{n(n-1)x^2} (\Sigma y^2 - 2\hat{R} \Sigma yx + \hat{R}^2 \Sigma x^2) =$$

$$= \frac{1}{34 \cdot 33 \cdot (4,9235)^2} [564 - 1,05137 \cdot 822,2 +$$

$$+ 0,27635 \cdot 1661,92] = 0,00584.$$

Аналогично находим $v(\hat{R}') = 0,00240$.

$$\text{cov}(\hat{R}\hat{R}') = \frac{1}{n(n-1)xx'} (\Sigma yy' - \hat{R} \Sigma y'x - \hat{R}' \Sigma yx' + \hat{R}\hat{R}' \Sigma xx') =$$

$$= \frac{497 - 0,52569 \cdot 844,6 - 0,34786 \cdot 1397,4 + 0,52569 \cdot 0,34786 \cdot 2690,8}{34 \cdot 33 \cdot 4,9235 \cdot 8,3706} = 0,00127.$$

Следовательно,

$$\text{стандартная ошибка } (\hat{R} - \hat{R}') = \sqrt{0,00584 + 0,00240 - 0,00254} = 0,0754.$$

Поскольку $\hat{R} - \hat{R}' = 0,1778$, разность (с некоторой натяжкой) можно считать значимой на 5% -ном уровне (при таком объеме выборки распределение $\hat{R} - \hat{R}'$ может быть несколько асимметричным). Возможное объяснение различия отношений состоит в том, что дети из группы непривитых могут иметь более сильный естественный иммунитет, чем дети из группы плацебо.

Проблема сравнения отношений может возникнуть при расслоенном отборе, когда области изучения пересекаются со слоями. Если ожидается, что \hat{R}_h, \hat{R}'_h изменяются от слоя к слою, то скорее всего сравнение будет основано на анализе значений $\hat{R}_h - \hat{R}'_h$ в отдельных слоях. Находя стандартные ошибки разностей $\hat{R}_h - \hat{R}'_h$, можно определить, значимы ли различия этих разностей от слоя к слою, и, если они не значимы, вычислить нужную общую разность.

Если \hat{R}_h и \hat{R}'_h не обнаруживают значимых различий от слоя к слою, то может оказаться достаточным сравнение совместных оценок \hat{R}_c и \hat{R}'_c . Как и ранее

$$v(\hat{R}_c - \hat{R}'_c) = v(\hat{R}_c) + v(\hat{R}'_c) - 2\text{cov}(\hat{R}_c \hat{R}'_c),$$

где, полагая $d_{hi} = (y_{hi} - \bar{y}_h) - \hat{R}_c(x_{hi} - \bar{x}_h)$,

$$v(\hat{R}_c) = \frac{1}{\bar{x}_{st}^2} \sum_h \frac{N_h^2}{n_h(n_h-1)} \sum_i d_{hi}^2;$$

$$\text{cov}(\hat{R}_c \hat{R}'_c) = \frac{1}{\bar{x}_{st} \bar{x}'_{st}} \sum_h \frac{N_h^2}{n_h(n_h-1)} \sum_i d_{hi} d'_{hi}.$$

Более подробно сравнение отношений проведено Кишем и Хессом (Kish and Hess, 1959). Они приводят также упрощенные вычислительные формулы для тех случаев, когда выборка это допускает.

6.17. МНОГОМЕРНЫЕ ОЦЕНКИ ПО ОТНОШЕНИЮ

Олкин (Olkin, 1958) исследовал оценку по отношению в случае, когда имеется p вспомогательных переменных x_i (x_1, x_2, \dots, x_p). Для суммарного значения совокупности предложенная им оценка, обозначим её \hat{Y}_{MR} [MR — от английского «multivariate ratio» — многомерное отношение], имеет вид:

$$\hat{Y}_{MR} = W_1 \frac{\bar{y}}{\bar{x}_1} X_1 + W_2 \frac{\bar{y}}{\bar{x}_2} X_2 + \dots + W_p \frac{\bar{y}}{\bar{x}_p} X_p =$$

$$= W_1 \hat{Y}_{R_1} + W_2 \hat{Y}_{R_2} + \dots + W_p \hat{Y}_{R_p},$$

где W_i — веса, которые выбираются с целью максимизировать точность \hat{Y}_{MR} , причем $\Sigma W_i = 1$. Оценка такого типа, по-видимому,

пригодна в том случае, когда регрессия y по x_1, x_2, \dots, x_p линейна и выражается прямыми, проходящими через начало координат. Суммарные значения для совокупности, X_i , должны быть известны.

Мы опишем этот метод для случая двух переменных x , который должен быть наиболее распространенным на практике. Имеем

$$\hat{Y}_{MR} - Y = W_1(\hat{Y}_{R1} - Y) + W_2(\hat{Y}_{R2} - Y).$$

Следовательно,

$$V(\hat{Y}_{MR}) = W_1^2 V(\hat{Y}_{R1}) + 2W_1 W_2 \text{cov}(\hat{Y}_{R1}, \hat{Y}_{R2}) + W_2^2 V(\hat{Y}_{R2}) = W_1^2 V_{11} + 2W_1 W_2 V_{12} + W_2^2 V_{22},$$

где $V_{11} = V(\hat{Y}_{R1})$ и т. д. Значения W_1, W_2 , минимизирующие дисперсию и подчиненные условию $W_1 + W_2 = 1$, оказываются равными:

$$W_1 = \frac{V_{22} - V_{12}}{V_{11} + V_{22} - 2V_{12}}; \quad W_2 = \frac{V_{11} - V_{12}}{V_{11} + V_{22} - 2V_{12}},$$

а минимальное значение дисперсии

$$V_{\min}(\hat{Y}_{MR}) = \frac{V_{11} V_{22} - V_{12}^2}{V_{11} + V_{22} - 2V_{12}}.$$

Для случая p переменных необходимо вычислить обратную к матрице V_{ij} матрицу V^{ij} . Тогда оптимальные $W_i = \Sigma_i / \Sigma$, где Σ_i — сумма элементов i -го столбца матрицы V^{ij} и Σ — сумма всех p^2 элементов V^{ij} . Минимальная дисперсия равна $1/\Sigma$.

На практике веса определяются на основании оценок дисперсий и ковариаций v_{ij} . Из формулы (6.7) параграфа 6.3

$$v_{11} = \frac{(1-f)\hat{Y}^2}{n} (c_{yy} + c_{11} - 2c_{y1});$$

$$v_{22} = \frac{(1-f)\hat{Y}^2}{n} (c_{yy} + c_{22} - 2c_{y2}),$$

где $c_{yy} = s_y^2/\bar{y}^2$ и т. д. Ковариацию можно выразить в виде

$$v_{12} = \frac{(1-f)\hat{Y}^2}{n} (c_{yy} + c_{12} - c_{y1} - c_{y2}).$$

Удобный метод вычислений состоит в том, чтобы сначала найти матрицу

$$C = \begin{pmatrix} c_{yy} & c_{y1} & c_{y2} \\ c_{y1} & c_{11} & c_{12} \\ c_{y2} & c_{12} & c_{22} \end{pmatrix}.$$

Если обозначить $v'_{ij} = n v_{ij} / (1-f)\hat{Y}^2$, то матрицу v'_{ij} легко получить, взяв диагональное дополнение в C , т. е.

$$v'_{11} = c_{yy} + c_{11} - c_{y1} - c_{y1};$$

$$v'_{12} = c_{yy} + c_{12} - c_{y1} - c_{y2}, \text{ и т. д.}$$

Множитель $(1-f)\hat{Y}^2/n$ не нужен при вычислении оценок весов w_i , но он должен быть введен при вычислении минимальной дисперсии. Таким образом,

$$v_{\min}(\hat{Y}_{MR}) = \frac{(1-f)\hat{Y}^2}{n} \frac{(v'_{11} v'_{22} - v_{12}^2)}{(v'_{11} + v'_{22} - 2v'_{12})}.$$

Ввиду большого объема необходимых вычислений применение рассмотренной оценки, вероятно, будет ограничено небольшими обследованиями специализированного характера. По сравнению с отдельными оценками \hat{Y}_{R1} или \hat{Y}_{R2} описанный метод может давать заметное увеличение точности.

У п р а ж н е н и я

6.1. Имеются следующие данные пробного обследования 21 домохозяйства о числе их членов (x), числе детей (y_1), количестве автомобилей (y_2) и телевизоров (y_3):

x	y_1	y_2	y_3	x	y_1	y_2	y_3	x	y_1	y_2	y_3
5	3	1	3	2	0	0	1	6	3	2	0
2	0	1	1	3	1	1	1	4	2	1	1
4	1	2	0	2	0	2	0	4	2	1	1
4	2	1	1	6	4	2	1	3	1	0	1
6	4	1	1	3	1	0	0	2	0	2	1
3	1	1	2	4	2	1	1	4	2	1	1
5	3	1	1	5	3	1	1	3	1	1	1

Если известно суммарное значение для совокупности, X , рекомендовали бы вы применение оценок по отношению вместо простого распространения при оценивании общего числа детей, общего количества автомобилей и общего количества телевизоров?

6.2. На ячменном поле взвешивалось зерно (y_1) и зерно вместе с соломой (x_1) для каждого из большого числа участков, служивших единицами отбора и размещенных на поле случайным образом. Кроме того, была взвешена общая масса (зерно плюс солома) со всего поля. Были получены следующие данные: $c_{yy} = 1,13$; $c_{yx} = 0,78$; $c_{xx} = 1,11$. Определите выигрыш в точности, получаемый при оценивании урожая зерна со всего поля с помощью отношения веса зерна к общей массе, а не с помощью среднего значения веса зерна на один участок.

За 20 минут можно сжать ячмень, обмолотить его и взвесить зерно на отдельном участке, за 2 минуты — взвесить на отдельном участке солому и за 2 часа собрать и взвесить общую массу на всем поле. Сколько участков нужно отобрать на поле для того, чтобы оценка по отношению была более экономичной, чем оценка по среднему на один участок?

6.3. Для данных из табл. 6.1 $\hat{Y}_R = 28\,367$; $c_{yy} = 0,0142068$; $c_{yx} = 0,0146541$; $c_{xx} = 0,0156830$. Вычислите 95%-ные квадратичные доверительные границы для Y и сравните их с границами, получающимися при аппроксимации нормальным распределением.

6.4. Значения y, x в некоторой совокупности с $N = 6$ таковы:

y	3	5	7	6	8	13
x	1	2	2	3	3	3

Проверьте, что регрессия y по x выражается прямой, проходящей через начало координат. Вычислив \bar{R} для всех 15 простых случайных выборок объемом $n = 2$ и всех 20 простых случайных выборок объемом $n = 3$, проверьте в обоих случаях утверждение теоремы 6.2 о том, что \bar{R} есть несмещенная оценка.

6.5. Значения y и x наблюдаются для каждой единицы в простой случайной выборке из некоторой совокупности. Если известно \bar{X} , среднее значение x для совокупности, какой из следующих способов вы рекомендуете для оценивания \bar{Y}/\bar{X} ? (а) Всегда применять \bar{y}/\bar{x} . (б) Иногда применять \bar{y}/\bar{x} , а иногда — \bar{y}/\bar{X} . (в) Всегда применять \bar{y}/\bar{x} . Обоснуйте ваш ответ.

6.6. Имеются следующие данные о небольшой искусственной совокупности с $N = 8$ и двумя слоями равного объема:

Слой 1		Слой 2	
x_{11}	y_{11}	x_{21}	y_{21}
2	0	10	7
5	3	18	15
9	7	21	10
15	10	25	16

Для расслоенной случайной выборки, в которой $n_1 = n_2 = 2$, сопоставьте СКО оценок \hat{Y}_{Rz} и \hat{Y}_{Rc} , вычислив нужные величины для всех возможных выборок. В какой степени различие между СКО обусловлено смещением оценок?

6.7. Для упражнения 6.6 вычислите дисперсию, получающуюся при применении внутри каждого слоя метода извлечения выборки, предложенного Лакхири, и раздельной оценки по отношению.

6.8. Сорок пять штатов США (за исключением пяти наиболее крупных) распределены по девяти слоям, по пять штатов в каждом так, что штаты из одного слоя имеют приблизительно одно и то же отношение численности населения в 1950 г. к численности населения в 1940 г. Расслоенная случайная выборка при $n_h = 2$ дала следующие значения численности населения в 1960 г. (y) и в 1950 г. (x) (в миллионах):

	С л о й								
	1	2	3	4	5	6	7	8	9
y_{h1}	0,23	0,63	0,97	2,54	4,67	4,32	4,56	1,79	2,18
x_{h1}	0,13	0,50	0,91	2,01	3,93	3,96	4,06	1,91	1,90
y_{h2}	4,95	2,85	0,61	6,07	3,96	1,41	3,57	1,86	1,76
x_{h2}	2,78	2,38	0,53	4,84	3,44	1,33	3,29	2,01	1,32

При условии, что общая численность населения X в 1950 г. составила 97,94 млн., оцените ее в 1960 г. с помощью совместной оценки по отношению. Найдите стандартную ошибку вашей оценки с помощью упрощенного метода Кейфица (параграф 6.13). Правильная численность населения в 1960 г. была 114,99 млн. Сопоставьте ли ваша оценка с этой величиной в пределах ошибки выборки?

6.9. В примере двумерной оценки по отношению, приводимом Олкином, рассматривается выборка 50 городов из совокупности, содержащей 200 больших городов. Переменные y , x_1 , x_2 представляют собой числа жителей в городе соответственно в 1950, 1940 и 1930 гг. Для совокупности $\bar{Y} = 1699$, $\bar{X}_1 = 1482$, $\bar{X}_2 = 1420$ (в сотнях человек) и для выборки $y = 1896$, $x_1 = 1693$, $x_2 = 1643$. Матрица C , определенная в параграфе 6.17, имеет вид

	y	x_1	x_2
y	1,213	1,241	1,256
x_1	1,241	1,302	1,335
x_2	1,256	1,335	1,381

Оцените \hat{Y} с помощью (а) выборочного среднего, (б) отношения числа жителей в 1950 г. к числу жителей в 1940 г., (в) двумерной оценки по отношению. Вычислите выборочную оценку стандартной ошибки для каждой из этих оценок.

6.10. Докажите, что для метода получения выборки, предложенного Мидзуно (параграф 6.15), вероятность того, что будет извлечена некоторая конкретная выборка, равна

$$\frac{(n-1)!(N-n)!}{(N-1)!} \frac{\sum_{i=1}^n (x_i)}{X}.$$

ЛИТЕРАТУРА

- David F. N., Neyman J. (1938). Extension of the Markoff theorem of least squares. *Stat. Res. Mem.*, 2, 105.
- Durbin J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46, 477—480.
- Fieller E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, 24, 428—440.
- Finkner A. L. (1950). Methods of sampling for estimating commercial peach production in North Carolina. *North Carolina Agr. Exp. Sta. Tech. Bull.* 91.
- Goodman L. A. and Hartley H. O. (1958). The precision of unbiased ratio-type estimators. *Jour. Amer. Stat. Assoc.*, 53, 491—508.
- Hájek J. (1958). Some contributions to the theory of probability sampling. *Int. Stat. Inst. Bull.*, 36, 3, 127—134.
- Hansen, M. H., Hurwitz W. N. and Gurney M. (1946). Problems and methods of the sample survey of business. *Jour. Amer. Stat. Assoc.*, 41, 173—189.
- Hansen M. H., Hurwitz W. N. and Madow W. G. (1953). *Sample survey methods and theory*. John Wiley & Sons. New York.
- Hartley H. O. and Ross A. (1954). Unbiased ratio estimates. *Nature*, 174, 270—271.
- Jessen R. J. et al. (1947). On a population sample for Greece. *Jour. Amer. Stat. Assoc.*, 42, 357—384.
- Keyfitz N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Jour. Amer. Stat. Assoc.*, 52, 503—510.
- Kish L. and Hess I. (1959). On variances of ratios and their differences in multistage samples. *Jour. Amer. Stat. Assoc.*, 54, 416—446.
- Lahiri D. B. (1951). A method for sample selection providing unbiased ratio estimates. *Int. Stat. Inst. Bull.*, 33, 2, 133—140.
- Midzuno H. (1951). On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Stat. Math.*, 2, 99—108.
- Olkin I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45, 154—165.
- Paulson E. (1942). A note on the estimation of some mean values for a bivariate distribution. *Ann. Math. Stat.*, 13, 440—444.
- Quenouille M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353—360.
- Robson D. S. (1957). Applications of multivariate polykeys to the theory of unbiased ratio type estimation. *Jour. Amer. Stat. Assoc.*, 52, 511—522.
- Sukhatme P. V. (1954). *Sampling theory of surveys with applications*. Iowa State College Press, Ames, Iowa.

ОЦЕНКИ ПО РЕГРЕССИИ

7.1. ЛИНЕЙНЫЕ ОЦЕНКИ ПО РЕГРЕССИИ

Как и оценки по отношению линейные оценки по регрессии предназначены для того, чтобы увеличить точность благодаря применению некоторой вспомогательной переменной x_i , коррелированной с y_i . При изучении зависимости между y_i и x_i может оказаться, что хотя эта зависимость приблизительно линейна, соответствующая прямая не проходит через начало координат. Это значит, что может быть построена оценка, основанная не на отношении этих двух переменных, а на линейной регрессии y_i по x_i .

Предположим, что для каждой единицы в выборке получены значения y_i и x_i и известно среднее значение переменной x_i для совокупности, \bar{X} . Линейная оценка по регрессии \bar{Y} , среднего значения переменной y_i для совокупности, имеет вид

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}), \quad (7.1)$$

где индекс lr означает *линейную регрессию* [lr — от английского «linear regression»], а b — оценка изменения y при изменении x на единицу. Эта оценка основана на том соображении, что если \bar{x} меньше среднего значения x_i для совокупности, то можно ожидать, что вследствие регрессии y_i по x_i \bar{y} также будет меньше среднего y_i на величину $b(\bar{X} - \bar{x})$. В качестве оценки суммарного значения для совокупности, Y , мы принимаем $\hat{Y}_{lr} = N\bar{y}_{lr}$.

Уотсон (Watson, 1937) воспользовался регрессией площади листьев по их весу для оценивания средней площади листьев на дереве. В соответствии с этим все листья дерева взвешивались. Для небольшой выборки листьев определялись площадь и вес каждого листа. После этого выборочное среднее значение площади листа корректировалось с помощью регрессии по весу листа. Практический смысл этого приема заключался, разумеется, в том, что определить вес листа можно гораздо быстрее, чем найти его площадь.

Этот пример иллюстрирует общее положение, в котором применимы оценки по регрессии. Предположим, что можно очень быстро определить приближенное значение x_i некоторого признака для каждой единицы и можно также, каким-либо более дорогостоящим способом,

определить правильное значение y_i этого признака для единиц из простой случайной выборки. Так, эксперт по крысам может быстро, на глаз, оценить число крыс в каждом городском квартале и затем путем отлова определить действительное число крыс в каждом из кварталов, попавших в простую случайную выборку. В другом примере, описанном Йейтсом (Yates, 1960), на каждой делянке из совокупности делянок площадью по 0,1 акра делалась глазомерная оценка объема древесины и для некоторой выборки этих делянок измерялся ее действительный объем. Оценка по регрессии

$$\bar{y} + b(\bar{X} - \bar{x})$$

корректирует выборочное среднее точных измерений с помощью регрессии точных измерений по глазомерным оценкам. При этом не обязательно, чтобы глазомерная оценка не имела смещения. Примем $x_i - y_i = D$, так что если исключить постоянное смещение D , то примерная оценка совпадает с точной. Тогда при $b = 1$ оценка по регрессии принимает вид

$$\bar{y} + (\bar{X} - \bar{x}) = \bar{X} + (\bar{y} - \bar{x}) = (\text{среднее значение примерной оценки для совокупности}) + (\text{поправка на смещение}).$$

О свойствах оценки по регрессии мы знаем столько же, сколько об оценке по отношению. Оценка по регрессии состоятельна в том тривиальном смысле, что если выборка охватывает всю совокупность, то $\bar{x} = \bar{X}$ и оценка по регрессии сводится к \bar{Y} . Как будет показано далее, оценка по регрессии, вообще говоря, — смещенная оценка, но для больших выборок отношение смещения к стандартной ошибке становится малым. Имеется формула дисперсии этой оценки для больших выборок, но о характере распределения оценки для небольших выборок и о величине n , начиная с которой можно практически применять результаты, относящиеся к большим выборкам, нам пока известно немного.

При соответствующих значениях b частными случаями оценки по регрессии становятся как оценка по среднему на единицу, так и оценка по отношению. Очевидно, что если положить b равным нулю, то $\bar{y}_{lr} = \bar{y}$. Если $b = \bar{y}/\bar{x}$, то

$$\bar{y}_{lr} = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{Y}_R. \quad (7.2)$$

7.2. ОЦЕНКИ ПО РЕГРЕССИИ ПРИ ЗАДАННОМ b

Хотя в большинстве приложений способа оценки по регрессии b оценивается по результатам выборки, иногда целесообразно выбрать значение b заранее. При многократных обследованиях одной и той же совокупности предыдущие вычисления могли показать, что значения b , получаемые по выборке, остаются довольно стабильными; или, если x — это значение y по последней переписи, то на основе общих сведений о совокупности мы можем заключить, что b близко к единице, так что выбирается $b = 1$. Поскольку для оценок по регрессии при за-

данном b теория выборочного метода одновременно и проста и информативна, то сначала мы рассмотрим этот случай.

Теорема 7.1. При простом случайном отборе, когда b_0 — заданная постоянная, линейная оценка по регрессии

$$\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x})$$

есть несмещенная оценка с дисперсией

$$V(\bar{y}_{lr}) = \frac{1-f}{n} \frac{\sum_{i=1}^N [(y_i - \bar{y}) - b_0(x_i - \bar{X})]^2}{N-1} = \quad (7.3)$$

$$= \frac{1-f}{n} (S_y^2 - 2b_0 S_{yx} + b_0^2 S_x^2). \quad (7.4)$$

Заметим, что не требуется никаких предположений относительно зависимости между y и x в конечной совокупности.

Доказательство. Так как при многократном отборе b_0 постоянно, то по теореме 2.1

$$E(\bar{y}_{lr}) = E(\bar{y}) + b_0 E(\bar{X} - \bar{x}) = \bar{y}.$$

Далее, \bar{y}_{lr} представляет собой выборочное среднее величин $y_i - b_0(x_i - \bar{X})$, а их среднее для совокупности равно \bar{y} . Тогда по теореме 2.2

$$V(\bar{y}_{lr}) = \frac{1-f}{n} \cdot \frac{\sum_{i=1}^N [(y_i - \bar{y}) - b_0(x_i - \bar{X})]^2}{N-1} =$$

$$= \frac{1-f}{n} (S_y^2 - 2b_0 S_{yx} + b_0^2 S_x^2).$$

Следствие. Несмещенная выборочная оценка $V(\bar{y}_{lr})$ есть

$$v(\bar{y}_{lr}) = \frac{1-f}{n} \cdot \frac{\sum_{i=1}^n [(y_i - \bar{y}) - b_0(x_i - \bar{x})]^2}{n-1} =$$

$$= \frac{1-f}{n} (s_y^2 - 2b_0 s_{yx} + b_0^2 s_x^2).$$

Для доказательства достаточно применить теорему 2.4 к переменной $y_i - b_0(x_i - \bar{X})$.

Теперь возникает естественный вопрос: каково наилучшее значение b_0 ? Ответ на него дает теорема 7.2.

Теорема 7.2. Значение b_0 , при котором $V(\bar{y}_{lr})$ минимальна, есть

$$b_0 = B = \frac{S_{yx}}{S_x^2} = \frac{\sum_{i=1}^N (y_i - \bar{y})(x_i - \bar{X})}{\sum_{i=1}^N (x_i - \bar{X})^2}. \quad (7.5)$$

Назовем его коэффициентом линейной регрессии y по x для конечной совокупности. Соответствующая минимальная дисперсия есть

$$V_{min}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1 - \rho^2),$$

где ρ — коэффициент корреляции между y и x для совокупности.

Доказательство. В выражении (7.4) для $V(\bar{y}_{lr})$ положим

$$b_0 = B + d = \frac{S_{yx}}{S_x^2} + d.$$

Это дает

$$V(\bar{y}_{lr}) = \frac{1-f}{n} \left[S_y^2 - 2S_{yx} \left(\frac{S_{yx}}{S_x^2} + d \right) + S_x^2 \cdot \left(\frac{S_{yx}^2}{S_x^4} + \right. \right.$$

$$\left. \left. + 2d \frac{S_{yx}}{S_x^2} + d^2 \right) \right] = \frac{1-f}{n} \left[\left(S_y^2 - \frac{S_{yx}^2}{S_x^2} \right) + d^2 S_x^2 \right]. \quad (7.6)$$

Очевидно, что последнее выражение минимально при $d = 0$. Поскольку $\rho^2 = S_{yx}^2 / S_y^2 S_x^2$,

$$V_{min}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1 - \rho^2). \quad (7.7)$$

Теми же самыми выкладками можно воспользоваться для того, чтобы определить, насколько b_0 может отклоняться от B , не вызывая существенной потери в точности. Из (7.6) и (7.7) имеем

$$V(\bar{y}_{lr}) = \frac{1-f}{n} [S_y^2 (1 - \rho^2) + (b_0 - B)^2 S_x^2] =$$

$$= V_{min}(\bar{y}_{lr}) \left[1 + \frac{(b_0 - B)^2 S_x^2}{S_y^2 (1 - \rho^2)} \right].$$

Поскольку $BS_x = \rho S_y$, это выражение можно записать в виде

$$V(\bar{y}_{lr}) = V_{min}(\bar{y}_{lr}) \left[1 + \left(\frac{b_0}{B} - 1 \right)^2 \frac{\rho^2}{(1 - \rho^2)} \right].$$

Таким образом, если мы хотим, чтобы относительное увеличение дисперсии было меньше α , нужно чтобы

$$\left| \frac{b_0}{B} - 1 \right| < \sqrt{\alpha (1 - \rho^2) / \rho^2}. \quad (7.8)$$

Например, при $\rho = 0,7$ дисперсия увеличится меньше чем на 10% ($\alpha = 0,1$), если

$$\left| \frac{b_0}{B} - 1 \right| < \sqrt{0,1 \cdot 0,51 / 0,49} = 0,32.$$

Как следует из выражения (7.8), для того чтобы относительное увеличение дисперсии было небольшим, b_0/B должно быть близко к 1 при очень больших значениях ρ , но может существенно отличаться от 1, если ρ сравнительно невелико.

7.3. ОЦЕНКА ПО РЕГРЕССИИ, КОГДА b ВЫЧИСЛЯЕТСЯ ПО ВЫБОРКЕ

Из теоремы 7.2 можно сделать вывод о том, что если b должно быть вычислено по выборке, по-видимому, наиболее эффективной оценкой будет знакомая нам оценка B , полученная по методу наименьших квадратов, т. е.

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (7.9)$$

Теория линейной регрессии играет в статистической методологии важную роль. Однако обычные утверждения этой теории не вполне применимы к выборочным обследованиям, поскольку они справедливы при предположениях, что регрессия y по x во всей совокупности линейна, что остаточная дисперсия y относительно линии регрессии постоянна и что совокупность бесконечна. Если первые два предположения совершенно ошибочны, линейную оценку по регрессии, видимо, применять не следует. Однако в тех обследованиях, где предполагается, что регрессия y по x приблизительно линейна, применение y_{lr} , если оно возможно, может быть полезным и без предположения о точной линейности или о постоянстве остаточной дисперсии.

Поэтому мы изложим подход, при котором не требуется, чтобы регрессия y по x в совокупности была линейна. Выводы справедливы только для больших выборок. Они аналогичны соответствующим утверждениям, справедливым при больших выборках для оценки по отношению.

Покажем сначала, что для выборок объема n величина $(b - B)$ имеет порядок $1/\sqrt{n}$. Введем переменную e_i с помощью равенства

$$e_i = y_i - \bar{Y} - B(x_i - \bar{X}). \quad (7.10)$$

Отсюда следует, что согласно определению B

$$\sum_{i=1}^n e_i (x_i - \bar{X}) = \sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{X}) - B \sum_{i=1}^n (x_i - \bar{X})^2 = 0. \quad (7.11)$$

Далее,

$$b = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

и на основании (7.10)

$$b = \frac{\sum_{i=1}^n [\bar{Y} + B(x_i - \bar{X}) + e_i](x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = B + \frac{\sum_{i=1}^n e_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (7.12)$$

По теореме 2.3, $\sum_{i=1}^n e_i (x_i - \bar{x})/(n-1)$ есть несмещенная оценка выражения $\sum_{i=1}^N e_i (x_i - \bar{X})/(N-1)$, которое согласно (7.11) равно нулю. Следовательно, при многократных выборках объема n распределение выборочной ковариации $\sum_{i=1}^n e_i (x_i - \bar{x})/(n-1)$ имеет среднее, равное нулю. Известно, что стандартная ошибка выборочной ковариации представляет собой величину порядка $1/\sqrt{n}$. Таким образом, для выборок объема n выражение $\sum_{i=1}^n e_i (x_i - \bar{x})/(n-1)$ будет величиной порядка $1/\sqrt{n}$. Но выражение $\sum_{i=1}^n (x_i - \bar{x})^2/(n-1) = s_x^2$ для выборок объема n будет величиной порядка 1. Следовательно, согласно (7.12) $(b - B)$ — величина порядка $1/\sqrt{n}$.

Теорема 7.3. Если b есть оценка B по методу наименьших квадратов и

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}), \quad (7.13)$$

то для простых случайных выборок объема n

$$V(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1 - \rho^2) \quad (7.14)$$

при условии, что n достаточно велико, для того чтобы членами порядка $1/\sqrt{n}$ можно было пренебречь.

Доказательство. Беря среднее (7.10) по всем единицам выборки, имеем

$$\bar{e} = \bar{y} - \bar{Y} - B(\bar{x} - \bar{X}).$$

Подставляя отсюда \bar{y} в выражение (7.13) для \bar{y}_{lr} , получаем

$$\bar{y}_{lr} = \bar{Y} + (b - B)(\bar{X} - \bar{x}) + \bar{e}. \quad (7.15)$$

Из (7.10) с очевидностью вытекает, что среднее e_i для совокупности равно нулю. Следовательно, \bar{e} — величина порядка $1/\sqrt{n}$. Но мы уже показали, что $(b - B)$ — величина порядка $1/\sqrt{n}$. Так как $(\bar{X} - \bar{x})$ также порядка $1/\sqrt{n}$, то их произведение $(b - B)(\bar{X} - \bar{x})$ имеет порядок $1/n$. Значит, если членами порядка $1/\sqrt{n}$ можно пренебречь, то это произведение по сравнению с \bar{e} пренебрежимо мало. Таким образом,

$$\bar{y}_{lr} - \bar{Y} \approx \bar{e}.$$

Поскольку согласно (7.10) и теореме 2.1 $E(\bar{e})$ равно нулю, дисперсия среднего значения величин e_i при простом случайном отборе равна $E(\bar{e}^2)$. Следовательно, по теореме 2.2

$$V(\bar{y}_{lr}) = \frac{1-f}{n} S_e^2 = \frac{1-f}{n} \frac{\sum_{i=1}^n e_i^2}{N-1}. \quad (7.16)$$

Далее,

$$\begin{aligned}\sum e_i^2 &= \sum [(y_i - \bar{Y}) - B(x_i - \bar{X})]^2 = \sum (y_i - \bar{Y})^2 - \\ &- 2B \sum (y_i - \bar{Y})(x_i - \bar{X}) + B^2 \sum (x_i - \bar{X})^2 = \\ &= \sum (y_i - \bar{Y})^2 - B^2 \sum (x_i - \bar{X})^2\end{aligned}$$

по определению B , равенство (7.5). Но при этом

$$\rho = \frac{\sum (y_i - \bar{Y})(x_i - \bar{X})}{\sqrt{\sum (y_i - \bar{Y})^2 \sum (x_i - \bar{X})^2}} = B \sqrt{\frac{\sum (x_i - \bar{X})^2}{\sum (y_i - \bar{Y})^2}}.$$

Таким образом,

$$\sum e_i^2 = \sum (y_i - \bar{Y})^2 (1 - \rho^2),$$

так что окончательно

$$V(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1 - \rho^2). \quad (7.17)$$

В качестве выборочной оценки $V(\bar{y}_{lr})$, справедливой для больших выборок, можно принять

$$v(\bar{y}_{lr}) = \frac{1-f}{n(n-2)} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 = \quad (7.18)$$

$$= \frac{1-f}{n(n-2)} \left\{ \sum (y_i - \bar{y})^2 - \frac{[\sum (y_i - \bar{y})(x_i - \bar{x})]^2}{\sum (x_i - \bar{x})^2} \right\}. \quad (7.19)$$

Последнее выражение представляет собой обычную, удобную для вычислений формулу. Приведем вывод этой оценки.

В ходе доказательства теоремы 7.3 мы получили равенство (7.16)

$$V(\bar{y}_{lr}) \approx \frac{(1-f)}{n} S_e^2.$$

Согласно теореме 2.4 несмещенной оценкой S_e^2 служит

$$s_e^2 = \frac{1}{n-1} \sum_{i=1}^n (e_i - \bar{e})^2.$$

Далее, из равенства (7.10) следует, что

$$\begin{aligned}e_i - \bar{e} &= (y_i - \bar{y}) - B(x_i - \bar{x}) = \\ &= [(y_i - \bar{y}) - b(x_i - \bar{x})] + (b - B)(x_i - \bar{x}).\end{aligned}$$

Второй член справа имеет порядок $1/\sqrt{n}$ и им можно пренебречь по сравнению с первым членом, который имеет порядок 1. Следова-

тельно, для больших выборок в качестве оценки S_e^2 можно принять выражение

$$\frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2.$$

В формулах (7.18) и (7.19) вместо $(n-1)$ был взят делитель $(n-2)$, поскольку он применяется в обычной теории регрессии, и, как известно, в случае бесконечной совокупности и линейной регрессии дает несмещенную оценку S_e^2 .

7.4. ДОСТОВЕРНОСТЬ ФОРМУЛЫ ДЛЯ $V(\bar{y}_{lr})$ ПРИ БОЛЬШИХ ВЫБОРКАХ

Содержание предыдущих параграфов не дает ответа на вопрос: насколько большой должна быть выборка? Исчерпывающего ответа на него пока дать нельзя, но некоторое представление об этом можно получить, рассматривая бесконечную совокупность.

Согласно равенству (7.15) из теоремы 7.3

$$\bar{y}_{lr} - \bar{Y} = \bar{e} + (b - B)(\bar{X} - \bar{x}).$$

Подставляя выражение для b из (7.12), имеем

$$\bar{y}_{lr} - \bar{Y} = \bar{e} + \frac{\sum e_i (x_i - \bar{x})(\bar{X} - \bar{x})}{\sum (x_i - \bar{x})^2}. \quad (7.20)$$

Этим выражением для ошибки оценки по регрессии мы воспользуемся для нахождения главных членов дисперсии и смещения оценки \bar{y}_{lr} . Для дисперсии запишем

$$\bar{y}_{lr} - \bar{Y} = \sum_i e_i \left[\frac{1}{n} + \frac{(x_i - \bar{x})(\bar{X} - \bar{x})}{\sum (x_i - \bar{x})^2} \right].$$

Следовательно, при условии, что x_i неизменны, причем предполагается, что S_e^2 одинакова для всех x ,

$$V(\bar{y}_{lr}|x_i) = \frac{S_e^2}{n} \left[1 + \frac{n(\bar{x} - \bar{X})^2}{\sum (x_i - \bar{x})^2} \right].$$

Среднее значение этой величины по возможным выборочным значениям x_i зависит от характера распределения частот x . Если x имеет нормальное распределение, то среднее равно

$$\frac{S_e^2}{n} \left(1 + \frac{1}{n-3} \right).$$

Для случая произвольного распределения x можно показать (Cochran, 1942), что с точностью до величины порядка $1/n^3$

$$V(\bar{y}_{lr}) \approx \frac{S_e^2}{n} \left(1 + \frac{1}{n} + \frac{3 + 2\gamma_2^2}{n^2} \right), \quad (7.21)$$

где $\gamma_2^2 = k_2^2/S_x^2$ есть мера относительной асимметрии распределения x .

Возвращаясь к (7.20), заметим, что смещение \bar{y}_{lr} обусловлено вторым членом в правой части равенства, так как среднее значение \bar{e} при простом случайном отборе равно нулю. Чтобы получить главный член смещения, мы можем заменить выражение $\sum (x_i - \bar{x})^2$ его главным членом, nS_x^2 . Запишем также

$$\sum e_i (x_i - \bar{x}) = \sum e_i (x_i - \bar{X}) + n\bar{e}(\bar{X} - \bar{x}).$$

Таким образом, главный член смещения представляет собой среднее значение величины

$$\frac{\sum e_i (x_i - \bar{X})(\bar{X} - \bar{x})}{nS_x^2} + \frac{\bar{e}(\bar{X} - \bar{x})^2}{S_x^2}. \quad (7.22)$$

Обозначим переменную $e_i(x_i - \bar{X})$ через u_i . Согласно (7.11) ее среднее для совокупности $\bar{U} = 0$. Поэтому, применяя теорему 2.3 (с. 39) и полагая, что $N \rightarrow \infty$, среднее значение первого члена в (7.22) можно записать в виде

$$-\frac{E(\bar{u} - \bar{U})(\bar{x} - \bar{X})}{S_x^2} = -\frac{E(u - \bar{U})(x - \bar{X})}{nS_x^2} = -\frac{Ee(x - \bar{X})^2}{nS_x^2}. \quad (7.23)$$

Нетрудно показать, что среднее значение второго члена в (7.22) есть величина порядка $1/n^2$. Таким образом, главный член смещения возникает из-за ковариации e и $(x - \bar{X})^2$ для совокупности; он отражает влияние квадратичной регрессии y по x и исчезает, когда зависимость между y и x линейна.

Если через ρ_2 обозначить коэффициент корреляции e и $(x - \bar{X})^2$, то главный член смещения можно выразить иначе в виде

$$-\frac{\rho_2 S_e \sqrt{2 + \gamma_2}}{n}, \quad (7.23')$$

поскольку известно, что дисперсия $(x - \bar{X})^2$ равна $S_x^2(2 + \gamma_2)$, где $\gamma_2 = k_2^2/S_x^2$ есть фишера мера относительного эксцесса.

Окончательно, объединяя смещение из (7.23') и дисперсию из (7.21), получаем средний квадрат ошибки \bar{y}_{lr} , который с точностью до членов порядка $1/n^2$ равен

$$\frac{S_e^2(1 - \rho^2)}{n} \left[1 + \frac{1 + \rho_2^2(2 + \gamma_2)}{n} \right]. \quad (7.24)$$

Этот результат показывает, что если распределение x обладает умеренным эксцессом, то при выборках объемом в 50 и более для дисперсии оценки по регрессии можно применять формулу для больших выборок.

Из наших рассуждений следует также, что при n , достаточно больших для того, чтобы можно было пренебречь членами порядка $1/n^2$, в качестве оценки B можно воспользоваться не оценкой по методу наименьших квадратов, а менее эффективной оценкой b' , поскольку ошибка в b влияет только на члены порядка $1/n^2$. Например, b можно вычислять по некоторой подвыборке из исходных данных. Если же единицы можно условно разбить согласно значениям x на три одинаковые по объему группы (из малых, средних и больших), то можно, как показал Бартлет (Bartlett, 1949), применить оценку

$$b' = \frac{\bar{y}_{\text{больш}} - \bar{y}_{\text{мал}}}{\bar{x}_{\text{больш}} - \bar{x}_{\text{мал}}},$$

эффективность которой составляет приблизительно 8/9.

7.5. ДОПОЛНИТЕЛЬНЫЕ ЗАМЕЧАНИЯ О СМЕЩЕНИИ

В случае конечной совокупности выражение для главного члена смещения можно получить из формулы (7.23) параграфа 7.4 с тем изменением, что, применяя теорему 2.3, не нужно полагать, что $N \rightarrow \infty$. Основное отличие от прежнего выражения состоит в том, что появляется член, соответствующий пкс, так что

$$\text{смещение } \bar{y}_{lr} \approx -\frac{1-f}{(n-1)S_x^2} \left[\frac{\sum e_i (x_i - \bar{X})^2}{N-1} \right].$$

В предыдущем параграфе мы показали также, что если в бесконечной совокупности регрессия линейна, то главный член смещения исчезает. В действительности для бесконечных совокупностей, как хорошо известно, если регрессия линейна, то исчезает все смещение, а не только его главный член. Соответствующий результат справедлив и для конечных совокупностей, если регрессия y по x линейна в обычном смысле этого термина. Мы считаем, что в некоторой конечной совокупности регрессия линейна, если

$$y_i = \bar{Y} + \beta(x_i - \bar{X}) + e_i \quad (7.25)$$

при $E(e_i|x_i) = 0$ для каждого неизменного x_i . В частности, если какое-то определенное значение x появляется только для одной единицы в совокупности, то значение e для этой единицы должно быть равно нулю.

Теорема 7.4. Если регрессия y_i по x_i линейна, то

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}),$$

где $b = \sum (y_i - \bar{y})(x_i - \bar{x}) / \sum (x_i - \bar{x})^2$ есть коэффициент регрессии y по x по методу наименьших квадратов, есть несмещенная оценка \bar{Y} .

Требуется также, чтобы $\sum (x_i - \bar{x})^2 > 0$ в каждой выборке. (Если v_{\max} — наибольшее число единиц в совокупности, имеющих одно и то же значение x , то условие $\sum (x_i - \bar{x})^2 > 0$ выполняется для любого $n > v_{\max}$).

Доказательство. Если (7.25) выполняется при $E(e_i|x_i) = 0$, то нетрудно показать, что $\beta = B$, где B — коэффициент регрессии y по x для совокупности, задаваемый формулой (7.5). Действительно, согласно (7.25)

$$\sum x_i (y_i - \bar{y}) = \beta \sum x_i (x_i - \bar{x}) + \sum e_i x_i.$$

Так как $E(e_i|x_i) = 0$, то $\sum e_i x_i = 0$, так что $\beta = B$. Следовательно, согласно (7.15) и (7.12)

$$\bar{y}_{lr} - \bar{y} = \frac{(\bar{x} - \bar{x}) \sum e_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} + \bar{e}. \quad (7.26)$$

Рассмотрим теперь выборки с одним и тем же набором значений x_i .

Для всех таких выборок $(\bar{x} - \bar{x})$ и $\sum (x_i - \bar{x})^2$ остаются постоянными. Далее, если какое-то значение x встречается m раз в выборке и v раз в совокупности, то каждая из v единиц будет с одинаковой частотой попадать в выборки рассматриваемого типа. Отсюда следует, что средние значения как $\sum e_i (x_i - \bar{x})$, так и \bar{e} по всем этим выборкам равны нулю. Таким образом, $E(\bar{y}_{lr} - \bar{y}) = 0$ по всем этим выборкам. Следовательно, \bar{y}_{lr} есть несмещенная оценка по всем простым случайным выборкам объема n .

7.6. СРАВНЕНИЕ ОЦЕНКИ ПО РЕГРЕССИИ С ОЦЕНКОЙ ПО ОТНОШЕНИЮ И ПО СРЕДНЕМУ НА ЕДИНИЦУ

Для такого сравнения объем выборки n должен быть достаточно велик для того, чтобы были справедливы приближенные формулы дисперсии оценок по отношению и по регрессии. Три сравниваемые дисперсии оценок среднего для совокупности \bar{y} имеют вид:

$$V(\bar{y}_{lr}) = \frac{N-n}{Nn} S_y^2 (1-\rho^2) \quad (\text{по регрессии});$$

$$V(\bar{y}_R) = \frac{N-n}{Nn} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x) \quad (\text{по отношению});$$

$$V(\bar{y}) = \frac{N-n}{Nn} S_y^2 \quad (\text{по среднему на единицу}).$$

Очевидно, что дисперсия оценки по регрессии меньше, чем дисперсия оценки по среднему на единицу, за исключением случая $\rho = 0$, когда обе дисперсии совпадают.

Дисперсия оценки по регрессии меньше и дисперсия оценки по отношению, если

$$-\rho^2 S_y^2 < R^2 S_x^2 - 2R\rho S_y S_x.$$

Это условие эквивалентно неравенствам

$$(\rho S_y - R S_x)^2 > 0 \text{ или } (B - R)^2 > 0.$$

Таким образом, оценка по регрессии более точна, чем оценка по отношению, если только B не равно R . Эти величины равны, когда зависимость между y_i и x_i выражается прямой, проходящей через начало координат.

Пример. Точность оценок по регрессии, по отношению и по среднему на единицу для простой случайной выборки можно сравнить, пользуясь данными сплошного учета персиковых деревьев, описанного в примере из параграфа 6.14 (с. 193). В этом примере y_i — оценка урожая персиков в саду и x_i — число персиковых деревьев в этом саду. Мы сравним оценки общего урожая в 256 садах, полученные по некоторой выборке объемом в 100 садов. Правда, сомнительно, что такая выборка будет достаточно велика, чтобы можно было с полным основанием применить формулы для дисперсий, поскольку коэффициенты вариации как \bar{y} , так и \bar{x} несколько больше 10%. Однако этот пример служит лишь для иллюстрации вычислений. Основные данные выглядят так:

$$S_y^2 = 6409; S_{yx} = 4434; S_x^2 = 3898; R = 1,270; \rho = 0,887;$$

$$n = 100; N = 256;$$

$$V(\bar{y}_{lr}) = \frac{N(N-n)}{n} S_y^2 (1-\rho^2) =$$

$$= \frac{256 \cdot 156}{100} \cdot 6409 (1 - 0,787) = 545\,000;$$

$$V(\bar{y}_R) = \frac{N(N-n)}{n} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x) =$$

$$= \frac{256 \cdot 156}{100} [6409 + 1,613 \cdot 3898 - 2 \cdot 1,270 \cdot 4434] = 573\,000;$$

$$V(\bar{y}) = \frac{N(N-n)}{n} S_y^2 = 2\,559\,000.$$

Между дисперсиями оценки по регрессии и оценки по отношению не оказывается большого различия, как можно было бы ожидать, судя по характеру переменных. Оценка по среднему на единицу, однако, значительно уступает в точности обеим этим оценкам.

7.7. ОЦЕНКИ ПО РЕГРЕССИИ ПРИ РАССЛОЕННОМ ОТБОРЕ

Как и в случае оценки по отношению, при расслоенном отборе можно построить оценки по регрессии двух типов. Для первой — раздельной оценки \bar{y}_{lrh} — оценка по регрессии вычисляется отдельно для среднего значения в каждом слое [s — от английского «separate» — раздельный], т. е.

$$\bar{y}_{lrh} = \bar{y}_h + b_h (\bar{x}_h - \bar{x}). \quad (7.27)$$

После чего полагаем

$$\bar{y}_{lrs} = \sum_h W_h \bar{y}_{lrh}. \quad (7.28)$$

Эта оценка уместна, если есть основания считать, что истинные коэффициенты регрессии B_h меняются от слоя к слою.

Вторая — совместная оценка по регрессии \bar{y}_{lre} [с — от английского «combined» — совместный] — целесообразна, если можно предположить, что B_h одинаковы во всех слоях. Для того чтобы вычислить \bar{y}_{lre} , нужно сначала найти

$$\bar{y}_{st} = \sum_h W_h \bar{y}_{sh}; \quad \bar{x}_{st} = \sum_h W_h \bar{x}_{sh}.$$

Тогда

$$\bar{y}_{lre} = \bar{y}_{st} + b(\bar{X} - \bar{x}_{st}). \quad (7.29)$$

Мы рассмотрим сначала эти оценки для случая, когда b_h и b выбраны заранее, поскольку в этой ситуации свойства оценок проверяются чрезвычайно просто. Согласно параграфу 7.2, \bar{y}_{lrh} есть несмещенная оценка \bar{Y}_h , так что \bar{y}_{lrs} — несмещенная оценка \bar{Y} . Далее, поскольку отбор в разных слоях производится независимо, из теоремы 7.1 следует, что

$$V(\bar{y}_{lrs}) = \sum_h \frac{W_h^2(1-f_h)}{n_h} (S_{ph}^2 - 2b S_{yph} + b^2 S_{xh}^2). \quad (7.30)$$

Как было показано в теореме 7.2, $V(\bar{y}_{lrs})$ минимальна при $b_h = B_h$ — истинному коэффициенту регрессии в слое h . Минимальное значение дисперсии можно записать так:

$$V_{min}(\bar{y}_{lrs}) = \sum_h \frac{W_h^2(1-f_h)}{n_h} \left(S_{ph}^2 - \frac{S_{yph}^2}{S_{xh}^2} \right). \quad (7.31)$$

Возвращаясь к совместной оценке с заданным b , заметим, что из (7.29) вытекает, что \bar{y}_{lre} — также несмещенная оценка \bar{Y} . Поскольку \bar{y}_{lre} — обычная оценка по расслоенной выборке для переменной $y_{hi} + b(\bar{X} - x_{hi})$, то к этой переменной можно применить теорему 5.3 и получить

$$V(\bar{y}_{lre}) = \sum_h \frac{W_h^2(1-f_h)}{n_h} (S_{ph}^2 - 2b S_{yph} + b^2 S_{xh}^2). \quad (7.32)$$

Значение b , при котором эта дисперсия минимальна, равно

$$B_c = \sum_h \frac{W_h^2(1-f_h) S_{yph}}{n_h} / \sum_h \frac{W_h^2(1-f_h) S_{xh}^2}{n_h}. \quad (7.33)$$

Величина B_c представляет собой взвешенное среднее коэффициентов регрессии $B_h = S_{yph}/S_{xh}^2$ для отдельных слоев. Если обозначить

$$a_h = \frac{W_h^2(1-f_h)}{n_h} S_{xh}^2,$$

то $B_c = \sum a_h B_h / \sum a_h$.

Из (7.31) и (7.32), подставив B_c вместо b , получаем

$$V_{min}(\bar{y}_{lre}) - V_{min}(\bar{y}_{lrs}) = \sum a_h B_c^2 - (\sum a_h) B_c^2 = \sum a_h (B_h - B_c)^2. \quad (7.34)$$

Из этого равенства вытекает, что наилучшая раздельная оценка имеет меньшую дисперсию, чем наилучшая совместная оценка, если только B_h не одинаковы во всех слоях.

7.8. ВЫБОРОЧНЫЕ ОЦЕНКИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Предыдущий анализ помогает нам указать, какого типа выборочные оценки b_h и b могут дать хорошие результаты, будучи применены в оценках по регрессии. Из этого анализа следует, что для раздельной оценки нужно взять

$$b_h = \frac{\sum_i (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_i (x_{hi} - \bar{x}_h)^2}$$

как оценку B_h внутри слоя по методу наименьших квадратов. Применяя теорему 7.3 к каждому слою, получаем

$$V(\bar{y}_{lrs}) = \sum_h \frac{W_h^2(1-f_h)}{n_h} S_{ph}^2 (1 - \rho_h^2) \quad (7.35)$$

при условии, что объем выборки n_h велик во всех слоях. Для того чтобы получить выборочную оценку дисперсии, подставим в выражение (7.35) вместо $S_{ph}^2 (1 - \rho_h^2)$

$$s_{y \cdot x}^2 = \frac{1}{n_h - 2} \left[\sum_i (y_{hi} - \bar{y}_h)^2 - b_h^2 \sum_i (x_{hi} - \bar{x}_h)^2 \right].$$

Оценка \bar{y}_{lrs} страдает тем же недостатком, что и соответствующая оценка по отношению, а именно отношение смещения к стандартной ошибке может стать довольно значительным. Из параграфа 7.5 следует, что оценки по регрессии \bar{y}_{lrh} в отдельных слоях могут иметь смещения порядка $1/n_h$, причем смещения во всех слоях могут иметь один и тот же знак, так что общее смещение в \bar{y}_{lrs} также может быть величиной порядка $1/n_h$. Поскольку, как показано в параграфе 7.5, главный член смещения возникает из-за квадратичной регрессии y_{hi} по x_{hi} ,

эта опасность особенно велика, когда форма зависимости между переменными ближе к квадратичной, а не к линейной. Мы видели, что в случае совместной оценки дисперсия минимальна, когда $b = B_c$ определяется формулой (7.33). Поэтому в качестве выборочной оценки B_c естественно взять

$$b_c = \sum_h \frac{W_h^2 (1-f_h)}{n_h (n_h - 1)} \sum_i (y_{hi} - \bar{y}_h) (x_{hi} - \bar{x}_h) / \sum_h \frac{W_h^2 (1-f_h)}{n_h (n_h - 1)} \times \\ \times \sum_i (x_{hi} - \bar{x}_h)^2.$$

Если при расслоенном отборе выборка размещена пропорционально и мы можем в b_c принять n_h вместо $(n_h - 1)$, то b_c сводится к знакомой нам объединенной оценке по методу наименьших квадратов:

$$b'_c = \sum_h \sum_i (y_{hi} - \bar{y}_h) (x_{hi} - \bar{x}_h) / \sum_h \sum_i (x_{hi} - \bar{x}_h)^2.$$

При определенных условиях предпочтительнее b_c или b'_c могут оказаться другие оценки. Например, если истинные коэффициенты регрессии B_h одинаковы во всех слоях, но остаточные дисперсии относительно линии регрессии существенно меняются от слоя к слою, то может оказаться более точным другое взвешенное среднее b_h , в котором веса обратно пропорциональны оценкам дисперсий. Однако при этом выигрыш в точности для \bar{y}_{irc} будет, вероятно, небольшим.

Имеем

$$\bar{y}_{irc} - \bar{Y} = \bar{y}_{st} - \bar{Y} + b_c (\bar{X} - \bar{x}_{st}) = [\bar{y}_{st} - \bar{Y} + B_c (\bar{X} - \bar{x}_{st})] + \\ + (b_c - B_c) (\bar{X} - \bar{x}_{st}).$$

Из этого равенства следует, что если пренебречь ошибками выборки в значении b_c , то

$$V(\bar{y}_{irc}) = \sum_h \frac{W_h^2 (1-f_h)}{n_h} (S_{y|h}^2 - 2B_c S_{yxh} + B_c^2 S_{xh}^2).$$

Дальнейший анализ структуры b_c и b'_c показывает, что, вообще говоря, это — смещенные оценки. Если выборка размещена пропорционально и остаточная дисперсия относительно линии регрессии приблизительно одинакова во всех слоях, то возникающее смещение \bar{y}_{irc} представляет величину порядка $1/n$, как и слагаемое дисперсии \bar{y}_{irc} , содержащее член $V(b_c)$. Если, однако, участие дисперсии какого-то слоя, скажем h -го, в образовании общей дисперсии оценки преобладает, то слагаемое $V(\bar{y}_{irc})$, содержащее $V(b_c)$, может быть величиной порядка $1/n_h$.

В качестве оценки $V(\bar{y}_{irc})$ можно взять

$$v(\bar{y}_{irc}) = \sum_h \frac{W_h^2 (1-f_h)}{n_h (n_h - 1)} \sum_i [(y_{hi} - \bar{y}_h) - b_c (x_{hi} - \bar{x}_h)]^2.$$

Сумму по i можно, конечно, вычислять как

$$\sum_i (y_{hi} - \bar{y}_h)^2 - 2b_c \sum_i (y_{hi} - \bar{y}_h) (x_{hi} - \bar{x}_h) + b_c^2 \sum_i (x_{hi} - \bar{x}_h)^2.$$

7.9. СРАВНЕНИЕ ДВУХ ВИДОВ ОЦЕНОК ПО РЕГРЕССИИ

Для определения того, какая из оценок по регрессии, раздельная или совместная, будет лучше в каждом конкретном случае, нельзя указать жесткого правила. Для того чтобы сделать между ними выбор, нужно представить себе их достоинства и недостатки. Недостатки раздельной оценки состоят в том, что она более подвержена смещению при небольших объемах выборок внутри отдельных слоев и что ошибки выборки в коэффициентах регрессии составляют более значительную долю ее дисперсии. Недостаток совместной оценки состоит в том, что ее дисперсия преувеличена, если коэффициенты регрессии для совокупности различаются от слоя к слою.

Если мы уверены в том, что регрессии линейны и если B_h , по-видимому, одинаковы во всех слоях, то, насколько можно судить, предпочтительнее совместная оценка. Если регрессии кажутся линейными (так что смещение не должно быть большим), но B_h , по-видимому, меняется от слоя к слою, то рекомендуется раздельная оценка. Если регрессии в некоторой степени отличаются от линейных, а применяется линейная оценка по регрессии, то, вероятно, более надежной будет совместная оценка, если только не велики объемы выборок во всех слоях.

Микки (Mickey, 1959) исследовал несмещенные оценки типа оценок по регрессии. Они, однако, еще не были испытаны в широких масштабах.

У п р а ж н е н и я

7.1. Опытный фермер оценивает на глаз урожай персиков, x_i , с каждого дерева в саду с $N = 200$ деревьев. Он определил, что их общий вес $X = 11\,600$ фунтов. Для некоторой простой случайной выборки объемом в 10 деревьев все плоды были собраны и взвешены, что дало следующие результаты:

	Номер дерева										Всего
	1	2	3	4	5	6	7	8	9	10	
Действительный вес y_i	61	42	50	58	67	45	39	57	71	53	543
Оценка веса x_i	59	47	52	60	67	48	44	58	76	58	569

В качестве оценки действительного общего веса Y берем

$$\hat{Y} = N [\bar{X} + (\bar{y} - \bar{x})].$$

Вычислите оценку и найдите ее стандартную ошибку.

7.2. Установите, не будет ли более точной оценкой для примера 7.1 линейная оценка по регрессии, коэффициент b которой определен по методу наименьших квадратов.

7.3. По данным выборки в табл. 6.1 (с. 174) вычислите оценку по регрессии общего числа жителей 196 больших городов в 1930 г. Найдите приближенную стандартную ошибку этой оценки и сравните ее точность с точностью оценки по отношению.

7.4. Для упражнения 7.3 найдите оценку общего числа жителей и ее стандартную ошибку при $b = 1$.

7.5. Для следующей совокупности с $N = 5$ проверьте, что (а) регрессия y по x линейна и (б) линейная оценка по регрессии есть несмещенная оценка для простых случайных выборок объема 3. Совокупность содержит следующие пары (y, x) : (3,0); (5,0); (8,2); (8,3); (12,3).

7.6. Приближенное значение переменной x , получаемое для каждой единицы, связано с истинным значением y соотношением

$$x = y + e + d,$$

где d — постоянное смещение и e — ошибка измерения, некоррелированная с y и имеющая среднее нуль и дисперсию S_e^2 для совокупности, которая предполагается бесконечной. Для простых случайных выборок объема n сравните дисперсии (а) «разностной» оценки $[y + (\bar{X} - \bar{x})]$ среднего \bar{Y} и (б) линейной оценки по регрессии с коэффициентом b , дающим минимальную дисперсию. (Эти дисперсии могут зависеть от S_y^2 .)

7.7. Рассмотрев все возможные случаи, сравните точность раздельной и совместной оценок по регрессии суммарного значения Y для следующей совокупности, если из каждого слоя извлекается простая случайная выборка объема 2:

Слой 1		Слой 2	
x_{1i}	y_{1i}	x_{2i}	y_{2i}
4	0	5	7
6	3	6	12
7	5	8	13

Воспользуйтесь обычными оценками B_h и B_c по методу наименьших квадратов, b_h и b_c , как на с. 221, 222.

ЛИТЕРАТУРА

- Bartlett M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, 5, 207—212.
 Cochran W. G. (1942). Sampling theory when the sampling units are of unequal sizes. *Jour. Amer. Stat. Assoc.*, 37, 199—212.
 Mickey M. R. (1959). Some finite population unbiased ratio and regression estimators. *Jour. Amer. Stat. Assoc.*, 54, 594—612.
 Watson D. J. (1937). The estimation of leaf areas. *Jour. Agr. Sci.*, 27, 474.
 Yates F. (1960). *Sampling methods for censuses and surveys*. Charles Griffin and Co., London, Third edition. Есть русский перевод: Йейтс Ф. Выборочный метод в переписях и обследованиях. М., «Статистика», 1965.

ГЛАВА 8

СИСТЕМАТИЧЕСКИЙ ОТБОР

8.1. ОПИСАНИЕ

Этот способ отбора на первый взгляд сильно отличается от простого случайного отбора. Предположим, что все N единиц совокупности перенумерованы от 1 до N в некотором порядке. Для получения выборки объемом в n единиц мы сначала случайным образом отбираем какую-либо одну из первых k единиц совокупности и после этого — каждую k -ю единицу. Например, если k равно 15 и первой извлеченной оказалась единица с номером 13, то следующими будут отобраны единицы с номерами 28, 43, 58 и т. д. Извлечение первой единицы определяет всю выборку. Такая выборка называется систематической выборкой *каждой k -й единицы*.

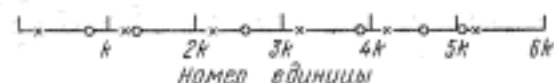
Укажем некоторые очевидные преимущества этого способа отбора по сравнению с простым случайным отбором.

1. Выборку легче извлекать и часто легче соблюдать правила отбора. Это особенно важно, когда отбор происходит непосредственно в ходе обследования. Иногда можно получить значительную экономию времени, даже если выборка извлекается до начала собственно обследования. Например, если данные обо всех единицах занесены на карточки одинакового размера, находящиеся в ящиках стандартной картотеки, то можно извлекать карточки из ящика через каждый дюйм, отмеряя расстояние линейкой. Эту операцию в отличие от простого случайного отбора можно произвести очень быстро. Конечно, такой метод несколько отличается от отбора строго каждой k -й карточки.

2. Интуитивно систематический отбор кажется более точным, чем простой случайный отбор. В сущности, при нем происходит расслоение совокупности на n слоев, которые состоят из первых k единиц, из вторых k единиц и т. д. Мы могли бы, следовательно, ожидать, что систематическая выборка обладает приблизительно той же точностью, что и соответствующая расслоенная выборка с *одной* единицей в каждом слое. Различие между ними состоит в том, что при систематическом отборе единица в каждом слое стоит на одном и том же месте относительно других, в то время, как при расслоенном случайном отборе ее место в слое определяется случайным образом отдельно для каждого слоя (см. рис. 8.1). Систематическая выборка распределена по совокупности более равномерно, и это обстоятельство делает иногда систе-

матический отбор значительно более точным, чем расслоенный случайный отбор.

В одном из вариантов систематического отбора каждая единица отбирается в центре слоя или около него, т. е. вместо того, чтобы начинать последовательность номеров некоторым случайным числом между 1 и k , мы принимаем номер первой единицы равным $(k+1)/2$, если k — нечетное, и $k/2$ или $(k+2)/2$, если k — четное число. Такой прием доводит идею систематического отбора до ее логического завершения. В том случае, когда y_i можно рассматривать как значения непрерывной функции непрерывно меняющегося аргумента i , есть основания ожидать, что такая центрально расположенная выборка будет более точной, чем случайно расположенная. Однако эффективность цент-



x — систематическая выборка
o — расслоенная случайная выборка

Рис. 8.1. Систематический отбор и расслоенный случайный отбор

рально расположенных выборок для обычно встречающихся при выборочных обследованиях типов совокупностей изучена мало, и мы ограничимся случайно расположенными выборками.

Поскольку, вообще говоря, N не есть целое кратное числа k , объемы разных систематических выборок из одной и той же совокупности могут на единицу отличаться один от другого. Так, для $N = 23$, $k = 5$ в табл. 8.1 указаны номера единиц для пяти систематических выборок. Первые три выборки имеют $n = 5$, а последние две $n = 4$. Это обстоятельство вносит некоторое осложнение в теорию систематического отбора. Если n превышает 50, то этим осложнением можно, по-видимому, пренебречь, и мы не будем для простоты принимать его во внимание при дальнейшем изложении теории. Даже при малых n вносимое им искажение вряд ли будет значительным.

Таблица 8.1

ВОЗМОЖНЫЕ СИСТЕМАТИЧЕСКИЕ ВЫБОРКИ ПРИ $N = 23$; $k = 5$

Номер систематической выборки				
I	II	III	IV	V
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23		

8.2. СВЯЗЬ СИСТЕМАТИЧЕСКОГО ОТБОРА С ГНЕЗДОВЫМ

На систематический отбор можно взглянуть и с другой стороны. В табл. 8.2 для $N = nk$ указаны k возможных систематических выборок. Из этой таблицы с очевидностью вытекает, что в этом случае вся совокупность разделена на k больших единиц отбора, каждая из которых содержит n исходных единиц. Извлечение случайно расположенной систематической выборки в точности совпадает с извлечением случайным образом одной из этих больших единиц. Таким образом, систематический отбор равносителен извлечению единственной составной единицы отбора, которая и образует всю выборку. Систематическая выборка — это простая случайная выборка, содержащая одну гнездовую единицу из совокупности k гнездовых единиц.

Таблица 8.2

СТРОЕНИЕ k СИСТЕМАТИЧЕСКИХ ВЫБОРОК

Номер выборки					
1	2	...	i	...	k
y_1	y_2		y_i		y_k
y_{k+1}	y_{k+2}		y_{k+i}		y_{2k}
...
$y_{(n-1)k+1}$	$y_{(n-1)k+2}$		$y_{(n-1)k+i}$		y_{nk}
Средние	\bar{y}_2		\bar{y}_i		\bar{y}_k

8.3. ДИСПЕРСИЯ ОЦЕНКИ СРЕДНЕГО

Для дисперсии \bar{y}_{sy} , среднего значения систематической выборки [sy — от английского «systematic» — систематический], существует несколько формул. Три формулы, приведенные далее, справедливы для любого вида гнездового отбора, когда гнезда содержат по n элементов и выборка состоит из одного гнезда.

Если $N = nk$, то для случайно расположенной систематической выборки \bar{y}_{sy} , как нетрудно проверить, есть несмещенная оценка \bar{Y} . Если $N \neq nk$, то это утверждение перестает быть справедливым, хотя смещение, по-видимому, незначительно. Смещения можно избежать, придавая некоторым выборкам большую вероятность быть извлеченными. Рассмотрим пример, приведенный в табл. 8.1. Если каждой из первых трех выборок придать вероятность быть отобранной, равную $5/23$, а двум оставшимся — вероятность, равную $4/23$, то выборочное среднее не будет иметь смещения. Способ извлечения выборки, обладающий этим свойством, заключается в том, чтобы извлечь случайное число между 1 и N и после этого взять каждую k -ю единицу, считая как до, так и после соответствующего номера. Так, для табл. 8.1 систематическая выборка I будет получена, если этим случайным числом окажется одно из пяти чисел 1, 6, 11, 16

или 21. Очевидно, что вероятность получить каждую из выборок I, II или III равна $5/23$, а каждую из выборок IV или V равна $4/23$.

При дальнейшем изложении символ y_{ij} будет обозначать j -й член i -й систематической выборки, так что $j = 1, 2, \dots, n$, $i = 1, 2, \dots, k$. Среднее i -й выборки будет обозначаться через \bar{y}_i .

Теорема 8.1. Дисперсия среднего систематической выборки есть

$$V(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2, \quad (8.1)$$

где

$$S_{wsy}^2 = \frac{1}{k(n-1)} \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$$

есть дисперсия единиц, принадлежащих одной и той же систематической выборке (*wsy* — от английского «within» — внутри и «systematic» — систематический). Знаменатель этой дисперсии $k(n-1)$ строится по обычным правилам дисперсионного анализа: каждая из k выборок вносит в сумму квадратов в числителе $(n-1)$ степеней свободы.

Доказательство. Согласно обычному тождеству дисперсионного анализа

$$(N-1)S^2 = \sum_i \sum_j (y_{ij} - \bar{Y})^2 = n \sum_i (\bar{y}_i - \bar{Y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_i)^2.$$

Но дисперсия \bar{y}_{sy} по определению равна:

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{i=1}^k (\bar{y}_i - \bar{Y})^2.$$

Следовательно,

$$(N-1)S^2 = nkV(\bar{y}_{sy}) + k(n-1)S_{wsy}^2.$$

Отсюда легко вывести утверждение теоремы.

Следствие. Среднее значение для систематической выборки более точно, чем среднее для простой случайной выборки, в том и только в том случае, когда

$$S_{wsy}^2 > S^2. \quad (8.2)$$

Доказательство. Для среднего \bar{y} простой случайной выборки объема n

$$V(\bar{y}) = \frac{N-n}{N} \frac{S^2}{n}.$$

Из (8.1) вытекает, что $V(\bar{y}_{sy}) < V(\bar{y})$ в том и только в том случае, когда

$$\frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2 < \frac{N-n}{N} \frac{S^2}{n}.$$

т. е. когда

$$k(n-1) S_{wsy}^2 > \left(N-1 - \frac{N-n}{n} \right) S^2 = k(n-1) S^2.$$

В этом важном результате, справедливом для гнездового отбора вообще, утверждается, что систематический отбор более точен, чем простой случайный отбор, если дисперсия внутри систематических выборок больше дисперсии всей совокупности. Систематический отбор точен, когда единицы внутри одной и той же выборки неоднородны, и неточен, когда они однородны. К этому можно прийти и интуитивно. Если внутри систематической выборки вариация по сравнению с вариацией в совокупности невелика, то последовательно отбиравшиеся единицы выборки несут более или менее одинаковую информацию. Другое выражение для дисперсии приводится в теореме 8.2.

Теорема 8.2.

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1) \rho_w], \quad (8.3)$$

где ρ_w — коэффициент корреляции между парами единиц, принадлежащих одной и той же систематической выборке. Этот коэффициент определяется формулой

$$\rho_w = \frac{E(y_{ij} - \bar{Y})(y_{in} - \bar{Y})}{E(y_{ij} - \bar{Y})^2},$$

где числитель представляет собой среднее по всем $kn(n-1)/2$ различным парам, а знаменатель — среднее по всем N значениям y_{ij} . Поскольку знаменатель равен $(N-1)S^2/N$, это дает

$$\rho_w = \frac{2}{(n-1)(N-1)S^2} \sum_{i=1}^k \sum_{j < n} (y_{ij} - \bar{Y})(y_{in} - \bar{Y}).$$

Доказательство

$$\begin{aligned} n^2 k V(\bar{y}_{sy}) &= n^2 \sum_{i=1}^k (\bar{y}_i - \bar{Y})^2 = \\ &= \sum_{i=1}^k [(y_{i1} - \bar{Y}) + (y_{i2} - \bar{Y}) + \dots + (y_{in} - \bar{Y})]^2. \end{aligned}$$

Сумма квадратов выражений, стоящих в круглых скобках, совпадает с общей суммой квадратов отклонений от \bar{Y} , т. е. $(N-1)S^2$. Отсюда

$$\begin{aligned} n^2 k V(\bar{y}_{sy}) &= (N-1)S^2 + 2 \sum_i \sum_{j < n} (y_{ij} - \bar{Y})(y_{in} - \bar{Y}) = \\ &= (N-1)S^2 + (n-1)(N-1)S^2 \rho_w. \end{aligned}$$

Следовательно,

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1) \rho_w].$$

Это выражение показывает, что положительная корреляция значений наблюдаемой переменной у единиц в одной и той же выборке увеличивает дисперсию выборочного среднего. Причем из-за множителя $(n-1)$ даже малая положительная корреляция может увеличить ее довольно сильно.

Две предыдущие теоремы выражали $V(\bar{y}_{st})$ через S^2 , т. е. соотносили эту дисперсию с дисперсией для простой случайной выборки. Существует аналог теоремы 8.2, в котором $V(\bar{y}_{st})$ выражена через дисперсию расслоенной случайной выборки, где слои были составлены из первых k единиц, вторых k единиц и т. д. В наших обозначениях индекс j при y_{ij} соответствует номеру слоя. Среднее для слоя будем записывать как $\bar{y}_{.j}$.

Теорема 8.3.

$$V(\bar{y}_{st}) = \frac{S_{st}^2}{n} \left(\frac{N-n}{N} \right) [1 + (n-1)\rho_{st}], \quad (8.4)$$

где

$$S_{st}^2 = \frac{1}{n(k-1)} \sum_{j=1}^n \sum_{i=1}^k (y_{ij} - \bar{y}_{.j})^2.$$

Это — дисперсия единиц, принадлежащих одному и тому же слою. В знаменателе стоит $n(k-1)$, потому что каждый из n слоев вносит $(k-1)$ степеней свободы. Далее,

$$\rho_{st} = \frac{E(y_{ij} - \bar{y}_{.j})(y_{in} - \bar{y}_{.n})}{E(y_{ij} - \bar{y}_{.j})^2}.$$

Эта величина представляет собой коэффициент корреляции между отклонениями от среднего значения для слоя по всем парам единиц, принадлежащих одной и той же систематической выборке.

$$\rho_{st} = \frac{2}{n(n-1)(k-1)} \sum_{i=1}^k \sum_{j < n} (y_{ij} - \bar{y}_{.j})(y_{in} - \bar{y}_{.n}) / S_{st}^2. \quad (8.5)$$

Доказательство аналогично доказательству теоремы 8.2.

Следствие. Если $\rho_{st} = 0$, то систематическая выборка имеет ту же точность, что и соответствующая расслоенная случайная выборка с одной единицей из каждого слоя. Это утверждение вытекает из того, что для расслоенной случайной выборки такого вида $V(\bar{y}_{st})$ (согласно следствию 2 из теоремы 5.3) равна:

$$V(\bar{y}_{st}) = \left(\frac{N-n}{N} \right) \frac{S_{st}^2}{n}.$$

Другие формулы для $V(\bar{y}_{st})$, относящиеся к автокоррелированной совокупности, были получены У. и Л. Мэдоу (W. G. and L. H. Madow, 1944), которые провели первое теоретическое исследование точности систематического отбора.

Пример. В табл. 8.3 приведены данные для небольшой искусственной совокупности, показывающей тенденцию [тренд] к довольно

устойчивому росту значений признака у последовательных единиц. Имеем $N = 40$, $k = 10$, $n = 4$. Каждый столбец соответствует некоторой систематической выборке, а строки представляют собой слои. Пример иллюстрирует ситуацию, когда корреляция «внутри слоев» положительна. Например, в первой выборке каждое из четырех чисел (0, 6, 18, 26) меньше среднего значения в слое, к которому оно принадлежит. Это справедливо, с небольшими исключениями, для первых пяти систематических выборок. В последних пяти выборках отклонения от средних значений для слоев в основном положительны. Таким образом, члены суммы в выражении для ρ_{st} преимущественно положительны. В соответствии с теоремой 8.3 можно ожидать, что систематический отбор будет менее точным, чем расслоенный случайный отбор с одной единицей в каждом слое.

Таблица 8.3

ДАННЫЕ ПО 10 СИСТЕМАТИЧЕСКИМ ВЫБОРКАМ ПРИ $n=4$; $N=kn=40$

Слой	Номер систематической выборки										Среднее значение для слоя
	1	2	3	4	5	6	7	8	9	10	
I	0	1	1	2	5	4	7	7	8	6	4,1
II	6	8	9	10	13	12	15	16	16	17	12,2
III	18	19	20	20	24	23	25	28	29	27	23,3
IV	26	30	31	31	33	32	35	37	38	38	33,1
Суммарные значения	50	58	61	63	75	71	82	88	91	88	72,7

Дисперсию $V(\bar{y}_{st})$ находим непосредственно по суммарным значениям систематических выборок:

$$V(\bar{y}_{st}) = V_{st} = \frac{1}{k} \sum_{i=1}^k (\bar{y}_{.i} - \bar{Y})^2 = \frac{1}{n^2 k} \sum_{i=1}^k (n\bar{y}_{.i} - n\bar{Y})^2 =$$

$$= \frac{1}{160} [(50)^2 + (58)^2 + \dots + (88)^2 - \frac{(727)^2}{10}] = 11,63.$$

Для того чтобы определить дисперсии при случайном и расслоенном случайном отборе, нам нужна таблица дисперсионного анализа для совокупности, выделяющая вариацию «между строками» и «внутри строк».

Таблица 8.4

ДИСПЕРСИОННЫЙ АНАЛИЗ

	Степени свободы	Суммы квадратов	Средние квадраты
Между слоями (строками)	3	4828,3	
Внутри слоев	36	485,5	13,49 = S_{st}^2
Вся совокупность	39	5313,8	136,25 = S^2

Такой анализ представлен в табл. 8.4. Из таблицы следует, что дисперсии оценок средних для простой случайной и расслоенной случайной выборки имеют значения:

$$V_{\text{ran}} = \left(\frac{N-n}{N} \right) \frac{S^2}{n} = \frac{9}{10} \cdot \frac{136,25}{4} = 30,66;$$

$$V_{\text{ст}} = \left(\frac{N-n}{N} \right) \frac{S_{\text{уст}}^2}{n} = \frac{9}{10} \cdot \frac{13,49}{4} = 3,04.$$

Как расслоенный случайный отбор, так и систематический отбор оказались гораздо более эффективными, чем простой случайный отбор, причем, как и ожидалось, систематический отбор менее точен, чем расслоенный случайный отбор.

В табл. 8.5 приведены те же данные, но значения наблюдений во втором и четвертом слоях расположены в обратном порядке. Из-за этого значение $\rho_{\text{уст}}$ становится отрицательным, поскольку становятся отрицательными большинство попарных произведений отклонений от средних значений слоев для пар наблюдений, лежащих в одной и той же систематической выборке. Например, для первой систематической выборки отклонения от средних значений слоев принимают теперь значения — 4,1; +4,8; —5,3; +4,9. Значит, из шести попарных произведений отклонений четыре отрицательны. Приблизительно так же обстоит дело в каждой систематической выборке.

Таблица 8.5
ДАННЫЕ ИЗ ТАБЛ. 8.3, РАСПОЛОЖЕННЫЕ В СЛОЯХ II И IV
В ОБРАТНОМ ПОРЯДКЕ

Слой	Номер систематической выборки										Средние значения для слоев
	1	2	3	4	5	6	7	8	9	10	
I	0	1	1	2	5	4	7	7	8	6	4,1
II	17	16	16	15	12	13	10	9	8	6	12,2
III	18	19	20	20	24	23	25	28	29	27	23,3
IV	38	38	37	35	32	33	31	31	30	26	33,1
Суммарные значения	73	74	74	72	73	73	73	75	75	65	72,7

Описанное изменение не влияет на V_{ran} и $V_{\text{ст}}$. В случае же систематического отбора оно приносит поразительное увеличение точности, как это можно видеть при сравнении суммарных значений для систематических выборок из табл. 8.5 и 8.3. Теперь мы имеем

$$V_{\text{ст}} = \frac{1}{160} \left[(73)^2 + (74)^2 + \dots + (65)^2 - \frac{(727)^2}{10} \right] = 0,46.$$

Иногда удастся воспользоваться этой особенностью систематического отбора, нумеруя единицы так, чтобы создать отрицательную корреляцию внутри слоев. Для этого нужно точно знать тенденцию [тренд],

которой следуют значения переменных внутри совокупности. Однако, как мы позднее увидим, в табл. 8.5 представлен один из тех случаев, для которых по выборке трудно получить хорошую оценку стандартной ошибки $\bar{y}_{\text{ст}}$.

8.4. СРАВНЕНИЕ СИСТЕМАТИЧЕСКОГО ОТБОРА СО СЛУЧАЙНЫМ РАССЛОЕННЫМ ОТБОРОМ

Эффективность систематического отбора по сравнению с расслоенным или простым случайным отбором очень сильно зависит от особенностей совокупности. Существуют такие совокупности, систематический отбор из которых дает высокую точность, и такие, применительно к которым он менее точен, чем простой случайный отбор. Для некоторых совокупностей и некоторых значений n дисперсия среднего систематической выборки, $V(\bar{y}_{\text{ст}})$, ведет себя поразительно плохо — она может даже *расти* при увеличении объема выборки. Поэтому трудно указать общие условия, при которых рекомендуется применять систематический отбор. Во всяком случае, чтобы его применение было эффективным, необходимо знать строение совокупности, из которой производится отбор.

В исследовании этой проблемы существуют два направления. При одном из них сравниваются различные типы отбора из искусственных совокупностей, для которых y_i представляет собой некоторую простую функцию i . При другом аналогичное сравнение производится для реальных совокупностей. Некоторые наиболее важные результаты таких исследований излагаются в последующих параграфах.

8.5. СОВОКУПНОСТИ СО «СЛУЧАЙНЫМ» ПОРЯДКОМ РАСПОЛОЖЕНИЯ ЕДИНИЦ

Систематический отбор, поскольку он удобен, применяется иногда к совокупностям, в которых единицы действительно расположены случайным образом. Так, например, обстоит дело при отборе из картотеки, составленной в алфавитном порядке фамилий, если измеряется признак, никак не связанный с фамилией обследуемого. В этом случае не будет ни какой-либо тенденции или расслоения по y_i в расположении карточек, ни корреляции между соседними единицами.

В такой ситуации мы могли бы ожидать, что систематический отбор будет, в сущности, равносильным простому случайному отбору и будет иметь ту же дисперсию. Для конкретной конечной совокупности с заданными значениями n и k это не всегда справедливо, потому что $V_{\text{ст}}$, имеющая только k степеней свободы, при малых k весьма неустойчива и может оказаться и больше и меньше, чем V_{ran} . Однако имеются две теоремы, показывающие, что в среднем эти дисперсии равны.

Теорема 8.4. Рассмотрим все $N!$ конечных совокупностей, которые образуются с помощью $N!$ перестановок некоторого набора чисел y_1, y_2, \dots, y_N . Тогда в среднем по всем этим конечным совокупностям

$$E(V_{\text{ст}}) = V_{\text{ran}}.$$

Заметим, что V_{ran} для всех перестановок одинакова.

Это теорема, доказанная У. и Л. Мэдоу (W. G. and L. H. Madow, 1944), утверждает, что если перестановку, определяющую порядок значений в некоторой конкретной конечной совокупности, можно считать выбранной случайным образом из возможных $N!$ перестановок, то в среднем систематический отбор эквивалентен простому случайному отбору.

При другом подходе конечную совокупность считают извлеченной случайным образом из некоторой бесконечной надсовокупности, обладающей определенными свойствами. Доказываемое утверждение относится не к какой-либо отдельной конечной совокупности (т. е. не к некоторому конкретному набору значений y_1, y_2, \dots, y_N), а к среднему по всем конечным совокупностям, которые могут быть извлечены из данной бесконечной совокупности. На первый взгляд может показаться, что такой подход не имеет прямого отношения к практике выборочного исследования, но это впечатление ошибочно. Всякий способ отбора применяется на практике к целому ряду конечных совокупностей. Один из способов описания класса конечных совокупностей, к которым применим данный способ отбора, состоит как раз в описании бесконечной надсовокупности, из которой эти конечные совокупности могут быть случайным образом извлечены.

Символ $\bar{\sigma}$ обозначает среднее по всем конечным совокупностям, которые могут быть получены из данной надсовокупности.

Теорема 8.5. Если переменные y_i ($i = 1, 2, \dots, N$) получены с помощью случайного отбора из надсовокупности, для которой

$$\bar{\sigma} y_i = \mu, \quad \bar{\sigma} (y_i - \mu)(y_j - \mu) = 0 \quad (i \neq j), \quad \bar{\sigma} (y_i - \mu)^2 = \sigma_i^2,$$

то

$$\bar{\sigma} V_{sy} = \bar{\sigma} V_{ran}.$$

Главную роль играют два условия — что все y_i имеют одно и то же среднее μ , т. е. в их изменении отсутствует какой-либо тренд, и что между значениями y_i и y_j в двух различных точках нет линейной корреляции. Дисперсия σ_i^2 может быть различной для разных i .

Доказательство. Для любой определенной конечной совокупности

$$V_{ran} = \frac{N-n}{Nn} \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1}.$$

Далее,

$$\begin{aligned} \sum_{i=1}^N (y_i - \bar{Y})^2 &= \sum_{i=1}^N [(y_i - \mu) - (\bar{Y} - \mu)]^2 = \\ &= \sum_{i=1}^N (y_i - \mu)^2 - N(\bar{Y} - \mu)^2. \end{aligned}$$

Поскольку y_i и y_j некоррелированы ($i \neq j$),

$$\bar{\sigma} (\bar{Y} - \mu)^2 = \frac{1}{N^2} \sum_{i=1}^N \sigma_i^2.$$

Следовательно,

$$\bar{\sigma} V_{ran} = \frac{N-n}{Nn(N-1)} \left(\sum_{i=1}^N \sigma_i^2 - N \frac{\sum \sigma_i^2}{N^2} \right).$$

Отсюда

$$\bar{\sigma} V_{ran} = \frac{N-n}{N^2 n} \sum_{i=1}^N \sigma_i^2.$$

Возвращаясь к V_{sy} , обозначим через \bar{y}_u среднее значение признака для u -й систематической выборки. Для любой определенной конечной совокупности

$$V_{sy} = \frac{1}{k} \sum_{u=1}^k (\bar{y}_u - \bar{Y})^2 = \frac{1}{k} \left[\sum_{u=1}^k (\bar{y}_u - \mu)^2 - k(\bar{Y} - \mu)^2 \right].$$

По теореме о дисперсии среднего для некоррелированной выборки из бесконечной совокупности

$$\bar{\sigma} V_{sy} = \frac{1}{k} \left(\frac{\sum_{i=1}^N \sigma_i^2}{n^2} - \frac{k \sum_{i=1}^N \sigma_i^2}{N^2} \right) = \frac{N-n}{N^2 n} \sum_{i=1}^N \sigma_i^2 = \bar{\sigma} V_{ran}.$$

8.6. СОВОКУПНОСТИ С ЛИНЕЙНЫМ ТРЕНДОМ

Если совокупность содержит только линейный тренд, как показано на рис. 8.2, то характер результатов представить себе довольно легко. Из рис. 8.2 видно, что V_{sy} и V_{st} (при выборке с одной единицей из каждого слоя) будут меньше, чем V_{ran} . Кроме того, V_{sy} будет больше, чем V_{st} , поскольку если в некотором слое значение наблюдения меньше среднего для этого слоя, то при систематическом отборе оно будет меньше и во всех остальных слоях, в то время как при случайном расслоенном отборе ошибки внутри слоев могут взаимно уничтожаться.

Для теоретической проверки этих результатов достаточно рассмотреть случай $y_i = i$. Имеем

$$\sum_{i=1}^N i = \frac{N(N+1)}{2}; \quad \sum_{i=1}^N i^2 = \frac{N(N+1)(2N+1)}{6}.$$

Дисперсия совокупности, S^2 , равна:

$$\begin{aligned} S^2 &= \frac{1}{N-1} (\sum y_i^2 - N\bar{Y}^2) = \frac{1}{N-1} \left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4} \right] = \\ &= \frac{N(N+1)}{12}. \end{aligned} \quad (8.6)$$

Следовательно, дисперсия среднего для простой случайной выборки равна:

$$V_{ran} = \frac{N-n}{N} \cdot \frac{S^2}{n} = \frac{n(n-1)}{N} \cdot \frac{N(N+1)}{12n} = \frac{(n-1)(N+1)}{12}. \quad (8.7)$$

Для того чтобы найти дисперсию внутри слоев, S_w^2 , достаточно лишь подставить в формуле (8.6) k вместо N . Это дает

$$V_{st} = \frac{N-n}{N} \cdot \frac{S_w^2}{n} = \frac{n(k-1)}{nk} \cdot \frac{k(k+1)}{12n} = \frac{(k^2-1)}{12n}. \quad (8.8)$$

При систематическом отборе среднее значение для второй выборки превосходит среднее для первой на 1; среднее для третьей выборки пре-

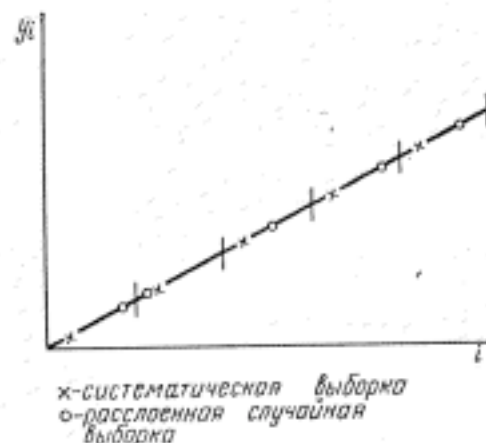


Рис. 8.2. Систематический отбор из совокупности с линейным трендом

восходит среднее для второй на 1 и т. д. Поэтому при вычислении дисперсии средние \bar{y}_n можно заменить числами 1, 2, ..., k . Следовательно, снова исходя из (8.6), получаем

$$\sum_{n=1}^k (\bar{y}_n - \bar{Y})^2 = \frac{k(k^2-1)}{12}.$$

Отсюда

$$V_{st} = \frac{1}{k} \sum (\bar{y}_n - \bar{Y})^2 = \frac{k^2-1}{12}. \quad (8.9)$$

Из формул (8.7), (8.8) и (8.9) заключаем, как и ожидалось, что

$$V_{st} = \frac{k^2-1}{12n} \leq V_{st} = \frac{k^2-1}{12} \leq V_{ran} = \frac{(k-1)(N+1)}{12}.$$

Дисперсии для разных способов отбора равны только при $n = 1$. Таким образом, если мы хотим устранить влияние линейного тренда, предполагаемого или неожиданного, то для этой цели систематическая выборка гораздо более эффективна, чем простая случайная выборка, но менее эффективна, чем расслоенная случайная выборка.

Эффект применения систематического отбора при наличии линейного тренда можно повысить несколькими способами. Один из них сос-

тоит в том, чтобы применить центрально расположенную выборку. Другой — в том, чтобы при вычислении оценки вместо невзвешенного среднего брать взвешенное, в котором всем внутренним членам выборки придаются веса, равные единице (до деления на n), а первому и последнему членам — другие веса. Если число, отобранное случайным образом из чисел 1, 2, ..., k , оказалось равным i , то эти веса будут равны

$$1 \pm \frac{n(2i-k-1)}{2(n-1)k},$$

причем вес, придаваемый первому члену, имеет знак «+», а последнему — знак «-». Очевидно, что при любом i сумма этих двух весов равна 2. Читатель может проверить: если совокупность содержит только линейный тренд и $N = nk$, то взвешенное таким способом выборочное среднее равно истинному среднему для совокупности. Эти *концевые поправки* были введены Йейтсом (Yates, 1948), который исследовал также эффект их применения.

8.7. СОВОКУПНОСТИ С ПЕРИОДИЧЕСКОЙ ВАРИАЦИЕЙ

Если совокупность содержит периодический тренд, например обычную синусоиду, то эффективность систематической выборки зависит от значения k . Это можно наглядно видеть на рис. 8.3. Высота кривой

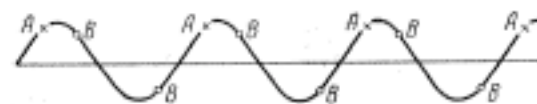


Рис. 8.3. Периодическая вариация

на нем соответствует наблюдению y_t . Выборочные точки A представляют наименее благоприятный для систематической выборки случай. Он имеет место, если k равно периоду синусоиды или целому числу, кратному этому периоду. Каждое наблюдение в систематической выборке будет одинаковым, так что выборка будет не более точной, чем единичное наблюдение, полученное из совокупности случайным образом.

Наиболее благоприятен тот случай (выборка B), когда k — нечетное число, кратное полупериоду. Среднее значение каждой систематической выборки будет в точности равно среднему для совокупности, поскольку отклонения вверх и вниз от прямой на рис. 8.3 взаимно уравниваются. Следовательно, дисперсия среднего выборки будет равна нулю. В промежуточных между этими двумя случаями эффективность выборки будет зависеть от соотношения между k и длиной волны.

Совокупности, которые можно описать точной синусоидой, на практике, по-видимому, не встречаются. Однако совокупности с более или менее выраженным периодическим трендом — не редкость. Примерами могут служить транспортный поток на определенном участке дороги в течение суток и объем продаж в магазине в течение семи дней недели. Для оценивания среднего за некоторый период времени было бы оче-

видно, нецелесообразным формировать систематическую выборку, производя наблюдения ежедневно в 4 часа дня или в каждый четверг. Напротив, нужно рассредоточивать выборку вдоль периодической кривой, в случае продаж, например, следя за тем, чтобы каждый день недели был одинаково представлен в выборке.

В некоторых совокупностях встречаются менее заметные периодические колебания. Например, если имеется ряд еженедельных платежных ведомостей для небольшого участка предприятия, то список рабочих в каждой из них может быть составлен в одном и том же порядке и содержать от 19 до 23 фамилий. Тогда систематическая выборка каждого 20-го рабочего за период в несколько недель может включать записи, относящиеся к одному и тому же рабочему или к двум или трем рабочим, принадлежащим к наиболее высоко оплачиваемой группе. Аналогично систематическая выборка фамилий из городского справочника, где под той же фамилией сначала значится глава домохозяйства, а потом его дети, может содержать слишком много глав домохозяйств или слишком много детей. Если есть время исследовать характер этой периодичности, то обычно систематическую выборку можно построить таким образом, чтобы воспользоваться ее особенностями. В противном случае, когда периодичность предполагается, но характер ее не известен, лучше применять простую или расслоенную случайную выборку.

В некоторых реальных совокупностях может присутствовать квазипериодическая вариация, которую трудно предвидеть заранее. Л. Мэдоу (L. H. Madow, 1946) обнаружила существование такой вариации в гряде саженцев деревьев лиственной породы в весьма небольшой совокупности ($N = 420$). Финни (Finney, 1950) рассмотрел подобное явление при изучении объема древесины на делянке в лесу Dehra Dun, хотя при повторном анализе данных Милн (Milne, 1959) высказал предположение о том, что кажущаяся периодичность могла быть вызвана процессом измерения. Влияние квазипериодичности приводит к тому, что систематический отбор может оказаться мало пригодным при одних значениях n и особенно удачным при других. Насколько часто проявляется такое влияние, сказать трудно. Матерн (Matérn, 1960) ссылается на примеры, в которых та или иная пространственная периодическая вариация может быть вызвана естественными причинами (например, приливами), но он считает, что при обследовании леса достоверных случаев такой вариации найдено не было.

8.8. АВТОКОРРЕЛИРОВАННЫЕ СОВОКУПНОСТИ

Для многих реальных совокупностей есть основания ожидать, что два наблюдения y_i и y_j будут более сходными, если единицы i и j расположены в ряду недалеко одна от другой, а не на большом расстоянии. Так обстоит дело всякий раз, когда какие-либо естественные причины обуславливают медленное изменение значений по мере продвижения вдоль ряда. В математической модели такой ситуации можно считать, что между y_i и y_j существует положительная корреляция, которая зависит только от расстояния между ними, $i - j$, и стремится к нулю при увеличении этого расстояния. Эта модель, хотя она и слиш-

ком упрощена, помогает выразить одну из рельефных особенностей многих реальных совокупностей.

Для того чтобы выяснить, применима ли эта модель к конкретной совокупности, мы можем вычислить коэффициенты корреляции ρ_u между парами наблюдений, находящимися на расстоянии u единиц одно от другого, и построить график соответствующих значений как функцию u . Этот график, или функция, которую он представляет, называется *коррелограммой*. Даже если модель применима к какой-либо конечной совокупности, коррелограмма для нее не будет гладкой функцией из-за неправильностей, обусловленных конечным характером совокупности. При сравнении систематического и расслоенного случайного отборов из совокупностей, описываемых моделью, эти неправильности затрудняют получение результатов для какой-либо одной конечной совокупности. Такое сравнение можно провести, рассматривая среднее из целого ряда конечных совокупностей, полученных случайным образом из некоторой бесконечной надсовокупности, к которой применима эта модель. Такой прием уже был применен в теореме 8.5.

Таким образом, мы предполагаем, что наблюдения y_i ($i = 1, 2, \dots, N$) извлечены из надсовокупности, для которой

$$E(y_i) = \mu; E(y_i - \mu)^2 = \sigma^2; E(y_i - \mu)(y_{i+u} - \mu) = \rho_u \sigma^2, \quad (8.10)$$

где

$$\rho_u \geq \rho_v \geq 0 \text{ при любых } u < v.$$

Извлечение одного набора значений y_i из этой надсовокупности приводит к образованию некоторой конечной совокупности объема N .

Средняя дисперсия по всем конечным совокупностям при систематическом отборе обозначается через

$$E V_{sy} = E(\bar{y}_{sy} - \bar{Y})^2.$$

Для этого класса совокупностей нетрудно показать, что расслоенный случайный отбор предпочтительнее простого случайного отбора, однако относительно систематического отбора общего утверждения сформулировать нельзя. Внутри этого класса существуют надсовокупности, для которых систематический отбор предпочтительнее расслоенного случайного отбора, но существуют и такие надсовокупности, для которых, при определенных значениях k , систематический отбор уступает простому случайному отбору.

Если дополнительно предположить, что коррелограмма есть выпуклая вверх функция, то можно доказать одну общую теорему.

Теорема 8.6. Если, в дополнение к условиям (8.10), выполняется

$$\delta_i^2 = \rho_{i+1} + \rho_{i-1} - 2\rho_i \geq 0 \quad [i = 2, 3, \dots, (kn - 2)],$$

то при любом объеме выборки

$$E V_{sy} \leq E V_{st} \leq E V_{ran}.$$

Далее, за исключением случая $\delta_i^2 = 0$, $i = 2, 3, \dots, (kn - 2)$, выполняется

$$E V_{sy} < E V_{st}.$$

Теорема была доказана Кокреном (Cochran, 1946).

Приведем набросок доказательства при $n = 2$, который показывает, какую роль играет условие выпуклости вверх. Члены пары, образующие систематическую выборку, всегда отстоят один от другого на k единиц. Следовательно,

$$\mathcal{E}V(\bar{y}_{st}) = \frac{1}{4}(\sigma^2 + \sigma^2 + 2\rho_k \sigma^2) = \frac{1}{2}\sigma^2(1 + \rho_k).$$

В случае расслоенной выборки для каждой единицы, извлекаемой из соответствующего слоя, существует k возможных мест, образующих k^2 возможных комбинаций расположения выборки. Числа комбинаций, для которых расстояние между единицами составляет 1, 2, ..., $(2k - 1)$, будут следующими:

Расстояние	1	2	...	$(k-1)$	k	$(k+1)$...	$(2k-1)$	Итого
Число комбинаций	1	2	...	$(k-1)$	k	$(k-1)$...	1	k^2

Следовательно, среднее значение $V(\bar{y}_{st})$, взятое по всем k^2 комбинациям, может быть записано в виде

$$\mathcal{E}V(\bar{y}_{st}) = \frac{\sigma^2}{2k^2} \left[\sum_{i=1}^{k-1} i(2 + \rho_i + \rho_{2k-i}) + k(1 + \rho_k) \right].$$

Аналогично $\mathcal{E}V(\bar{y}_{sy})$ можно выразить в виде

$$\mathcal{E}V(\bar{y}_{sy}) = \frac{\sigma^2}{2k^2} \left[\sum_{i=1}^{k-1} i(2 + 2\rho_k) + k(1 + \rho_k) \right].$$

Следовательно,

$$\mathcal{E}V(\bar{y}_{st}) - \mathcal{E}V(\bar{y}_{sy}) = \frac{\sigma^2}{2k^2} \left[\sum_{i=1}^{k-1} i(\rho_i + \rho_{2k-i} - 2\rho_k) \right].$$

Если

$$\rho_{i+1} + \rho_{i-1} \geq 2\rho_i \quad (i = 2, 3, \dots),$$

то нетрудно показать, что каждый член внутри скобок положителен. Тем самым доказательство закончено. Коротко говоря, дело в том, что среднее расстояние между единицами равно k как для систематической, так и для расслоенной выборки, но из-за условия выпуклости вверх расслоенная выборка больше проигрывает в точности, когда расстояние между единицами меньше k , чем выигрывает, когда это расстояние больше k .

Кенуй (Queiroz, 1949) показал, что неравенства, содержащиеся в утверждении теоремы 8.6, остаются справедливыми, если сделать менее жесткими два условия из (8.10), а именно

$$\mathcal{E}(y_i) = \mu_i; \quad \mathcal{E}(y_i - \mu_i)^2 = \sigma_i^2.$$

В этом случае каждая из трех средних дисперсий для надсовокупности увеличивается в одинаковой степени.

Что касается практических применений, то выпуклые вверх кореллограммы были предложены несколькими авторами в качестве моделей для конкретных реальных совокупностей. Фишер и Маккензи (Fisher and Mackenzie, 1922) предложили функцию $\rho_u = \tanh(u^{-3/2})$ в качестве корреляционной функции для еженедельных уровней осадков на двух метеорологических станциях, находящихся на расстоянии u одна от другой; Осборн (Osborne, 1942) и Матерн (Matérn, 1947) — функцию $\rho_u = e^{-\lambda u}$ для исследования лесо- и землепользования; Уолд (Wold, 1938) — функцию $\rho_u = (1 - u)/l$ для некоторых видов экономических временных рядов.

8.9. РЕАЛЬНЫЕ СОВОКУПНОСТИ

Исследования были произведены для разнообразных реальных совокупностей. Некоторые из этих исследований указаны в табл. 8.6. Первые три исследования проводились с помощью географических карт. В первом из них совокупность состоит из 288 значений высот точек, находящихся на расстоянии 0,1 мили одна от другой в холмистой местности. В двух последующих данными служат доли длин отрезков прямых, проведенных на карте с раскраской, приходящиеся на области с определенным покрытием (под травой, лесом и т. д.). Эти примеры можно считать наиболее близкими к моделям с непрерывной в строгом смысле вариацией.

Следующие три исследования основаны на показаниях температуры в течение 192 последовательных дней в следующих точках: (а) 12 дюймов под поверхностью травы, (б) 4 дюйма под поверхностью земли, (в) в воздухе. Эти три исследования отражают три различные степени влияния (в направлении увеличения) на изучаемую характеристику неустойчивых ежедневных изменений погоды и медленных сезонных изменений.

В остальных обследованиях наблюдались растения или деревья, растущие в последовательных точках, расположенных вдоль некоторой линии. В обследовании картофеля, типичном для этой группы, конечная совокупность состоит из значений урожая на 96 грядках некоторого поля. Мы не проводили тщательного изучения литературы, но, по-видимому, можно найти и другие конкретные примеры.

В некоторых исследованиях V_{st} сравнивали с V_{st2} для расслоенной случайной выборки со слоями объема $2k$ и двумя единицами в каждом слое. Такое сравнение представляет интерес, поскольку по данным выборки можно получить несмещенную оценку V_{st2} . Для V_{st1} (со слоями объема k и одной единицей в каждом слое) или для V_{sy} ее получить нельзя. В других работах сообщается о сравнении V_{sy} как с V_{st1} , так и с V_{st2} . В большинстве источников непосредственное сравнение с V_{ran} в явном виде не производится, но в общем V_{st2} , по-видимому, дает выигрыш в точности по сравнению с V_{ran} .

В работах Йейтса и Финни сравнение производится относительно целого ряда значений n и k для каждой конечной совокупности.

Таблица 8.6
РЕАЛЬНЫЕ СОВОКУПНОСТИ, ИЗУЧЕННЫЕ ПРИ АНАЛИЗЕ
СИСТЕМАТИЧЕСКОГО ОТБОРА

Ссылка	N	Вид данных
Yates (1948), табл. 13	288	Значения высот в точках, отстоящих на расстоянии в 0,1 мили, полученные по карте английского государственного картографического управления
Osborne (1942)	*	Процент площади под (а) возделываемой землей, (б) кустарником, (в) травой, (г) лесом на параллельных прямых, проведенных на карте с раскраской
Osborne (1942)	*	Процент площади под елью Дугласа, подсчитанный с помощью параллельных прямых, проведенных на карте с раскраской
Yates (1948)	192	Температура почвы (12 дюймов под поверхностью травы) в течение 192 последовательных дней
Yates (1948)	192	Температура почвы (4 дюйма под поверхностью земли) в течение 192 дней
Yates (1948)	96	Температура воздуха в течение 192 дней
Finney (1948)	160	Урожай картофеля на 96 грядках
Finney (1948)	288	Объем леса, годного к продаже, в расчете на делянку шириной в 3 ряда и переменной длины (Mt. Stuart forest)
Finney (1950)	292	Объем подрастающего леса на делянку шириной в 2,5 ряда и длиной в 80 рядов (Black's Mountain forest)
Johnson (1943)	400**	Объем леса на делянку шириной в 2 ряда и переменной длины (Dehra Dun forest)
Johnson (1943)	400**	Число саженцев на 1 фут длины гряды для 4 гряд саженцев лиственных пород
Johnson (1943)	400**	Число саженцев на 1 фут длины гряды для 3 гряд саженцев хвойных пород
Johnson (1943)	400**	Число пересаженных деревьев хвойных пород на 1 фут длины гряды для 6 гряд

* Теоретически N бесконечно, если считать, что толщина прямых бесконечно мала.
** Приближенно. В действительности это число выналось от гряд к гряде.

Для этих случаев данные табл. 8.7 представляют собой геометрические средние из отношений дисперсий для отдельных значений k . Другие авторы производили сравнение только для одного значения k в каждой совокупности, но иногда приводили данные для разных признаков или для нескольких реальных совокупностей одного и того же характера. При этом опять брались геометрические средние из отношений дисперсий.

Хотя эти данные и ограничены по масштабам, результаты производят впечатление. В тех исследованиях, где возможно сравнение с V_{st1} , систематическая выборка неизменно дает, хотя и умеренный, но вполне

Таблица 8.7
ОТНОСИТЕЛЬНАЯ ТОЧНОСТЬ СИСТЕМАТИЧЕСКОГО И РАССЛОЕННОГО
СЛУЧАЙНОГО ОТБОРА

Данные	Размах значений k	Относительная точность систематического отбора по сравнению с расслоенным	
		V_{st1}/V_{sy}	V_{st2}/V_{sy}
Высоты	2—20	2,99	5,68
Процент площади (4 типа покрытия)	—	—	4,42
Процент площади под елью Дугласа	—	—	1,83
Температура почвы (12 дюймов)	2—24	2,42	4,23
Температура почвы (4 дюйма)	4—24	1,45	2,07
Температура воздуха	4—24	1,26	1,65
Картофель	3—16	1,37	1,90
Объем леса (Mt. Stuart)	2—32	1,07	1,35
Объем леса (Black's Mt.)	2—24	1,19	1,44
Объем леса (Dehra Dun)	2—32	1,39	1,89
Лиственные саженцы	14	—	1,89
Хвойные саженцы	14—24	—	2,22
Пересаженные хвойные деревья	12—22	—	0,93

ощутимый выигрыш в точности. Медианное значение отношений V_{st1}/V_{sy} равно 1,4. Выигрыш в точности по сравнению с V_{st2} существеннее, здесь медианное значение отношений равно 1,9.

Характер полученных результатов в общем соответствует ожидаемому, хотя ввиду небольшого числа обследований трудно было рассчитывать на получение определенных выводов. Выигрыш оказался наибольшим для тех видов данных, относительно которых можно было предположить, что их вариация наиболее близка к непрерывной. С этой точки зрения и при переходе от почвенных температур к температурам воздуха можно было ожидать, что отношение V_{st1}/V_{sy} уменьшится. Из последних трех признаков (данные о лесных питомниках) выигрыша в точности не оказалось лишь для одного — пересаженных хвойных деревьев, которые старше и более однородны, чем молодые саженцы.

8.10. ОЦЕНИВАНИЕ ДИСПЕРСИИ ПО ОТДЕЛЬНОЙ ВЫБОРКЕ

Согласно результатам, относящимся к простым случайным выборкам с $n > 1$, мы можем вычислить несмещенную оценку дисперсии выборочного среднего, причем оценка будет несмещенной независимо от вида совокупности. Однако для систематической выборки это полезное свойство не сохраняется, поскольку ее можно рассматривать лишь как простую случайную выборку с $n = 1$. Проиллюстрируем это на примере с изменением «по синусоиде». Пусть

$$y_i = m + a \sin \frac{\pi i}{2},$$

где $k = 4$ и $i = 1, 2, \dots, 4n$. Последовательные наблюдения в совокупности будут

$$(m + a), m, (m - a), m, (m + a), m, (m - a), m, \dots$$

Если в качестве первого члена выбрано значение $i = 1$, то все члены систематической выборки имеют значение $(m + a)$. При трех других возможных значениях первого члена все члены принимают значения соответственно m , $(m - a)$ или m . Таким образом, по отдельной выборке мы никак не можем оценить величину a . В то же время истинное значение дисперсии выборочного среднего систематической выборки равно $a^2/2$. Этот пример показывает, что при существовании периодической вариации несмещенную оценку дисперсии по выборке построить невозможно.

Из сказанного не следует, что вообще ничего нельзя сделать. За исключением случая периодической вариации, мы можем располагать достаточными сведениями о структуре совокупности, чтобы построить математическую модель, адекватно представляющую существующий в ней тип вариации. После этого мы могли бы вывести формулу для оценки дисперсии, которая для этой модели была бы приближенно несмещенной, хотя, возможно, для других моделей смещение было бы большим. Решать, какую из моделей следует применить, должен тот, кто организует обследование.

Далее представлены некоторые простые модели и соответствующие им оценки дисперсий. Доказательства не приводятся.

Наиболее простая модель относится к совокупности, в которой y_i содержит некоторый тренд плюс «случайное» слагаемое. Тогда

$$y_i = \mu_i + e_i,$$

где μ_i — некоторая функция i . Относительно случайного слагаемого мы предполагаем, что существует надсовокупность, для которой

$$E(e_i) = 0; E(e_i^2) = \sigma^2; E(e_i e_j) = 0 \quad (i \neq j).$$

Формула оценки дисперсии, s_{sp}^2 , называется несмещенной, если

$$E(s_{sp}^2) = \sigma^2.$$

т. е. если она не смещена относительно среднего по всем конечным совокупностям, которые могут быть получены из этой надсовокупности

Совокупность, единицы которой расположены в «случайном» порядке

$$\mu_i = \text{постоянной} \quad (i = 1, 2, \dots, N);$$

$$s_{sp1}^2 = \frac{N-n}{Nn} \frac{\sum (y_i - \bar{y}_{sp})^2}{n-1}.$$

Эта модель применяется, если мы уверены в том, что порядок расположения единиц имеет в основном случайный характер относительно наблюдаемого признака. Формула дисперсии та же, что и для простой случайной выборки, и ее оценка не имеет смещения, если наша модель верна.

Расслоенная совокупность, единицы которой в слоях расположены в случайном порядке

$$\mu_i = \text{постоянной} \quad (rk + 1 \leq i \leq rk + k);$$

$$s_{sp2}^2 = \frac{N-n}{Nn} \frac{\sum (y_i - \bar{y}_{sp})^2}{2(n-1)}.$$

В этом случае среднее значение постоянно внутри каждого слоя из k единиц. Оценка s_{sp2}^2 , основанная на среднем квадрате последовательных разностей, не будет несмещенной. В ее образовании принимают нежелательное участие разности значений μ соседних слоев n , кроме того, при оценивании случайного слагаемого дисперсии первый и последний слои имеют слишком малые веса. Если наша модель верна, то для достаточно больших выборок эта оценка будет, вообще говоря, преувеличивать дисперсию.

Линейный тренд

$$\mu_i = \mu + \beta i \quad (1 \leq i \leq n-2);$$

$$s_{sp3}^2 = \frac{N-n}{N} \frac{n'}{n^2} \frac{\sum (y_i - 2y_{i+k} + y_{i+2k})^2}{6(n-2)}.$$

Оценка основана на квадратах последовательных разностей, образуемых тремя соседними значениями y_i, y_{i+k}, y_{i+2k} в выборке. Сумма квадратов содержит $(n-2)$ членов. В случае линейного тренда, как мы видели (параграф 8.6), его можно исключить, вводя конечные поправки. Член n'/n^2 равен сумме квадратов весов в выражении \bar{y}_{wsp} . Если только n не мало, n'/n^2 можно заменить обычным множителем $1/n$. Из-за того, что крайним слоям придан слишком малый вес, оценка смещена, за исключением случая, когда σ^2 постоянна, но если n велико и наша модель верна, то оценка будет вполне удовлетворительной.

Если совокупность характеризуется непрерывной вариацией более сложного типа, указанные только что формулы могут давать неудовлетворительные результаты. В табл. 8.8 приведены данные о применении второй и третьей формул к наблюдениям на шести грядках лес-

Таблица 8.8
ДИСПЕРСИЯ ВЫБОРОЧНОГО СРЕДНЕГО ЧИСЛА САЖЕНЦЕВ

	Гряда	Фактическая V_{sp}	s_{sp2}^2	s_{sp3}^2
Клен серебристый	1	0,91	2,8	2,5
	2	0,74	3,6	2,9
Ильм американский	1	4,8	28,4	12,6
	2	15,5	22,6	18,6
Ель сизая	1	5,5	17,2	11,2
	2	2,0	11,6	6,4
Сосна горная	1	8,2	21,0	21,9

питомника (Johnson, 1943). Квадратичная формула оказалась несколько лучше формулы, основанной на последовательных разностях, но обе они сильно преувеличивают дисперсию.

Можно предложить много других формул. Если μ_i меняется непрерывно и не очень быстро, то может оказаться эффективным оценивание на основе «остатков» после приближения с помощью полиномов высоких степеней; для этого метода соответствующие таблицы были составлены Делюри (De Lury, 1950 г.).

Формулы, полученные при простых предположениях о характере коррелограммы, были рассмотрены Осборном (Osborne, 1942), Кокремом (Cochran, 1946) и Матерном (Matérn, 1947). Йейтс (Yates, 1960) исследовал оценку, основанную на величине вида

$$(y_u + y_{u+2h} + y_{u+4h} + \dots) - (y_{u+h} + y_{u+3h} + \dots).$$

Последовательным членам выборки приписаны попеременно знаки «+» и «-». Если это выражение берется по всей выборке, то имеется только одна степень свободы. Для того чтобы обеспечить большее число степеней свободы, данные выборки можно разбить на части, каждая из которых (по предложению Йейтса) должна насчитывать девять наблюдений. Если обозначить последовательные наблюдения в систематической выборке через y'_1, y'_2 и т. д. и придать первому и последнему члену веса, равные $1/2$, то можно записать:

$$d_1 = \left(\frac{1}{2} y'_1 + y'_3 + y'_5 + y'_7 + \frac{1}{2} y'_9 \right) - (y'_2 + y'_4 + y'_6 + y'_8).$$

Следующая разность, d_2 , должна начаться с y'_5 и т. д. Тогда в качестве оценки дисперсии \bar{y}_{sy} принимается

$$s_{sy}^2 = \frac{N-n}{Nn} \sum_{u=1}^g \frac{d_u^2}{7,5g}.$$

Множитель 7,5 — сумма квадратов коэффициентов для каждого d_u , а g — число разностей, образующих оценку (g приблизительно равно $n/9$). Для реальных совокупностей, исследованных Йейтсом, формула такого типа оказалась лучше, чем формула s_{sy}^2 , основанная на последовательных разностях, но и она преувеличивает фактическую дисперсию \bar{y}_{sy} .

Заметим в заключение, что в формулах для оценки дисперсии нет недостатка, но, по-видимому, все они имеют ограниченную область применения.

8.11. РАССЛОЕННЫЙ СИСТЕМАТИЧЕСКИЙ ОТБОР

Мы уже видели, что если единицы соответствующим образом упорядочены, то систематический отбор обеспечивает некоторого рода расслоение с равными долями отбора. Если расслоение произведено по некоторому другому критерию, то в каждом слое можно извлечь отдельную систематическую выборку, определяя точки отсчета независимо. Такой подход удобен, если мы хотим получить раздельные оценки для каждого слоя или если применяются неравные доли отбора. Этот метод

будет, конечно, более точным, чем расслоенный случайный отбор, если систематический отбор внутри слоев более точен, чем случайный отбор внутри слоев.

Если \bar{y}_{syh} — среднее значение для систематической выборки в слое h , то оценка среднего для совокупности \bar{Y} и ее дисперсия равны:

$$\bar{y}_{sy} = \sum W_h \bar{y}_{syh}; \quad V(\bar{y}_{sy}) = \sum W_h^2 V(\bar{y}_{syh}).$$

Если число слоев невелико, то задача нахождения оценки этой дисперсии по выборке сводится к уже рассмотренной задаче нахождения по выборке удовлетворительной оценки $V(\bar{y}_{syh})$ в каждом слое.

Если слоев более многочисленны, то может оказаться предпочтительнее оценка по методу «совмещенных слоев» (параграф 5A.11). Из сказанного в этом параграфе вытекает, что оценка

$$v(\bar{y}_{sy}) = \sum' W_h^2 (\bar{y}_{syh} - \bar{y}_{sy})^2,$$

где суммирование происходит по всем парам слоев, в среднем преувеличивает дисперсию, даже если вариация периодического характера существует внутри слоев.

Несмещенную оценку дисперсии ошибки можно получить, если в каждом слое извлекаются две систематические выборки с разными точками отсчета, выбранными случайным образом, и с интервалом отбора $2k$. При этом каждый слой обеспечивает одну степень свободы. Если систематический отбор эффективен, то такой прием приведет к некоторой потере в точности. Если слоев много, то в большинстве их можно взять только по одной систематической выборке, а по две выборки для оценивания по ним ошибки извлечь лишь в части слоев, отобрав эту часть случайным образом.

8.12. ДВУМЕРНЫЙ СИСТЕМАТИЧЕСКИЙ ОТБОР

При отборе из совокупности, представляющей собой некоторую территорию, простейшим обобщением одномерного систематического отбора на этот случай будет отбор по схеме квадратной решетки, изображенной на рис. 8.4 (а). Выборка полностью определяется парой случайных чисел, задающих координаты левой верхней единицы. Характеристики схемы квадратной решетки были исследованы на примерах как теоретических, так и реальных совокупностей. Матерн (Matérn, 1960) исследовал наилучший тип выборки для случая, когда корреляция наблюдений в любых двух точках выражается монотонно убывающей выпуклой вверх функцией расстояния между ними, d . Для коррелограмм вида $e^{-1,4d}$ отбор по квадратной решетке оказывается достаточно пригодным и превосходит простой или расслоенный случайный отбор с одной единицей в каждом слое, хотя Матерн и указывает причины, по которым можно ожидать, что наилучшей схемой для этой ситуации окажется отбор по треугольной решетке, образованной вершинами равносторонних треугольников.

В 14 сельскохозяйственных исследованиях на однородность Хейнс (Haynes, 1948) нашел, что отбор по квадратной решетке дает почти ту же точность, что и простой случайный отбор по двум измерениям. Милл

(Milne, 1959) изучал отбор по «центральной» схеме квадратной решетки, когда выборка определяется точкой, лежащей в центре квадрата, в 50 исследованиях на однородность. Такой способ отбора оказался лучше простого случайного отбора и, возможно, несколько лучше, чем расслоенный случайный отбор, хотя последнее преимущество не было статистически значимым. Эти результаты указывают на то, что, по крайней мере, для данных такого типа, автокорреляция выра-

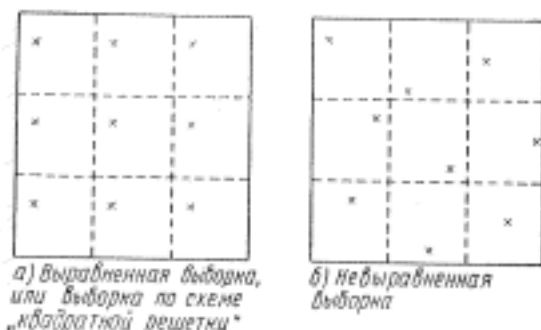


Рис. 8.4. Два типа двумерной систематической выборки

жена слабо. При оценивании по карте площади, занимаемой лесом или водой, Матерн в двух примерах обнаружил, что квадратная решетка превосходит случайные методы отбора.

На рис. 8.4 (б) показана систематическая выборка другого типа, называемая *невыравненной выборкой*. Сначала, извлекая пару случайных чисел, задают координаты левой верхней единицы. Еще два случайных числа определяют горизонтальные координаты оставшихся двух единиц в первом столбце слоев. Наконец, еще два случайных числа нужны, чтобы фиксировать вертикальные координаты оставшихся единиц в первой строке слоев. После этого постоянный интервал k (равный сторонам квадратов) задает расположение всех остальных точек. Исследования Кенуа (Kenouille, 1949) и Даса (Das, 1950) для простых двумерных коррелограмм указывают на то, что невыравненная схема часто дает лучшие результаты, чем квадратная решетка и чем расслоенный случайный отбор.

Еще одно свидетельство превосходства невыравненной выборки дает опыт планирования экспериментов, который обнаружил, что для размещения наблюдений в прямоугольной области вполне применима схема латинского квадрата. Можно считать, что латинский квадрат (5×5), показанный на рис. 8.5 (а), задает разбиение области на пять систематических выборок, каждая из которых соответствует определенной букве. Есть некоторые данные о том, что этот особый квадрат, называемый латинским квадратом «ходом коня», будет несколько более точным, чем случайно выбранный квадрат (5×5). Причина этого, вероятно, в том, что у первого никакая выборка не содержит двух элементов не только из одной строки или одного столбца, но и из каждой диагонали.

Принципом построения латинских квадратов воспользовались Хомейер и Блэк (Homeyer and Black, 1946) при отборе на прямоугольных полях овса. Каждое поле содержало 21 участок. Три возможные систематические выборки, обозначенные соответственно буквами А, В и С, показаны на рис. 8.5 (б). Такое размещение, когда на каждом поле выбирается случайным образом одна из букв, увеличило точность приблизительно на 25% по сравнению с расслоенным случайным отбором со строками в качестве слоев. Поскольку каждая буква встречается трижды в одном столбце и по два раза в других, такое размещение не совсем точно удовлетворяет определению латинского квадрата, но, насколько это возможно, соответствует ему.

А В С D E	А В С
D E A B C	B C A
B C D E A	C A B
E A B C D	A B C
C D E A B	B C A
	C A B
	A B C

(а) Латинский квадрат «ходом коня»

(б) Схема систематического отбора для прямоугольного поля 3×7

Рис. 8.5. Две схемы систематического отбора, основанные на латинских квадратах

Йейтс (Yates, 1960), который назвал размещения такого типа *отбором по решетке*, рассматривает их применение для дву- и трехмерного отбора. В случае трех измерений каждая строка, каждый столбец и каждая вертикаль могут быть представлены в выборке путем отбора p единиц из p^3 единиц совокупности. Если выборка содержит p^2 единиц, то в ней могут быть представлены каждое из p^2 сочетаний строк и столбцов или строк и вертикалей, или столбцов и вертикалей. Паттерсон (Patterson, 1954) исследовал размещения, которые дают несмещенную оценку ошибки.

8.13. РЕЗЮМЕ

Систематические выборки удобно намечать и извлекать. В большинстве исследований, упоминавшихся в этой главе как по искусственным, так и по реальным совокупностям, они выигрывали в точности по сравнению с расслоенными случайными выборками. Недостатки систематической выборки заключаются в том, что ее точность может оказаться невысокой, если существует неожиданная периодичность, и в том, что неизвестен надежный метод оценивания $V(\bar{y}_{sy})$ по данным выборки. В свете сказанного ранее систематический отбор может быть без опасения рекомендован в следующих ситуациях:

1. Когда единицы совокупности расположены в основном в случайном порядке или когда расслоение в совокупности намечено очень слабо. В этом случае систематический отбор применяется из-за его удобства, поскольку нельзя рассчитывать на выигрыш в точности. Имеются выборочные оценки ошибки, смещение которых находится в допустимых пределах (параграф 8.10).

2. Когда применяется расслоение с большим числом слоев и систематическая выборка извлекается независимо в каждом слое. В этом случае влияние скрытой периодичности имеет тенденцию нейтрализоваться и можно получить оценку ошибки, которая заведомо преувеличена (параграф 8.11). При другом способе можно воспользоваться лишь половиной числа слоев и извлечь из каждого слоя по две систематические выборки с независимым случайным началом отсчета каждая. Такой способ обеспечивает несмещенную оценку ошибки.

3. При подборе единиц (гл. 10). В этом случае оказывается, что в большинстве практических приложений можно получить несмещенную оценку ошибки выборки.

4. При выборочном изучении совокупностей с вариацией непрерывного характера при условии, что оценка ошибки выборки обычно не требуется. Если проводится ряд обследований такого типа, то может оказаться достаточным проверять ошибки выборки лишь от случая к случаю. Йейтс (Yates, 1948) указывает, что можно делать такую проверку с помощью дополнительных наблюдений.

Упражнения

8.1. В таблице на с. 251 указано число саженцев на каждом футе длины гряды, общей длиной в 200 футов.

Найдите дисперсию среднего систематической выборки, включающей каждый двадцатый фут гряды. Сравните ее с дисперсиями: (а) простой случайной выборки, (б) расслоенной случайной выборки с двумя единицами в каждом слое, (в) расслоенной случайной выборки с одной единицей в каждом слое. Для всех выборок $n = 10$. [$\sum (y_i - \bar{Y})^2 = 23\ 601$.]

8.2. Совокупность 360 домохозяйств Балтимора (перенумерованных от 1 до 360) размещена в картотеке в алфавитном порядке по фамилиям глав хозяйств. Домохозяйства, где глава небелый, имеют следующие номера: 28, 31—33, 36—41, 44, 45, 47, 55, 56, 58, 68, 69, 82, 83, 85, 86, 89—94, 98, 99, 101, 107—110, 114, 154, 156, 178, 223, 224, 296, 298—300, 302—304, 306—323, 325—331, 333, 335—339, 341, 342. (Среди небелых иногда встречаются «скопления» домохозяйств из-за связи между фамилией и цветом кожи.)

Сравните точность систематической выборки каждого восьмого домохозяйства с простой случайной выборкой того же объема при оценивании доли домохозяйств, в которых глава небелый.

8.3. Жители некоторого района принадлежат к трем компактно расселившимся общинам, состоящим из людей англо-саксонского, польского и итальянского происхождения. Имеется справочник с новейшими данными. В нем жители каждого дома записаны в следующем порядке: муж, жена, дети (по возрасту), прочие. Дома записаны по порядку вдоль улиц. Среднее число жителей в одном доме равно пяти. Можно взять либо систематическую выборку каждого пятого лица из справочника, либо 20%-ную простую случайную выборку. Для каких из следующих переменных систематическая выборка будет, по вашему мнению, более точной: (а) доля лиц польского происхождения, (б) доля мужчин, (в) доля детей? Объясните почему.

8.4. Имеется следующий список жителей 13 домов некоторой улицы. М — взрослый мужчина, Ж — взрослая женщина, м — мальчик, ж — девочка.

Семья												
1	2	3	4	5	6	7	8	9	10	11	12	13
М	М	М	М	М	М	М	М	М	М	М	М	М
Ж	Ж	Ж	Ж	Ж	Ж	Ж	Ж	Ж	Ж	Ж	Ж	Ж
ж	ж	ж	ж	ж	ж	ж	ж	ж	ж	ж	ж	ж
м	м	ж	м	м	ж	ж	ж	ж	ж	ж	ж	ж
ж	ж	ж	ж	ж	ж	ж	ж	ж	ж	ж	ж	ж

Число саженцев

Футы длины гряды										Итого систематический выборки	
1—20	21—40	41—60	61—80	81—100	101—120	121—140	141—160	161—180	181—200		
1	2	3	4	5	6	7	8	9	10		
8	20	26	34	31	24	18	16	36	10	223	
6	19	26	21	23	19	13	12	8	35	182	
6	25	10	27	41	28	7	8	29	7	188	
23	11	41	25	18	18	9	10	33	9	197	
25	31	30	32	15	29	11	12	14	12	211	
16	26	55	43	21	24	20	20	13	7	245	
28	29	34	33	8	33	16	17	18	6	222	
21	19	56	45	22	37	9	12	20	14	255	
22	17	39	23	11	32	14	7	13	12	190	
18	28	41	27	3	26	15	17	24	15	214	
26	16	27	37	4	36	20	21	29	18	234	
28	9	20	14	5	20	21	26	18	4	165	
11	22	25	14	11	43	15	16	16	4	177	
16	26	39	24	9	27	14	18	20	9	202	
7	17	24	18	25	20	13	11	6	8	149	
22	39	25	17	16	21	9	19	15	8	191	
44	21	18	14	13	18	25	27	4	9	193	
26	14	44	38	22	19	17	29	8	10	227	
31	40	55	36	18	24	7	31	8	5	255	
26	30	39	29	9	30	30	29	10	3	235	
Итого для слоев	410	459	674	551	325	528	303	358	342	205	4 155

Сравните дисперсии для систематической выборки каждого пятого человека и 20%-ной простой случайной выборки при оценивании: (а) доли лиц мужского пола, (б) доли детей, (в) доли людей в семьях специалистов (к ним относятся семьи в домах с номерами 1, 2, 3, 12 и 13). Подтверждают ли эти результаты ваши ответы к упражнению 8.3? В случае систематической выборки ведите отчет в каждом столбце сверху вниз и далее с верха следующего столбца.

8.5. В упражнении 8.1 можно оценить $V(\bar{y}_{sy})$: (а) рассматривая каждую систематическую выборку как простую случайную выборку, (б) считая, что каждая систематическая выборка, включающая каждый 20-й фут, образована двумя систематическими выборками, включающими каждый 40-й фут гряды и имеющими разные случайные начала отсчета. Сравните для каждого способа средние значения оценок дисперсий с фактической дисперсией \bar{y}_{sy} .

8.6. Покажите, что для совокупности, содержащей только линейный тренд (параграф 8.5), систематическая выборка менее точна, чем расслоенная случайная выборка со слоями объема $2k$ и двумя единицами в каждом слое, если $n > (4k + 2)/(k + 1)$.

8.7. Двумерная совокупность с линейным трендом может быть представлена следующим образом:

$$y_{ij} = i + j \quad (i, j = 1, 2, \dots, nk),$$

где y_{ij} — значение признака в i -й строке и j -м столбце. Совокупность содержит $N^2 = n^2 k^2$ единиц.

Систематическая выборка по квадратной решетке определяется выбранными случайным образом двумя независимыми начальными координатами i_0, j_0 , между 1 и k каждая. Выборка объема n^2 содержит все единицы, координаты которых имеют вид

$$i_0 + \gamma k; j_0 + \delta k,$$

где γ, δ — любые два целых числа между 0 и $(n-1)$ включительно.

Покажите, что среднее этой выборки имеет ту же точность, что и среднее простой случайной выборки объема n^2 .

8.8. Если такое же сравнение, как в упражнении 8.7, произвести для трехмерной совокупности с линейным трендом, то какого можно ожидать результата?

ЛИТЕРАТУРА

- Cochran W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Stat.*, 17, 164—177.
- Das A. C. (1950). Two-dimensional systematic sampling and the associated stratified and random sampling. *Sankhya*, 10, 95—108.
- DeLury D. B. (1950). *Values and integrals of the orthogonal polynomials up to $n=26$* . University of Toronto Press.
- Finney D. J. (1948). Random and systematic sampling in timber surveys. *Forestry*, 22, 1—36.
- Finney D. J. (1950). An example of periodic variation in forest sampling. *Forestry*, 23, 96—111.
- Fisher R. A. and Mackenzie W. A. (1922). The correlation of weekly rainfall. *Quart. Jour. Roy. Met. Soc.*, 48, 234—245.
- Haynes J. D. (1948). An empirical investigation of sampling methods for an area. M. S. thesis, University of North Carolina.
- Hornberger P. G. and Black C. A. (1946). Sampling replicated field experiments on oats for yield determinations. *Proc. Soil. Sci. Soc. America*, 11, 341—344.
- Johnson F. A. (1943). A statistical study of sampling methods for tree nursery inventories. *Jour. Forestry*, 41, 674—689.
- Madow L. H. (1946). Systematic sampling and its relation to other sampling designs. *Jour. Amer. Stat. Assoc.*, 41, 207—214.
- Madow W. G. and Madow L. H. (1944). On the theory of systematic sampling. *Ann. Math. Stat.*, 15, 1—24.
- Matérn B. (1947). Methods of estimating the accuracy of line and sample plot surveys. *Medd. fr. Statens Skogsforsknings Institut*, 36, 1—138.
- Matérn B. (1960). Spatial variation. *Medd. fr. Statens Skogsforsknings Institut*, 49, 5, 1—144.
- Milne A. (1959). The centric systematic area sample treated as a random sample. *Biometrics*, 15, 270—297.
- Osborne J. G. (1942). Sampling errors of systematic and random surveys of cover-type areas. *Jour. Amer. Stat. Assoc.*, 37, 256—264.
- Patterson H. D. (1954). The errors of lattice sampling. *Jour. Roy. Stat. Soc. B.*, 16, 140—149.
- Quenouille M. H. (1949). Problems in plane sampling. *Ann. Math. Stat.*, 20, 355—375.
- Wold H. (1938). *A study of the analysis of stationary time series*. Uppsala.
- Yates F. (1948). Systematic sampling. *Phil. Trans. Roy. Soc. London*, A241, 345—377.
- Yates F. (1960). *Sampling methods for censuses and surveys*. Charles Griffin and Co., London. Third edition. Есть русский перевод: Яйтс Ф. Выборочный метод в переписях и обследованиях. М., «Статистика», 1965.
- Buckland W. R. (1951). A review of the literature of systematic sampling. *Jour. Roy. Stat. Soc.*, B 13, 208—215.

ОДНОСТУПЕНЧАТЫЙ ГНЕЗДОВОЙ ОТБОР

9.1. ПОЧЕМУ НЕОБХОДИМ ГНЕЗДОВОЙ ОТБОР

В предыдущих главах мы уже сделали несколько замечаний относительно обследований, в которых единица отбора состоит из группы или гнезда более мелких единиц, которые мы называли элементами. Широкое применение гнездового отбора обусловлено двумя главными причинами. Хотя, казалось бы, в качестве единиц отбора лучше принять сами элементы, обнаруживается, что для многих обследований достоверного списка элементов совокупности не существует, а составление такого списка обошлось бы слишком дорого. Во многих странах для некоторых больших географических районов нет полных и неустаревших списков населения, домов или ферм. Однако по картам эти районы могут быть разделены на территориальные единицы, такие, как кварталы в городах и участки земли с легко идентифицируемыми границами в сельской местности. В США выбирают такие гнезда потому, что таким образом решается задача построения списка единиц отбора.

Даже если имеется список отдельных домов, экономические соображения могут диктовать выбор каких-либо более крупных единиц отбора. При данном объеме выборки малые единицы отбора обычно обеспечивают более точные результаты, чем большие единицы. Например, простая случайная выборка объемом в 600 домов распределена по городу более равномерно, чем 20 городских кварталов, содержащих в среднем по 30 домов каждый. Однако определение местонахождения 600 домов и разъезды между ними потребуют больших расходов, чем определение местонахождения 20 кварталов и посещение всех домов в этих кварталах. Если наряду с точностью учитываются и издержки, то более крупные единицы могут оказаться предпочтительнее.

Разумный выбор между двумя типами или размерами единиц может быть сделан и на основании знакомого нам правила, согласно которому выбирается единица, обеспечивающая наименьшую дисперсию при заданных издержках или наименьшие издержки при заданной дисперсии. Как это часто бывает на практике, решение может зависеть здесь от факторов, не поддающихся количественной оценке: какой-либо тип единиц может обладать такими специфическими удобствами или недостатками, которые трудно учесть при расчете издержек.

жёл. При выборочном исследовании урожая на корню, как показывает некоторый опыт, малые единицы могут давать смещенные оценки из-за того, что трудно установить точные границы этих единиц. Хомейер и Блэк (Hornseyer and Black, 1946) обнаружили, что при отборе участков 2×2 фута урожай овса преувеличивается приблизительно на 8% по сравнению с участками 3×3 фута. Возможно, это объясняется тем, что в сомнительных случаях обследователи были склонны включать в участок растения, расположенные на его границе. Сукхатм (Sukhatme, 1947) приводит аналогичные примеры для пшеницы и риса.

9.2. ОДНО ПРОСТОЕ ПРАВИЛО

При сравнении нескольких конкретных типов или размеров единиц полезен следующий результат.

Теорема 9.1. Теорема относится к простому случайному отбору в случае, когда пкс можно пренебречь. Оценивается суммарное значение для совокупности. Пусть для единиц u -го типа

M_u — относительный размер единицы;

S_u^2 — дисперсия суммарных значений для единиц;

C_u — относительные издержки на наблюдение одной единицы.

Тогда относительные издержки при определенной точности или относительная дисперсия при определенных издержках пропорциональны $C_u S_u^2 / M_u^2$.

Доказательство. Пусть V_u — дисперсия оценки суммарного значения для совокупности, получающаяся при применении единиц u -го типа. Тогда

$$V(\bar{Y}) = V_u = \frac{N_u^2 S_u^2}{n_u}.$$

Издержки на исследование этих единиц равны $C_u n_u$. Поскольку для разных единиц $N_u M_u$ — величина постоянная, то как относительные издержки при определенной дисперсии, так и относительная дисперсия при определенных издержках пропорциональны

$$C_u n_u V_u = C_u N_u^2 S_u^2 \propto \frac{C_u S_u^2}{M_u^2}.$$

Тем самым теорема доказана.

Следствие 1. Если мы назовем *относительной чистой точностью* единицы величину, обратно пропорциональную дисперсии, получаемой при неизменных издержках, то теорему 9.1 можно сформулировать так:

$$\text{относительная чистая точность} \propto \frac{M_u^2}{C_u S_u^2}. \quad (9.1)$$

Следствие 2. В дисперсионном анализе дисперсии для единиц разных размеров часто вычисляются на так называемом общем основа-

нии, обычно в расчете на единицу наименьшего размера. Для того чтобы привести дисперсии к общему основанию, дисперсию S_u^2 суммарных значений для единиц размера M_u нужно разделить на M_u . Пусть

$S_u'^2 = \frac{S_u^2}{M_u}$ есть дисперсия суммарных значений единиц (на общем основании);

$C_u' \propto \frac{C_u}{M_u}$ есть относительные издержки на получение данной массы (количества элементов) выборки.

Теперь теорему 9.1 и следствие 1 можно сформулировать следующим образом:

относительные издержки при равной точности $\propto \frac{C_u S_u^2}{M_u^2} \propto C_u' S_u'^2$;

$$\text{относительная чистая точность} \propto \frac{1}{C_u' S_u'^2}. \quad (9.2)$$

Это утверждение показывает, что если не учитывать различий в издержках на получение выборки (т. е. считать C_u' постоянной), то относительная чистая точность при применении u -й единицы пропорциональна $1/S_u'^2$.

Таким образом, для того чтобы сравнить различные единицы при одной и той же массе выборки, нужно знать дисперсии единиц, приведенные к общему основанию.

Пример. Простым примером служат данные Джонсона (Johnson, 1941), относящиеся к гряде саженцев горных сосен. Гряда состояла из шести рядов, каждый 434 фута длиной. Гряде можно разделить на единицы отбора многими способами. В табл. 9.1 приведены данные для четырех типов единиц. Поскольку гряда была обследована полностью, приведенные данные есть правильные значения для совокупности.

Таблица 9.1

ДАННЫЕ ПО ЧЕТЫРЕМ ТИПАМ ЕДИНИЦ ОТБОРА

Исходные данные	Тип единицы			
	1 фут ряда	2 фута ряда	1 фут гряды	2 фута гряды
M_u — относительный размер единицы	1	2	6	12
N_u — число единиц в совокупности	2 604	1 302	434	217
S_u^2 — дисперсия для совокупности на единицу	2,537	6,746	23,094	68,558
Число футов ряда, на которых можно подсчитать число саженцев в течение 15 мин	44	62	78	108

Единицами служили:

один фут длины отдельного ряда;
два фута длины отдельного ряда;
один фут длины всей гряды;
два фута длины всей гряды.

Для единиц первых двух типов предполагалось, что отбор будет расслоенным по рядам, так что S_n^2 соответствует дисперсии внутри рядов. Для единиц двух последних типов предполагался простой случайный отбор.

Поскольку основные затраты были связаны с определением местоположения единиц и подсчетом их числа, издержки измерялись в единицах времени (последняя строка табл. 9.1). Для более крупных единиц в течение 15 мин можно обследовать большую массу выборки, так как меньше времени затрачивается на переход от одной единицы к другой.

Оцениваемой величиной служит общее число саженцев на гряде. В обозначениях теоремы 9.1 в табл. 9.1 указаны значения M_n и S_n^2 . Относительные величины C_n , выраженные временем, нужным для подсчета саженцев на одной единице, имеют следующие значения:

	1 фут ряда	2 фута ряда	1 фут гряды	2 фута гряды
C_n (в четвертях часа)	$\frac{1}{44}$	$\frac{2}{62}$	$\frac{6}{78}$	$\frac{12}{108}$

Значения относительной чистой точности выведены в табл. 9.2 согласно следствию 1 из теоремы 9.1.

Таблица 9.2
ОТНОСИТЕЛЬНАЯ ЧИСТАЯ ТОЧНОСТЬ ЧЕТЫРЕХ ЕДИНИЦ

	1 фут ряда	2 фута ряда	1 фут гряды	2 фута гряды
$\frac{M_n^2}{C_n S_n^2}$	$\frac{44}{2,537} = 17,34$ 100	$\frac{4 \cdot 62}{2 \cdot 6,746} = 18,38$ 106	$\frac{36 \cdot 78}{6 \cdot 23,094} = 20,27$ 117	$\frac{144 \cdot 108}{12 \cdot 68,558} = 18,90$ 109

В последней строке табл. 9.2 приведены значения относительной точности при относительной точности наименьшей единицы, принятой за 100. Наилучшей единицей оказался 1 фут гряды.

Заслуживают также внимания значения дисперсий для единиц, приведенных к общему основанию. Значения $S_n^2 = S_n^2/M_n$ в расчете на один фут длины ряда равны соответственно 2,537; 3,373; 3,849; 5,713. Заметим, что с увеличением размера единицы эти дисперсии постепенно увеличиваются. Так обычно и происходит (хотя бывают и исключения). Поскольку относительная чистая точность пропорциональна $1/C_n S_n^2$, издержки на получение данной массы выборки при более крупных единицах должны быть меньше, если оказывается, что эти единицы более экономичны.

Теорема 9.1 и ее следствия остаются справедливыми для расслоенного отбора с пропорциональным размещением, если все слои одинакового размера и $S_n^2, S_n'^2$ представляют собой средние значения дисперсий внутри слоев. Действительно, при указанных условиях дисперсия оценки суммарного значения для совокупности, если пренебречь пкс, равна $N^2 S_n^2/n$, и, следовательно, имеет тот же вид, что и при простом случайном отборе. Для более сложных видов отбора теорема 9.1 силы не имеет.

Сказанное ранее имело цель проиллюстрировать следующее общее правило. Сравнение единиц должно всегда производиться для того способа отбора, который должен быть применен на практике, или, если это еще не решено, для тех способов отбора, которые рассматриваются в качестве возможных. Изменения в методе отбора или способе оценивания повлияют и на относительную чистую точность различных единиц. Даже при неизменных способах отбора и оценивания значения относительной чистой точности меняются в зависимости от объема выборки, если издержки не есть линейная функция объема выборки или если этот объем достаточно велик, так что пкс пренебречь нельзя.

Обычно изучается несколько признаков. Один из подходов в этом случае состоит в том, что задают общие издержки и находят значения относительной чистой точности для каждого типа единиц и для каждого признака. Если ни один из типов единиц не дает преимуществ для всех признаков, то принимается некоторое компромиссное решение, при котором предпочтение отдается наиболее важным признакам.

Ввиду того, что на окончательный выбор влияют многочисленные факторы, определение оптимального размера единиц для обширного обследования представляет собой сложную задачу. Интересный пример сравнения единиц для выборочного обследования ферм описан Джессеном (Jessen, 1942). Выдержки из его работы приведены в табл. 9.3. В ней сравниваются четыре размера единиц — четверть участка ($S/4$), половина участка ($S/2$), участок (S) и блок, состоящий из двух смежных участков ($2S$). Участок представляет собой территорию площадью в 1 квадратную милю и содержит в среднем несколько меньше четырех ферм. При этом сравнении общие издержки на собственно обследование (1000 долл.), продолжительность опроса одного фермера (в общей сложности не более 60 мин) и путевые расходы (5 центов на милю) были определены заранее, потому что относительная чистая точность меняется при изменении любой из этих переменных. Издержки приняты на уровне 1939 г.

В таблице приведены относительные стандартные ошибки (в процентах) оценок средних на одну ферму по 18 признакам. Ни одна из единиц не была наилучшей для всех признаков. Однако половина участка и четверть участка оказались выгоднее более крупных единиц для всех признаков кроме двух, причем между половиной участка и четвертью участка не оказалось в этом смысле большого различия. Половина участка, вероятно, все же предпочтительнее, потому что для этой единицы проще идентифицировать ее границы на местности.

Таблица 9.3

ОЦЕНКИ СТАНДАРТНЫХ ОШИБОК (В %) ДЛЯ ЕДИНИЦ ЧЕТЫРЕХ РАЗМЕРОВ
ПРИ ПРОСТОМ СЛУЧАЙНОМ ОТБОРЕ

Признаки	S/4	S/2	S	2S	Лучшая единица
Число свиней	5,0	4,9	5,3	6,2	S/2
Число лошадей	3,4	3,3	3,6	4,2	S/2
Число овец	17,4	15,7	14,9	14,3	2S
Число цыплят	3,0	3,0	3,3	3,8	S/4, S/2
Число яиц за предыдущий день	5,7	5,2	4,9	4,7	2S
Число голов рогатого скота	4,7	4,6	4,8	5,5	S/2
Число дойных коров	3,7	3,6	3,8	4,4	S/2
Число галлонов молока	4,4	4,2	4,4	4,9	S/2
Объем полученной молочной продукции	5,5	5,2	5,4	6,0	S/2
Площадь земли на ферме	2,9	2,8	3,0	3,5	S/2
Площадь под зерновыми	3,7	3,5	3,8	4,4	S/2
Площадь под овсом	4,6	4,8	5,6	7,0	S/4
Урожай зерновых	1,6	1,7	2,0	2,5	S/4
Урожай овса	1,6	1,5	1,6	1,8	S/2
Расходы на откорм скота	12,6	13,6	16,7	21,8	S/4
Общие расходы владельца	7,8	8,1	9,6	12,0	S/4
Общий доход владельца	6,2	6,5	7,7	9,8	S/4
Чистая прибыль	6,8	6,9	7,8	9,5	S/4

9.3. СРАВНЕНИЕ ТОЧНОСТИ ПО ДАННЫМ ВЫБОРОЧНОГО
ОБСЛЕДОВАНИЯ

В примере с саженцами дисперсии для различных типов единиц были получены по результатам сплошного обследования всей гряды. Однако, за исключением небольших совокупностей, редко оказывается возможным провести обследование только для того, чтобы сравнить разные единицы. Чаще всего сведения, необходимые для определения оптимальной единицы, получают каким-либо специальным приемом как дополнительный результат обследования, главная цель которого состоит в получении оценок тех или иных характеристик совокупности.

Предположим, что в некотором обследовании каждую единицу можно разделить на M меньших единиц. Вместо того чтобы измерить лишь суммарное значение для каждой «большой» единицы выборки, мы можем получить отдельно данные для каждой из M малых единиц. В этом случае можно произвести сравнение точности, какую дают большие и малые единицы. Предположим сперва, что извлекается простая случайная выборка объема n .

По данным выборки можно составить таблицу дисперсионного анализа 9.4.

Оценка дисперсии для большой единицы на основании малой равна s_b^2 . Можно полагать, что соответствующей оценкой дисперсии для

Таблица 9.4

ДИСПЕРСИОННЫЙ АНАЛИЗ ПО ДАННЫМ ВЫБОРКИ
(НА ОСНОВАНИИ МАЛОЙ ЕДИНИЦЫ)

	Число степеней свободы	Средний квадрат
Между большими единицами	$(n-1)$	s_b^2
Между малыми единицами внутри больших единиц	$n(M-1)$	s_w^2
Между малыми единицами в выборке	$nM-1$	$s^2 = \frac{(n-1)s_b^2 + n(M-1)s_w^2}{nM-1}$

[b —от английского «between» — между; w —от английского «within» — внутри.]

малой единицы будет средний квадрат отклонений для всех малых единиц в выборке, а именно

$$s^2 = \frac{(n-1)s_b^2 + n(M-1)s_w^2}{nM-1} \quad (9.3)$$

Эта оценка, зачастую вполне удовлетворительная, имеет небольшое смещение, так как рассматриваемая выборка не будет простой случайной выборкой малых единиц, поскольку они извлекались целыми группами, состоящими из M малых единиц.

Несмещенную оценку по выборке можно получить, проведя, в соответствии с табл. 9.5, дисперсионный анализ для всей совокупности, которая содержит N больших и NM малых единиц.

Таблица 9.5

ДИСПЕРСИОННЫЙ АНАЛИЗ ДЛЯ ВСЕЙ СОВОКУПНОСТИ
(НА ОСНОВАНИИ МАЛОЙ ЕДИНИЦЫ)

	Число степеней свободы	Средний квадрат
Между большими единицами	$N-1$	S_b^2
Между малыми единицами внутри больших единиц	$N(M-1)$	S_w^2
Между малыми единицами в совокупности	$NM-1$	$S^2 = \frac{(N-1)S_b^2 + N(M-1)S_w^2}{NM-1}$

По определению, дисперсия признака в совокупности между малыми единицами задается выражением в последней строке таблицы, а именно

$$S^2 = \frac{(N-1)S_b^2 + N(M-1)S_w^2}{NM-1}$$

При простом случайном отборе s_b^2 из табл. 9.4 есть несмещенная оценка S_b^2 (это следует из результатов параграфа 2.6). Нетрудно показать, что s_w^2 есть несмещенная оценка S_w^2 . Следовательно, несмещенная оценка дисперсии S^2 среди всех малых единиц в совокупности есть

$$\hat{S}^2 = \frac{(N-1)s_b^2 + N(M-1)s_w^2}{NM-1}. \quad (9.4)$$

Очевидно, что это выражение почти совпадает с более простым выражением

$$\hat{S}^2 \approx \frac{s_b^2 + (M-1)s_w^2}{M}. \quad (9.5)$$

При $n > 50$ (9.3) для s^2 также сводится к (9.5), так что s^2 служит удовлетворительным приближением к S^2 при $n > 50$.

Эти две оценки, s_b^2 (для большой единицы) и \hat{S}^2 (для малой единицы), приведены к общему основанию и их можно подставить в формулу (9.2) из следствия 2 теоремы 9.1.

Если выборка велика, то данные для малых единиц можно получить по случайной подвыборке больших единиц (скажем, 100 из 600). Иным способом можно наблюдать по две малые единицы, отобранные случайным образом из каждой большой единицы. Одновременно можно исследовать несколько размеров малых единиц при условии, что получаемые данные обеспечивают несмещенную оценку S_w^2 для каждой малой единицы.

В случае расслоенного отбора дисперсии для больших и малых единиц можно оценивать этими же методами отдельно в каждом слое и после этого подставлять оценки в соответствующую формулу дисперсии оценки для расслоенной выборки.

Пример. Данные получены по выборке ферм штата Северная Каролина, взятой в 1942 г. при обследовании занятости на фермах (Finkner, Morgan and Mongee, 1943). Способ извлечения выборки состоял в том, что на карту случайным образом наносили точки и единицами отбора служили три фермы, наиболее близкие к каждой такой точке. Вообще говоря, этот способ не рекомендуется, потому что большую вероятность попасть в выборку имеет крупная ферма по сравнению с небольшой и изолированная ферма — по сравнению с фермой, находящейся в тесном соседстве с другими. Здесь возникающего при этом смещения мы не учитываем.

По выборочным данным для отдельных ферм можно сравнить в качестве единиц отбора группу из трех ферм и отдельную ферму. Взятым для сравнения признаком служило число наемных работников. Выборка была расслоенной, причем слоем служила группа таунишпов, сходных по плотности сельского населения и по отношению посевной площади к общей площади ферм. Так как доля отбора составляла 1,9%, пкс можно пренебречь.

Дисперсия оценки суммарного значения для совокупности есть

$$V(\hat{Y}_{st}) = \sum_h \frac{N_h^2 s_h^2}{n_h}.$$

Правильным было бы вычислять $N_h^2 s_h^2 / n_h$ отдельно внутри каждого слоя для двух типов единиц, применяя дисперсионный анализ и выражение (9.5). Мы воспользуемся приближенным, но более простым приемом.

Большинство слоев содержало от 300 до 450 ферм и для того чтобы сделать отбор приблизительно пропорциональным, в каждом слое отбирались либо две, либо три единицы, состоящие из трех ферм. Предполагая, что размещение было пропорциональным, т. е. что $n_h/N_h = n/N$, можно записать

$$V(\hat{Y}_{st}) = \frac{N}{n} \sum_h N_h s_h^2 = \frac{N^2}{n} \bar{S}_h^2,$$

если считать, кроме того, что S_h^2 мало меняются от слоя к слою, так что вместо них можно подставить их среднее значение \bar{S}_h^2 .

Оценки \bar{S}_h^2 получены из дисперсионного анализа на основании отдельной фермы, представленного в табл. 9.6.

Таблица 9.6
ВЫБОРОЧНЫЙ АНАЛИЗ ДИСПЕРСИИ ЧИСЛА НАЕМНЫХ РАБОТНИКОВ
(НА ОСНОВАНИИ ОТДЕЛЬНОЙ ФЕРМЫ)

	Число степеней свободы	Средний квадрат
Между единицами внутри слоев	825	6,218
Между фермами внутри единиц	2 768	2,918
Между фермами внутри слоев	3 593	3,676

Для группы из трех ферм средний квадрат, $\bar{s}_{h3}^2 = 6,218$, служит оценкой S_h^2 , вычисленной на основании отдельной фермы. Для отдельной фермы, пользуясь (9.5), имеем

$$\hat{S}^2 = \frac{6,218 + 2 \cdot 2,918}{3} = 4,018.$$

По теореме 9.1, следствие 2, два значения, 6,218 — для группы из трех ферм и 4,018 — для отдельной фермы, указывают значения относительной точности при неизменном общем объеме выборки. Группа ферм как единица отбора обеспечивает точность, составляющую приблизительно две трети точности, получаемой при единице отбора — отдельной ферме. С учетом издержек более предпочтительной могла бы оказаться единица отбора, состоящая из трех ферм.

9.4. ДИСПЕРСИЯ, ВЫРАЖЕННАЯ ЧЕРЕЗ ВНУТРИГНЕЗДОВУЮ КОРРЕЛЯЦИЮ

Иногда дисперсию выражают через коэффициент корреляции ρ между элементами одного и того же гнезда. Такой подход уже применялся в случае систематического отбора (параграф 8.3).

Пусть y_{ij} — наблюдаемое значение признака у j -го элемента из i -й единицы и пусть y_i — суммарное значение для этой единицы. При гнездовом отборе нужно различать средние двух видов: среднее на единицу $\bar{Y} = \sum y_i / N$ и среднее на элемент $\bar{y} = \sum y_i / NM = \bar{Y} / M$. Дисперсия признака между элементами равна:

$$S^2 = \frac{\sum_{i,j} (y_{ij} - \bar{Y})^2}{NM - 1}.$$

Коэффициент внутригнездовой корреляции ρ был определен (параграф 8.3) равенством:

$$\rho = \frac{E(y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{E(y_{ij} - \bar{Y})^2} = \frac{2 \sum_i \sum_{j < k} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{(M-1)(NM-1)S^2}, \quad (9.6)$$

так как число членов (попарных произведений) в числителе средней части равенства равно $NM(M-1)/2$, а математическое ожидание в знаменателе равно $(NM-1)S^2/NM$.

Теорема 9.2. Из совокупности, содержащей N гнезд, извлекается простая случайная выборка объемом n гнезд, каждое из которых состоит из M элементов. Тогда выборочное среднее на элемент \bar{y} есть несмещенная оценка \bar{Y} с дисперсией

$$V(\bar{y}) = \frac{1-f}{n} \cdot \frac{NM-1}{M^2(N-1)} S^2 [1 + (M-1)\rho] \approx \frac{1-f}{nM} S^2 [1 + (M-1)\rho], \quad (9.7)$$

где ρ — коэффициент внутригнездовой корреляции.

Доказательство. Пусть y_i обозначает суммарное значение для i -го гнезда и $\bar{y} = \sum y_i / n$. По теоремам 2.1 и 2.2 \bar{y} есть несмещенная оценка \bar{Y} с дисперсией

$$V(\bar{y}) = \frac{(1-f)}{n} \frac{\sum (y_i - \bar{Y})^2}{N-1}.$$

Но $\bar{y} = M\bar{y}$ и $\bar{Y} = M\bar{y}$. Следовательно, \bar{y} есть несмещенная оценка \bar{Y} с дисперсией

$$V(\bar{y}) = \frac{1-f}{nM^2} \frac{\sum (y_i - \bar{Y})^2}{N-1}. \quad (9.8)$$

Но

$$(y_i - \bar{Y}) = (y_{i1} - \bar{Y}) + (y_{i2} - \bar{Y}) + \dots + (y_{iM} - \bar{Y}).$$

Возведем в квадрат и просуммируем по всем N гнездам.

$$\begin{aligned} \sum_i (y_i - \bar{Y})^2 &= \sum_i \sum_j (y_{ij} - \bar{Y})^2 + 2 \sum_i \sum_{j < k} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) = \\ &= (NM-1)S^2 + (M-1)(NM-1)\rho S^2 = \\ &= (NM-1)S^2 [1 + (M-1)\rho], \end{aligned} \quad (9.8a)$$

пользуясь определением ρ в (9.6). Подставим полученное выражение в формулу (9.8) для $V(\bar{y})$. Это дает

$$V(\bar{y}) = \frac{1-f}{n} \cdot \frac{NM-1}{M^2(N-1)} S^2 [1 + (M-1)\rho].$$

Теорема доказана.

Если извлечена простая случайная выборка объемом nM элементов, то формула для $V(\bar{y})$ будет той же, что и (9.7), за исключением члена в квадратных скобках. Множитель

$$1 + (M-1)\rho$$

показывает, во сколько раз меняется дисперсия, если в качестве единиц отбора принимаются гнезда, а не элементы. Если $\rho > 0$, то при данной массе выборки гнездо дает менее точные результаты. Если $\rho < 0$, как иногда бывает, то гнездо дает более точные результаты. Полученное утверждение представляет собой простое обобщение теоремы 8.2.

Величина ρ может быть представлена и в другом виде. Пусть S_y^2 обозначает дисперсию суммарных значений для гнезд в расчете на одну малую единицу. Тогда

$$\sum (y_i - \bar{Y})^2 = (N-1)MS_y^2.$$

Равенство (9.8a) может быть переписано в виде

$$(N-1)MS_y^2 = (NM-1)S^2 [1 + (M-1)\rho],$$

так что

$$\rho = \frac{(N-1)MS_y^2 - (NM-1)S^2}{(NM-1)(M-1)S^2} \approx \frac{S_y^2 - S^2}{(M-1)S^2}.$$

Содержательный анализ численных значений ρ для различных признаков и разных размеров гнезд приводят Хансен, Хервиц и Мэддоу (Hansen, Hurwitz and Madow, 1953), которые рассматривают ρ в качестве «меры однородности» гнезда.

9.5. ДИСПЕРСИЯ КАК ФУНКЦИЯ РАЗМЕРА ЕДИНИЦЫ ОТБОРА

Для некоторых видов обследований, например, при выборочном исследовании почв, урожайности, при сельскохозяйственных обследованиях, где применяются территориальные единицы отбора, вариация размера гнездовых единиц может иметь почти непрерывный

характер. При этом нахождение наилучшей единицы отбора состоит не в выборе из двух или трех испытывавшихся конкретных размеров, а в определении оптимальной величины M , рассматриваемой как непрерывная переменная. При таком подходе нужно иметь некоторый метод выражения дисперсии S_b^2 между гнездовыми единицами в совокупности как функции M . Посредством дисперсионного анализа S_b^2 может быть найдена, если мы знаем (а) дисперсию S^2 между всеми элементами совокупности и (б) дисперсию S_w^2 между элементами, принадлежащими одной и той же гнездовой единице. Наш метод и состоит в том, чтобы указать S_w^2 и S^2 и найти S_b^2 с помощью дисперсионного анализа.

По данным выборки мы получаем оценки S^2 и S_w^2 для единиц того размера, который был принят в действительности. Поскольку S^2 — дисперсия признака между элементами, размер гнездовой единицы на нее не влияет. Он, однако, влияет на S_w^2 . Можно ожидать, что при увеличении размера большой единицы эта дисперсия будет увеличиваться. Если большие единицы, которые нужно исследовать, мало отличаются своими размерами от действительно применяемых, то в качестве первого приближения можно считать S_w^2 постоянной, пользуясь ее оценкой, полученной по данным выборки. Как показывает исследование Маквея (McVay, 1947), такое приближение часто может быть удовлетворительным.

Для того чтобы лучше аппроксимировать S_b^2 , предпринимались попытки (Jessen, 1942; Mahalanobis, 1944; Hendricks, 1944) вывести общую формулу ее изменения при изменении размера единицы. По данным нескольких сельскохозяйственных обследований оказалось, что S_b^2 связано с M эмпирической формулой

$$S_b^2 = AM^g \quad (g > 0), \quad (9.9)$$

где A и g — постоянные, не зависящие от M . Согласно этой формуле S_b^2 постепенно увеличивается при увеличении M . Обычно g бывает невелико. Зависимость такого вида можно ожидать, если существуют факторы, оказывающие на рядом расположенные элементы приблизительно одинаковое влияние. Например, климат, характер почвы, ландшафт и близость к рынкам сбыта имеют тенденцию придавать соседним фермам сходные черты.

С теоретической точки зрения эта формула не безупречна, поскольку при увеличении M величина S_b^2 может возрастать неограниченно. Если предположить, а это кажется оправданным, что между далеко отстоящими элементами нет корреляции, то более подходящей формулой будет такая, при которой S_b^2 с увеличением M стремится к некоторой верхней границе. Впрочем нас устроит любая формула, если она дает хорошее согласие для исследуемой области изменения M .

Если формула (9.9) справедлива, то график зависимости $\log S_b^2$ от $\log M$ имеет вид прямой. Для того чтобы оценить постоянные $\log A$ и g , нужно знать значения S_b^2 , по крайней мере, для двух значений M , а чтобы проверить каким-либо образом, имеет ли зависимость линейный характер — по крайней мере, для трех значений M .

Согласно табл. 9.5 дисперсионного анализа (с. 259) находим

$$S_b^2 = \frac{(NM-1)S^2 - N(M-1)S_w^2}{N-1} = \frac{(NM-1)S^2 - N(M-1)AM^g}{N-1} \approx \quad (9.10)$$

$$\approx MS^2 - (M-1)AM^g. \quad (9.11)$$

Хендрикс (Hendricks, 1944) заметил, что всю совокупность можно рассматривать как отдельную большую единицу отбора, содержащую NM элементов. Если выполняется (9.9), то $S^2 = A(NM)^g$. Преимущество такого подхода заключается в том, что значения A и g можно теперь оценить по данным обследования, в котором применялось только одно значение M . Два уравнения, приводящие к этим оценкам, имеют вид

$$\log S_w^2 = \log A + g \log M;$$

$$\log S^2 = \log A + g \log (NM).$$

Формула для S_b^2 согласно (9.10) принимает вид

$$S_b^2 = \frac{AM^g [(NM-1)N^g - N(M-1)]}{N-1}.$$

Этот способ, однако, не дает возможности проверить правильность (9.9). Может случиться, что достаточно хорошее согласие с этой формулой имеется при малых значениях M и неудовлетворительное при таком большом его значении, как NM . В этом случае следует применять более общие формулы (9.10) и (9.11).

Формула (9.9) приводится скорее как методический пример, а не как общее правило. Читатель, который сталкивается с подобной проблемой, должен сам вывести формулу, наиболее подходящую к его материалу, и проверить ее. В некоторых случаях какой-либо простой функцией M может быть $\log S_b^2$.

9.6. ФУНКЦИЯ ИЗДЕРЖЕК

Для обширного обследования характер затрат на проведение собственно обследования играет значительную роль в определении оптимальной единицы. Для того чтобы проиллюстрировать роль фактора издержек, мы опишем одну функцию издержек, построенную Джессеном (Jessen, 1942) для сельскохозяйственных обследований, в которых большими единицами служили группы, образованные соседними фермами.

Различаются два слагаемых издержек на собственно обследование. Слагаемое $c_1 M^{\frac{1}{2}}$ включает издержки, меняющиеся в непосредственной зависимости от общего числа элементов (ферм). Таким образом, c_1 содержит затраты на опрос и затраты на переезд от фермы к ферме внутри гнезда.

Второе слагаемое, $c_2 \sqrt{N}$, измеряет путевые расходы при переезде между гнездами. Проверка по карте показала, что эти расходы для данной совокупности меняются приблизительно как квадратный ко-

рель из числа гнезд. Следовательно, суммарные расходы на собственное обследование равны:

$$C = c_1 M n + c_2 \sqrt{n}. \quad (9.12)$$

Если предположить простой случайный отбор и пренебречь пкс, то дисперсия среднего на элемент, \bar{y} , равна S_y^2/nM . Согласно (9.11), эта величина равна:

$$V(\bar{y}) = \frac{S^2 - (M-1) AM^{g-1}}{n}. \quad (9.13)$$

Для того чтобы определить оптимальный размер единицы, мы найдем M и вместе с тем n , минимизирующие V при неизменном C . Решение в общем виде довольно сложно, хотя численное нахождение решения не составляет особой трудности.

С помощью некоторых преобразований можно получить уравнение, решение которого дает оптимальное M . Сначала решим уравнение издержек (9.12) как квадратное уравнение относительно \sqrt{n} . Это дает

$$\frac{2c_1 M \sqrt{n}}{c_2} = \left(1 + \frac{4Cc_1 M}{c_2^2}\right)^{1/2} - 1. \quad (9.14)$$

Для нахождения оптимального M мы должны минимизировать выражение

$$C + \lambda V = c_1 M n + c_2 \sqrt{n} + \lambda V.$$

Дифференцируя по n и по M и замечая, что $\partial V/\partial n = -V/n$, приходим к уравнениям:

$$\text{по } n: c_1 M + \frac{1}{2} c_2 n^{-1/2} = -\frac{\lambda \partial V}{\partial n} = \frac{\lambda V}{n}; \quad (9.15)$$

$$\text{по } M: c_1 n = -\frac{\lambda \partial V}{\partial M}. \quad (9.16)$$

Разделим (9.16) на (9.15), чтобы исключить λ . Получаем

$$\frac{n}{V} \frac{\partial V}{\partial M} = -\frac{c_1 n}{c_1 M + \frac{1}{2} c_2 n^{-1/2}},$$

или

$$\frac{M}{V} \frac{\partial V}{\partial M} = -\frac{1}{1 + c_2/2c_1 M \sqrt{n}}. \quad (9.17)$$

Если теперь подставить \sqrt{n} из (9.14), то после некоторых упрощений получим

$$\frac{M}{V} \frac{\partial V}{\partial M} = \left(1 + \frac{4Cc_1 M}{c_2^2}\right)^{-1/2} - 1.$$

Выписывая полностью левую часть этого равенства и меняя знак в обеих частях равенства, приходим к уравнению

$$\frac{AM^{g-1} [gM - (g-1)]}{S^2 - (M-1) AM^{g-1}} = 1 - \left(1 + \frac{4Cc_1 M}{c_2^2}\right)^{-1/2}.$$

Из этого уравнения определяется оптимальное M . Левая часть равенства не содержит постоянных величин, характеризующих издержки, и зависит только от вида функции дисперсии. Можно показать, что обе части равенства — монотонно возрастающие функции M при $g > 0$, $M \geq 1$, т. е. для интересующих нас значений. Предположим, что было найдено решение для конкретных значений C , c_1 и c_2 , и мы хотим выяснить, как изменится это решение при увеличении c_1 . Левая часть равенства не зависит от c_1 , а правая по мере увеличения c_1 растет. Следовательно, оптимальное значение M уменьшится. Такой же эффект дает уменьшение c_2^* .

Величина c_1 возрастает, если увеличивается продолжительность опроса, в то время как c_2 убывает, если проезд становится дешевле или если фермы на данной территории располагаются теснее. Эти обстоятельства приводят к выводу, что оптимальный размер единицы становится меньше, если:

- увеличивается продолжительность опроса;
- дешевле становится проезд;
- более тесно располагаются элементы (фермы);
- увеличивается размер отпущенной суммы (C).

Этот вывод зависит от вида функции издержек и для иного, чем рассмотренный здесь, ее вида потребует нового анализа. Таким образом, оптимальная единица — это не некоторая неизменная характеристика совокупности, а характеристика, зависящая также от типа обследования и от уровня цен и заработной платы.

Превосходное изложение проблем, связанных с построением функции издержек для обследований, в которых применяется гнездовой отбор, содержится в книге Хансена, Хервица и Мэдоу (Hansen, Hurwitz and Madow, 1953).

9.7. ГНЕЗДОВОЙ ОТБОР ДЛЯ ОЦЕНИВАНИЯ ДОЛЕЙ

Те же методы применимы и к гнездовому отбору для оценивания долей. Предположим, что M элементов каждого гнезда можно разделить на два класса и что $p_i = a_i/M$ есть доля элементов из класса C в i -м гнезде. Извлекается простая случайная выборка объемом n гнезд и в качестве оценки P , доли для совокупности, принимается p , среднее значение наблюдений p_i в выборке.

* Это рассуждение не точно, так как из монотонности обеих частей равенства еще не следует, что решение (корень уравнения) сдвигается в нужную сторону. Для доказательства нужно провести дополнительное рассуждение, учитывающее, что уравнения (9.15) и (9.16) описывают точку минимума, а не просто точку обращения в нуль производной. — *Примеч. пер.*

Как было отмечено ранее (параграф 3.12), при нахождении $V(p)$ мы не можем считать, что p подчиняется биномиальному распределению, а должны применять к p_i формулу для непрерывных переменных. Это дает

$$V(p) = \frac{N-n}{Nn} \cdot \frac{\sum_{i=1}^N (p_i - P)^2}{N-1} \approx \frac{N-n}{N^2 n} \sum (p_i - P)^2.$$

Если же мы извлекаем простую случайную выборку объемом в nM элементов, то дисперсия p получается в соответствии с предположением о биномиальном распределении (теорема 3.2), т. е.

$$V_{\text{bin}}(p) = \frac{(NM-nM)}{NM-1} \frac{PQ}{nM} \approx \frac{N-n}{N} \frac{PQ}{nM},$$

если N велико. [*bin* — от английского «binomial» — биномиальный.] Следовательно, дробь

$$\frac{V(p)}{V_{\text{bin}}(p)} \approx \frac{M \sum (p_i - P)^2}{NPQ} \quad (N \text{ велико}) \quad (9.18)$$

показывает относительное изменение дисперсии, вызванное применением гнезд. Знание численных значений этого множителя полезно для отыскания предварительных оценок объема выборки в случае гнездового отбора. Сначала требуемый объем выборки оценивается по формуле, соответствующей биномиальному распределению, и затем умножается на эту дробь, чтобы получить объем выборки, необходимый при гнездовом отборе. Примеры приводятся в работе Корифилда (Cornfield, 1951).

Если размер гнезда M_i — переменная величина, то оценка $p = \sum a_i / \sum M_i$ представляет собой оценку по отношению. Ее приближенная дисперсия выражается формулой (параграф 3.12)

$$V(p) \approx \frac{N-n}{Nn \bar{M}^2} \frac{\sum_{i=1}^N M_i^2 (p_i - P)^2}{N-1},$$

где $\bar{M} = \sum M_i / N$ есть средний размер гнезда.

Если эту выборку сравнить с простой случайной выборкой объемом в $n\bar{M}$ элементов, то находим, как обобщение (9.18),

$$\frac{V(p)}{V_{\text{bin}}(p)} \approx \frac{\sum M_i^2 (p_i - P)^2}{N \bar{M} PQ}. \quad (9.19)$$

Как и в случае непрерывных переменных, зависимость между размером гнезда и межгнездовой дисперсией можно исследовать, либо представляя множители (9.18) и (9.19) как функции \bar{M} , либо устанавливая связь между внутригнездовой дисперсией и \bar{M} . Если мы придадим значение 1 любой единице, принадлежащей классу C , и значение

0 любой из остальных единиц, то основное тождество дисперсионного анализа при заданном M принимает вид

$$NMP(1-P) = M \sum (p_i - P)^2 + M \sum p_i (1 - p_i)$$

общая сумма квадратов = сумма квадратов между гнездами + сумма квадратов внутри гнезда.

На основании этого равенства можно вычислить средний квадрат отклонений внутри гнезд и представить его как функцию M . Маквей (McVay, 1947) описывает применение такого анализа для исследования оптимального размера гнезда.

9.8. ГНЕЗДОВЫЕ ЕДИНИЦЫ НЕОДИНАКОВОГО РАЗМЕРА

Существует несколько способов оценивания суммарных и средних значений для совокупности в случае, когда гнездовые единицы содержат разное число элементов. Эти способы будут рассмотрены в оставшейся части этой главы. Пусть M_i — число элементов в i -й единице. Отметим, как важный практический момент, что в некоторых обследованиях значения всех M_i для совокупности точно или почти точно известны заранее, например, если элементы — это работники фирмы, в которой хорошо налажен текущий учет, а гнездовые единицы — отделения фирмы. В других обследованиях M_i не известны, за исключением единиц, попавших в выборку, для которых мы узнаем значения M_i в ходе обследования. В любом случае исследователь должен быть уверен в том, что ему известны значения M_i , необходимые для вычисления рассматриваемой оценки.

Сначала рассмотрим оценивание суммарного значения Y величин Y_{ij} по простой случайной выборке объемом в n гнездовых единиц.

Несмещенная оценка

Пусть, как и ранее,

$$y_i = \sum_{j=1}^{M_i} y_{ij} = M_i \bar{y}_i$$

обозначает суммарное значение признака для i -й единицы. Согласно следствию из теоремы 2.1 несмещенная оценка Y есть

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^n y_i. \quad (9.20)$$

Согласно следствию 2 из теоремы 2.2 дисперсия этой оценки есть

$$V(\hat{Y}) = \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1}, \quad (9.21)$$

где $\bar{Y} = Y/N$ есть среднее на единицу для совокупности.

Часто оказывается, что оценка \hat{Y} имеет невысокую точность. Это происходит в тех случаях, когда \bar{y}_i (среднее на элемент) мало меняется от единицы к единице, а M_i меняются значительно. Тогда $y_i = M_i \bar{y}_i$ также значительно меняется от единицы к единице и дисперсия (9.21) велика.

Оценка по отношению с размером в знаменателе

$$M_0 = \sum_{i=1}^N M_i \text{ есть общее число элементов в совокупности.}$$

Если M_0 известно, то суммарное значение можно оценить иначе, с помощью оценки по отношению, в которой роль вспомогательной переменной x_i играет M_i .

$$\hat{Y}_R = M_0 \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} = M_0 \times (\text{выборочное среднее на элемент}).$$

В обозначениях, применявшихся для оценки по отношению, отношение для совокупности $R = Y/X = Y/M_0 = \bar{Y}$, т. е. среднему на элемент для совокупности. По теореме 6.1 в предположении, что число гнезд в выборке велико,

$$V(\hat{Y}_R) \approx \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^N (y_i - M_i \bar{Y})^2}{N-1} \approx \quad (9.22)$$

$$\approx \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^N M_i^2 (\bar{y}_i - \bar{Y})^2}{N-1}. \quad (9.23)$$

Как следует из (9.23), дисперсия \hat{Y}_R зависит от изменчивости средних на элемент и часто оказывается значительно меньше, чем $V(\hat{Y})$.

Соответствующие оценки среднего на элемент для совокупности есть

$$\hat{\bar{Y}} = \frac{\hat{Y}}{M_0} = \frac{N}{nM_0} \sum_{i=1}^n y_i; \quad \hat{\bar{Y}}_R = \frac{\hat{Y}_R}{M_0} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} \text{ есть выборочное среднее на элемент.}$$

Заметим, что несмещенная оценка $\hat{\bar{Y}}$ требует знания M_0 , в то время как, применяя оценку по отношению, нужно знать только M_i для единиц, попавших в выборку.

Среднее значение средних по единицам

Третья возможность заключается в том, чтобы применять невзвешенное среднее из средних по единицам, т. е.

$$\bar{y}' = \frac{1}{n} (\bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n). \quad (9.24)$$

Если M_i неодинаковы, эта оценка будет не только смещенной, но и несостоятельной. Смещение может быть, однако, небольшим, если \bar{y}_i не коррелированы с M_i , и эта оценка иногда полезна. Ее свойства исследовал Сукхатм (Sukhatme, 1954).

9.9. ОТБОР С ВЕРОЯТНОСТЯМИ, ПРОПОРЦИОНАЛЬНЫМИ РАЗМЕРУ ЕДИНИЦЫ

Если все M_i известны, то другой метод отбора, предложенный Хансеном и Хервицем (Hansen and Hurwitz, 1943), заключается в том, чтобы извлекать единицы с вероятностями, пропорциональными их размеру M_i . Этот метод применяется в основном в обследованиях, где отбор производится в две ступени (см. гл. 11), но его можно применять также и в данном случае. Проиллюстрируем способ отбора с вероятностями, пропорциональными размеру единицы, на следующем примере небольшой совокупности, состоящей из семи единиц:

Номер единицы	Размер единицы M_i	ΣM_i	Интервал накопленных значений размера
1	3	3	1—3
2	1	4	4
3	11	15	5—15
4	6	21	16—21
5	4	25	22—25
6	2	27	26—27
7	3	30	28—30

Подсчитываются накопленные суммы M_i и выписываются соответствующие им интервалы накопленных значений размера. Для того чтобы извлечь единицу, берем случайное число между 1 и 30: предположим, что им оказалось число 19. Оно попадает в интервал накопленной суммы, соответствующий 4-й единице и содержащий значения этой суммы от 16 до 21 включительно. При таком способе отбора вероятность того, что какая-либо единица отбирается, пропорциональна размеру этой единицы.

Если нужно отобрать еще одну единицу, то описанный процесс повторяется, причем берется новое случайное число между 1 и 30. Однако в отличие от рассмотренных ранее приемов извлечение 4-й единицы вторично не запрещается. При n , превосходящем 1, отбор с возвращением необходим, чтобы сохранить вероятность извлечения пропорциональными размеру единицы. В этом можно убедиться, рассмотрев крайний случай, когда $n = 7$. Если производить отбор без возвращения, то все единицы окажутся отобранными автоматически, даже если мы откажемся от методики отбора с вероятностями, пропорциональными размеру. При значении n между 1 и 7 отбор без возвращения приведет к вероятностям, промежуточным между равными вероятностями и вероятностями, пропорциональными размеру. Преимущество отбора с возвращением состоит в том, что формулы для истинных дисперсий оценок и выборочных оценок этих дисперсий принимают простой вид. Вообще говоря, отбор с возвращением менее то-

чен, чем отбор без возвращения. Однако, когда n/N мало, вероятность того, что одна и та же единица появится в выборке дважды, мала, и отбор с возвращением почти эквивалентен отбору без возвращения. В последние годы было сделано много, чтобы разработать приемлемые методы отбора без возвращения и с неравными вероятностями для случаев, когда N малы, как, например, при расслоенном отборе (см. параграф 9.14).

9.10. ТЕОРИЯ ДЛЯ ОТБОРА С ПРОИЗВОЛЬНЫМИ ВЕРОЯТНОСТЯМИ

Мы покажем, что если i -я единица извлекается с вероятностью $z_i = M_i/M_0$ и с возвращением, то несмещенная оценка суммарного значения для совокупности, Y , есть

$$\hat{Y}_{pps} = \frac{M_0}{n} (\bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n) = M_0 \times (\text{среднее из средних значений на элемент по единицам}), \quad (9.25)$$

где $M_0 = \sum M_i$ — общему числу элементов в совокупности [*pps* — от английского «probability proportional to size» — вероятность, пропорциональная размеру]. Далее,

$$V(\hat{Y}_{pps}) = \frac{M_0}{n} \sum_{i=1}^N M_i (\bar{y}_i - \bar{Y})^2, \quad (9.26)$$

так что дисперсия оценки \hat{Y}_{pps} так же, как и оценки \hat{Y}_R , зависит от изменчивости средних на элемент по единицам.

В некоторых приложениях размер единиц известен только приближенно. В других «размером» служит не число элементов в единице, а просто некоторая другая характеристика ее величины, которая считается тесно коррелированной с суммарным значением для единицы y_i . Например, «размер» больницы можно характеризовать общим числом коек или же средним числом занятых коек. Аналогично можно придумать разные характеристики «размера» ресторана, банка или фермы. Следовательно, мы будем рассматривать отбор с вероятностями, пропорциональными некоторой оценке или характеристике размера M'_i . Мы покажем, что если $z_i = M'_i/M'_0$, где $M'_0 = \sum M'_i$, то

$$\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i} \quad (9.27)$$

служит несмещенной оценкой Y с дисперсией

$$V(\hat{Y}_{pps}) = \frac{1}{n} \sum_{i=1}^N z_i \left(\frac{y_i}{z_i} - Y \right)^2 \quad (9.28)$$

[*pps* — от английского «probability proportional to an estimate of size» — вероятность, пропорциональная оценке размера]. Формулы (9.27) и (9.28) представляют собой обобщения формул (9.25) и (9.26).

В доказательстве мы воспользуемся приемом, уже примененным в параграфе 2.8. Пусть t_i — случайная величина, определенная для

каждой выборки объема n и показывающая, сколько раз i -я единица появляется в этой выборке, t_i может принимать любое из значений $0, 1, 2, \dots, n$. Рассмотрим совместное распределение частот t_i для всех N единиц в совокупности.

Метод извлечения выборки эквивалентен известной вероятностной задаче, в которой n шаров распределяются по N урнам, причем вероятность того, что некоторый шар попадет в i -ю урну, равна z_i и не зависит от того, где находятся остальные шары. Таким образом, совместное распределение t_i задается полиномиальным выражением

$$\frac{n!}{t_1! t_2! \dots t_N!} z_1^{t_1} z_2^{t_2} \dots z_N^{t_N}.$$

Хорошо известны следующие свойства этого распределения:

$$\begin{aligned} E(t_i) &= nz_i; \quad V(t_i) = nz_i(1 - z_i); \\ \text{Cov}(t_i, t_j) &= -nz_i z_j. \end{aligned} \quad (9.29)$$

Теорема 9.3. Если извлекается выборка объемом в n единиц с вероятностями z_i и с возвращением, то

$$\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i} \quad (9.27)$$

есть несмещенная оценка Y с дисперсией

$$V(\hat{Y}_{pps}) = \frac{1}{n} \sum_{i=1}^N z_i \left(\frac{y_i}{z_i} - Y \right)^2. \quad (9.28)$$

Доказательство. Мы можем записать

$$\hat{Y}_{pps} = \frac{1}{n} \left(t_1 \frac{y_1}{z_1} + t_2 \frac{y_2}{z_2} + \dots + t_N \frac{y_N}{z_N} \right) = \frac{1}{n} \sum_{i=1}^N t_i \frac{y_i}{z_i},$$

где суммирование распространяется на все единицы совокупности. При многократном отборе все t_i выступают как случайные переменные, а y_i и z_i — как некоторый набор неизменных чисел. Следовательно, так как согласно (9.29) $E(t_i) = nz_i$,

$$E(\hat{Y}_{pps}) = \frac{1}{n} \sum_{i=1}^N (nz_i) \frac{y_i}{z_i} = \sum_{i=1}^N y_i = Y.$$

Далее,

$$\begin{aligned} V(\hat{Y}_{pps}) &= \frac{1}{n^2} \left[\sum_{i=1}^N \left(\frac{y_i}{z_i} \right)^2 V(t_i) + 2 \sum_{i < j} \frac{y_i}{z_i} \frac{y_j}{z_j} \text{Cov}(t_i, t_j) \right] = \\ &= \frac{1}{n} \left[\sum_{i=1}^N \left(\frac{y_i}{z_i} \right)^2 z_i (1 - z_i) - 2 \sum_{i < j} \frac{y_i}{z_i} \frac{y_j}{z_j} z_i z_j \right] = \\ &= \frac{1}{n} \left(\sum_{i=1}^N \frac{y_i^2}{z_i} - Y^2 \right) = \frac{1}{n} \sum_{i=1}^N z_i \left(\frac{y_i}{z_i} - Y \right)^2, \end{aligned}$$

поскольку $\sum z_i = 1$.

Полагая в теореме 9.3 $z_i = M_i/M_0$, получаем соответствующий результат для отбора с вероятностью, пропорциональной размеру.

Теорема 9.4. Если выборка объемом в n единиц извлекается с вероятностями $z_i = M_i/M_0$ и с возвращением, то

$$\hat{Y}_{prs} = \frac{M_0}{n} (\bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n) \quad (9.25)$$

есть несмещенная оценка Y с дисперсией

$$V(\hat{Y}_{prs}) = \frac{M_0}{n} \sum_{i=1}^N M_i (\bar{y}_i - \bar{Y})^2. \quad (9.26)$$

Доказательство. Полагая в теореме 9.3 $z_i = M_i/M_0$, мы получаем

$$\hat{Y}_{prs} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i} = \frac{M_0}{n} \sum_{i=1}^n \frac{y_i}{M_i} = \frac{M_0}{n} \sum_{i=1}^n \bar{y}_i = \hat{Y}_{prs};$$

$$\begin{aligned} V(\hat{Y}_{prs}) &= \frac{1}{n} \sum_{i=1}^n z_i \left(\frac{y_i}{z_i} - Y \right)^2 = \frac{1}{n} \sum_{i=1}^n \frac{M_i}{M_0} \left(\frac{M_0 y_i}{M_i} - Y \right)^2 = \\ &= \frac{M_0}{n} \sum_{i=1}^N M_i (\bar{y}_i - \bar{Y})^2, \end{aligned}$$

поскольку $\bar{Y} = Y/M_0$.

Следствие. Несмещенная оценка \bar{Y} есть

$$\bar{\bar{Y}}_{prs} = \frac{1}{n} (\bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n)$$

с дисперсией

$$V(\bar{\bar{Y}}_{prs}) = \frac{1}{n M_0} \sum_{i=1}^N M_i (\bar{y}_i - \bar{Y})^2. \quad (9.30)$$

Следующие две теоремы показывают, как оценивать дисперсию по выборке.

Теорема 9.5. В условиях теоремы 9.3 несмещенная оценка $V(\hat{Y}_{prs})$ есть

$$v(\hat{Y}_{prs}) = \sum_{i=1}^n \frac{[(y_i/z_i) - \hat{Y}_{prs}]^2}{n(n-1)}. \quad (9.31)$$

Доказательство. С помощью обычного алгебраического тождества можно записать

$$\sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{prs} \right)^2 = \sum_{i=1}^n \left(\frac{y_i}{z_i} - Y \right)^2 - n(\hat{Y}_{prs} - Y)^2.$$

Следовательно,

$$E \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{prs} \right)^2 = E \sum_{i=1}^n \left(\frac{y_i}{z_i} - Y \right)^2 - nV(\hat{Y}_{prs}),$$

поскольку, по определению $V(\hat{Y}_{prs})$, среднее значение второго члена в правой части равно $-nV(\hat{Y}_{prs})$. Вводя переменные t_i , имеем

$$\begin{aligned} E \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{prs} \right)^2 &= E \sum_{i=1}^n t_i \left(\frac{y_i}{z_i} - Y \right)^2 - \\ &- nV(\hat{Y}_{prs}) = n \sum_{i=1}^N z_i \left(\frac{y_i}{z_i} - Y \right)^2 - nV(\hat{Y}_{prs}) \end{aligned}$$

или, на основании (9.28) из теоремы 9.3,

$$n(n-1)E[v(\hat{Y}_{prs})] = n^2 V(\hat{Y}_{prs}) - nV(\hat{Y}_{prs}) = n(n-1)V(\hat{Y}_{prs}).$$

Теорема доказана.

Теорема 9.6. Если единицы извлекаются с вероятностями $z_i = M_i/M_0$ и с возвращением, то

$$v(\hat{Y}_{prs}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 \quad (9.32)$$

есть несмещенная оценка $V(\hat{Y}_{prs})$, где \bar{y} — невзвешенное среднее \bar{y}_i .

Для того чтобы получить этот результат, достаточно подставить $z_i = M_i/M_0$ в (9.31). Поскольку $\hat{Y}_{prs} = M_0 \bar{y}$, оценка дисперсии (не считая множителя) представляет собой знакомую нам сумму квадратов отклонений \bar{y}_i от их среднего.

9.11. ОПТИМАЛЬНАЯ ХАРАКТЕРИСТИКА РАЗМЕРА ЕДИНИЦЫ

В тех случаях, когда характеристикой размера M_i служит некоторая оценка величины единицы, теоретический интерес представляет вопрос: какая характеристика размера минимизирует дисперсию оценки \hat{Y}_{prs} ? Имеем

$$V(\hat{Y}_{prs}) = \frac{1}{n} \sum_{i=1}^N z_i \left(\frac{y_i}{z_i} - Y \right)^2 = \frac{1}{n} \left(\sum_{i=1}^N \frac{y_i^2}{z_i} - Y^2 \right).$$

Это выражение обращается в нуль, если z_i пропорциональны y_i , т. е. $z_i = y_i/Y$. Если все y_i положительны, то соответствующий им набор z_i представляет собой некоторый подходящий набор вероятностей для единиц быть отобранными. Следовательно, наилучшей характеристикой размера единицы будут числа, пропорциональные суммарным значениям признака y_i для единиц.

Этот результат не имеет практического значения, потому что если бы y_i были известны заранее, то выборка оказалась бы ненужной.

Однако из него следует, что если y_i сравнительно устойчивы во времени, то наилучшими значениями характеристики размера единицы относительно этого признака могли бы служить последние по времени из прошлых значений y_i , которыми мы располагаем. На практике при извлечении выборки для всех признаков должна применяться, конечно, какая-либо одна характеристика размера единиц. Если мы можем выбирать из нескольких характеристик размера, то, вероятно, наилучшей будет та из них, значения которой ближе всего к пропорциональности суммарным значениям признака для единиц по основным признакам.

9.12. СРАВНИТЕЛЬНАЯ ТОЧНОСТЬ СПОСОБОВ ОТБОРА И ОЦЕНИВАНИЯ

При отборе гнездовых единиц неодинакового размера мы можем выбирать, по крайней мере, из четырех методов отбора и оценивания (в предположении, что M_i известны, если метод этого требует):

1. Отбор с равными вероятностями. Оценка \hat{Y} или $\hat{\bar{Y}}$ (несмещенная).
2. Отбор с равными вероятностями. Оценка \hat{Y}_R или $\hat{\bar{Y}}_R$ (по отношению).
3. Отбор с вероятностями, пропорциональными размеру. Оценка \hat{Y}_{pps} или $\hat{\bar{Y}}_{pps}$.
4. Расслоение единиц по размеру. Отбор с равными вероятностями внутри слоев. Оценка обычная для расслоенного отбора.

Сначала мы сравним первые три метода. Простого общего правила для того, чтобы решить, какой из них наиболее точен, не существует. Решение зависит от характера связи между \bar{y}_i и M_i (если такая связь есть) и от вида дисперсии \bar{y}_i как функции M_i . Наиболее благоприятным для оценки по отношению и оценки с вероятностью, пропорциональной размеру, будет случай, когда среднее на элемент не связано с размером гнезда (\bar{y}_i некоррелировано с M_i). Для того чтобы учесть также совокупности, для которых \bar{y}_i могут увеличиваться или уменьшаться при увеличении M_i , мы исследуем модель вида

$$\bar{y}_i = \frac{\alpha}{M_i} + \beta + e_i, \quad (9.33)$$

где $E(e_i | M_i) = 0$.

Нужно сделать также некоторые предположения относительно дисперсии e_i для гнезд данного размера. Как уже говорилось в параграфе 9.4, между значениями признака у элементов одного гнезда часто существует положительная, обычно небольшая, корреляция ρ , которая уменьшается при увеличении M_i . Если записать $\rho = \rho_0 + \rho_1/M_i$, то согласно (9.7) из параграфа 9.4

$$V(e_i) = V(\bar{y}_i) = \frac{S^2}{M_i} [1 + (M_i - 1)\rho] = S^2 \left(\rho_0 + \frac{1 - \rho_0 + \rho_1}{M_i} - \frac{\rho_1}{M_i^2} \right).$$

В качестве упрощения примем, что $V(e_i) = v/M_i^g$, где $g > 0$. Поскольку ρ при изменении M_i меняется, по-видимому, сравнительно медленно, g , вероятно, заключено между 0 и 1. Однако есть переменные, для которых суммарные значения для единиц не связаны с M_i и для них может быть приемлемым значение $g = 2$.

Анализ, который будет проведен с помощью этой модели, основан на работах Йейтса (Yates, 1960), Дес Раджа (Des Raj, 1954, 1958), Жарковича (Žarković, 1960), и Кокрена (Cochran, 1953), пользовавшихся подобными моделями.

Мы ограничимся случаем тех обследований, для которых можно пренебречь пкс и n достаточно велико для того, чтобы была справедливой приближенная формула дисперсии для оценки по отношению.

Из (9.33) вытекает, что

$$\begin{aligned} y_i &= \alpha + \beta M_i + e_i M_i; \\ \bar{Y} &= \alpha + \beta \bar{M} \quad (\bar{M} = \sum M_i / N); \\ \bar{\bar{Y}} &= \frac{\alpha}{\bar{M}} + \beta. \end{aligned}$$

Для того чтобы оценить среднее на элемент для совокупности, воспользуемся (9.21), откуда, после деления на $M_i^2 = N^2 \bar{M}^2$, имеем

$$nV(\hat{\bar{Y}}) = \frac{E(y_i - \bar{Y})^2}{\bar{M}^2} = \frac{E[\beta(M_i - \bar{M}) + e_i M_i]^2}{\bar{M}^2} = \beta^2 c^2 + \frac{vE(M_i^{2-g})}{\bar{M}^2},$$

где $c^2 = E(M_i - \bar{M})^2 / \bar{M}^2$ есть квадрат коэффициента вариации размера M_i . Для оценки по отношению согласно (9.23)

$$\begin{aligned} nV(\hat{\bar{Y}}_R) &= \frac{EM_i^2 (\bar{y}_i - \bar{\bar{Y}})^2}{\bar{M}^2} = \frac{EM_i^2 [\alpha(1/M_i - 1/\bar{M}) + e_i]^2}{\bar{M}^2} = \\ &= \frac{\alpha^2 c^2}{\bar{M}^2} + \frac{vE(M_i^{2-g})}{\bar{M}^2}. \end{aligned}$$

Для оценки при отборе с вероятностями, пропорциональными размеру, согласно (9.26)

$$\begin{aligned} nV(\hat{\bar{Y}}_{pps}) &= \frac{EM_i (\bar{y}_i - \bar{\bar{Y}})^2}{\bar{M}} = \frac{EM_i [\alpha(1/M_i - 1/\bar{M}) + e_i]^2}{\bar{M}} = \\ &= \frac{\alpha^2}{\bar{M}} E\left(\frac{1}{M_i} - \frac{1}{\bar{M}}\right) + \frac{vE(M_i^{1-g})}{\bar{M}} \approx \frac{\alpha^2 c^2}{\bar{M}^2} + \frac{vE(M_i^{1-g})}{\bar{M}}. \end{aligned}$$

Результаты приведены в табл. 9.7 отдельно для $g = 0, 1, 2$. Член, содержащий v , стоит первым.

Рассмотрим сперва случай $\alpha = 0$, т. е. когда \bar{y}_i не зависит от размера гнезда. Вторые члены у nV_R и nV_{pps} исчезают. Очевидно, что

$$(1) \quad V_R < V_{pps} \text{ при } g = 0, 1, 2$$

(\bar{y}_i — от английского «unbiased» — несмещенный). Если β (оно в этом случае равно $\bar{\bar{Y}}$) велико, то превосходство оценки по отношению может

Таблица 9.7
СРАВНИМЫЕ ЗНАЧЕНИЯ nV_{un} , nV_R и nV_{prs}

g	$V(\bar{y}_i)$	Равные вероятности. Несмещенная оценка	Равные вероятности. Оценка по отношению	Отбор с вероятностями, пропорциональными размеру
0	v	$v(1+c^2)+c^2\beta^2$	$v(1+c^2)+\frac{c^2\alpha^2}{M^2}$	$v+\frac{\alpha^2}{M}\left[E\left(\frac{1}{M_i}\right)-\frac{1}{M}\right]$
1	$\frac{v}{M_i}$	$\frac{v}{M}+c^2\beta^2$	$\frac{v}{M}+\frac{c^2\alpha^2}{M^2}$	$\frac{v}{M}+\frac{\alpha^2}{M}\left[E\left(\frac{1}{M_i}\right)-\frac{1}{M}\right]$
2	$\frac{v}{M_i^2}$	$\frac{v}{M^2}+c^2\beta^2$	$\frac{v}{M^2}+\frac{c^2\alpha^2}{M^2}$	$\frac{v}{M}E\left(\frac{1}{M_i}\right)+\frac{\alpha^2}{M}\times$ $\times\left[E\left(\frac{1}{M_i}\right)-\frac{1}{M}\right]$

стать весьма большим. Отметим, между прочим, что случай $g=1$, $\alpha=0$ соответствует положению, когда элементы объединены в гнезда случайным образом, т. е. когда гнездовая единица столь же результативна, как и отдельный элемент. Кроме того, в этом случае оценка по отношению оказывается наилучшей несмещенной линейной оценкой. Далее, если $\alpha=0$, то

$$(2) \quad V_{prs} < V_{un} \text{ при } g=0, 1.$$

При $g=2$ сравниваемые дисперсии имеют вид

$$V_{un} = \frac{v}{M^2} + c^2\beta^2; \quad V_{prs} = \frac{v}{M} E\left(\frac{1}{M_i}\right) \approx \frac{v(1+c^2)}{M^2}.$$

Следовательно, отбор с вероятностями, пропорциональными размеру, предпочтительнее, если $v/M^2 > \beta^2$.

Когда α не равно нулю, т. е. когда \bar{y}_i увеличивается или уменьшается при увеличении M_i , эффект оценки по отношению и оценки при отборе с вероятностями, пропорциональными размеру, по сравнению с простым распространением зависит от соотношения величин α и β . При $\beta=0$, т. е. когда суммарное значение для единицы y_i не коррелировано с M_i , несмещенная оценка всегда превосходит оценку по отношению и оценку при отборе с вероятностями, пропорциональными размеру, за исключением, возможно, случая, когда $g=0$.

Что касается сравнения V_R и V_{prs} , то коэффициенты при α^2 в обоих выражениях приблизительно равны. Следовательно, имеем приблизительно:

$$V_R > V_{prs}, \text{ если } g=0;$$

$$V_R = V_{prs}, \text{ если } g=1;$$

$$V_R < V_{prs}, \text{ если } g=2.$$

При расслоении по размеру единицы реалистичное сравнение дисперсий произвести трудно, поскольку результат зависит от того, насколько велика вариация значений M_i , сохраняющаяся внутри образованных слоев. Для случая, когда M_i внутри слоев одинаковы (это наиболее благоприятный для расслоенного отбора случай), в табл. 9.8 приведены значения nV_{st} для пропорционального (V_{prop}) и нейманова оптимального размещения, сравнимые с соответствующими значениями в табл. 9.7.

Таблица 9.8
СРАВНИМЫЕ ЗНАЧЕНИЯ nV_{st}

g	$V(\bar{y}_i)$	Пропорциональное размещение	Оптимальное размещение
0	v	$v(1+c^2)$	v
1	$\frac{v}{M_i}$	$\frac{v}{M}$	$\frac{v(E\sqrt{M_i})^2}{M^2} \approx \frac{v}{M} \left(1 - \frac{c^2}{8}\right)^2$
2	$\frac{v}{M_i^2}$	$\frac{v}{M^2}$	$\frac{v}{M^2}$

Получились следующие результаты. (Заметим, что для расслоенных выборок значение α не влияет на дисперсию.) При $\alpha=0$ $V_{prop} = V_R$ для всех трех значений g . Далее, V_{prop} превосходит дисперсию оценки при отборе с вероятностями, пропорциональными размеру, при $g=2$, одинакова с ней при $g=1$ и уступает ей при $g=0$. Если α значительно отличается от нуля, то расслоенный отбор превосходит как отбор с оценкой по отношению, так и отбор с вероятностями, пропорциональными размеру. Если можно произвести оптимальное размещение, то расслоенный отбор никогда не уступает и почти всегда превосходит другие методы (в предположении, что M_i одинаковы внутри слоев).

Окончательные выводы таковы. Если \bar{y}_i не обнаруживает тренда или обнаруживает незначительный тренд при увеличении M_i , то способ отношения и отбор с вероятностями, пропорциональными размеру, более точны, чем несмещенное оценивание при отборе с равными вероятностями, причем они могут быть значительно более точными. Несмещенная оценка оказывается наилучшей, если суммарное значение для единицы y_i некоррелировано с M_i . При этом точность оценки по отношению и оценки при отборе с вероятностями, пропорциональными размеру, приблизительно одинакова. Поскольку ожидается, что g в большинстве случаев заключено между 0 и 1, по-видимому, в целом более точна оценка при отборе с вероятностями, пропорциональными размеру. С другой стороны, оценку по отношению легче вычислять и она обходится дешевле, если затраты на получение данных для больших единиц выше, чем для малых, поскольку при отборе с вероятностями, пропорциональными размеру, больших единиц будет в выборке относительно больше. Расслоение дает значительные преимущества,

если можно образовать слои, внутри которых M_i меняются мало, и особенно если осуществимо оптимальное размещение. Одно из преимуществ способа отношения и отбора с вероятностями, пропорциональными размеру, состоит в том, что они дают возможность применить расслоение для каких-либо других целей.

9.13. ОБОБЩЕНИЕ НА СЛУЧАЙ РАССЛОЕННОГО ОТБОРА

Отбор внутри слоев с вероятностью, пропорциональной некоторой характеристике размера, полезен, по-видимому, в том случае, когда само расслоение проведено по некоторой переменной, иной, чем размер. Если выборки внутри каждого слоя малы и не очень велик общий объем выборки, то, как мы уже видели (параграф 6.10), имеющиеся для оценок по отношению формулы не очень надежны, и отдельная оценка по отношению может обладать смещением, которым нельзя пренебречь.

При отборе с вероятностями, пропорциональными оценке размера (p_{res}), оценка суммарного значения для совокупности есть сумма оценок по отдельным слоям:

$$\hat{Y}_{pres} = \sum_h \hat{Y}_h = \sum_h \frac{1}{n_h} \sum_i^{n_h} \frac{y_{hi}}{z_{hi}}.$$

Из теорем 9.3 и 9.5 вытекает:

$$V(\hat{Y}_{pres}) = \sum_h \frac{1}{n_h} \sum_i^{n_h} z_{hi} \left(\frac{y_{hi}}{z_{hi}} - Y_h \right)^2;$$

$$v(\hat{Y}_{pres}) = \sum_h \frac{1}{n_h(n_h-1)} \sum_i^{n_h} \left(\frac{y_{hi}}{z_{hi}} - \hat{Y}_h \right)^2.$$

9.14. ОТБОР С НЕРАВНЫМИ ВЕРОЯТНОСТЯМИ БЕЗ ВОЗВРАЩЕНИЯ

Много интересных исследований было проведено в отношении методов отбора единиц с неравными вероятностями, но без возвращения. На практике эти проблемы связаны в основном с многоступенчатым расслоенным отбором (гл. 11), когда на первой ступени отбираются большие гнездовые единицы. Расслоение больших единиц может быть доведено до такой степени, что слои будут содержать лишь по несколько единиц и поэтому долей отбора n_h/N_h на первой ступени нельзя не учитывать. Однако большая часть методов была разработана сначала для одноступенчатого отбора, для которого все выкладки проще.

Предположим, что из слоя должно быть извлечено две единицы. Первая единица отбирается с вероятностью, пропорциональной размеру. Пусть при первом извлечении была получена i -я единица и ее относительный размер равен z_i , где $\sum z_i = 1$. При втором извлечении из оставшихся единиц отбирается одна с вероятностью, пропорциональной ее относительному размеру, т. е. j -я единица с вероятностью

$z_j/(1-z_i)$. Следовательно, вероятность того, что i -я единица будет отобрана либо при первом, либо при втором извлечении, равна:

$$\pi_i = z_i + \sum_{j \neq i} \frac{z_j z_i}{1-z_j} = \quad (9.34)$$

$$= z_i + \sum_{j=1}^N \frac{z_i z_j}{1-z_j} - \frac{z_i^2}{1-z_i} = z_i \left(1 + A - \frac{z_i}{1-z_i} \right), \quad (9.35)$$

где $A = \sum z_j/(1-z_j)$, взятой по всем N единицам.

Математическое ожидание числа извлекаемых единиц равно $\sum \pi_i$. Поскольку обязательно извлекаются две единицы, мы должны иметь $\sum \pi_i = 2$, что легко проверить и алгебраически. Таким образом, относительная вероятность (скажем, z'_i) того, что i -я единица появится в выборке равна $\pi_i/2 = z'_i$. При таком способе извлечения z'_i различаются между собой всегда меньше, чем z_i . В примере, приводимом Йейтсом и Гранди (Yates and Grundy, 1953) при $N = 4$, $z_i = 0.1; 0.2; 0.3$ и 0.4 , z'_i оказались равными $0.1173; 0.2206; 0.3042$ и 0.3579 . Искажение вероятностей не очень значительно, если иметь в виду, что извлекается половина единиц.

Предположим теперь, что с помощью этого или какого-либо другого метода извлекается выборка объемом в n единиц без возвращения. Пусть

π_i — вероятность того, что i -я единица содержится в выборке;
 π_{ij} — вероятность того, что как i -я, так и j -я единицы содержатся в выборке.

Справедливы следующие равенства:

$$\sum_i \pi_i = n; \quad \sum_{j \neq i} \pi_{ij} = (n-1) \pi_i; \quad \sum_i \sum_{j > i} \pi_{ij} = \frac{1}{2} n(n-1). \quad (9.36)$$

Для того чтобы проверить второе равенство, обозначим через $P(s)$ вероятность получения выборки, состоящей из n определенных единиц [s — от английского «sample» — выборка]. Тогда $\pi_{ij} = \sum P(s)$ по всем выборкам, содержащим i -ю и j -ю единицы, и $\pi_i = \sum P(s)$ по всем выборкам, содержащим i -ю единицу. При подсчете $\sum \pi_{ij}$ по всем $j \neq i$ каждая $P(s)$ для выборки, содержащей i -ю единицу, будет участвовать в этой сумме $(n-1)$ раз, поскольку всего имеется $(n-1)$ отличных от i значений j . Тем самым второе равенство доказано. Третье равенство следует из первых двух.

Покажем теперь, как получить несмещенную оценку Y , суммарного значения для слоя, ее дисперсию и оценку этой дисперсии. Если положить $z'_i = \pi_i/n$, то эта оценка имеет вид

$$\hat{Y}_U = \frac{1}{n} \sum_i \frac{y_i}{z'_i}, \quad (9.37)$$

где y_i — значение признака для i -й единицы. Пусть $t_i (i = 1, 2, \dots, N)$ будет случайной переменной, принимающей значение 1, если i -я

единица попадает в выборку, и нуль в противном случае. Тогда t_i подчиняется биномиальному распределению при одном испытании с вероятностью успеха π_i . Таким образом,

$$E(t_i) = \pi_i = n z_i'; \quad V(t_i) = \pi_i(1 - \pi_i).$$

Нам понадобится также значение $\text{Cov}(t_i, t_j)$. Поскольку $t_i t_j = 1$, только если обе единицы попали в выборку, то

$$\text{Cov}(t_i, t_j) = E(t_i t_j) - E(t_i)E(t_j) = \pi_{ij} - \pi_i \pi_j.$$

Следовательно, считая y_i неизменными числами, а t_i случайными переменными, имеем

$$\begin{aligned} E(\hat{Y}_U) &= E \frac{1}{n} \left(\sum_{i=1}^N \frac{t_i y_i}{z_i'} \right) = \sum_{i=1}^N y_i = Y; \\ V(\hat{Y}_U) &= \frac{1}{n^2} \left[\sum_{i=1}^N \left(\frac{y_i}{z_i'} \right)^2 V(t_i) + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{y_i}{z_i'} \frac{y_j}{z_j'} \text{Cov}(t_i, t_j) \right] = \\ &= \frac{1}{n^2} \left[\sum_{i=1}^N \left(\frac{y_i}{z_i'} \right)^2 \pi_i(1 - \pi_i) + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{y_i}{z_i'} \frac{y_j}{z_j'} (\pi_{ij} - \pi_i \pi_j) \right]. \quad (9.38) \end{aligned}$$

Эти результаты были получены Хорвицем и Томпсоном (Horvitz and Thompson, 1952). Другое выражение для дисперсии можно получить, пользуясь первыми двумя равенствами из (9.36).

Эти равенства дают

$$\sum_{i \neq j} (\pi_{ij} - \pi_i \pi_j) = (n-1) \pi_i - \pi_i(n - \pi_i) = -\pi_i(1 - \pi_i).$$

Отсюда, заменяя $\pi_i(1 - \pi_i)$ в (9.38), получаем

$$V(\hat{Y}_U) = \frac{1}{n^2} \sum_{i=1}^N \sum_{j>i}^N \left\{ (\pi_i \pi_j - \pi_{ij}) \left[\left(\frac{y_i}{z_i'} \right)^2 + \left(\frac{y_j}{z_j'} \right)^2 - 2 \frac{y_i}{z_i'} \frac{y_j}{z_j'} \right] \right\},$$

что можно представить как

$$V(\hat{Y}_U) = \frac{1}{n^2} \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{z_i'} - \frac{y_j}{z_j'} \right)^2. \quad (9.39)$$

Отсюда следует, что несмещенная оценка этой дисперсии по выборке есть

$$v(\hat{Y}_U) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{z_i'} - \frac{y_j}{z_j'} \right)^2 \quad (9.40)$$

при условии, что π_{ij} не обращаются в нуль ни для какой пары единиц. Эта оценка была получена Йейтсом и Гранди (Yates and Grundy, 1953).

Полученные соотношения подкрепляют теоретическое обоснование отбора без возвращения. При их практическом применении возникают некоторые трудности. По мере возрастания n при любом методе отбора становится все труднее сохранить z_i' близкими к исходным

z_i . Вычисление π_i и π_{ij} усложняется. Оценка дисперсии (9.40) становится все более неустойчивой величиной из-за того, что величины $(\pi_i \pi_j - \pi_{ij})/\pi_{ij}$ меняются в довольно широких пределах и иногда могут принимать отрицательные значения. Некоторые остроумные приемы, предназначенные для того чтобы преодолеть эти трудности, описываются в следующем параграфе.

9.15. ДРУГИЕ ПОДХОДЫ

Нарейн (Narain, 1951) нашел такие исходные вероятности извлечения, при которых окончательные вероятности становятся пропорциональными размеру единицы. Рассмотрим, например, случай, когда $n = 2$. Если мы хотим, чтобы $\pi_i = 2z_i$, то, как следует из формулы (9.35) параграфа 9.14, исходные вероятности Z_i должны удовлетворять уравнениям

$$2z_i = Z_i \left(1 + \sum_{j=1}^N \frac{Z_j}{1 - Z_j} - \frac{Z_i}{1 - Z_i} \right).$$

Методы решения этих уравнений были разработаны Нарейном и Йейтсом и Гранди. При $n > 2$ вычисления довольно громоздки, а при достаточно больших n метод, как и все другие подходы, становится непригодным.

Мерти (Murthy, 1957), следуя работе Дес Раджа (Des Raj, 1956a), пользуется в качестве оценки взвешенной суммой

$$\hat{Y}_M = \frac{\sum_{i=1}^n P(s|i) y_i}{P(s)},$$

где $P(s|i)$ — условная вероятность получения набора единиц, составляющих извлеченную выборку при условии, что *первой* была извлечена i -я единица;

$P(s)$ — безусловная вероятность получения набора единиц, составляющих извлеченную выборку.

Этот метод применим к любой схеме отбора, при которой вероятность извлечения последующих единиц выборки не зависит от порядка, в котором были извлечены предыдущие единицы, хотя, конечно, она может зависеть от размера отдельных единиц. При этих условиях оценка будет несмещенной и Мерти получил общие выражения для ее дисперсии и оценки дисперсии. Эта схема обладает тем преимуществом, что при $n = 2$ оценка дисперсии всегда положительна. В этом случае оценка принимает вид

$$\hat{Y}_M = \frac{1}{2 - z_i - z_j} \left[(1 - z_j) \frac{y_i}{z_i} + (1 - z_i) \frac{y_j}{z_j} \right],$$

а оценка ее дисперсии

$$v(\hat{Y}_M) = \frac{(1 - z_i)(1 - z_j)(1 - z_i - z_j)}{(2 - z_i - z_j)^2} \left(\frac{y_i}{z_i} - \frac{y_j}{z_j} \right)^2.$$

В методе Дес Раджа (Des Raj, 1956 b) предполагается, что нам известны значения некоторой вспомогательной переменной x_i (ею может служить и размер единицы), такой, что зависимость между y_i и x_i линейна. С помощью методов линейного программирования Дес Радж нашел, для $n = 2$, такие значения π_{ij} , что π_i пропорциональны x_i и что $V(\hat{Y}_D)$, задаваемая (9.38), минимальна.

Два остальных метода основаны на уже рассмотренных приемах. Первый из них, разработанный Хартли и Рао (Hartley and Rao, 1962), состоит в том, чтобы расположить единицы в случайном порядке, подсчитать накопленные суммы их размера и затем извлечь по этому накопленному размеру систематическую выборку «каждого k -го». Если нужно получить выборку объемом в n единиц, то мы полагаем $k = M_0/n$, находим случайное число r между 1 и k и отбираем единицы, содержащиеся в соответствующих интервалах накопленного размера числа $r, r+k, r+2k$ и т. д. Если размер какой-либо единицы больше, чем M_0/n , то она имеет шанс быть включенной в выборку дважды, но в остальном при такой схеме отбора вероятность извлечения единицы пропорциональна ее исходному размеру. Среднее y_i/z_i есть несмещенная оценка Y .

Хартли и Рао получили выражения для дисперсии и для оценки дисперсии при такой схеме отбора, представленные в виде разложения по отрицательным степеням N . Результативность этого метода, по-видимому, близка к результативности метода Нарейна, при котором вероятности также остаются пропорциональными исходному размеру единицы. Метод, предусматривающий систематический отбор, позволяет избежать вычисления новых исходных вероятностей извлечения.

Наконец, мы можем разделить совокупность на n групп и из каждой группы извлекать по одной единице с вероятностями, пропорциональными относительному размеру для этой группы, как описано в параграфе 9.9. Если i -я единица оказалась в первой группе, то вероятность ее извлечения равна z_i/Z_1 , где $Z_1 = \sum z_i$, взятой по единицам первой группы. Следовательно, чтобы сохранить свойство отбора с вероятностями, пропорциональными размеру, группы должны быть образованы таким образом, чтобы, насколько это возможно, $Z_1 = Z_2 = \dots = Z_n$ и т. д. Несмещенной оценкой Y служит

$$\hat{Y}_G = \sum_{j=1}^n Z_j \frac{y_j}{z_j},$$

где y_j, z_j — значение признака и размер единицы, извлеченной из группы j [G — от английского «group» — группа]. Несмещенной оценки дисперсии не найдено, но преувеличенную оценку можно получить методом совмещенных слоев (параграф 5A.11).

Один из вариантов описанной только что схемы состоит в том, чтобы объединять единицы в группы случайным образом, стремясь сделать число единиц в группах, по возможности, одинаковым. Если $N = Qn + k$, где Q — некоторое целое число и $k < n$, то мы образуем k групп, содержащих по $(Q+1)$ единиц. Остальные $(n-k)$ групп содержат по Q единиц. Оценка \hat{Y}_G будет несмещенной. По-

скольку при этой схеме отбора Z_j , вообще говоря, не равны, то вероятности уже не обязательно пропорциональны z_i . К достоинствам этой схемы, помимо простоты ее реализации, относится и то, что найдены точные выражения для дисперсий. Рао, Хартли и Кокрен (Rao, Hartley and Cochran, 1962) показали, что

$$V(\hat{Y}_G) = \frac{1}{n} \left(1 - \frac{n-1}{N-1} + \frac{k(n-k)}{N(N-1)} \right) \sum_{i=1}^n z_i \left(\frac{y_i}{z_i} - Y \right)^2 = \left(1 - \frac{n-1}{N-1} + \frac{k(n-k)}{N(N-1)} \right) V(\hat{Y}_{pps}),$$

где \hat{Y}_{pps} — оценка (параграф 9.10) для отбора с вероятностями, пропорциональными z_i , и с возвращением. Первый член в скобках играет роль своего рода пкс. Выражение, дающее несмещенную оценку дисперсии, имеет вид

$$v(\hat{Y}_G) = \frac{N^2 + k(n-k) - Nn}{N^2(n-1) - k(n-k)} \cdot \left[\sum_{j=1}^n z_j \left(\frac{y_j}{z_j} - \hat{Y}_G \right)^2 \right].$$

9.16. НЕКОТОРЫЕ СРАВНЕНИЯ ПРИ $n=2$

Случай, когда $n = 2$, вероятно, наиболее распространенный, а также наиболее простой. Выбирая метод отбора, нужно учитывать следующие факторы: (а) легкость извлечения выборки, (б) простоту оценки, (в) достоверность оценки и (г) возможность получения некоторой оценки дисперсии этой оценки.

Подробного сравнения эффективности описанных методов не производилось, но некоторые из них были применены к трем небольшим совокупностям с $N = 4, n = 2$, построенным Йейтсом и Гранди (Yates and Grundy, 1953). Далее сравниваются шесть методов применительно к трем совокупностям с $N = 5, n = 2$, образованным следующим образом. Размер единиц z_i одинаков для всех трех совокупностей (A, B, C). Для совокупности A среднее на элемент, пропорцио-

Таблица 9.9

ТРИ НЕБОЛЬШИЕ ИСКУССТВЕННЫЕ СОВОКУПНОСТИ

Относительный размер единиц z_i		0,1	0,1	0,2	0,3	0,3
Совокупность А	y_i	0,3	0,5	0,8	0,9	1,5
	y_i/z_i	3	5	4	3	5
Совокупность В	y_i	0,3	0,3	0,8	1,5	1,5
	y_i/z_i	3	3	4	5	5
Совокупность С	y_i	0,5	0,5	0,8	0,9	0,9
	y_i/z_i	5	5	4	3	3

нальное y_i/z_i , некоррелировано с z_i . Для совокупности B среднее на элемент растет по мере увеличения размера единицы и для совокупности B среднее на элемент убывает по мере увеличения размера.

Сравниваются следующие схемы отбора. Все они дают несмещенные оценки.

1. Первая единица извлекается с вероятностью, пропорциональной z_i , вторая — с вероятностью, пропорциональной размеру оставшихся единиц. Оценка:

$$\hat{Y}_U = \frac{1}{2} \sum y_i/z_i.$$

2. Исходные вероятности выбираются, как предлагал Нарейн, так, чтобы $\pi_i = 2z_i$. Оценка: $\hat{Y}_N = \frac{1}{2} \sum y_i/z_i$.

3. Единицы расположены случайным образом и извлекается систематическая выборка. Оценка:

$$\hat{Y}_{SYS} = \frac{1}{2} \sum y_i/z_i.$$

4. Совокупность разделена на две группы, суммарный размер единиц в которых одинаков. Одна группа содержит единицы, имеющие размер 0,1; 0,1; 0,3; другая — имеющие размер 0,2; 0,3. (Какая из больших единиц включена в группу с малыми единицами, роли не играет.) Оценка: $\hat{Y}_{G1} = \sum Z_j (y_j/z_j)$.

5. Единицы случайным образом объединены в группы, содержащие три единицы и две единицы. Оценка: $\hat{Y}_{G2} = \sum Z_j (y_j/z_j)$.

6. Единицы отбираются с вероятностями, пропорциональными размеру, и с возвращением. Оценка: $\hat{Y}_{PPS} = \frac{1}{2} \sum y_i/z_i$.

В табл. 9.10 приведены дисперсии соответствующих оценок. В среднем первые пять методов не обнаружили между собой особых различий и все оказались превосходящими отбор с возвращением. Средние значения дисперсий по трем совокупностям могут ввести в заблуждение, потому что совокупность A , пожалуй, более типична для ситуаций, при которых применяется отбор с вероятностями, пропорцио-

Таблица 9.10
ДИСПЕРСИИ ОЦЕНОК СУММАРНЫХ ЗНАЧЕНИЙ ДЛЯ СОВОКУПНОСТИ

Совокупность	Оценка					
	\hat{Y}_U	\hat{Y}_N	\hat{Y}_{SYS}	\hat{Y}_{G1}	\hat{Y}_{G2}	\hat{Y}_{PPS}
A	0,279	0,244	0,233	0,220	0,320	0,400
B	0,434	0,252	0,273	0,300	0,256	0,320
B	0,120	0,252	0,273	0,300	0,256	0,320
Среднее	0,278	0,249	0,260	0,273	0,277	0,347

нальными размеру, чем совокупности B и B , хотя такой отбор применяется и для случаев, когда y_i/z_i коррелированы с z_i .

Для совокупности A схемы (\hat{Y}_U и \hat{Y}_{G2}), при которых вероятности извлечения искажаются, оказались менее точными, чем три схемы, при которых эти вероятности сохраняются.

Несмещенные оценки ошибки имеются согласно (9.40) для \hat{Y}_U и \hat{Y}_N с тем, однако, недостатком, что нужно вычислять π_{ij} и оценка может быть весьма неустойчивой. Оценка \hat{Y}_{G2} (разбиение на группы случайным образом) применительно к оценке ее ошибки находится по сравнению с остальными в наилучшем положении.

У п р а ж н е н и я

9.1. Сравните по данным табл. 9.1 относительные издержки при применении четырех типов единиц, если цель обследования состоит в том, чтобы оценить общее число саженцев на гряде со стандартной ошибкой в 200 саженцев. (Заметьте, что пкс учитывается.)

9.2. По данным табл. 3.5 (с. 81) оцените сравнительную точность как единицы отбора домохозяйства и отдельного лица при оценивании отношения числа мужчин к числу женщин и доли людей, посетивших врача в течение прошедших 12 месяцев, предполагая, что производится простой случайный отбор.

9.3. Совокупность, состоящая из 2500 элементов, разделена на 10 слоев, каждый из которых содержит 50 больших единиц, включающих по пять элементов. Анализ дисперсии, на основании элемента, некоторого признака для совокупности имеет вид:

	Число степеней свободы	Средний квадрат
Между слоями	9	30,6
Между единицами внутри слоев	490	3,0
Между элементами внутри больших единиц	2000	1,6

Будет ли относительная точность большой единицы по сравнению с малой единицей при простом случайном отборе больше, чем при расслоенном случайном отборе (с пропорциональным размещением)? Пкс можно пренебречь.

9.4. Совокупность, содержащая LNM элементов, разделена на L слоев, в каждом по N больших единиц, причем каждая из них состоит из M малых единиц. Из анализа дисперсии на основании элемента для этой совокупности известны следующие величины:

S_1^2 — средний квадрат между слоями;

S_2^2 — средний квадрат между большими единицами внутри слоев;

S_3^2 — средний квадрат между элементами внутри слоев.

Покажите, что если N велико и пкс можно пренебречь, то относительная точность большой единицы по сравнению с малой единицей (элементом) увеличивается при расслоении, если

$$\frac{(M-1)}{S_1^2} < \frac{M}{S_2^2} - \frac{1}{S_3^2}.$$

9.5. В сельскохозяйственном обследовании, в котором единицей отбора служит гнездо, состоящее из M ферм, издержки на получение выборки объемом в n единиц равны:

$$C = 4t M n + 60 \sqrt{n},$$

где t — время в часах, затрачиваемое на получение ответов от одного фермера. При условии, что на обследование отпущено 2000 долл., были получены следующие значения n для $M = 1, 5, 10$; $t = 1/2, 2$:

	M		
	1	5	10
$t = \frac{1}{2}$ часа	400	131	75
$t = 2$ часа	153	40	21

Проверьте два из этих значений, чтобы убедиться в том, что вы умеете применять формулу.

Дисперсия выборочного среднего (если пренебречь пкс) равна

$$\frac{S^2}{Mn} [1 + (M-1) \rho].$$

Если $\rho = 0,1$ для всех M между 1 и 10, то какой размер единиц обеспечит наибольшую точность при (а) $t = 1/2$ часа, (б), $t = 2$ часам? Объясните различие в результатах.

9.6. Если на то же обследование отпущено 5000 долл., увеличится ли, по вашему мнению, оптимальный размер единицы или уменьшится (по сравнению со случаем, когда издержки равны 2000 долл.)? По каким причинам это произойдет? Вы можете, если хотите, найти оптимальный размер, чтобы подтвердить ваши доводы.

9.7. Хорвиц и Томпсон (Horvitz and Thompson, 1952) приводят следующие данные о глазомерных оценках числа домохозяйств M_i и о действительных значениях y_i по 20 городским кварталам Эймса, штат Айова:

M_i	y_i	\bar{y}_i	y_i^2/M_i	M_i	y_i	\bar{y}_i	y_i^2/M_i
9	9	1,0000	9,000	19	19	1,0000	19,000
9	13	1,4444	18,778	21	25	1,1905	29,762
12	12	1,0000	12,000	23	27	1,1739	31,696
12	12	1,0000	12,000	24	21	0,8750	18,375
12	14	1,1667	16,333	24	35	1,4583	51,042
14	17	1,2143	20,643	25	22	0,8800	19,360
14	15	1,0714	16,071	26	25	0,9615	24,038
17	20	1,1765	23,529	27	27	1,0000	27,000
18	19	1,0556	20,056	30	47	1,5667	73,633
18	18	1,0000	18,000	40	37	0,9250	34,225

Для того чтобы облегчить вычисления, приводятся также значения \bar{y}_i и y_i^2/M_i . Извлекается выборка, состоящая из $n = 1$ квартала. Вычислите дисперсию следующих оценок общего числа домохозяйств, Y : (а) несмещенной оценки при отборе с равными вероятностями, (б) оценки по отношению при отборе с равными вероятностями, (в) оценки при отборе с вероятностями, пропорциональными M_i . (Для оценки по отношению вычислите истинный средний квадрат ошибки, а не его приближенное значение.) Согласуются ли ваши результаты с тем, что говорилось в параграфе 9.12?

9.8. По некоторой выборке средних школ рассылается анкета, чтобы выяснить, в каких из этих школ имеется определенное оборудование, например, для обучения русскому языку или для плавания. Если M_i — число учащихся в i -й

школе, то для того или иного вида оборудования величинной, подлежащей оценке, служит доля P учащихся тех школ, которые имеют это оборудование, т. е.

$$P = \frac{\sum_{i=1}^N M_i}{\sum_{i=1}^N M_i},$$

где \sum — сумма по всем школам, имеющим это оборудование.

Извлекается выборка объемом n школ с вероятностью, пропорциональной M_i , с возвращением. Оказалось, что определенный вид оборудования имеется в a школах из n . (а) Покажите, что $\hat{P} = a/n$ есть несмещенная оценка P и что истинная дисперсия этой оценки равна $P(1-P)/n$. (Указание. В следствии из теоремы 9.4 положите $y_i = M_i$, если школа имеет это оборудование, и 0 в противном случае.) (б) Покажите, что несмещенная оценка $V(\hat{P})$ есть $v(\hat{P}) = \hat{P}(1-\hat{P})/(n-1)$.

9.9. В некоторой совокупности большие единицы сгруппированы по размерам в конечное число групп: все единицы из группы h содержат M_h малых единиц. (а) При каких условиях отбор с вероятностями, пропорциональными размеру, даст, в среднем, то же распределение групп по размеру в выборке, что и отбор с расслоением по размеру единицы с оптимальным размещением при неизменном объеме выборки? (б) Если дисперсия признака среди больших единиц в группе h равна kM_h , где k — одно и то же для всех групп, то при каких значениях вероятностей для единиц быть отобранными распределение групп по размеру в выборке будет приблизительно таким же, как и для расслоенной случайной выборки с оптимальным размещением при неизменном объеме выборки?

9.10. Из некоторой совокупности с $N = 3$, $z_1 = 1/2$, $1/3$, $1/6$, $y_1 = 7$, 5 , 2 извлекаются без возвращения две единицы, первая — с вероятностью, пропорциональной z_i , вторая — с вероятностью, пропорциональной размеру оставшихся единиц. (а) Проверьте, что в обозначениях параграфа 9.14 $\pi_1 = 51/60$, $\pi_2 = 44/60$, $\pi_3 = 25/60$ и что $\pi_{12} = 35/60$, $\pi_{13} = 16/60$, $\pi_{23} = 9/60$. (б) Сравните дисперсии или средние квадраты ошибки оценок \hat{Y}_U , \hat{Y}_{SYS} , \hat{Y}_{G2} , определенных в параграфе 9.16. (Для \hat{Y}_U и \hat{Y}_{G2} или постройте все возможные оценки или примените формулы дисперсии. Для \hat{Y}_{SYS} постройте все возможные оценки.)

ЛИТЕРАТУРА

- Cornfield J. (1951). The determination of sample size. *Amer. Jour. Pub. Health*, 41, 654—661.
 Des Raj (1954). On sampling with probabilities proportional to size. *Ganita*, 5, 175—182.
 Des Raj (1956a). Some estimators in sampling with varying probabilities without replacement. *Jour. Amer. Stat. Assoc.*, 51, 269—284.
 Des Raj (1956b). A note on the determination of optimum probabilities in sampling without replacement. *Sankhyā*, 17, 197—200.
 Des Raj (1958). On the relative accuracy of some sampling techniques. *Jour. Amer. Stat. Assoc.*, 53, 98—101.
 Finkner A. L., Morgan J. J. and Monroe R. J. (1943). Methods of estimating farm employment from sample data in North Carolina. *N. C. Agr. Exp. Sta. Tech. Bull.* 75.
 Hansen M. H. and Hurwitz W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Stat.*, 14, 333—362.
 Hansen M. H., Hurwitz W. N. and Madow W. G. (1953). *Sample survey methods and theory*. Vol. I. John Wiley and Sons. New York.
 Hartley H. O. and Rao J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Stat.*, 33, 350—374.

- Hendricks W. A. (1944). The relative efficiencies of groups of farms as sampling units. *Jour. Amer. Stat. Assoc.*, 39, 367—376.
- Homeyer P. G. and Black C. A. (1946). Sampling replicated field experiments on oats for yield determinations. *Proc. Soil. Sci. Soc. America*, 11, 341—344.
- Horvitz D. G. and Thompson D. J. (1952). A generalization of sampling without replacement from a finite universe. *Jour. Amer. Stat. Assoc.*, 47, 663—685.
- Jessen R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agr. Exp. Sta. Res. Bull.* 304.
- Johnson F. A. (1941). A statistical study of sampling methods for tree nursery inventories. M. S. thesis, Iowa State College.
- McVay F. E. (1947). Sampling methods applied to estimating numbers of commercial orchards in a commercial peach area. *Jour. Amer. Stat. Assoc.*, 42, 533—540.
- Mahalanobis P. C. (1944). On large-scale sample surveys. *Phil. Trans. Roy. Soc. London*, B231, 329—451.
- Murthy M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, 18, 379—390.
- Narain R. D. (1951). On sampling without replacement with varying probabilities. *Jour. Ind. Soc. Agric. Stat.*, 3, 169—174.
- Rao J. N. K., Hartley H. O. and Cochran W. G. (1962). A simple procedure of unequal probability sampling without replacement. *Jour. Roy. Stat. Soc.*, B, 24.
- Sukhatme P. V. (1947). The problem of plot size in large-scale yield surveys. *Jour. Amer. Stat. Assoc.*, 42, 297—310.
- Sukhatme P. V. (1954). *Sampling theory of surveys, with applications*, Iowa State College Press, Ames, Iowa.
- Yates F. (1960). *Sampling methods for censuses and surveys*. Charles Griffin & Sons, London, third edition. Есть русский перевод: Йейтс Ф. Выборочный метод в переписях и обследованиях. М., «Статистика», 1965.
- Yates F. and Grundy P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Jour. Roy. Stat. Soc.*, B15, 253—261.
- Zarković S. S. (1960). On the efficiency of sampling with various probabilities and the selection of units with replacement. *Metrika*, 3, 53—60.

ПОДОБОР ПРИ ЕДИНИЦАХ ОДИНАКОВОГО РАЗМЕРА

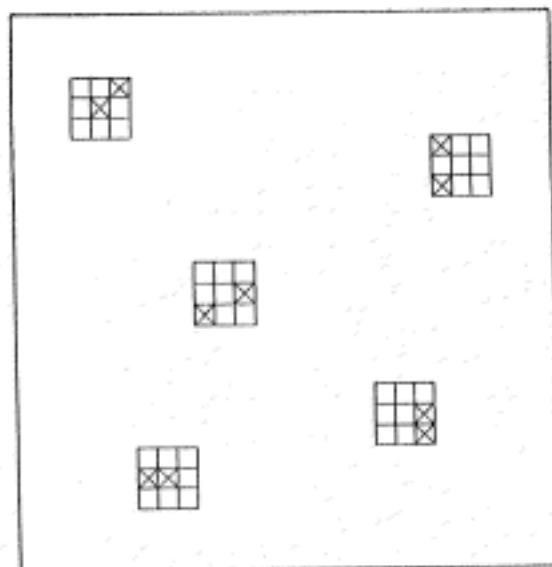
10.1. ДВУХСТУПЕНЧАТЫЙ ОБОР

Предположим, что каждую единицу совокупности можно разделить на некоторое число меньших единиц, или элементов. Извлекается выборка объемом n единиц. Если элементы в некоторой отобранной единице дают сходные результаты, то, по-видимому, неэкономично наблюдать все эти элементы. Обычно из каждой отобранной единицы извлекается и наблюдается некоторая выборка элементов. Такой прием называется *подбором*, поскольку единица не наблюдается полностью, а из нее самой производится отбор. Махаланобис дал ему другое название — *двухступенчатый отбор*, поскольку выборка берется в два этапа. Первый этап состоит в извлечении некоторой выборки единиц, часто называемых *исходными единицами*, а второй — в извлечении некоторой выборки элементов из каждой отобранной исходной единицы.

Подбор имеет разнообразное применение, далеко выходящее за рамки непосредственно выборочных исследований. Когда речь идет о химических, физических или биологических исследованиях, которые могут проводиться на небольшом количестве материала, как правило, этот материал можно получать как некоторую подвыборку из большего количества материала, в свою очередь, представляющего собой какую-то выборку.

В этой главе мы рассмотрим простейший случай, когда каждая единица содержит одно и то же число M элементов, m из которых извлекается, если из этой единицы производится подбор. Схематически двухступенчатая выборка при $M = 9$ и $m = 2$ изображена на рис. 10.1.

Основное преимущество двухступенчатого отбора заключается в том, что он дает большую свободу действий, чем одноступенчатый отбор. Когда $m = M$, двухступенчатый отбор сводится к одноступенчатому, однако, если такое значение m не наилучшее, мы можем взять некоторое меньшее значение m , которое кажется более результативным. Как обычно окончательное решение определяется соотношением между статистической точностью и издержками. Если элементы внутри одной и той же единицы различаются очень мало, то соображения точности подсказывают, что величина m должна быть небольшой. С другой стороны, иногда издержки на наблюдение всей единицы и некоторой под-



⊗ обозначает отобранный элемент

Рис. 10.1. Схематическое изображение двухступенчатой выборки ($N=81$; $n=5$; $M=9$; $m=2$)

выборки из нее почти одинаковы, как, например, в случае, когда единицей служит домохозяйство и один человек может дать точные сведения обо всех членах этого домохозяйства.

10.2. ДВЕ ПОЛЕЗНЫЕ ТЕОРЕМЫ

При двухступенчатом отборе нужно находить математические ожидания не только по всем возможным выборкам объемом n из исходных единиц, но также и по всем возможным подвыборкам, которые могут быть получены из отобранных исходных единиц. К счастью, существует тесная связь между дисперсиями при двухступенчатом отборе и соответствующими (уже изученными) дисперсиями при одноступенчатом отборе. Мы докажем две общие теоремы, полученные Дербинном (Durbin, 1953).

Если каждая исходная единица содержит M подъединиц, m которых отбираются, то простейшие оценки суммарного значения и среднего на подъединицу для совокупности есть соответственно

$$\hat{Y} = \frac{NM}{n} (\bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n); \quad \hat{\bar{Y}} = \frac{1}{n} (\bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n),$$

где \bar{y}_i — выборочное среднее на подъединицу в i -й исходной единице. Обе эти оценки имеют вид

$$y' = y'_1 + y'_2 + \dots + y'_n,$$

где y'_i — оценка, полученная по подвыборке, извлеченной из i -й исходной единицы. Пусть

$$Y'_i = E(y'_i | i),$$

где символ $E(\cdot | i)$ обозначает среднее, взятое по всем подвыборкам, которые можно извлечь из i -й исходной единицы. Если бы эти средние были известны, то мы могли бы построить оценку

$$\hat{Y}' = Y'_1 + Y'_2 + \dots + Y'_n.$$

Это одноступенчатый аналог y' .

Две теоремы, которые будут доказаны, относятся к случаю, когда исходные единицы как одинаковы, так и неодинаковы по размеру. Они справедливы также тогда, когда исходные единицы отбираются с неравными вероятностями. Символ π_i обозначает вероятность для i -й исходной единицы быть отобранной.

Теорема 10.1. Если исходные единицы извлекаются без возвращения и подвыборки в разных единицах извлекаются независимо, то y' есть несмещенная оценка

$$Y' = \sum_i \pi_i Y'_i.$$

с дисперсией

$$V(y') = V(\hat{Y}') + \sum_i \pi_i \sigma_{i1}^2, \quad (10.1)$$

где

$$\sigma_{i1}^2 = E[(y'_i - Y'_i)^2 | i]$$

— дисперсия y'_i при многократном подотборе из i -й единицы.

Доказательство. Для того чтобы найти $E(y')$, сначала возьмем среднее по всем выборкам, содержащим один и тот же набор исходных единиц. Обозначим это среднее через $E(y' | pu)$ (pu — от английского «primary unit» — исходная единица). Очевидно, что

$$E(y' | pu) = Y'_1 + Y'_2 + \dots + Y'_n = \hat{Y}'.$$

Если мы возьмем далее среднее по всем извлечениям n исходных единиц, то относительная частота появления члена Y'_i будет π_i . Следовательно,

$$E(y') = \sum_i \pi_i Y'_i = Y'.$$

Для дисперсии, по определению, имеем

$$\begin{aligned} V(y') &= E(y'^2) - [E(y')]^2 = \\ &= E\left(\sum_i y_i'^2 + 2 \sum_i \sum_{j > i} y'_i y'_j\right) - [E(y')]^2. \end{aligned} \quad (10.2)$$

Сначала возьмем среднее по выборкам, содержащим один и тот же набор l исходных единиц. Имеем

$$E(y_i^2 | l) = Y_l^2 + \sigma_{2l}^2. \quad (10.3)$$

Далее, если подотбор в разных единицах производится независимо, то

$$E(y_i' y_j' | ij) = Y_i' Y_j'. \quad (10.4)$$

Подставляя полученные соотношения в (10.2) и \hat{Y}' вместо $E(y' | pu)$, получаем

$$V(y' | pu) = \sum_i^n Y_i'^2 + 2 \sum_i^n \sum_{j>i}^n Y_i' Y_j' + \sum_i^n \sigma_{2i}^2 - [E(\hat{Y}')]^2. \quad (10.5)$$

Эта условная дисперсия может быть переписана в виде

$$V(y' | pu) = \hat{Y}'^2 - [E(\hat{Y}')]^2 + \sum_i^n \sigma_{2i}^2.$$

Теперь возьмем среднее по всем извлечениям исходных единиц. Получаем

$$V(y') = V(\hat{Y}') + \sum_i^n \pi_i \sigma_{2i}^2.$$

Теорема доказана.

Этот результат можно сформулировать следующим образом. При двухступенчатом отборе дисперсия оценки вида y' состоит из двух частей. Первая, $V(\hat{Y}')$ — это дисперсия, получаемая при замене оценки y_i' , полученной по подвыборке в i -й единице, ее средним значением Y_i' . Эта часть общей дисперсии обусловлена вариацией Y_i' между исходными единицами. Вторая часть, $\sum_i^n \pi_i \sigma_{2i}^2$ — это сумма внутри-единичных дисперсий значений y_i' , взвешенных вероятностями их извлечения.

Следствие 1. Теорема 10.1 представляет собой обобщение на случай двухступенчатого отбора формулы (9.38) из параграфа 9.14 для дисперсии оценки при отборе единиц с произвольными вероятностями и без возвращения. Вернемся к (10.5) и возьмем среднее по всем извлечениям исходных единиц. Если π_{ij} — вероятность того, что выборка содержит как i -ю, так и j -ю единицы, то это среднее можно записать в виде

$$\begin{aligned} V(y') &= \sum_i^n \pi_i Y_i'^2 + 2 \sum_i^n \sum_{j>i}^n \pi_{ij} Y_i' Y_j' + \sum_i^n \pi_i \sigma_{2i}^2 - \\ &- \left(\sum_i^n \pi_i Y_i' \right)^2 = \sum_i^n \left[\pi_i (1 - \pi_i) Y_i'^2 + \right. \\ &+ \left. 2 \sum_{j>i}^n (\pi_{ij} - \pi_i \pi_j) Y_i' Y_j' \right] + \sum_i^n \pi_i \sigma_{2i}^2. \end{aligned} \quad (10.6)$$

Если положить $Y_i' = y_i/z_i'$, то слагаемое дисперсии, отражающее вариацию между единицами, сводится к (9.38).

Следствие 2. Читатель может проверить, что теорема 10.1 справедлива также для случая, когда исходные единицы извлекаются с *возвращением* при условии, что если какая-то исходная единица попадает в выборку более одного раза, то подвыборка из этой единицы извлекается каждый раз и независимо. Это условие гарантирует, что сохраняется участвующее в доказательстве равенство (10.4).

Теорема 10.2 указывает несмещенную оценку $V(y')$ по выборке, если для $V(\hat{Y}')$ имеется несмещенная [при всевозможных значениях Y_i' — *Примеч. пер.*] оценка по выборке $v(\hat{Y}')$. В наиболее общем виде оценка $v(\hat{Y}')$ будет квадратичной формой вида

$$v(\hat{Y}') = \sum_i^n a_{ijk} \dots Y_i'^2 + 2 \sum_i^n \sum_{j>i}^n b_{ijk} \dots Y_i' Y_j',$$

где индексы при a означают, что в общем случае коэффициент при $Y_i'^2$ может зависеть от других единиц, попавших в выборку вместе с i -й единицей; то же означают индексы при b .

Пусть $v_c(y')$ будет «аналогом» $v(\hat{Y}')$ [с — от английского «сорус» — аналог], получаемым путем замены Y_i' везде, где Y_i' фигурирует, величиной y_i' , т. е.

$$v_c(y') = \sum_i^n a_{ijk} \dots y_i'^2 + 2 \sum_i^n \sum_{j>i}^n b_{ijk} \dots y_i' y_j'.$$

Теорема 10.2. В условиях теоремы 10.1 несмещенная оценка $V(y')$ есть

$$v(y') = v_c(y') + \sum_i^n \pi_i \hat{\sigma}_{2i}^2, \quad (10.7)$$

где $\hat{\sigma}_{2i}^2$ — та или иная несмещенная оценка σ_{2i}^2 по выборке.

Доказательство. Из определения $v_c(y')$ и равенств (10.3) и (10.4) имеем

$$\begin{aligned} E[v_c(y') | pu] &= \sum_i^n a_{ijk} \dots Y_i'^2 + 2 \sum_i^n \sum_{j>i}^n b_{ijk} \dots Y_i' Y_j' + \sum_i^n a_{ijk} \dots \sigma_{2i}^2 = \\ &= v(\hat{Y}') + \sum_i^n a_{ijk} \dots \sigma_{2i}^2. \end{aligned}$$

Если среднее берется по всем извлечениям исходных единиц, то коэффициент при σ_{2i}^2 равен $E(a_{ijk} \dots)$. Для того чтобы $v(y')$ была несмещенной оценкой $V(\hat{Y}')$, это среднее согласно (10.6) должно быть равно $\pi_i (1 - \pi_i)$. Отсюда

$$E[v_c(y')] = V(\hat{Y}') + \sum_i^n \pi_i (1 - \pi_i) \sigma_{2i}^2.$$

Но по теореме 10.1

$$V(y') = V(\hat{Y}') + \sum_i^n \pi_i \sigma_{2i}^2.$$

Следовательно, для того чтобы получить несмещенную оценку $V(y')$, мы должны к $\sum_i \pi_i \hat{\sigma}_{2i}^2$ прибавить несмещенную оценку $\sum_i \pi_i^2 \sigma_{2i}^2$. Тогда

$$E\left(\sum_i \pi_i \hat{\sigma}_{2i}^2\right) = \sum_i \pi_i^2 \sigma_{2i}^2.$$

Теорема доказана.

Из этой теоремы вытекает рабочее правило: для того чтобы найти несмещенную выборочную оценку $V(y')$, получите несмещенную оценку величины $V(\hat{Y}')$ для одноступенчатого отбора, $v(\hat{Y}')$; вычислите ее аналог, $v_c(y')$, подставляя везде y'_i вместо Y'_i ; после этого прибавьте член $\sum_i \pi_i \hat{\sigma}_{2i}^2$, где $\hat{\sigma}_{2i}^2$ — несмещенная выборочная оценка дисперсии y'_i внутри единиц.

Теоремы 10.1 и 10.2 широко применяются, поэтому мы приводим их здесь, хотя в настоящей главе нам понадобятся лишь частные случаи этих теорем.

10.3. ДИСПЕРСИЯ ОЦЕНКИ СРЕДНЕГО ПРИ ДВУХСТУПЕНЧАТОМ ОТБОРЕ

Применяются следующие обозначения:

y_{ij} — значение, получаемое для j -го элемента в i -й исходной единице;

$\bar{y}_i = \sum_{j=1}^m \frac{y_{ij}}{m}$ — выборочное среднее на элемент в i -й исходной единице;

$\bar{y} = \sum_{i=1}^N \frac{\bar{y}_i}{n}$ — среднее на элемент для всей выборки;

$S_1^2 = \frac{\sum_{i=1}^N (\bar{y}_i - \bar{y})^2}{N-1}$ — дисперсия средних значений исходных единиц;

$S_2^2 = \frac{\sum_{i=1}^N \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2}{N(M-1)}$ — дисперсия значений элементов внутри исходных единиц.

Теорема 10.3. Если путем простого случайного отбора извлекаются n единиц и m подъединиц из каждой отобранной единицы, то \bar{y} есть несмещенная оценка \bar{Y} с дисперсией

$$V(\bar{y}) = \left(\frac{N-n}{N}\right) \frac{S_1^2}{n} + \left(\frac{M-m}{M}\right) \frac{S_2^2}{mn}. \quad (10.8)$$

Доказательство. В обозначениях теоремы 10.1 положим $y'_i = \bar{y}_i/n$. Тогда

$$\bar{y} = y'; \quad Y'_i = \frac{1}{n} \bar{y}_i; \quad \hat{Y}' = \frac{1}{n} \sum \bar{y}_i; \quad \pi_i = \frac{n}{N}.$$

По теореме 10.1

$$E(\bar{y}) = E(y') = \sum_i \pi_i Y'_i = \frac{1}{N} \sum_i \bar{y}_i = \bar{Y}.$$

Поскольку \hat{Y}' — среднее n величин \bar{Y}_i , то согласно теореме 2.2 (с. 37) для одноступенчатого отбора имеем

$$V(\hat{Y}') = \frac{N-n}{Nn} \frac{\sum_i (\bar{y}_i - \bar{Y})^2}{N-1} = \frac{N-n}{N} \frac{S_1^2}{n}. \quad (10.9)$$

Так как из M элементов i -й единицы извлекаются m элементов, то согласно той же теореме дисперсия величины $y'_i = \bar{y}_i/n$, рассматриваемой в качестве оценки $Y'_i = \bar{Y}_i/n$, равна:

$$\sigma_{2i}^2 = \frac{M-m}{Mn^2} \frac{S_{2i}^2}{m},$$

где S_{2i}^2 — дисперсия значений признака между подъединицами в i -й исходной единице. Следовательно, по теореме 10.1

$$\begin{aligned} V(\bar{y}) &= V(\hat{Y}') + \sum_i \pi_i \sigma_{2i}^2 = \\ &= \frac{(N-n)}{N} \frac{S_1^2}{n} + \frac{1}{n^2} \sum_i \pi_i \left(\frac{M-m}{M}\right) \frac{S_{2i}^2}{m}. \end{aligned}$$

Но $S_2^2 = \sum_i S_{2i}^2/N$. Тогда

$$V(\bar{y}) = \frac{(N-n)}{N} \frac{S_1^2}{n} + \left(\frac{M-m}{M}\right) \frac{S_2^2}{mn}.$$

Если $f_1 = n/N$ и $f_2 = m/M$ есть доли отбора на первой и на второй ступенях, то полученный результат легче запомнить в виде

$$V(\bar{y}) = \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} S_2^2. \quad (10.10)$$

10.4. ОЦЕНИВАНИЕ ДИСПЕРСИИ

Если бы n средних значений признака у исходных единиц, \bar{Y}_i , были известны, то несмещенной оценкой дисперсии их среднего, \bar{Y}' , служило бы

$$v(\hat{Y}') = \frac{(1-f_1)}{n} \frac{\sum_i (\bar{Y}_i - \bar{Y}')^2}{n-1}.$$

Аналог $v(\hat{Y}')$ для двухступенчатой выборки имеет вид

$$v_c(y') = v_c(\bar{y}) = \frac{(1-f_1)}{n} \frac{\sum_{i=1}^n (\bar{y}_i - \bar{y})^2}{n-1}. \quad (10.11)$$

Согласно теореме 10.2, необходимо иметь также несмещенную оценку σ_{2i}^2 из формулы (10.1). Поскольку подвыборки извлекаются путем простого случайного отбора, эта оценка имеет вид

$$\hat{\sigma}_{2i}^2 = \frac{M-m}{M} \frac{s_{2i}^2}{m} = \frac{(1-f_2)}{mn} s_{2i}^2, \quad (10.12)$$

где

$$s_{2i}^2 = \frac{\sum_{j=1}^m (y_{ij} - \bar{y}_i)^2}{m-1}.$$

Теорема 10.4. В условиях теоремы 10.3 несмещенная оценка $V(\bar{y})$ есть

$$v(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{mn} s_2^2, \quad (10.13)$$

где

$$s_1^2 = \frac{\sum_{i=1}^n (\bar{y}_i - \bar{y})^2}{n-1}; \quad s_2^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(y_{ij} - \bar{y}_i)^2}{n(m-1)}. \quad (10.14)$$

Доказательство. Согласно теореме 10.2 несмещенная оценка $V(\bar{y})$ есть

$$v(\bar{y}) = v_c(\bar{y}) + \sum_{i=1}^n \pi_i \hat{\sigma}_{2i}^2.$$

Пользуясь (10.11), (10.12) и принимая $\pi_i = n/N$, получаем отсюда

$$v(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{1}{n^2} \sum_{i=1}^n \frac{n}{N} \frac{1-f_2}{m} s_{2i}^2.$$

Но s_1^2 , определенное в (10.14), равно $\sum s_{2i}^2/n$. Следовательно,

$$v(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{mn} s_2^2.$$

Следствие. Справедливо равенство (мы воспользуемся им в дальнейшем):

$$E(s_1^2) = S_1^2 - \frac{S_1^2}{M} + \frac{S_1^2}{m}. \quad (10.15)$$

Доказательство. Поскольку $v(y^*)$ — несмещенная оценка $V(y^*)$, то (10.13) дает

$$\begin{aligned} \frac{1-f_1}{n} E(s_1^2) &= \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} S_2^2 - \frac{f_1(1-f_2)}{mn} S_2^2 = \\ &= \frac{1-f_1}{n} \left(S_1^2 - \frac{S_1^2}{M} + \frac{S_1^2}{m} \right). \end{aligned}$$

Отсюда вытекает, что несмещенной оценкой S_1^2 будет $[s_1^2 - s_2^2(1-f_2)/m]$.

Замечания к теореме 10.4. Если $m = M$, т. е. $f_2 = 1$, то формула (10.13) превращается в соответствующую формулу для простого случайного отбора единиц. Если $n = N$, то мы получаем формулу для пропорционального расслоенного случайного отбора, поскольку в этом случае исходные единицы можно считать слоями, в каждом из которых производится отбор. Таким образом, двухступенчатый отбор можно рассматривать как своего рода неполное расслоение, при котором слоями служат единицы.

В обычных условиях, когда величиной $f_1 = n/N$ можно пренебречь, мы получаем полезный результат:

$$v(\bar{y}) \approx \frac{s_1^2}{n} = \frac{\sum_{i=1}^n (\bar{y}_i - \bar{y})^2}{n(n-1)}. \quad (10.16)$$

Таким образом, оценку дисперсии можно вычислить, зная только выборочные средние значения для единиц. Этот результат особенно важен тогда, когда применяется систематический подотбор, потому что в этом случае мы не можем вычислить несмещенную оценку S_2^2 . Формула же (10.16) может быть применена при условии, что n/N мало. Если это условие не выполняется, то, как легко видеть, (10.16) дает преувеличенную оценку дисперсии.

10.5. ОЦЕНИВАНИЕ ДОЛЕЙ

Если все элементы разделены на два класса и мы оцениваем долю элементов, принадлежащих первому классу, только что полученные формулы можно применять, пользуясь обычным приемом — считать y_{ij} равным 1, если соответствующий элемент относится к этому классу, и 0 — в противном случае. Пусть $p_i = a_i/m$ — доля элементов, принадлежащих первому классу в подвыборке из i -й единицы. Две оценки дисперсии, s_1^2 и s_2^2 , участвующие в формулировке теоремы 10.4, принимают вид:

$$s_1^2 = \frac{\sum_{i=1}^n (p_i - \bar{p})^2}{n-1};$$

$$s_2^2 = \frac{m}{n(m-1)} \sum_{i=1}^n p_i q_i,$$

где $\bar{p} = \sum p_i / n$. Следовательно, по теореме 10.4,

$$v(\bar{p}) = \frac{1-f_1}{n(n-1)} \sum_i^n (p_i - \bar{p})^2 + \frac{f_1(1-f_2)}{n^2(m-1)} \sum_i^n p_i q_i.$$

Пример. При изучении одного из заболеваний растений растения были высажены на 160 небольших участках по 9 растений на каждом. Была взята случайная выборка 40 участков и с каждого полученного участка отбирались и проверялись на наличие заболевания по три растения. Оказалось, что 22 участка не имели больных растений (среди трех отобранных), 11 имеют одно, 4 имеют два и 3 участка имеют три больных растения. Оцените долю заболевших растений и ее стандартную ошибку. Символ ϕ обозначает частоты 22, 11, 4 и 3.

Имеем $N = 160$, $M = 9$, $n = 40$, $m = 3$. При нахождении s_1^2 и s_2^2 удобно оперировать с числом больных растений ($3p_i$) и числом здоровых растений ($3q_i$). Вычисления представляются следующим образом:

$3p_i$	Частота ϕ	$9p_i q_i$	$9\phi p_i q_i$	$3\phi p_i$	$9\phi p_i^2$
0	22	0	0	0	0
1	11	2	22	11	11
2	4	2	8	8	16
3	3	0	0	9	27
	40		30	28	54

$$\bar{p} = \frac{3\sum \phi p_i}{3\sum \phi} = \frac{28}{120} = 0,233;$$

$$\sum \phi (p_i - \bar{p})^2 = \frac{1}{9} \left(54 - \frac{(28)^2}{40} \right) = 3,822;$$

$$\sum \phi p_i q_i = \frac{30}{9} = 3,333.$$

Следовательно, по формуле, приведенной непосредственно перед примером,

$$v(\bar{p}) = \frac{3 \cdot 3,822}{4 \cdot 40 \cdot 39} + \frac{2 \cdot 3,333}{4 \cdot 3 \cdot 1600 \cdot 2} = 0,00201.$$

Доля больных растений составляет 0,233 со стандартной ошибкой 0,045. Приближенная формула для s_1/\sqrt{n} , согласно (10.16), дает значение 0,049, т. е. вполне удовлетворительную оценку, если учесть, что $f_1 = \frac{1}{4}$.

10.6. ОПТИМАЛЬНЫЕ ДОЛИ ОТБОРА И ПОДОТБОРА

Эти доли зависят от вида функции издержек. Для случая, когда расходы на передвижение от одной единицы к другой незначительны, оказалась удобной функция вида

$$C = c_1 n + c_2 m.$$

Первое слагаемое издержек, $c_1 n$, пропорционально числу исходных единиц в выборке; второе, $c_2 m$ — общему числу единиц второй ступени или элементов. Согласно теореме 10.3 $V(\bar{y})$ можно записать в виде

$$V(\bar{y}) = \frac{1}{n} \left(S_1^2 - \frac{S_2^2}{M} \right) + \frac{1}{mn} S_2^2 - \frac{1}{N} S_1^2. \quad (10.17)$$

Последний член в правой части равенства не зависит от выбора n и m . Минимизация V при неизменном C или C при неизменном V эквивалентна минимизации произведения

$$\left(V + \frac{1}{N} S_1^2 \right) C = c_1 \left(S_1^2 - \frac{S_2^2}{M} \right) + c_2 S_2^2 + \frac{1}{m} c_1 S_2^2 + m c_2 \left(S_1^2 - \frac{S_2^2}{M} \right).$$

Обратите внимание на то, что первые два члена постоянны, в то время как два последних зависят от m , но не от n . Значение m , при котором достигается минимум, можно найти с помощью дифференцирования. Но поскольку на практике m должно быть целым и часто оказывается небольшим, применяется более точный подход, предложенный Айзенхартом (Cameron, 1951). Запишем

$$a = c_1 S_2^2; \quad b = c_2 \left(S_1^2 - \frac{S_2^2}{M} \right).$$

Мы хотим найти целое m , такое, что

$$\frac{a}{m} + bm \leq \frac{a}{m+1} + b(m+1), \quad \text{т. е. } m(m+1) \geq \frac{a}{b};$$

$$\frac{a}{m} + bm \leq \frac{a}{m-1} + b(m-1), \quad \text{т. е. } m(m-1) \leq \frac{a}{b}.$$

Из этих соотношений вытекает следующее правило. Нужно вычислить

$$m_{opt} = \sqrt{a/b} = \frac{S_2}{\sqrt{S_1^2 - S_2^2/M}} \sqrt{c_1/c_2}. \quad (10.18)$$

Если m_{opt} заключено между целыми числами m , $m+1$, то при $m_{opt} > m(m+1)$ нужно выбрать значение $m+1$, т. е. округлить с избытком; в противном случае нужно округлить с недостатком. Так, например, если m_{opt} заключено между 1,414 = $\sqrt{2}$ и 2, то мы округляем до 2. Если m_{opt} больше M или если S_1^2 меньше, чем S_2^2/M , то мы полагаем $m = M$ и применяем одноступенчатый отбор.

ОТНОСИТЕЛЬНАЯ ДИСПЕРСИЯ И ОТНОСИТЕЛЬНАЯ ТОЧНОСТЬ ПРИ
РАЗЛИЧНЫХ ЗНАЧЕНИЯХ m

m	1	2	3	4	5	6	7	8	9	10
Относительная дисперсия	29,59	22,14	20,32	19,92	20,07	20,51	21,10	21,80	22,56	23,38
Относительная точность	0,67	0,90	0,98	1,00	0,99	0,97	0,94	0,91	0,88	0,85

Для любого значения m между 2 и 9 потеря в точности по сравнению с оптимальной составляет менее 12%.

На практике для выбора m требуется оценить c_1/c_2 , а также S_2/S_u или что, эквивалентно, S_2/S_u . Из-за того, что оптимум выражен слабо, эти отношения не обязательно получать с большой достоверностью. Если c_1/c_2 известно достаточно точно и выбрано некоторое значение m , скажем m_0 , то можно воспользоваться удобной таблицей (Brooks, 1955), указывающей интервал значений S_2^2/S_u^2 , внутри которого m_0 дает точность выборки не менее 90% оптимальной.

Эта таблица была построена следующим образом. При заданных издержках, считая N большим, относительную точность m_0 по сравнению с m_{opt} находили по формуле

$$\frac{V(\bar{y} | m_{opt})}{V(\bar{y} | m_0)} = \frac{(S_u \sqrt{c_1} + S_2 \sqrt{c_2})^2}{S_u^2 c_1 + S_2^2 c_2 + m_0 c_2 S_u^2 + c_1 S_2^2 / m_0} \quad (10.20)$$

Значения S_2/S_u , для которых это выражение превышает некоторый определенный уровень L , располагаются между двумя корнями

$$\frac{S_2}{S_u} = \frac{\gamma \pm \sqrt{L(1-L)}(\sqrt{m_0} + \gamma^2/\sqrt{m_0})}{(L\gamma^2/m_0) - (1-L)} \quad (10.21)$$

где $\gamma^2 = c_1/c_2$.

В табл. 10.2, составленной на основе таблицы Брукса (Brooks, 1955), указаны нижние и верхние границы S_2^2/S_u^2 для $L = 0.9$. Почти во всех случаях интервал между верхней и нижней границей поразительно широк. Обратите внимание на то, что интервал между значениями m_0 в разных местах таблицы неодинаков.

Если мы имеем общее представление о величинах S_2^2/S_u^2 для основных изучаемых признаков при обследовании, то табл. 10.2 можно воспользоваться для того, чтобы выбрать значение m_0 . Заметим, что если ρ — коэффициент корреляции между элементами одной и той же исходной единицы, определенный в параграфе 9.4, то отношение S_2^2/S_u^2 приблизительно равно $(1-\rho)/\rho$. Значение S_2^2/S_u^2 , равное 1, соответствует $\rho = 0.5$. Однако такой уровень корреляции внутри единиц необычайно высок. Соответственно $\rho = 0.1$ дает $S_2^2/S_u^2 = 9$, а $\rho = 0.01$ дает $S_2^2/S_u^2 = 99$.

Общий вид (10.18) совпадает с тем, какого можно было ожидать. Если бы элементы были объединены в единицы случайным образом, т. е. если бы применение исходных единиц давало те же результаты, что и применение элементов, то дисперсия средних значений исходных единиц была бы S_2^2/M , так что $(S_1^2 - S_2^2/M)$ обращалось бы в нуль. Отсюда m_{opt} — бесконечности, т. е. нужно сплошное обследование исходных единиц. Наоборот, чем больше дисперсия средних значений исходных единиц S_1^2 по сравнению с дисперсией внутри исходных единиц, тем меньше значение m_{opt} . Чем больше c_1 , издержки на включение в выборку единицы, по сравнению с c_2 , издержками на обследование отдельного элемента в единице, тем выше оптимальное m .

Значение n получаем, решая или уравнение издержек или уравнение дисперсии, в зависимости от того, какая величина задана.

Для большинства практических ситуаций оптимум выражен довольно слабо. Ошибка в несколько единиц при выборе m приводит лишь к небольшой потере в точности. Это можно видеть на следующем примере.

В дисперсионном анализе величина $(S_1^2 - S_2^2/M)$, стоящая в знаменателе m_{opt} , известна как слагаемое дисперсии, выражающее вариацию средних значений единиц. Обозначим ее $\{u$ — от английского *units* — единица]

$$S_u^2 = S_1^2 - \frac{S_2^2}{M} \quad (10.19)$$

Пример. Пусть

$$c_1 = 10c_2; S_2 = 1.3S_u,$$

тогда

$$m_{opt} = 1.3 \sqrt{10} = 4.1.$$

Мы будем считать общие издержки неизменными и посмотрим, как меняется дисперсия \bar{y} при изменении m . Число единиц совокупности, N , предполагается большим. Согласно (10.17)

$$V(\bar{y}) = \frac{S_u^2}{n} + \frac{S_2^2}{nm}.$$

Исключая в этом равенстве n с помощью уравнения издержек, получаем

$$V(\bar{y}) = \left(S_u^2 + \frac{S_2^2}{m} \right) \frac{c_1 + mc_2}{C}.$$

Отсюда

$$V(\bar{y}) = \frac{S_u^2 c_2}{C} \left(1 + \frac{S_2^2}{m S_u^2} \right) \left(\frac{c_1}{c_2} + m \right) = \frac{S_u^2 c_2}{C} \left(1 + \frac{1.69}{m} \right) (10 + m).$$

Опуская постоянный множитель, можно вычислить относительную дисперсию для различных значений m . В табл. 10.1 указаны эти дисперсии и значения относительной точности (за основу сравнения принята максимальная точность при $m = 4$).

Таблица 10.2
ГРАНИЦЫ ДЛЯ S_1^2/S_2^2 , ВНУТРИ КОТОРЫХ m_0 ДАЕТ ТОЧНОСТЬ НЕ МЕНЕЕ 90% МАКСИМАЛЬНОЙ

$c_1/c_2 =$	$\frac{1}{2}$	1	$c_1/c_2 =$	2	4
m_0	L U	L U	m_0	L U	L U
1	0,0 11	0,0 4	2	0,5 8	0,2 4
2	2,0 98	1,1 22	3	1,2 21	0,5 8
3	4,1 >*	2,4 72	4	2,2 44	1,0 16
4	6,6 >	4,0 >	5	3,3 82	1,6 27
5	9,5 >	5,9 >	6	4,7 >	2,4 42
6	13 >	8,1 >	7	6,3 >	3,3 61
7	16 >	11 >	8	8,0 >	4,3 87
8	20 >	13 >	9	10 >	5,4 >
$c_1/c_2 =$	8	16	$c_1/c_2 =$	32	64
m_0	L U	L U	m_0	L U	L U
6	1,0 17	0,3 8	5	0,1 3	0,0 2
7	1,5 24	0,5 11	10	0,4 12	0,1 7
8	2,0 32	0,7 15	15	1,2 26	0,3 14
9	2,6 42	1,0 19	20	2,7 46	0,7 24
10	3,3 53	1,3 23	25	4,5 74	1,5 37
15	7,6 >	3,5 55	30	6,9 >	2,5 52
20	13,3 >	6,6 >	35	9,7 >	3,7 71
25	20,4 >	10,5 >	40	13 >	5,2 93

* > означает «>100».

Пример. Пусть c_1/c_2 приблизительно равно 1 и предполагается, что S_1^2/S_2^2 для основных признаков заключено между 5 и 100. Столбец $c_1/c_2 = 1$ указывает в качестве подходящего значение $m_0 = 4$, потому что соответствующие ему отношения дисперсий заключены между 4 и числом, большим 100 (в действительности равным 196). При $c_1/c_2 = 16$ и том же допустимом интервале таблица указывает значение m_0 , лежащее где-то между 15 и 20. Дальнейшие вычисления по формуле (10.21) показывают, что наилучшим будет $m_0 = 18$. Ему соответствует интервал от 5,2 до 84 — не столь широкий, как хотелось бы.

Когда расходы на передвижение от одной исходной единицы к другой существенны, то, возможно, более адекватной будет функция издержек вида

$$C = c_1 n + c_2 \sqrt{n} + c_3 n m, \quad (10.22)$$

поскольку эти расходы чаще всего пропорциональны \sqrt{n} . Если определено желательное значение $V(\bar{y})$, то по формуле (10.17) легко вычислить пары значений (n, m) , дающие это значение дисперсии. После этого по формуле (10.22) вычисляются издержки для различных сочетаний n и m и находится такое сочетание, для которого эти издержки

минимальны. Для случая, когда издержки заданы заранее, Хансен, Хервиг и Мэдоу (Hansen, Hurwitz and Madow, 1953) предлагают метод определения сочетаний (n, m) , минимизирующих дисперсию, и приводят таблицу, позволяющую быстро найти нужные значения. Заметим, что их n соответствует нашему m и наоборот.

10.7. ОЦЕНИВАНИЕ m_{opt} ПО ДАННЫМ ПРОБНОГО ОБСЛЕДОВАНИЯ

Иногда оценки S_1^2 и S_2^2 или S_n^2 получают по результатам пробного обследования, в котором отбирается n' исходных единиц и в каждой из этих единиц берется m' элементов. В этом параграфе рассматривается нахождение чисел n' и m' . Если s_1^2 — дисперсия средних значений единиц и s_2^2 — дисперсия значений признака у элементов внутри единиц, определенные в параграфе 10.4, то (10.15) дает

$$E(s_1^2) = \left(S_1^2 - \frac{S_2^2}{M} \right) + \frac{S_2^2}{m'} = S_n^2 + \frac{S_2^2}{m'}. \quad (10.23)$$

Для функции издержек простого вида $c_1 n + c_2 n m$, мы имели

$$m_{opt} = \frac{S_2}{\sqrt{S_1^2 - S_2^2/M}} \sqrt{c_1/c_2}.$$

Из (10.23) следует, что в качестве оценки m_{opt} по пробному обследованию можно взять

$$\hat{m}_{opt} = \frac{s_2}{\sqrt{s_1^2 - s_2^2/m'}} \sqrt{c_1/c_2} = \frac{\sqrt{m'}}{\sqrt{(m' s_1^2/s_2^2) - 1}} \sqrt{c_1/c_2}. \quad (10.24)$$

Оценка \hat{m}_{opt} подвержена ошибке выборки, зависящей от ошибки выборки отношения s_1^2/s_2^2 . Из дисперсионного анализа известно, что $m' s_1^2/s_2^2$ распределено как

$$F \left(1 + m' \frac{S_n^2}{S_2^2} \right),$$

где F имеет $(n' - 1)$ и $n' (m' - 1)$ степеней свободы при условии, что \bar{y}_{ij} подчиняются нормальному распределению. Отсюда следует, что \hat{m}_{opt} , получаемое по выборке при данных значениях n' и m' , распределено как

$$\hat{m}_{opt} = \frac{\sqrt{m'} c_1/c_2}{\sqrt{F \left(1 + \frac{m' S_n^2}{S_2^2} \right) - 1}}. \quad (10.25)$$

Пример. Для примера из параграфа 10.6, в котором

$$c_1 = 10 c_2; S_2 = 1,3 S_n; m_{opt} = 1,3 \sqrt{10} = 4,1,$$

рассмотрим, насколько хорошо оценивается m_{opt} по пробной выборке с $n' = 10$ и $m' = 4$. Согласно (10.25)

$$\hat{m}_{opt} = \frac{6,324}{\sqrt{F[1 + (4/1,69)] - 1}} = \frac{6,324}{\sqrt{3,367 F - 1}},$$

где F имеет 9 и 30 степеней свободы. Для того чтобы найти границы, в которых m_{opt} заключено в 80% случаев, имеем, при 10%-ном одностороннем доверительном уровне F :

$$F_{0,10}(9;30) = 1,8490; F_{0,90}(9;30) = 1/F_{0,10}(30;9) = 1/2,2547 = 0,4435.$$

Подставляя эти значения F , получаем:

нижняя граница для $\hat{m}_{opt} = 2,8$;

верхняя граница для $\hat{m}_{opt} = 9,0$.

Как было показано ранее в табл. 10.1, любое значение m в этом интервале обеспечивает точность, близкую к оптимальной. Таким образом, при $n' = 10$, $m' = 4$ в 8 случаях из 10 потеря в точности будет небольшой.

Таблица 10.3
нижние и верхние границы для \hat{m}_{opt}

n'	80%	95%
5	2,5; ∞	1,8; ∞
10	2,8; 9,0	2,3; ∞
20	3,1; 6,4	2,7; 9,1

80%-ные и 95%-ные границы при $n' = 5, 10, 20$ и $m' = 4$ приведены в табл. 10.3. При $n' = 20$ мы почти уверены в том, что полученное значение m_{opt} дает точность, близкую к оптимальной. Однако при $n' = 5$ такой уверенности уже не будет.

Если отношение c_1/c_2 для пробного и основного обследований одинаково, то издержки пробного обследования будут пропорциональны $c_1 n' + c_2 n' m'$. Брукс (Brooks, 1955) приводит таблицу значений (n' , m') для наиболее экономичного пробного обследования, которое обеспечивает при оценивании m_{opt} 90%-ную ожидаемую относительную точность. Табл. 10.4 представляет собой часть этой таблицы.

Таблица 10.4
СХЕМЫ ПРОБНОЙ ВЫБОРКИ, ОБЕСПЕЧИВАЮЩИЕ 90%-НУЮ ОЖИДАЕМУЮ ОТНОСИТЕЛЬНУЮ ТОЧНОСТЬ

c_1/c_2	≤ 1		2		4		8		16		32		64	
S_0^2/S_2^2	n'	m'	n'	m'	n'	m'	n'	m'	n'	m'	n'	m'	n'	m'
1	7	3	6	4	6	5	5	6	5	7	4	10	4	12
2	8	5	7	7	6	9	6	9	5	13	5	14	4	20
4	9	9	8	11	8	12	7	14	7	15	5	25	5	27
8	10	14	10	15	9	17	9	18	8	22	6	32	5	44
16	10	25	10	27	10	27	10	28	8	37	7	46	6	60
32	10	46	10	47	10	48	10	49	9	58	8	69	6	102
64	10	92	10	93	10	96	10	100	10	104	8	137	7	169

При составлении таблицы предполагалось, что N и M велики; если принять в расчет пкс, то окажется, что пробная выборка обеспечивает точность, большую, чем указанная. Обратите внимание на то, что в любом случае не требуется более 10 исходных единиц и что приводимые значения относительно мало чувствительны к величине отношения c_1/c_2 .

10.8. ТРЕХСТУПЕНЧАТЫЙ ОТБОР

Процесс подбора может быть распространен и на третью ступень, когда вместо сплошного наблюдения подъединиц (элементов) из них производится отбор. Например, при обследовании урожайности в Индии (Sukhatne, 1947) удобной единицей отбора была деревня. Внутри деревни отбиралась только часть полей с изучаемой культурой, так что подъединицей служило поле. На отобранных полях для определения урожая на один акр обследовались только определенные участки, так что производился отбор и из самих подъединиц. Если дополнительно производился физический или химический анализ зерна, то отбор мог применяться и еще раз, поскольку анализу часто подвергалась лишь часть выборки, полученной с некоторого поля.

Соответствующие результаты представляют собой непосредственное обобщение результатов для двухступенчатого отбора и далее излагаются кратко. Совокупность состоит из N единиц первой ступени, содержащих каждая по M единиц второй ступени, каждая из которых содержит K единиц третьей ступени. Соответствующие числа для выборки составляют n , m и k . Пусть y_{iju} — значение признака при наблюдении u -й единицы третьей ступени из j -й единицы второй ступени, отобранной из i -й исходной единицы. Соответствующие средние для совокупности на единицу третьей ступени имеют вид:

$$\bar{Y}_{ij} = \frac{\sum_u y_{iju}}{K}; \quad \bar{Y}_i = \frac{\sum_j \sum_u y_{iju}}{MK};$$

$$\bar{\bar{Y}} = \frac{\sum_i \sum_j \sum_u y_{iju}}{NMK}.$$

Нам понадобятся следующие дисперсии для совокупности:

$$S_1^2 = \frac{\sum_i (\bar{Y}_i - \bar{\bar{Y}})^2}{N-1};$$

$$S_2^2 = \frac{\sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_i)^2}{N(M-1)};$$

$$S_3^2 = \frac{\sum_i \sum_j \sum_u (y_{iju} - \bar{Y}_{ij})^2}{NM(K-1)}.$$

Теорема 10.5. Если на всех трех ступенях применяется простой случайный отбор, то выборочное среднее на единицу третьей ступени \bar{y} есть несмещенная оценка \bar{Y} с дисперсией

$$V(\bar{y}) = \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{nm} S_2^2 + \frac{1-f_3}{nmk} S_3^2, \quad (10.26)$$

где $f_1 = n/N$, $f_2 = m/M$, $f_3 = k/K$ есть доли отбора на трех ступенях.

Доказательство. Приведем только основные этапы доказательства. Запишем

$$\bar{y} - \bar{Y} = (\bar{y} - \bar{Y}_{nm}) + (\bar{Y}_{nm} - \bar{Y}_n) + (\bar{Y}_n - \bar{Y}),$$

где \bar{Y}_{nm} — среднее для совокупности по nm отобранным единицам второй ступени и \bar{Y}_n — среднее для совокупности по n отобранным исходным единицам. После того как мы возведем в квадрат и возьмем среднее, попарные произведения выражений в круглых скобках обратятся в нули. Соответствующие квадратичные члены окажутся равными:

$$E(\bar{y} - \bar{Y}_{nm})^2 = \frac{1-f_3}{nmk} S_3^2;$$

$$E(\bar{Y}_{nm} - \bar{Y}_n)^2 = \frac{1-f_2}{nm} S_2^2;$$

$$E(\bar{Y}_n - \bar{Y})^2 = \frac{1-f_1}{n} S_1^2.$$

Сложив эти три выражения, получим утверждение теоремы.

Теорема 10.6. Несмещенная оценка $V(\bar{y})$ по выборке есть

$$v(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{nm} s_2^2 + \frac{f_1 f_2 (1-f_3)}{nmk} s_3^2, \quad (10.27)$$

где s_1^2 , s_2^2 и s_3^2 — выборочные аналоги соответственно S_1^2 , S_2^2 и S_3^2 .

Доказательство. Его можно получить или с помощью методов, рассмотренных в параграфе 10.4, или, показав, что

$$E(s_1^2) = S_1^2 + \frac{1-f_2}{m} S_2^2 + \frac{1-f_3}{mk} S_3^2; \quad (10.28)$$

$$E(s_2^2) = S_2^2 + \frac{1-f_3}{k} S_3^2$$

и $E(s_3^2) = S_3^2$. Для того чтобы доказать первое равенство, обозначим через y_{iK} среднее по m единицам второй ступени в i -й исходной единице при условии, что на третьей ступени обследуются все K элементов. Пусть \bar{y}_K — среднее n значений \bar{y}_{iK} . Тогда из формулы (10.15) для

двухступенчатого отбора следует, что

$$E\left[\frac{\sum (\bar{y}_{iK} - \bar{y}_K)^2}{n-1}\right] = S_1^2 + \frac{1-f_2}{m} S_2^2.$$

Далее, если \bar{y}_i — выборочное среднее для i -й исходной единицы, запишем

$$(\bar{y}_i - \bar{y}) = (\bar{y}_{iK} - \bar{y}_K) + [(\bar{y}_i - \bar{y}_{iK}) - (\bar{y} - \bar{y}_K)].$$

Беря сначала среднее по выборкам, для которых единицы первой и второй ступеней считаются заданными, нетрудно показать, что

$$\frac{1}{(n-1)} E \sum [(\bar{y}_i - \bar{y}_{iK}) - (\bar{y} - \bar{y}_K)]^2 = \frac{S_3^2}{mk} (1-f_3)$$

и что сумма попарных произведений несовпадающих членов обращается в нуль. Отсюда вытекает формула для $E(s_1^2)$. Формула для $E(s_2^2)$ находится аналогично. Следовательно,

$$\begin{aligned} E[v(\bar{y})] &= \frac{1-f_1}{n} \left(S_1^2 + \frac{1-f_2}{m} S_2^2 + \frac{1-f_3}{mk} S_3^2 \right) + \\ &+ \frac{f_1(1-f_2)}{nm} \left(S_2^2 + \frac{1-f_3}{k} S_3^2 \right) + \frac{f_1 f_2 (1-f_3)}{nmk} S_3^2 = \\ &= \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{nm} S_2^2 + \frac{1-f_3}{nmk} S_3^2 = V(\bar{y}). \end{aligned}$$

Как и при двухступенчатом отборе, из (10.27) непосредственно следует, что если f_1 можно пренебречь, то $v(\bar{y})$ сводится к

$$v(\bar{y}) = \frac{s_1^2}{n} = \frac{\sum (\bar{y}_i - \bar{y})^2}{n(n-1)}. \quad (10.29)$$

Если f_1 пренебречь нельзя, то эта оценка дает преувеличенное значение дисперсии.

Для функции издержек вида

$$C = c_1 n + c_2 nm + c_3 nmk$$

оптимальные значения k и m равны:

$$\begin{aligned} k_{opt} &= \frac{S_3}{\sqrt{S_2^2 - S_3^2/K}} \sqrt{c_2/c_3}; \\ m_{opt} &= \frac{\sqrt{S_2^2 - S_3^2/K}}{\sqrt{S_1^2 - S_2^2/M}} \sqrt{c_1/c_2}. \end{aligned} \quad (10.30)$$

Как распространить результаты этого параграфа на случай дополнительных ступеней отбора должно быть ясно из структуры полученных формул.

Подотбор может сочетаться с любым видом отбора исходных единиц. В свою очередь подотбор может быть произведен с расслоением или как систематический отбор. Формулы дисперсии при таких модификациях отбора могут быть построены на основе формул для более простых способов отбора.

Далее излагаются результаты для случая двухступенчатого отбора с расслоением исходных единиц. Предполагается, что размер единиц внутри каждого слоя одинаков, но может меняться от слоя к слою. Такое положение возникает, когда расслоение исходных единиц производится по размеру, так что их размер внутри отдельного слоя становится или одинаковым, или почти одинаковым.

Пусть h -й слой содержит N_h исходных единиц с M_h единицами второй ступени в каждой; соответствующие объемы выборок — n_h и m_h . Оценка среднего для совокупности на единицу второй ступени есть

$$\bar{y}_{st} = \frac{\sum_h N_h M_h \bar{y}_h}{\sum_h N_h M_h} = \sum_h W_h \bar{y}_h, \quad (10.31)$$

где $W_h = N_h M_h / \sum_h N_h M_h$ есть относительный объем слоя, выраженный через единицы второй ступени, и \bar{y}_h — выборочное среднее в слое. Применяя теорему 10.3 к каждому слою, имеем

$$V(\bar{y}_{st}) = \sum_h W_h^2 \left(\frac{1-f_{1h}}{n_h} S_{1h}^2 + \frac{1-f_{2h}}{n_h m_h} S_{2h}^2 \right), \quad (10.32)$$

где $f_{1h} = n_h/N_h$, $f_{2h} = m_h/M_h$.

Согласно теореме 10.4, несмещенной выборочной оценкой этой дисперсии будет

$$v(\bar{y}_{st}) = \sum_h W_h^2 \left[\frac{1-f_{1h}}{n_h} s_{1h}^2 + \frac{f_{1h}(1-f_{2h})}{n_h m_h} s_{2h}^2 \right]. \quad (10.33)$$

Соответствующие дисперсия оценки суммарного значения для совокупности и оценка этой дисперсии получаются путем умножения формул (10.32) и (10.33) на $(\sum_h N_h M_h)^2$.

10.10. ОПТИМАЛЬНОЕ РАЗМЕЩЕНИЕ ПРИ РАССЛОЕННОМ ОТБОРЕ

Этот параграф посвящен тому, как наилучшим образом выбрать значения n_h и m_h . Если расходы на передвижение от единицы к единице не играют основной роли, то издержки могут быть достаточно хорошо представлены формулой

$$C = \sum_h c_{1h} n_h + \sum_h c_{2h} n_h m_h. \quad (10.34)$$

Согласно (10.32) дисперсию можно записать в виде

$$V(\bar{y}_{st}) = \sum_h W_h^2 \left[\frac{1}{n_h} \left(S_{1h}^2 - \frac{S_{2h}^2}{M_h} \right) + \frac{1}{n_h m_h} S_{2h}^2 - \frac{1}{N_h} S_{1h}^2 \right].$$

Выражение

$$V(\bar{y}_{st}) + \lambda \left(\sum_h c_{1h} n_h + \sum_h c_{2h} n_h m_h - C \right),$$

где λ — множитель Лагранжа, представляет собой функцию переменных n_h и $(n_h m_h)$. Следовательно, чтобы минимизировать V при неизменном C или наоборот, должны выполняться условия

$$n_h \sqrt{\lambda} = \frac{W_h}{c_{1h}} \sqrt{S_{1h}^2 - S_{2h}^2/M_h}, \quad (10.35)$$

$$n_h m_h \sqrt{\lambda} = \frac{W_h S_{2h}}{c_{2h}}. \quad (10.36)$$

Отсюда

$$m_h = \frac{S_{2h}}{\sqrt{S_{1h}^2 - S_{2h}^2/M_h}} \cdot \sqrt{c_{1h}/c_{2h}}.$$

Формула для оптимального значения m_h имеет в точности тот же вид, что и при нерасслоенном отборе [(10.18) в параграфе 10.6].

Поскольку $W_h \propto N_h M_h$, из (10.35) получаем

$$n_h \propto \frac{N_h M_h S_{uh}}{\sqrt{c_{1h}}}, \quad \text{где } S_{uh}^2 = S_{1h}^2 - \frac{S_{2h}^2}{M_h}. \quad (10.37)$$

Напомним, что при одноступенчатом расслоенном отборе (параграф 5.5) оптимальные n_h пропорциональны $N_h S_h / \sqrt{c_h}$, где S_h — среднее квадратичное отклонение суммарных значений признака у единиц и c_h — издержки в расчете на одну единицу. Обращаясь к (10.37), заметим, что, как было выяснено в параграфе 10.6, величина S_{uh}^2 представляет собой слагаемое дисперсии, отражающее вариацию средних значений признака у исходных единиц. Следовательно, $M_h S_{uh}$ в (10.37) можно рассматривать как своего рода среднее квадратичное отклонение суммарных значений исходных единиц, если не считать того, что теперь мы имеем дело со слагаемым дисперсии, а не с полной дисперсией.

Поскольку равновзвешенные оценки удобны, мы рассмотрим, при каких обстоятельствах оптимальное размещение приводит к равновзвешенной оценке. Из (10.31) следует, что оценка \bar{y}_{st} будет равновзвешенной, если $n_h m_h / N_h M_h = f_0 = \text{постоянной}$, так как в этом случае

$$\begin{aligned} \bar{y}_{st} &= \frac{\sum_h N_h M_h / n_h m_h \sum_i^{n_h} \sum_j^{m_h} y_{hij}}{\sum_h N_h M_h} = \\ &= \frac{\sum_h \sum_i \sum_j y_{hij}}{f_0 \sum_h N_h M_h} = \frac{\sum_h \sum_i \sum_j y_{hij}}{\sum_h n_h m_h} = \bar{y}. \end{aligned}$$

Искомое условие, как и следовало ожидать, состоит в том, что общая доля отбора f_0 должна быть одинаковой во всех слоях.

На основании (10.36) получаем, что при оптимальном размещении

$$\bar{f}_{0h} = \frac{n_h m_h}{N_h M_h} \propto \frac{S_{2h}}{\sqrt{c_{2h}}}.$$

Часто c_{2h} , издержки в расчете на одну единицу второй ступени, приблизительно одинаковы как для больших, так и для малых исходных единиц; но S_{2h} для больших единиц может быть больше, чем для малых. Однако, поскольку оптимум выражен слабо, равновзвешенная выборка будет часто почти столь же точна, что и выборка при оптимальном размещении. Отметим, что это утверждение остается в силе, даже если оптимальный отбор исходных единиц значительно отличается от пропорционального.

Упражнения

10.1. В картотеке, насчитывающей 400 ящиков, размещены 20 000 карточек по 50 карточек в каждом. Двухступенчатая выборка содержит по пять карточек, извлеченных случайным образом в каждом из 80 случайно отобранных ящиков. Для некоторого признака оценки дисперсий, как они определены в параграфе 10.4, составляют $s_1^2 = 362$, $s_2^2 = 805$. (а) Вычислите стандартную ошибку среднего значения признака на одну карточку по этой выборке. (б) Сравните ее со стандартной ошибкой, задаваемой приближенной формулой (10.16) из параграфа 10.4.

10.2. По результатам пробной двухступенчатой выборки, в которой в каждой из n' единиц отобрано m' подединиц, желательно уметь оценивать значение $V(\bar{y})$, которое имела бы выборка, содержащая m подединиц в каждой из n единиц. Покажите, что несмещенной оценкой $V(\bar{y})$ будет

$$V(\bar{y}) = \left(\frac{N-n}{N} \right) \frac{s_1^2}{n} + \frac{s_2^2}{mn} \left(1 - \frac{m}{m'} + \frac{mn}{m'N} - \frac{mn}{MN} \right),$$

где s_1^2 и s_2^2 вычисляются по результатам пробной выборки. Указание. Воспользуйтесь теоремой 10.3 и равенством

$$E(s_1^2) = S_1^2 - \frac{S_2^2}{M} + \frac{S_2^2}{m}.$$

10.3. По данным обследования урожайности пшеницы в штате Канзас, где исходной единицей служило поле, Кинг и Маккарти (King and McCarty, 1941) приводят следующие средние квадраты отклонений урожайности (в бушелях на акр): $s_1^2 = 165$, $s_2^2 = 66$. На каждом поле отбирались две подвыборки. Для выборки объемом в n полей сравните дисперсия выборочного среднего (а) при действительно полученной выборке, (б) при четырех подвыборках на каждом из n полей, (в) при полной уборке урожая на n полях.

Предполагается, что N и M достаточно велики и не меняются. В случае (в) считается, что полная уборка урожая эквивалентна одноступенчатому отбору (т. е. $m = M$).

10.4. В том же обследовании в случае двух подвыборок на каждом поле средние квадраты отклонений процента содержания белка составили $s_1^2 = 7.73$, $s_2^2 = 1.43$. Сколько нужно отобрать полей, чтобы оценить среднюю урожайность с точностью в пределах ± 1 бушеля и средний процент белка в пределах $\pm 1/4$ с вероятностью ошибиться в каждом случае, равной $1/20$. Произведите вычисления, предполагая, что в основном обследовании (а) на каждом поле отбираются две подвыборки, (б) на каждом отобранном поле производится полная уборка урожая.

10.5. По данным об урожайности пшеницы из упражнения 10.3 определите значение c_1/c_2 в случае линейной функции издержек, если оценка оптимального m равна 2.

10.6. Для случая, когда как m/M , так и n/N малы и функция издержек линейна, покажите, что $m = 2$ дает меньшее значение $V(\bar{y})$, чем $m = 1$, если

$$\frac{c_1}{c_2} > 2 \frac{S_1^2}{S_2^2}.$$

10.7. В крупном отделении универсального магазина обрабатывается приблизительно 20 000 счетов, получаемых в течение месяца. Каждый месяц на протяжении двух лет ($n = 24$) проверялась 2%-ная выборка счетов ($m = 400$). 24 значения числа неверных счетов за месяц (из 400) оказались равными (в порядке возрастания: хронологический порядок здесь нарушен): 0, 0, 1, 1, 2, 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9, 9, 10, 10, 13, 14, 17. В соответствии с результатами параграфа 10.5 вычислите s_1^2 и s_2^2 . После этого вычислите стандартную ошибку величины \bar{p} , рассматриваемой в качестве оценки процента неверных счетов в течение года, которая была бы получена при проверке: (а) 1200 счетов в течение одного месяца, отобранного случайным образом, (б) 300 счетов в каждый из четырех месяцев, отобранных случайным образом, (в) 100 счетов каждый месяц. Указание. Воспользуйтесь формулой из упражнения 10.2 при $m' = 400$ или же получите несмещенные оценки S_1^2 и S_2^2 , примените теорему 10.3.

10.8. При планировании двухступенчатого обследования предполагалось, что c_1/c_2 будет приблизительно равно 4 и что S_1^2/S_2^2 будет заключено между 5 и 50. (а) Какое значение m выбрали бы вы по табл. 10.2? (б) Предположим, что после окончания обследования найдено, что c_1/c_2 близко к 8, а величина S_1^2/S_2^2 близка к 25. Вычислите относительную точность, соответствующую выбранному нами m , по сравнению с оптимальным m . (в) Проведите те же вычисления при $c_1/c_2 = 4$, $S_1^2/S_2^2 = 100$.

10.9. Обозначим через ρ коэффициент корреляции между единицами второй ступени в одной и той же исходной единице. Докажите, что

$$\frac{1-\rho}{\rho} = \frac{S_1^2}{[(N-1)/N] S_1^2 - S_2^2/M} = \frac{S_2^2}{S_n^2}.$$

(Тем самым доказывалось равенство, применявшееся в параграфе 10.6).

10.10. Покажите, что если (в обозначениях параграфа 10.6) $S_n^2 > 0$, то простая случайная выборка объемом в n исходных единиц, где на каждой единице отбиралось по 1 элементу, более точна, чем простая случайная выборка объемом в n элементов ($n \geq 1$, $M \geq 1$). Покажите, что точность обоих методов одинакова, если n/N можно пренебречь. Можно ли было ожидать этого интуитивно?

ЛИТЕРАТУРА

- Brooks S. (1955). The estimation of an optimum subsampling number. *Jour. Amer. Stat. Assoc.*, 50, 398—415.
Cameron J. M. (1951). Use of variance components in preparing schedules for the sampling of baled wool. *Biometrics*, 7, 83—96.
Durbin J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Jour. Roy. Stat. Soc. B15*, 262—269.
Hansen M. H., Hurwitz W. N. and Madow W. G. (1953). *Sample survey methods and theory*. Vol. I. John Wiley and Sons, New York.
King A. J. and McCarty D. E. (1941). Application of sampling to agricultural statistics with emphasis on stratified samples. *Jour. Marketing*, April, 462—474.
Sukhatme P. V. (1947). The problem of plot size in large-yield surveys. *Jour. Amer. Stat. Assoc.*, 42, 297—310.

ПОДОТБОР ПРИ ЕДИНИЦАХ
НЕОДИНАКОВОГО РАЗМЕРА

11.1. ВВЕДЕНИЕ

При выборочном исследовании обширных совокупностей часто оказывается, что исходные единицы различаются своим размером. Кроме того, во многих случаях экономические соображения диктуют необходимость применения многоступенчатого отбора. Таким образом, с проблемами, рассматриваемыми в этой главе, приходится сталкиваться довольно часто. Если размер варьирует не очень сильно, то один из приемов состоит в том, чтобы провести расслоение по размеру исходных единиц, так чтобы размер единиц внутри слоя стал одинаковым или почти одинаковым. В этом случае можно воспользоваться в качестве приближенных формулами из параграфа 10.9. Однако часто между размером исходных единиц внутри некоторых слоев сохраняются значительные различия, а иногда и само расслоение целесообразно производить по другим переменным. Так, Грей и Корлетт (Gray and Corlett, 1950) указывают в обзоре Британских социальных обследований (British Social Surveys), которые представляют собой выборочные обследования в масштабе страны с округами в качестве исходных единиц, что сначала размер был включен в число переменных, по которым должно было производиться расслоение, но после более тщательного изучения характеристик населения предпочтение было отдано другому фактору.

Из-за большого разнообразия возможных приемов многоступенчатого отбора в случае, когда размер единиц варьирует, нужно приложить довольно много усилий, чтобы овладеть этими приемами практически. Единицы можно отбирать либо с равными вероятностями, либо с вероятностями, пропорциональными размеру или некоторой его оценке. Можно разработать различные правила для определения долей отбора и подотбора и применять различные методы оценивания. Преимущества тех или иных приемов зависят от характера совокупности, от затрат на собственно обследование и от того, какие дополнительные сведения имеются в нашем распоряжении.

Первая часть этой главы посвящена описанию основных применяемых методов. Сначала мы рассмотрим совокупности, состоящие только из одного слоя. Обобщение на случай расслоенного отбора может

быть сделано, как и в предыдущих главах, посредством суммирования соответствующих формул дисперсии по всем слоям. Для простоты мы предположим сначала, что отбирается лишь одна исходная единица, т. е. что $n = 1$. Это далеко не столь отвлеченный случай, как может показаться с первого взгляда, потому что, когда число слоев велико, можно получить удовлетворительную точность даже при $n_h = 1$. В серии ежемесячных обследований, предпринимаемых Бюро переписи США для оценки занятости, исходной единицей служит графство или группа соседних графств. Это довольно большие единицы, но организационные удобства их применения уменьшают общие издержки. Поскольку графства по своим характеристикам весьма неоднородны, расслоение доводится до такой степени, когда из каждого слоя извлекается только одна единица. Излагаемая далее теория применима при такой схеме отбора к отдельному слою.

Как и в предыдущих главах, величинами, которые нужно оценить, могут быть суммарное значение для совокупности, Y , среднее для совокупности (обычно среднее на элемент, \bar{Y}) или отношение двух переменных.

Обозначения. Значение наблюдения для j -го элемента в i -й единице обозначается через y_{ij} . Для i -й единицы применяются следующие символы:

	Совокупность	Выборка
Число элементов	M_i	m_i
Среднее на элемент	\bar{Y}_i	\bar{y}_i
Суммарное значение	$Y_i = M_i \bar{Y}_i$	$y_i = m_i \bar{y}_i$

Для совокупности или для выборки в целом применяются символы:

	Совокупность	Выборка
Число элементов	$M_0 = \sum^N M_i$	$\sum^n m_i$
Суммарное значение	$Y = \sum^N Y_i$	$\sum^n y_i$
Среднее на элемент	$\bar{Y} = Y/M_0$	$\bar{y} = \sum y_i / \sum m_i$
Среднее на исходную единицу	$\bar{Y} = Y/N$	$\bar{y} = \sum y_i / n$

11.2. МЕТОДЫ ОТБОРА ПРИ $n = 1$

Предположим, что отобрана i -я единица, содержащая M_i элементов, из которых случайным образом извлечены m_i элементов. Рассмотрим три метода оценивания \bar{Y} , среднего значения на элемент.

1. Единицы отбираются с равными вероятностями

$$\text{Оценка} = \bar{y}_1 = \bar{y}_i.$$

Оценкой служит выборочное среднее на элемент. Оценка будет смещенной, так как при многократном отборе из той же единицы сред-

нее значение \bar{y}_i равно \bar{Y}_i , и, поскольку каждая единица имеет одинаковые шансы быть отобранной, среднее значение \bar{Y}_i , обозначим его через \bar{Y}_a , равно:

$$\frac{1}{N} \sum_{i=1}^N \bar{Y}_i = \bar{Y}_a.$$

Но среднее для совокупности равно:

$$\bar{Y} = \frac{\sum_{i=1}^N M_i \bar{Y}_i}{M_0}, \text{ где } M_0 = \sum_{i=1}^N M_i.$$

Следовательно, смещение равно $(\bar{Y}_a - \bar{Y})$. Поскольку оценка смещена, вычислим средний квадрат ошибки (СКО) относительно \bar{Y} . Запишем

$$\bar{y}_i - \bar{Y} = (\bar{y}_i - \bar{Y}_i) + (\bar{Y}_i - \bar{Y}_a) + (\bar{Y}_a - \bar{Y}).$$

Возведем в квадрат и возьмем среднее по всем возможным выборкам. Все члены, содержащие попарные произведения, обратятся в нуль. Математические ожидания квадратов легко получить с помощью приемов, изложенных в гл. 10. Получаем

$$\text{СКО}(\bar{y}_i) = \underbrace{\frac{1}{N} \sum_{i=1}^N \frac{(M_i - m_i)}{M_i} \cdot \frac{S_{2i}^2}{m_i}}_{\text{внутри единиц}} + \underbrace{\frac{1}{N} \sum_{i=1}^N (\bar{Y}_i - \bar{Y}_a)^2}_{\text{между единицами}} + \underbrace{(\bar{Y}_a - \bar{Y})^2}_{\text{смещение}}, \quad (11.1)$$

где

$$S_{2i}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2$$

— дисперсия между элементами в i -й единице.

СКО (\bar{y}_i) состоит из трех слагаемых: первое возникает вследствие вариации значений наблюдений внутри единиц, второе — вследствие вариации истинных средних по единицам и третье — вследствие смещения.

Мы еще ничего не сказали о значениях m_i . Обычно либо берут все m_i одинаковыми, либо полагают их пропорциональными M_i , т. е. из каждой единицы извлекают подвыборку с одинаковой долей отбора. Выбор m_i влияет только на первое из трех слагаемых дисперсии — на слагаемое, обусловленное вариацией наблюдений внутри единиц.

II. Единицы отбираются с равными вероятностями

$$\text{Оценка} = \bar{y}_{II} = \frac{NM_1 \bar{y}_I}{M_0}.$$

Эта оценка будет несмещенной. Поскольку \bar{y}_i есть несмещенная оценка \bar{Y}_i , произведение $M_i \bar{y}_i$ будет несмещенной оценкой Y_i , суммарного значения для единицы. Следовательно, $NM_i \bar{y}_i$ есть несмещенная оценка суммарного значения для совокупности, Y . Разделив ее на M_0 , т. е. на общее число элементов в совокупности, получаем несмещенную оценку \bar{Y} .

Для того чтобы найти $V(\bar{y}_{II})$, дисперсию этой оценки, которая, конечно, равна ее СКО, запишем

$$\bar{y}_{II} - \bar{Y} = \frac{NM_1 \bar{y}_I}{M_0} - \bar{Y} = \frac{NM_1}{M_0} (\bar{y}_I - \bar{Y}_I) + \left(\frac{NM_1}{M_0} \bar{Y}_I - \bar{Y} \right).$$

Далее, $M_i \bar{Y}_i = Y_i$, суммарному значению для единицы, и $\bar{Y} = N\bar{Y}/M_0$, где \bar{Y} — среднее для совокупности на единицу. Это дает

$$\bar{y}_{II} - \bar{Y} = \frac{NM_1}{M_0} (\bar{y}_I - \bar{Y}_I) + \frac{N}{M_0} (Y_i - \bar{Y}).$$

Следовательно,

$$V(\bar{y}_{II}) = \frac{N}{M_0^2} \sum_{i=1}^N M_i (M_i - m_i) \frac{S_{2i}^2}{m_i} + \frac{N}{M_0^2} \sum_{i=1}^N (Y_i - \bar{Y})^2. \quad (11.2)$$

Междоединичное слагаемое этой дисперсии (второй член в правой части равенства) характеризует вариацию Y_i , суммарных значений признака у единиц. На это слагаемое влияют как изменения M_i от единицы к единице, так и вариация средних на элемент, \bar{Y}_i . Если единицы сильно различаются своим размером, то это слагаемое велико, даже если средние на элемент почти одинаковы для всех единиц. Часто оказывается, что это слагаемое столь велико, что \bar{y}_{II} имеет гораздо больший СКО, чем смещенная оценка \bar{y}_I . Поэтому ни метод I, ни метод II не будут вполне удовлетворительными.

III. Единицы отбираются с вероятностями, пропорциональными размеру

Оценка $\bar{y}_{III} = \bar{y}_I =$ выборочному среднему.

Эта методика предложена Хансеном и Хервицем (Hansen and Hurwitz, 1943). При ней выборочное среднее не подвержено смещению и не имеет столь большой дисперсии, как в методе II.

При многократном отборе i -я единица появляется в выборках с относительной частотой M_i/M_0 . Следовательно,

$$E(\bar{y}_{III}) = \sum_{i=1}^N \frac{M_i}{M_0} \bar{Y}_i = \bar{Y}.$$

Далее,

$$\bar{y}_{III} - \bar{Y} = (\bar{y}_{III} - \bar{Y}_I) + (\bar{Y}_I - \bar{Y}).$$

Возьмем сначала среднее по всем выборкам, в которые попала i -я единица.

$$E(\bar{y}_{in} - \bar{Y})^2 = \left(\frac{M_i - m_i}{M_i} \right) \frac{S_{2i}^2}{m_i} + (\bar{Y}_i - \bar{Y})^2.$$

Теперь возьмем среднее по всем возможным извлечениям исходной единицы. Поскольку i -я единица извлекается с относительной частотой M_i/M_0 , то

$$V(\bar{y}_{in}) = \frac{1}{M_0} \left[\sum_{i=1}^N (M_i - m_i) \frac{S_{2i}^2}{m_i} + \sum_{i=1}^N M_i (\bar{Y}_i - \bar{Y})^2 \right]. \quad (11.3)$$

Заметим, что, как и в методе I, междуединичное слагаемое возникает из-за различия между средними на элемент, \bar{Y}_i , у разных единиц. Если средние на элемент приблизительно одинаковы, то это слагаемое мало.

Пример. Применим полученные результаты к небольшой, искусственно образованной совокупности. Данные приводятся в таблице 11.1. Имеются три единицы соответственно с 2, 4 и 6 элементами. Читатель может проверить приведенные для Y_i , S_{2i}^2 и \bar{Y}_i значения. Среднее для совокупности, \bar{Y} , равно 33/12, или 2,75. Незвешенное среднее величин \bar{Y}_i равно $2,167 = \bar{Y}_a$, так что для метода I смещение составляет — 0,583. Его квадрат, представляющий собой третье слагаемое СКО, равен 0,340.

Таблица 11.1

ИСКУССТВЕННАЯ СОВОКУПНОСТЬ С ЕДИНИЦАМИ НЕОДИНАКОВОГО РАЗМЕРА

Единица	y_{ij}	M_i	Y_i	S_{2i}^2	\bar{Y}_i	$\bar{Y}_i - \bar{Y}$
1	0, 1	2	1	0,500	0,5	-2,25
2	1, 2, 2, 3	4	8	0,667	2,0	-0,75
3	3, 3, 4, 4, 5, 5	6	24	0,800	4,0	+1,25
Итого		12	33			

Извлекается одна единица и из нее отбираются два элемента. Рассмотрим четыре метода, два из которых представляют собой варианты метода I.

Метод Ia

Способ отбора: вероятности извлечения единиц одинаковы, $m_i = 2$.

Оценка: \bar{y}_i (смещенная).

Метод Ib

Способ отбора: вероятности извлечения единиц одинаковы, $m_i = \frac{1}{2} M_i$.

Оценка: \bar{y}_i (смещенная).

Оценка: \bar{y}_i (смещенная).

Метод II

Способ отбора: вероятности извлечения единиц одинаковы, $m_i = 2$.

Оценка: $N M_i \bar{y}_i / M_0$ (несмещенная).

Метод III

Способ отбора: вероятности извлечения единиц равны M_i/M_0 , $m_i = 2$.

Оценка: \bar{y}_i (несмещенная).

Для метода Ib (пропорциональный подотбор) объем выборки не обязательно равен 2 (он может быть 1, 2 или 3), но средний объем выборки равен 2.

Применяя для ошибки выборки формулы (11.1), (11.2) и (11.3), получаем результаты, представленные в табл. 11.2.

Таблица 11.2

СРЕДНИЕ КВАДРАТЫ ОШИБКИ ОЦЕНОК \bar{Y} ПО ВЫБОРКЕ

Метод	Слагаемые СКО, обусловленные			Общий СКО
	вариацией внутри единиц	вариацией между единицами	смещением	
Ia	0,145	2,056	0,340	2,541
Ib	0,183	2,056	0,340	2,579
II	0,256	5,792	0,000	6,048
III	0,189	1,813	0,000	2,002

Хотя мы рассматривали искусственный пример, его результаты типичны для многих совокупностей, по которым производилось аналогичное сравнение. Наименьший СКО дает метод III, потому что у него наименьшее слагаемое, обусловленное вариацией между единицами. Метод II, хотя он и не дает смещения, оказался очень неудачным. Метод Ia (единый объем подвыборки) несколько лучше метода Ib (пропорциональный подотбор).

Сравнение этих методов производилось также на некоторых реальных совокупностях. Для шести признаков (общее число рабочих, общее число сельскохозяйственных рабочих, общее число несельскохозяйственных рабочих, причем каждое оценивалось отдельно для мужчин и женщин) Хансен и Хервиц (Hansen and Hurwitz, 1943) обнаружили, что метод III дает значительное уменьшение слагаемого дисперсии, обусловленного вариацией между единицами по сравнению с несмещенным методом II, а по сравнению с методом I это уменьшение составляет в среднем 30%. (Авторы предполагали, что слагаемым, возникающим из-за вариации внутри единиц, можно пренебречь.) Джебе (Jebe, 1952) нашел при оценивании типичных сельскохозяйственных признаков для штата Северная Каролина, что уменьшение общей дисперсии составляет величину порядка 15% по сравнению с методами типа I. В обоих обследованиях исходной единицей служило графство.

11.3. ОТБОР С ВЕРОЯТНОСТЯМИ, ПРОПОРЦИОНАЛЬНЫМИ ОЦЕНКЕ РАЗМЕРА

Как уже упоминалось в гл. 9, в некоторых обследованиях размер единицы, M_i , известен по прежним данным только приближенно, а в других имеется несколько возможных характеристик размера единицы. Пусть z_i — вероятность, или относительный размер, приписываемый i -й единице, причем значения z_i образуют некоторый набор положительных чисел, в сумме дающих 1. Мы по-прежнему предполагаем, что $n = 1$.

Метод IV. Несмещенная оценка \bar{Y} есть

$$\bar{y}_{IV} = \frac{M_i \bar{y}_i}{z_i M_0} \quad (11.4)$$

Это вытекает из того, что при многократном отборе i -я единица извлекается с относительной частотой z_i , так что

$$E(\bar{y}_{IV}) = \sum_{i=1}^N z_i \left(\frac{M_i \bar{y}_i}{z_i M_0} \right) = \sum_{i=1}^N \frac{M_i \bar{y}_i}{M_0} = \bar{Y}.$$

Дисперсия \bar{y}_{IV} вычисляется обычным способом. Запишем

$$\begin{aligned} \bar{y}_{IV} - \bar{Y} &= \frac{M_i \bar{y}_i}{z_i M_0} - \bar{Y} = \\ &= \frac{1}{M_0} \left[\frac{M_i}{z_i} (\bar{y}_i - \bar{Y}_i) + \left(\frac{M_i}{z_i} \bar{Y}_i - M_0 \bar{Y} \right) \right] \quad [\text{согласно (11.4)}]. \end{aligned}$$

При усреднении каждый квадрат входит в дисперсию с весом z_i . Следовательно,

$$\begin{aligned} V(\bar{y}_{IV}) &= \frac{1}{M_0^2} \left[\sum_{i=1}^N \frac{M_i (M_i - m_i)}{z_i} \frac{S_{y_i}^2}{m_i} + \right. \\ &\quad \left. + \sum_{i=1}^N z_i \left(\frac{M_i \bar{Y}_i}{z_i} - M_0 \bar{Y} \right)^2 \right]. \quad (11.5) \end{aligned}$$

Если $z_i = M_i/M_0$, то (11.5) сводится к формуле (11.3) для $V(\bar{y}_{III})$. Если $z_i = 1/N$ (исходные вероятности равны), то (11.5) сводится к формуле (11.2) для дисперсии несмещенной оценки в случае равных вероятностей.

За исключением случая, когда $z_i = M_i/M_0$, на междуединичное слагаемое в (11.5) влияют в некоторой степени как вариация в размере единиц M_i , так и вариация в средних значениях на элемент, \bar{Y}_i .

Пример. В табл. 11.3 представлены вычисления при нахождении $V(\bar{y}_{IV})$ для искусственной совокупности, приведенной в табл. 11.1. Значения z_i взяты равными 0,2; 0,4 и 0,4, а $m_i = 2$.

Таблица 11.3

ВЫЧИСЛЕНИЕ $V(\bar{y}_{IV})$

Единица	M_i	M_i/M_0	z_i	m_i	$\frac{M_i (M_i - m_i)}{z_i m_i}$	$S_{y_i}^2$	Y_i	$\frac{Y_i}{z_i}$	$\frac{Y_i}{z_i} - Y$
1	2	0,17	0,2	2	0	0,500	1	5	-28
2	4	0,33	0,4	2	10	0,667	8	20	-13
3	6	0,50	0,4	2	30	0,800	24	60	+27

Согласно (11.5) дисперсия складывается следующим образом:

$$\text{внутриединичное слагаемое} = \sum \frac{M_i (M_i - m_i) S_{y_i}^2}{z_i m_i} / M_0^2 = 0,213;$$

$$\text{междуединичное слагаемое} = \sum z_i \left(\frac{Y_i}{z_i} - Y \right)^2 / M_0^2 = 3,583.$$

Сравнение с табл. 11.2 показывает, что метод IV дает меньшую дисперсию, чем несмещенный метод II, при котором единицы отбираются с равными вероятностями. Однако метод IV определенно уступает методам I и III. В нашем примере несмещенность оценки, получаемой методом IV, дается слишком дорогой ценой.

Поэтому естественно стремление выяснить, не будет ли выборочное среднее (как в методе I) лучшей оценкой, чем оценка, получаемая методом IV.

V. Единицы отбираются с вероятностями, пропорциональными оценке размера

Оценка $= \bar{y}_V = \bar{y}_I =$ выборочному среднему.

Эта оценка будет смещенной, поскольку, например,

$$E(\bar{y}_I) = \sum z_i \bar{Y}_i = \bar{Y}_z.$$

Если z_i представляют собой хорошие оценки, то \bar{Y}_z близко к правильному среднему значению $\bar{Y} = \sum M_i \bar{Y}_i / M_0$ и смещение будет небольшим.

Если записать

$$\bar{y}_V - \bar{Y} = (\bar{y}_I - \bar{Y}_I) + (\bar{Y}_I - \bar{Y}_z) + (\bar{Y}_z - \bar{Y}),$$

то три слагаемых СКО принимают вид

$$\text{СКО}(\bar{y}_V) = \sum_{i=1}^N \frac{z_i (M_i - m_i)}{M_i} \frac{S_{y_i}^2}{m_i} + \sum_{i=1}^N z_i (\bar{Y}_I - \bar{Y}_z)^2 + (\bar{Y}_z - \bar{Y})^2.$$

Пример. Если взять значения z_i и m_i из табл. 11.3, то, как может проверить читатель, слагаемые дисперсии \bar{y}_V равны приведенным в табл. 11.4.

Таблица 11.4

Слагаемые СКО для метода V, обусловленные			Общий СКО
вариацией внутри единицы	вариацией между единицами	смещением	
0,173	1,800	0,062	2,035

Метод V превосходит все методы, за исключением метода III (отбор с вероятностями, пропорциональными размеру), и почти не уступает последнему.

Дисперсии этих пяти оценок могли быть получены как частные случаи теоремы 10.1, но при $n = 1$ их было легко найти непосредственно.

11.4. СВОДКА МЕТОДОВ ПРИ $n = 1$

Пять методов оценивания среднего на элемент, \bar{Y} , и соответствующие им СКО для рассмотренного числового примера сведены в табл. 11.5.

Таблица 11.5
МЕТОДЫ ДВУХСТУПЕНЧАТОГО ОТБОРА ($n = 1$)

Метод	Вероятности, с которыми извлекаются единицы	Оценка \bar{Y}	Смещенность оценки	СКО для примера
I	Одинаковые	\bar{y}_1	смещенная	Ia: 2,541 Iб: 2,579
II	Одинаковые	$\frac{N \sum M_i \bar{y}_i}{M}$	несмещенная	6,048
III	$\frac{M_i}{M_0}$ от размеру	\bar{y}_1	несмещенная	2,002
IV	z_i от оценке размера	$\frac{M_i \bar{y}_i}{z_i M_0}$	несмещенная	3,796
V	z_i от оценке размера	\bar{y}_1	смещенная	2,035

11.5. МЕТОДЫ ОТБОРА ПРИ $n > 1$

Основные методы отбора при $n > 1$ представляют собой непосредственные обобщения рассмотренных в параграфах 9.8—9.12 методов одноступенчатого отбора из гнездовых единиц неодинакового размера. Поэтому мы можем воспользоваться как формулами дисперсии, полученными в этих параграфах, так и результатами сравнения этих методов.

В последующих параграфах приводятся формулы вычисления истинных СКО и их оценок для наиболее важных методов отбора и оценивания при $n > 1$. Ввиду практического удобства равновзвешенных оценок для каждого метода указываются условия, при которых оценка принимает такой вид.

11.6. ОТБОР ЕДИНИЦ С РАВНЫМИ ВЕРОЯТНОСТЯМИ. ОЦЕНКА ПО ОТНОШЕНИЮ С РАЗМЕРОМ В ЗНАМЕНАТЕЛЕ

Метод I можно обобщить на случай оценивания среднего значения для совокупности, \bar{Y} , несколькими способами. Одной из наиболее удобных во всех отношениях будет, по-видимому, оценка

$$\hat{Y}_R = \frac{\sum M_i \bar{y}_i}{\sum M_i}.$$

Это — типичная оценка по отношению, поскольку как числитель, так и знаменатель меняются от выборки к выборке. Оценка будет смещенной, как это свойственно оценкам по отношению, однако при больших n смещением можно пренебречь. Для того чтобы найти приближенное значение СКО, запишем

$$\frac{\sum M_i \bar{y}_i}{\sum M_i} - \bar{Y} = \frac{\sum M_i (\bar{y}_i - \bar{Y})}{\sum M_i} \approx \frac{\sum M_i (\bar{y}_i - \bar{Y})}{n\bar{M}},$$

где $\bar{M} = M_0/N$ есть средний размер исходной единицы. Для того чтобы применить теорему 10.1 (с. 293), запишем

$$y'_i = \frac{M_i (\bar{y}_i - \bar{Y})}{n\bar{M}}; \quad Y'_i = \frac{M_i (\bar{Y}_i - \bar{Y})}{n\bar{M}}; \quad \hat{Y}' = \sum Y'_i; \quad \pi_i = \frac{n}{N}.$$

Отсюда вытекает, что \hat{Y}' — невзвешенное среднее переменных $M_i (\bar{Y}_i - \bar{Y})/\bar{M}$. Следовательно, по теореме 2.2

$$V(\hat{Y}') = \frac{1 - f_1}{n\bar{M}^2} \frac{\sum M_i^2 (\bar{Y}_i - \bar{Y})^2}{N - 1}.$$

Далее, в обозначениях теоремы 10.1,

$$\begin{aligned} \sigma_{2i}^2 &= E(y'_i - Y'_i)^2 = \frac{1}{n^2 \bar{M}^2} E[M_i^2 (\bar{y}_i - \bar{Y}_i)^2] = \\ &= \frac{1}{n^2 \bar{M}^2} \frac{M_i^2 (1 - f_{2i}) S_{2i}^2}{m_i}. \end{aligned}$$

Следовательно, по теореме 10.1

$$\begin{aligned} \text{СКО}(\hat{Y}_R) \approx V(\hat{Y}') &= V(\hat{Y}') + \sum \pi_i \sigma_{2i}^2 \approx \frac{1 - f_1}{n\bar{M}^2} \frac{\sum M_i^2 (\bar{Y}_i - \bar{Y})^2}{N - 1} + \\ &+ \frac{1}{nN\bar{M}^2} \sum \frac{M_i^2 (1 - f_{2i}) S_{2i}^2}{m_i}. \end{aligned} \quad (11.6)$$

Эта оценка сводится к выборочному среднему, т. е. становится равно-
взвешенной при

$$f_{2i} = \frac{m_i}{M_1} = \text{постоянной} = \frac{\bar{m}}{\bar{M}}.$$

Обозначим эту постоянную через f_2 . В этом случае слагаемое диспер-
сии, обусловленное вариацией внутри единиц (внутриединичное сла-
гаемое), можно выразить более просто, а именно

$$\text{СКО}(\bar{y}) \approx \frac{1-f_1}{n} \frac{\sum M_i^2 (\bar{y}_i - \bar{y})^2}{\bar{M}^2 (N-1)} + \frac{1-f_2}{nm} \sum \left(\frac{M_i}{M_1} \right) S_{2i}^2. \quad (11.7)$$

Можно отметить сходство с соответствующей формулой для случая,
когда исходные единицы имеют одинаковый размер. На основании
(10.10) из параграфа 10.3 получаем

$$V(\bar{y}) = \frac{1-f_1}{n} \frac{\sum (\bar{y}_i - \bar{y})^2}{N-1} + \frac{1-f_2}{nm} \sum \left(\frac{1}{N} \right) S_{2i}^2. \quad (11.8)$$

Различие состоит в том, что в (11.7) слагаемые СКО, отвечающие
разным исходным единицам, входят с разными весами.

Оценку по выборке приближенного значения СКО, выраженного
(11.6), можно получить с помощью теоремы 10.2. Для междуединич-
ного слагаемого $V(\hat{y}')$ обычная оценка по выборке (подверженная
смещению порядка $1/n$) есть

$$v(\hat{y}') = \frac{1-f_1}{n\bar{M}^2} \frac{\sum M_i^2 (\bar{y}_i - \bar{y}_n)^2}{n-1}, \quad (11.9)$$

где $\bar{y}_n = \sum M_i \bar{y}_i / \sum M_i$. Ее «аналогом» будет

$$v_c(\hat{y}') = \frac{1-f_1}{n\bar{M}^2} \frac{\sum M_i^2 (\bar{y}_i - \hat{y}_R)^2}{n-1}, \quad (11.10)$$

если заметить, что аналог \bar{y}_n есть $\hat{y}_R = \sum M_i \bar{y}_i / \sum M_i$.

Для $\hat{\sigma}_{2i}^2$ оценкой будет

$$\hat{\sigma}_{2i}^2 = \frac{1}{n^2 \bar{M}^2} \frac{M_i^2 (1-f_{2i}) s_{2i}^2}{m_i},$$

где, как обычно,

$$s_{2i}^2 = \frac{\sum (y_{ij} - \bar{y}_i)^2}{m_i - 1}.$$

Следовательно, по теореме 10.2, оценка СКО по выборке есть

$$\begin{aligned} v(\hat{y}_R) &= v(y') = v_c(\hat{y}') + \sum \pi_i \hat{\sigma}_{2i}^2 = \\ &= \frac{1-f_1}{n\bar{M}^2} \frac{\sum M_i^2 (\bar{y}_i - \hat{y}_R)^2}{n-1} + \\ &+ \frac{f_1}{n^2 \bar{M}^2} \sum \frac{M_i^2 (1-f_{2i}) s_{2i}^2}{m_i}. \end{aligned} \quad (11.11)$$

Проведя более подробное исследование, Сукхатм (Sukhatme, 1954)
нашел более точную оценку внутриединичного слагаемого.

В случае, когда выборка равновзвешенная, (11.11) принимает
более простой вид

$$\begin{aligned} v(\hat{y}_R) &= \frac{1-f_1}{n\bar{M}^2} \frac{\sum M_i^2 (\bar{y}_i - \hat{y}_R)^2}{n-1} + \\ &+ \frac{f_1(1-f_2)}{n^2 m \bar{M}} \sum M_i s_{2i}^2. \end{aligned} \quad (11.12)$$

Заметим, что если f_1 можно пренебречь, то как в (11.11), так и в (11.12)
оценки дисперсии сводятся к их первым членам.

Пример. Из сборника «American Men of Sciences» случайным об-
разом было отобрано 20 страниц. На каждой странице фиксировался
возраст двух ученых из двух, также отобранных случайным образом
биографий. Общее число биографий на странице меняется в справоч-
нике в пределах от 14 до 21. По данным табл. 11.6 нужно оценить сред-
ний возраст и найти стандартную ошибку оценки.

По данным крайнего правого столбца

$$\hat{y}_R = \frac{\sum M_i \bar{y}_i}{\sum M_i} = \frac{17121,5}{359} = 47,7 \text{ года.}$$

Так как n/N можно пренебречь, из (11.11) получаем

$$v(\hat{y}_R) \approx \frac{\sum M_i^2 (\bar{y}_i - \hat{y}_R)^2}{n\bar{M}^2 (n-1)}.$$

Числитель легче всего вычислить по формуле

$$\begin{aligned} \sum (M_i \bar{y}_i)^2 - 2\hat{y}_R \sum (M_i \bar{y}_i) M_i + \hat{y}_R^2 \sum M_i^2 = \\ = 15\,375\,020 - 95\,384,4 \cdot 309\,747,5 + 2274,55 \cdot 6481 = 571\,300. \end{aligned}$$

Так как по данным выборки $\bar{M} = 359/20$, то

$$v(\hat{y}_R) = \frac{20 \cdot 571\,300}{19 \cdot (359)^2} = 4,67;$$

$$s(\hat{y}_R) = 2,16 \text{ года.}$$

Таблица 11.6

ВОЗРАСТ 40 УЧЕНЫХ ИЗ СБОРНИКА «American Men of Science»
($n=20$; $m=2$)

Номер единицы	M_i	Возраст		Сумма y_i	$M_i \bar{y}_i$
		y_{i1}	y_{i2}		
1	15	47	30	77	577,5
2	19	38	51	89	845,5
3	19	43	45	88	836,0
4	16	55	41	96	768,0
5	16	59	45	104	832,0
6	19	39	38	77	731,5
7	18	43	43	86	774,0
8	18	49	51	100	900,0
9	18	45	35	80	720,0
10	18	46	59	105	945,0
11	20	71	64	135	1350,0
12	18	35	46	81	729,0
13	19	61	54	115	1092,5
14	19	45	87	132	1254,0
15	18	31	38	69	621,0
16	16	64	39	103	824,0
17	16	63	47	110	880,0
18	19	36	33	69	655,5
19	19	61	39	100	950,0
20	19	54	34	88	836,0
Итого	359			1904	17121,5

Если исходные единицы отбираются с равными вероятностями, то другая оценка среднего для совокупности есть

$$\frac{1}{n} (\bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n).$$

Эта оценка будет равновзвешенной, если $m_i = \text{постоянной}$, как в предыдущем примере. Если M_i и \bar{Y}_i некоррелированы, то эта оценка может оказаться удовлетворительной, но если M_i и \bar{Y}_i коррелированы, то она подвержена смещению, которое не исчезает даже при больших n .

11.7. ОТБОР ЕДИНИЦ С РАВНЫМИ ВЕРОЯТНОСТЯМИ. НЕСМЕЩЕННАЯ ОЦЕНКА

Несмещенная оценка (метод II из параграфа 11.2) есть

$$\hat{Y}_u = \frac{N}{nM_0} \sum M_i \bar{y}_i = \frac{1}{nM} \sum M_i \bar{y}_i.$$

Для того чтобы найти дисперсию, представим ошибку как обычно в виде суммы внутриединичного и междуединичного слагаемых, записав

$$\hat{Y}_u - \bar{Y} = \frac{1}{nM} \sum M_i (\bar{y}_i - \bar{Y}_i) + \frac{1}{nM} \sum (Y_i - \bar{Y}).$$

Здесь мы воспользовались тем, что $Y_i = M_i \bar{Y}_i$ и $\sum \bar{Y}_i / nM = \bar{Y} / M = \bar{Y}$. Возводя в квадрат и беря среднее, получаем (ставя на первое место междуединичное слагаемое)

$$V(\hat{Y}_u) = \frac{1-f_1}{nM^2} \frac{\sum (Y_i - \bar{Y})^2}{N-1} + \frac{1}{nNM^2} \sum \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{m_i}. \quad (11.13)$$

Как и оценка по отношению с размером в знаменателе, несмещенная оценка становится равновзвешенной при $f_{2i} = m_i / M_i = \text{постоянной} = f_2$. В этом случае

$$\hat{Y}_u = \frac{1}{nM} \sum \frac{M_i y_i}{f_2 M_i} = \frac{1}{nf_2 M} \sum_{i=1}^n \sum_{j=1}^m y_{ij}. \quad (11.14)$$

Для равновзвешенной оценки дисперсию (11.13) можно записать в виде

$$V(\hat{Y}_u) = \frac{1-f_1}{nM^2} \frac{\sum (Y_i - \bar{Y})^2}{N-1} + \frac{1-f_2}{nmN} \sum \frac{M_i}{M} S_{2i}^2. \quad (11.15)$$

Обычная методика получения несмещенной оценки (11.13) по выборке приводит к формуле

$$v(\hat{Y}_u) = \frac{1-f_1}{nM^2} \frac{\sum (M_i \bar{y}_i - \hat{Y}_u)^2}{n-1} + \frac{f_1}{n^2 M^2} \sum \frac{M_i^2 (1-f_{2i}) s_{2i}^2}{m_i}, \quad (11.16)$$

где $\hat{Y}_u = \sum M_i \bar{y}_i / n$.

Для равновзвешенной оценки (11.14) эта формула сводится к

$$v(\hat{Y}_u) = \frac{1-f_1}{nM^2} \frac{\sum (M_i \bar{y}_i - \hat{Y}_u)^2}{n-1} + \frac{f_1 (1-f_2)}{nmM} \sum M_i s_{2i}^2. \quad (11.17)$$

Пример. Для получения несмещенной оценки по данным таблицы 11.6 нужно знать также N (число страниц) и M_0 (число биографий в книге). Пусть $N = 2823$ и M_0 приблизительно равно 50 000.

Приняв эти условные значения, имеем

$$\hat{Y}_u = \frac{2823}{20 \cdot 50\,000} \cdot 17121,5 = 48,3 \text{ года.}$$

Согласно (11.16) при $\bar{M} = 50\,000/2823 = 17,712$ имеем

$$v(\hat{Y}_u) = \frac{1}{20 \cdot (17,712)^2 \cdot 19} \left[(577,5)^2 + \dots + (836,0)^2 - \frac{(17121,5)^2}{20} \right] = 6,021.$$

Стандартная ошибка оценки равна 2,45 года.

11.8. ОТБОР ЕДИНИЦ С ВЕРОЯТНОСТЯМИ, ПРОПОРЦИОНАЛЬНЫМИ ХАРАКТЕРИСТИКЕ РАЗМЕРА. НЕСМЕЩЕННАЯ ОЦЕНКА

Исходные единицы отбираются с вероятностями, пропорциональными z_i . Для простоты предполагается, что они отбираются с *возвращением*. Как частный случай будут получены результаты для $z_i = M_i/M_0$ (вероятность пропорциональна размеру).

Предполагается, что из i -й единицы отбирается без возвращения подвыборка объемом m_i подъединиц. Если i -я единица отобрана дважды, то мы считаем, что вся подвыборка возвращается, и вновь независимо извлекаются m_i подъединиц, также без возвращения.

Несмещенная оценка среднего для совокупности (обобщение метода IV) есть

$$\hat{Y}_{\text{pres}} = \frac{1}{nM_0} \sum_{i=1}^N \frac{M_i \bar{y}_i}{z_i}. \quad (11.18)$$

Для того чтобы найти дисперсию, запишем

$$\hat{Y}_{\text{pres}} - \bar{Y} = \frac{1}{nM_0} \sum_{i=1}^N \frac{M_i (\bar{y}_i - \bar{Y})}{z_i} + \frac{1}{nM_0} \sum_{i=1}^N \left(\frac{Y_i}{z_i} - \bar{Y} \right). \quad (11.19)$$

Междуединичное слагаемое можно записать в виде

$$\frac{1}{nM_0} \sum_{i=1}^N t_i \left(\frac{Y_i}{z_i} - \bar{Y} \right).$$

Его дисперсию можно получить по теореме 9.3 в виде

$$V_1 = \frac{1}{n^2 M_0^2} \sum_{i=1}^N z_i \left(\frac{Y_i}{z_i} - \bar{Y} \right)^2. \quad (11.20)$$

Внутриединичное слагаемое дисперсии для единицы, отобранной один раз, по теореме 2.2 равно

$$\frac{1}{n^2 M_0^2} \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{z_i^2 m_i}.$$

Если единица извлекается t_i раз, то каждое ее извлечение приводит к появлению дополнительного слагаемого той же величины, поскольку последовательные извлечения независимы. Отсюда [pu — от английского «primary unit» — исходная единица]

$$V_2(pu) = \frac{1}{n^2 M_0^2} \sum_{i=1}^N t_i \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{z_i^2 m_i}.$$

Следовательно,

$$V_2 = E[V_2(pu)] = \frac{1}{n^2 M_0^2} \sum_{i=1}^N \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{z_i m_i}. \quad (11.21)$$

Подвыборки могут извлекаться и другими способами. Если i -я единица отбирается t_i раз, то один из способов состоит в том, чтобы извлечь подвыборку объема $m_i t_i$ без возвращения при условии, конечно, что $M_i > m_i t_i$. Этот способ более точен, но обходится несколько дороже, потому что нужно наблюдать большее число подъединиц. Сукхатм (Sukhatme, 1954) показал, что внутриединичное слагаемое дисперсии для этого способа составляет

$$V_2 = \frac{n-1}{n M_0^2} \sum_{i=1}^N M_i S_i^2. \quad (11.22)$$

где V_2 задано формулой (11.21).

Другая возможность заключается в том, чтобы извлечь единственную выборку объема m_i независимо от того, сколько раз была извлечена i -я единица. Эта выборка при построении оценки получает вес t_i . Внутриединичное слагаемое дисперсии в этом случае равно

$$V_2 + \frac{n-1}{n M_0^2} \sum_{i=1}^N \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{m_i}.$$

Можно показать, что различие в точности между этими тремя способами невелико, если общая доля отбора мала.

Возвращаясь к нашему первому методу подотбора, мы получаем из (11.20) и (11.21)

$$V(\hat{Y}_{\text{pres}}) = \frac{1}{n M_0^2} \sum_{i=1}^N z_i \left(\frac{Y_i}{z_i} - \bar{Y} \right)^2 + \frac{1}{n M_0^2} \sum_{i=1}^N \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{z_i m_i}. \quad (11.23)$$

Для того чтобы выяснить, когда оценка становится равновзвешенной, запишем

$$\hat{Y}_{\text{pres}} = \frac{1}{n M_0} \sum_{i=1}^N \frac{M_i}{z_i m_i} \sum_{j=1}^{m_i} y_{ij}.$$

Следовательно, необходимым условием будет

$$\frac{M_i}{z_i m_i} = \text{постоянной}. \quad (11.24)$$

Обозначим эту постоянную через n/f_0 . При этом оценка принимает вид $\sum y_{ij}/f_0 M_0$. Если воспользоваться (11.24), то математическое ожидание числа подъединиц в выборке равно

$$E\left(\sum m_i\right) = E\left(\sum t_i m_i\right) = n \sum z_i m_i = f_0 M_0.$$

Поэтому величину f_0 можно определить как *ожидаемую общую долю отбора*.

Согласно (11.24) $m_i/M_i = f_0/nz_i$. Если f_0 определено заранее, то работникам, проводящим обследование, можно указать, какую долю отбора m_i/M_i им следует взять в каждой исходной единице. Предположим, например, что имеется в виду общая доля отбора, равная 2%, так что $f_0 = 0,02$ и что извлекается $n = 60$ исходных единиц. Если для какой-то единицы $z_i = 0,0026$, то мы должны положить $m_i/M_i = 0,02/60 \cdot 0,0026$, или 1 из 7,8.

Несмещенная оценка $V(\hat{Y}_{pps})$ имеет простой вид

$$v(\hat{Y}_{pps}) = \frac{1}{n(n-1)M_0^2} \sum (y_i' - \bar{y}')^2, \quad (11.25)$$

где $y_i' = M_i \bar{y}_i / z_i$ и \bar{y}' есть невзвешенное среднее величин y_i' . Для равновзвешенной выборки $y_i' = ny_i / f_0$, где y_i — суммарное значение для выборки в i -й единице.

Доказательство. Из теоремы 9.5, приведенной в параграфе 9.10, вытекает, что если $Y_i = M_i \bar{Y}_i$ известны, то несмещенная оценка междуединичного слагаемого дисперсии (после деления на M_0^2) есть

$$\frac{1}{n(n-1)M_0^2} \sum \left[\frac{Y_i}{z_i} - \left(\frac{\bar{Y}}{z} \right) \right]^2, \text{ где } \left(\frac{\bar{Y}}{z} \right) = \frac{1}{n} \sum \frac{Y_i}{z_i}.$$

Тогда, как обычно, можно записать

$$y_i' - \bar{y}' = \frac{Y_i}{z_i} - \left(\frac{\bar{Y}}{z} \right) + \left[y_i' - \frac{Y_i}{z_i} - \left[\bar{y}' - \left(\frac{\bar{Y}}{z} \right) \right] \right].$$

При усреднении по заданному набору исходных единиц внутриединичное слагаемое суммы $\sum (y_i' - \bar{y}')^2$ будет равно

$$\frac{n-1}{n} \sum \frac{t_i M_i^2 (1-f_{2i}) S_{2i}^2}{z_i^2 m_i}.$$

Среднее значение этого выражения по всем наборам равно:

$$E_2 \left[\sum (y_i' - \bar{y}')^2 \right] = (n-1) \sum \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{z_i m_i}.$$

Следовательно, после деления на $n(n-1)M_0^2$ мы получим правильное значение внутриединичного слагаемого дисперсии $V(\hat{Y}_{pps})$ из (11.23). Тем самым доказано утверждение (11.25). В случае, когда выборка равновзвешенная, (11.25) принимает более простой вид:

$$v(\hat{Y}_{pps}) = \frac{n}{(n-1)(f_0 M_0)^2} \sum (y_i - \bar{y})^2, \quad (11.25')$$

где y_i — суммарное значение для выборки в i -й единице.

11.9. ОТБОР ЕДИНИЦ С ВЕРОЯТНОСТЯМИ, ПРОПОРЦИОНАЛЬНЫМИ РАЗМЕРУ. НЕСМЕЩЕННАЯ ОЦЕНКА

Если $z_i = M_i/M_0$, то несмещенная оценка (11.18) принимает вид

$$\hat{Y}_{pps} = \frac{1}{n} (\bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n). \quad (11.26)$$

Очевидно, что при $m_i = n$ эта оценка будет невзвешенным средним для выборки на подъединицу, \bar{y} .

Согласно (11.23) дисперсия этой оценки есть

$$V(\hat{Y}_{pps}) = \frac{1}{n} \sum \frac{M_i}{M_0} (\bar{Y}_i - \bar{Y})^2 + \frac{1}{n} \sum \frac{M_i}{M_0} \frac{1-f_{2i}}{m_i} S_{2i}^2, \quad (11.27)$$

а согласно (11.25), поскольку в этом случае y_i' равно $M_i \bar{y}_i$, несмещенная оценка дисперсии есть

$$v(\hat{Y}_{pps}) = \frac{1}{n(n-1)} \sum (\bar{y}_i - \hat{Y}_{pps})^2. \quad (11.28)$$

Если $m_i = n$, то это выражение можно записать в виде

$$v(\hat{Y}_{pps}) = \frac{1}{n(n-1)n^2} \sum (y_i - \bar{y})^2, \quad (11.29)$$

где $y_i = n \bar{y}_i$ — суммарному значению для выборки в i -й единице.

11.10. ОТБОР ЕДИНИЦ С ВЕРОЯТНОСТЯМИ, ПРОПОРЦИОНАЛЬНЫМИ ХАРАКТЕРИСТИКЕ РАЗМЕРА. ОЦЕНКА ПО ОТНОШЕНИЮ С РАЗМЕРОМ В ЗНАМЕНАТЕЛЕ

В обследованиях, где отбор единиц с вероятностями, пропорциональными размеру, кажется весьма результативным, но в распоряжении исследователя имеются лишь оценки z_i относительного размера, несмещенная оценка, рассмотренная в параграфе 11.8, может дать слишком большое междуединичное слагаемое дисперсии, как это оказалось в числовом примере из параграфа 11.3. Другой оценкой (обобщение метода V) служит смещенная оценка

$$\hat{Y}_{Rpps} = \frac{\sum M_i \bar{y}_i / z_i}{\sum M_i / z_i}. \quad (11.30)$$

Числитель представляет собой несмещенную оценку $n\bar{Y}$, а знаменатель — смещенную оценку $n\bar{M}_0$. Эта оценка принимает вид невзвешенного выборочного среднего \bar{y} при $m_i/M_i = f_0/nz_i$, т. е. при том же условии, что и для несмещенной оценки.

Поскольку эта оценка представляет собой частный случай более общей оценки по отношению, которая будет рассмотрена в параграфе 11.14, формулы дисперсии будут выведены там же.

Предполагая, что n велико, имеем

$$V(\hat{Y}_{Rppes}) \approx \frac{1}{nM_0^2} \sum \frac{M_i^2}{z_i} (\bar{Y}_i - \bar{Y})^2 + \frac{1}{nM_0^2} \sum \frac{M_i^2(1-f_{2i})S_{2i}^2}{z_i m_i} \quad (11.31)$$

Оценка дисперсии (несколько смещенная) есть

$$v(\hat{Y}_{Rppes}) = \frac{1}{n(n-1)M_0^2} \sum \left[\frac{M_i}{z_i} (\bar{Y}_i - \hat{Y}_{Rppes})^2 \right] \quad (11.32)$$

Для случая равновзвешенной выборки ее можно записать в виде

$$v(\bar{y}) = \frac{n}{(n-1)(f_0 M_0)^2} \sum (y_i - m_i \bar{y})^2 \quad (11.33)$$

11.11. СРАВНЕНИЕ СХЕМ ОТБОРА И ОЦЕНИВАНИЯ

В параграфе 9.12 сравнивалась точность трех следующих схем отбора и оценивания для случая одноступенчатого отбора с единицами неодинакового размера:

- отбор единиц с равными вероятностями. Несмещенная оценка;
- отбор единиц с равными вероятностями. Оценка по отношению с размером в знаменателе;
- отбор единиц с вероятностями, пропорциональными размеру. Несмещенная оценка.

При двухступенчатом отборе заключения, сделанные в параграфе 9.12, останутся справедливыми для междуединичного слагаемого дисперсии, потому что это слагаемое совпадает с дисперсией для соответствующего одноступенчатого отбора. Выводы из параграфа 9.12 можно сформулировать следующим образом: если \bar{Y}_i некоррелировано с M_i или меняется при изменении M_i лишь незначительно, то оценка при отборе с вероятностями, пропорциональными размеру, и оценка по отношению с размером в знаменателе превосходят несмещенную оценку. Если M_i сильно варьирует, то превосходство может оказаться значительным. Напротив, если суммарные значения для единиц, Y_i , некоррелированы с M_i , то предпочтительнее несмещенная оценка.

Сравнительные характеристики оценки по отношению с размером в знаменателе и оценки при отборе с вероятностями, пропорциональными размеру, зависят от соотношения между дисперсией величин \bar{Y}_i и M_i . Если $V(\bar{Y}_i)$ пропорциональна M_i^{-g} , то оценка при отборе с вероятностями, пропорциональными размеру, более точна при $g < 1$ и менее точна при $g > 1$. Условие $g < 1$, вероятно, отражает положение в большинстве практических приложений.

Для каждой схемы отбора и оценивания в параграфах 11.5—11.10 был указан вид равновзвешенной оценки. За исключением случая, когда дисперсии внутри единиц, S_{2i}^2 , сильно отличаются одна от другой, применение схемы, приводящей к равновзвешенной оценке, не

влечет за собой существенной потери в точности. Как мы уже видели, выбор m_i влияет только на внутриединичное слагаемое дисперсии. Как показывают равенства (11.6) и (11.13), внутриединичные слагаемые приблизительно одинаковы для оценки по отношению с размером в знаменателе и для несмещенной оценки, т. е.

$$V_2 = \frac{1}{nNM^2} \sum \frac{M_i^2(1-f_{2i})S_{2i}^2}{m_i} = \frac{1}{nNM^2} \left(\sum \frac{M_i^2 S_{2i}^2}{m_i} - \sum M_i S_{2i}^2 \right)$$

Если m_i выбираются так, чтобы минимизировать V_2 при неизменном общем объеме выборки $\sum m_i$, то m_i должны быть пропорциональны $M_i S_{2i}$. Для того чтобы оценка была равновзвешенной, необходимо, чтобы m_i были пропорциональны M_i . Читатель может проверить, что для оценки при отборе с вероятностями, пропорциональными размеру, минимум V_2 достигается при m_i , пропорциональных S_{2i} , в то время как для равновзвешенной схемы нужны $m_i = \text{постоянной}$.

При сравнении дисперсий внутри единиц для различных схем мы предполагаем, что применяются равновзвешенные оценки. Из (11.7) и (11.27) следует, что члены V_2 имеют вид:

$$\text{при равных вероятностях: } V_2 = \frac{1}{nm} \sum \left(1 - \frac{\bar{m}}{M_i} \right) \frac{M_i}{M_0} S_{2i}^2$$

$$\text{при вероятностях, пропорциональных размеру: } V_2 = \frac{1}{nm} \sum \left(1 - \frac{\bar{m}}{M_i} \right) \frac{M_i}{M_0} S_{2i}^2$$

Эти два выражения имеют только одно несущественное различие. В случае равных вероятностей член \bar{m} одинаков для всех единиц, в то время как при отборе с вероятностями, пропорциональными размеру, отношение \bar{m}/M_i для более крупных единиц меньше и, следовательно, $(1 - \bar{m}/M_i)$ для них больше. Поскольку S_{2i}^2 часто больше в больших единицах, чем в малых, то, возможно, отбор с вероятностями, пропорциональными размеру, дает большее значение дисперсии внутри единиц. Однако при долях подотбора, которые обычно применяются на практике, такое различие должно быть незначительным. В примере из табл. 11.2 (параграф 11.2) слагаемое V_2 составляло 0,189 для отбора с вероятностями, пропорциональными размеру, и 0,183 для равновзвешенной оценки по отношению (метод 16).

Сравнивая три схемы, мы можем, следовательно, заключить, что в чистом виде влияние внутриединичного слагаемого должно ослаблять различия в дисперсиях, создаваемые междуединичными слагаемыми, так что относительные точности не будут отличаться столь значительно, как при одноступенчатом отборе. Например, если междуединичные слагаемые для двух схем составляют $V_2 = 2$ и $V_2 = 1$, а внутриединичные слагаемые $V_1 = 1$ для обеих схем, то относительная точность худшей схемы возрастает от $1/2$ при одноступенчатом отборе до $2/3$ при двухступенчатом.

При предыдущих сравнениях мы не рассматривали отбор с вероятностями, пропорциональными оценке размера. Если имеются достаточно хорошие оценки размера, то такой отбор при оценке по отношению с размером в знаменателе (обобщение метода V) должен дать приблизительно те же результаты, что и отбор с вероятностями, пропорциональными размеру. При несмещенной оценке (обобщение метода IV) точность должна занимать промежуточное место между точностью для отбора с вероятностями, пропорциональными размеру, и точностью при несмещенной оценке в случае, когда единицы отбираются с равными вероятностями.

Выбор схемы отбора и оценивания зависит также от того, какими сведениями о M_i мы должны располагать. При всех схемах отбора нужно, конечно, знать значения M_i для n исходных единиц, попавших в выборку, а отбор с вероятностями, пропорциональными размеру, требует знания всех M_i из совокупности. При оценивании среднего для совокупности для применения несмещенных оценок как в случае равных вероятностей, так и в случае вероятностей, пропорциональных оценке размера, нужно знать M_0 , общее число подъединиц в совокупности. Применение же соответствующих оценок по отношению с размером единицы в знаменателе этого не требует. При оценивании суммарного значения для совокупности положение обратное.

11.12. ОТНОШЕНИЯ С ДРУГОЙ ПЕРЕМЕННОЙ В ЗНАМЕНАТЕЛЕ

При двухступенчатом отборе величиной, которую нужно оценить, часто бывает отношение Y/X . Необходимость в этом возникает по двум различным причинам. Как упоминалось ранее, если x — значение y по данным последней переписи, то отношение y/x может быть относительно устойчивым. Оценка суммарного или среднего значения y для совокупности, основанная на таком отношении, может оказаться более точной, чем оценки, рассматриваемые в этой главе. Так, в частности, обстоит дело при изучении сельскохозяйственных признаков в штате Северная Каролина (L. H. Madow, 1950; Jebe, 1952).

Оценки по отношению такого вида встречаются также при оценивании долей или средних для подразделений совокупности. В обследовании городского населения, при котором исходной единицей служит городской квартал, примером доли такого вида может быть отношение:

$$\frac{\text{число работающих мужчин старше 16 лет}}{\text{число всех мужчин старше 16 лет}}$$

Если положить $y_{ij} = 1$ для каждого работающего мужчины старше 16 лет и $y_{ij} = 0$ в противном случае и $x_{ij} = 1$ для любого мужчины старше 16 лет и $x_{ij} = 0$ в противном случае, то искомая доля для совокупности будет Y/X . Другими примерами отношений для рассматриваемого вида обследования будет средний доход семей, выписывающих тот или иной журнал, или среднее количество карманных денег у одного подростка.

11.13. ДИСПЕРСИЯ ОЦЕНКИ ПО ОТНОШЕНИЮ ПРИ ОТБОРЕ С РАВНЫМИ ВЕРОЯТНОСТЯМИ

Формулы СКО и оценки дисперсии легко вывести из уже полученных результатов. Рассмотрим сначала отбор единиц с равными вероятностями. Оценка по отношению есть

$$\hat{R} = \frac{\sum M_i \bar{y}_i}{\sum M_i \bar{x}_i}$$

Далее, при $R = Y/X$

$$\hat{R} - R = \frac{\sum M_i (\bar{y}_i - R \bar{x}_i)}{\sum M_i \bar{x}_i} \approx \frac{1}{n\bar{X}} \sum M_i (\bar{y}_i - R \bar{x}_i)$$

Здесь, как обычно, мы переходим к приближенному значению, заменяя $\sum M_i \bar{x}_i$ в знаменателе ее математическим ожиданием nX/N или $n\bar{X}$.

Пусть $d_{ij} = y_{ij} - Rx_{ij}$. В соответствии с определением R как суммарное значение для совокупности, D , так и среднее для совокупности, \bar{D} , оба равны нулю. В случае оценки по отношению с размером в знаменателе (параграф 11.6) приближенное значение ошибки оценки имело вид $\sum M_i (\bar{y}_i - \bar{Y})/n\bar{M}$. В случае рассматриваемой оценки по отношению приближенное значение ошибки можно записать в виде $\sum M_i (\bar{d}_i - \bar{D})/n\bar{X}$. Следовательно, формулы дисперсии для \hat{R} можно получить из соответствующих формул параграфа 11.6, подставляя d_{ij} вместо y_{ij} и умножая на $(\bar{M}/\bar{X})^2$.

Поэтому для истинного СКО на основании (11.6) имеем

$$\text{СКО}(\hat{R}) \approx \frac{1-f_1}{n\bar{X}^2} \frac{\sum M_i^2 (\bar{Y}_i - R\bar{X}_i)^2}{N-1} + \frac{1}{nN\bar{X}^2} \sum \frac{M_i^2 (1-f_{2i})}{m_i} S_{22i}^2 \quad (11.34)$$

где

$$S_{22i}^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} [(y_{ij} - Rx_{ij}) - (\bar{Y}_i - R\bar{X}_i)]^2$$

Если $f_{2i} = m_i/M_i = f_2 = \text{постоянной}$, то \hat{R} сводится к отношению суммарных значений для выборки, $\sum y_{ij}/\sum x_{ij}$. В этом случае СКО принимает вид

$$\text{СКО}(\hat{R}) = \frac{1-f_1}{n\bar{X}^2} \frac{\sum M_i^2 (\bar{Y}_i - R\bar{X}_i)^2}{N-1} + \frac{1-f_2}{n\bar{m}\bar{X}^2} \sum \frac{M_i}{M_0} S_{22i}^2 \quad (11.35)$$

Для того чтобы получить оценку дисперсии, подставляем d_{ij} вместо y_{ij} и \bar{X} вместо \bar{M} в формулу (11.11) для $v(\bar{Y}_R)$. Получившееся выраже-

ние содержит R . Подставляя \hat{R} вместо R , замечаем, что член \hat{Y}_R в (11.11) обращается в нуль. Тогда

$$v(\hat{R}) \approx \frac{1-f_1}{n\bar{X}^2} \frac{\sum M_i^2 (\bar{y}_i - \hat{R}\bar{x}_i)^2}{n-1} + \frac{f_1}{n^2\bar{X}^2} \sum \frac{M_i^2 (1-f_{2i}) s_{d2i}^2}{m_i}. \quad (11.36)$$

11.14. ДИСПЕРСИЯ ОЦЕНКИ ПО ОТНОШЕНИЮ ПРИ ОТБОРЕ С ВЕРОЯТНОСТЯМИ, ПРОПОРЦИОНАЛЬНЫМИ ОЦЕНКЕ РАЗМЕРА

Если исходные единицы отбираются с вероятностями, пропорциональными z_i , с возвращением, то оценка есть

$$\hat{R} = \frac{\sum M_i \bar{y}_i / z_i}{\sum M_i \bar{x}_i / z_i}. \quad (11.37)$$

Числитель и знаменатель будут соответственно несмещенными оценками nY и nX . Для того чтобы найти дисперсию оценки, запишем

$$\hat{R} - R = \frac{\sum M_i (\bar{y}_i - R\bar{x}_i) / z_i}{\sum M_i \bar{x}_i / z_i} \approx \frac{1}{nX} \sum \frac{M_i (\bar{y}_i - R\bar{x}_i)}{z_i}.$$

Сравнивая приближенное значение ошибки \hat{R} с ошибкой несмещенной оценки $\hat{Y}_{прес}$ (параграф 11.8), замечаем, что $V(\hat{R})$ можно получить из $V(\hat{Y}_{прес})$, подставляя $d_{ij} = y_{ij} - Rx_{ij}$ вместо y_{ij} и X вместо M_0 . Таким образом, из (11.23) следует

$$V(\hat{R}) = \frac{1}{nX^2} \sum \frac{1}{z_i} (Y_i - RX_i)^2 + \frac{1}{nX^2} \sum \frac{M_i^2 (1-f_{2i}) S_{d2i}^2}{z_i m_i}. \quad (11.38)$$

Если

$$\frac{M_i}{z_i m_i} = \text{постоянной} = \frac{n}{f_2},$$

то оценка \hat{R} сводится к отношению суммарных значений для выборки. Это то же самое условие, что и для $\hat{Y}_{прес}$.

Из (11.25) следует, что оценка $V(\hat{R})$ по выборке есть

$$v(\hat{R}) = \frac{1}{n(n-1)X^2} \sum (y'_i - \hat{R}x'_i)^2, \quad (11.39)$$

где $y'_i = M_i \bar{y}_i / z_i$ и $x'_i = M_i \bar{x}_i / z_i$.

Эта оценка имеет небольшое смещение.

11.15. ОПРЕДЕЛЕНИЕ ДОЛЕЙ ОТБОРА И ПОДОТБОРА. ОТБОР С РАВНЫМИ ВЕРОЯТНОСТЯМИ

Сначала мы рассмотрим эту задачу для случая оценки по отношению с размером в знаменателе, когда единицы отбираются с равными вероятностями. Доля подотбора m_i/M_i предполагается постоянной, так что оценкой служит выборочное среднее на элемент.

Простейшая функция издержек содержит три члена:

c_n — неизменные издержки на одну исходную единицу;

c_2 — издержки на одну подединицу;

c_1 — издержки на составление перечня подединиц в отобранной единице в расчете на одну подединицу [l — от английского «listing» — перечисление].

Мы включили третий член, потому что обычно, для того чтобы извлечь подвыборку, обследователь должен составить перечень элементов и определить их число в каждой отобранной единице. Следовательно,

$$\text{издержки} = c_n n + c_2 \sum m_i + c_1 \sum M_i. \quad (11.40)$$

В таком виде формулой пользоваться нельзя, так как общие издержки будут зависеть от того, какие конкретно единицы отобраны. Поэтому рассмотрим средние издержки при условии, что отбирается n единиц. Они равны:

$$E(C) = c_n n + c_2 n\bar{m} + c_1 n\bar{M} = (c_n + c_1 \bar{M}) n + c_2 n\bar{m} = c_1 n + c_2 n\bar{m}, \quad (11.41)$$

где c_1 включает теперь средние издержки на составление перечня подединиц в единице.

Согласно формуле (11.7) из параграфа 11.6

$$\text{СКО}(\bar{y}) = \frac{1-f_1}{n} \frac{\sum M_i^2 (\bar{Y}_i - \bar{Y})^2}{\bar{M}^2 (N-1)} + \frac{1-f_2}{nm} \sum \frac{M_i}{M_0} S_{2i}^2.$$

Обозначим

$$S_b^2 = \frac{\sum M_i^2 (\bar{Y}_i - \bar{Y})^2}{\bar{M}^2 (N-1)}.$$

Это взвешенная дисперсия средних значений по единицам в расчете на элемент. Она аналогична дисперсии S_1^2 в параграфе 10.6 и сводится к S_1^2 , если все M_i одинаковы. Мы можем обозначить также

$$S_2^2 = \sum \frac{M_i}{M_0} S_{2i}^2.$$

Это взвешенное среднее дисперсий внутри единиц. Оно сводится к S_2^2 из параграфа 10.6, если все M_i одинаковы.

Пользуясь этими обозначениями, записываем

$$\text{СКО}(\bar{y}) = \frac{1}{n} \left(S_b^2 - \frac{S_2^2}{\bar{M}} \right) + \frac{1}{nm} S_2^2 - \frac{1}{N} S_b^2. \quad (11.42)$$

Функция издержек (11.41) и формула (11.42) имеют в точности тот же вид, что и соответствующие выражения в параграфе 10.6, за исключением того, что вместо m стоит m_i , вместо S_1^2 стоит S_{21}^2 и c_1 включает издержки на составление перечня подъединиц. Следовательно, согласно (10.18)

$$\bar{m}_{opt} \approx \frac{S_2}{\sqrt{S_2^2 - S_{21}^2/M}} \sqrt{\frac{c_1}{c_2}}. \quad (11.43)$$

Методы параграфа 10.6, позволяющие воспользоваться для нахождения \bar{m}_{opt} данными об отношениях S_2/S_1 и c_1/c_2 , применимы и в рассматриваемом случае. Аналогичные рассуждения можно провести и для несмещенной оценки в случае, когда единицы отбираются с равными вероятностями.

Следующий параграф посвящен анализу этой проблемы в более общем случае.

11.16. ДОЛИ ОТБОРА И ПОДОТБОРА ПРИ ОТБОРЕ С ВЕРОЯТНОСТЯМИ, ПРОПОРЦИОНАЛЬНЫМИ ОЦЕНКЕ РАЗМЕРА

Важное исследование, проведенное Хансеном и Хервицем (Hansen and Hurwitz, 1949), показывает, как определить одновременно оптимальные вероятности извлечения единиц и оптимальные доли отбора и подотбора. Полученные ими результаты излагаются далее для случая оценки по отношению, \hat{R} . Единицы отбираются с вероятностями, пропорциональными z_i . Предполагается, что доли подотбора выбраны такими, что \hat{R} сводится к $\sum \sum y_{ij} / \sum \sum x_{ij}$. Из параграфа 11.14 следует, что для этого нужно, чтобы $m_i = kM_i/z_i$, причем мы подставили k вместо прежнего f_0/n .

Как и в параграфе 11.15, функция издержек имеет вид

$$C = c_u n + c_2 \sum m_i + c_1 \sum M_i.$$

Такую функцию издержек можно применять только в том случае, когда для всех единиц совокупности имеются хорошие предварительные оценки размера, поскольку эта функция включает издержки на составление перечня подъединиц лишь для тех единиц, которые попадают в выборку. Если такие перечни должны быть составлены для всех единиц совокупности заранее, то для отдельного обследования отбор с вероятностями, пропорциональными размеру, редко будет экономичным, если не считать тех случаев, когда составление такого перечня крайне дешево.

Поскольку

$$E\left(\sum m_i\right) = \sum n z_i m_i = nk \sum M_i = nkM_0;$$

$$E\left(\sum M_i\right) = \sum n z_i M_i,$$

средние издержки на выборочное исследование n единиц есть

$$C = c_u n + c_2 nkM_0 + c_1 n \sum z_i M_i.$$

Для того чтобы найти минимальное значение $V(\hat{R})$ при неизменных средних издержках, мы можем придавать различные значения переменным n , k и вероятностям z_i .

Согласно формуле (11.38) из параграфа 11.14 дисперсия, которую следует минимизировать, имеет вид

$$V(\hat{R}) = \frac{1}{n X^2} \sum \left[\frac{1}{z_i} (Y_i - R X_i)^2 + \frac{M_i (M_i - m_i)}{z_i m_i} S_{22i}^2 \right].$$

Так как $d_{ij} = y_{ij} - R x_{ij}$, мы можем записать $(Y_i - R X_i) = M_i \bar{D}_i$. Замечая, что $M_i/z_i m_i = 1/k$, получаем

$$V(\hat{R}) = \frac{1}{n X^2} \sum \left[\frac{M_i^2}{z_i} \bar{D}_i^2 + \frac{M_i}{k} S_{22i}^2 - \frac{M_i}{z_i} S_{d2i}^2 \right].$$

Объединяя первый и третий члены внутри скобок, имеем

$$V = X^2 V(\hat{R}) = \sum \frac{1}{n} \left[\frac{M_i^2}{z_i} \left(\bar{D}_i^2 - \frac{S_{22i}^2}{M_i} \right) + \frac{M_i}{k} S_{22i}^2 \right].$$

Заметим, наконец, что n появляется только в произведениях $n z_i$ и $n k$. Введем переменные $z'_i = n z_i$ и $k' = nk$. Тогда

$$V = \sum \left[\frac{M_i^2}{z'_i} \left(\bar{D}_i^2 - \frac{S_{22i}^2}{M_i} \right) + \frac{M_i}{k'} S_{22i}^2 \right]. \quad (11.44)$$

Задача состоит в минимизации V по переменным n , k' и z'_i при условии, что средние издержки неизменны и что

$$\sum z'_i = 1, \text{ т. е. } \sum z'_i = n.$$

Вводя λ и μ в качестве неопределенных множителей, мы сводим задачу к минимизации

$$V + \lambda \left(c_u n + c_2 k' M_0 + c_1 \sum z'_i M_i - C \right) + \mu \left(n - \sum z'_i \right). \quad (11.45)$$

Дифференцируя

по n , получаем: $\lambda c_u + \mu = 0$;

по z'_i , получаем: $-\frac{M_i^2}{z_i'^2} \left(\bar{D}_i^2 - \frac{S_{22i}^2}{M_i} \right) + \lambda c_1 M_i - \mu = 0$,

т. е.

$$\lambda z_i'^2 = \frac{M_i^2 (\bar{D}_i^2 - S_{22i}^2/M_i)}{c_u + c_1 M_i}.$$

Поскольку $z_i = z'_i/n$ и $\sum z_i = 1$, отсюда следует, что

$$z_i = \frac{M_i D_{i\alpha} / \sqrt{c_\alpha + c_i M_i}}{\sum M_i D_{i\alpha} / \sqrt{c_\alpha + c_i M_i}}, \quad (11.46)$$

где

$$D_{i\alpha}^2 = \bar{D}_i^2 - \frac{S_{d2i}^2}{M_i},$$

причем предполагается, что $D_{i\alpha}^2$ — положительные числа. Формула (11.46) дает оптимальные вероятности извлечения единиц.

Исследуем теперь величину $D_{i\alpha}^2$, поскольку она может зависеть от размера единицы M_i . В параграфе 9.4 мы выражали дисперсию средних значений гнездовых единиц через коэффициент корреляции внутри единиц. Согласно формуле (9.7) при $n = 1$ и достаточно большом N дисперсия средних значений признака по какой-либо группе исходных единиц может быть выражена приближенно в виде

$$V(\bar{Y}_i) \approx \frac{S^2}{M} [1 + (\bar{M} - 1) \rho_{\bar{M}}], \quad (11.47)$$

где S^2 — дисперсия значений признака у подъединиц в совокупности и \bar{M} — средний размер исходной единицы. Коэффициент корреляции внутри единиц обозначен через $\rho_{\bar{M}}$, чтобы напомнить, что он зависит от размера единицы.

Применим этот результат к дисперсионному анализу переменной d_{ij} , рассматривая слагаемые дисперсии внутри единиц и между единицами. Символом S_d^2 обозначим дисперсию среди всех подъединиц в совокупности. Считая N большим и $M_i = \bar{M}$, имеем:

общая сумма квадратов: $N \bar{M} S_d^2$;

сумма квадратов между единицами: $\bar{M} \sum \bar{D}_i^2 = N S_d^2 [1 + (\bar{M} - 1) \rho_{\bar{M}}]$. (11.48)

Здесь мы воспользовались (11.47) и равенством $\bar{D} = 0$, вытекающим из определения R . Следовательно, вычитая, получаем

сумма квадратов внутри единиц: $N (\bar{M} - 1) S_d^2 (1 - \rho_{\bar{M}}) = N (\bar{M} - 1) S_{d2}^2$,

где S_{d2}^2 — дисперсия внутри единиц. Отсюда

$$S_{d2}^2 = S_d^2 (1 - \rho_{\bar{M}}). \quad (11.49)$$

Из (11.48) и (11.49) мы получаем среднее значение величин $D_{i\alpha}^2$ для исходных единиц размера \bar{M} , а именно

$$E(D_{i\alpha}^2) = \frac{1}{N} \sum \bar{D}_i^2 - \frac{S_{d2}^2}{\bar{M}} = \frac{S_d^2}{\bar{M}} [1 + (\bar{M} - 1) \rho_{\bar{M}} - (1 - \rho_{\bar{M}})] = \rho_{\bar{M}} S_d^2.$$

Если \bar{M} варьирует не очень сильно, то предположение, что $\rho_{\bar{M}}$ постоянно, а следовательно, что и $E(D_{i\alpha}^2)$ постоянно, часто вполне удовлетворительно. В общем, однако, можно ожидать, что $\rho_{\bar{M}}$ будет уменьшаться при увеличении \bar{M} , поскольку подъединицы, дальше отстоящие одна от другой, менее подвержены влиянию общих факторов. Как указывают Хансен и Хервиц (Hansen and Hurwitz, 1949), степень этого уменьшения обычно достаточно мала для того, чтобы при увеличении \bar{M} произведение $\bar{M} \rho_{\bar{M}}$ все же увеличивалось и, следовательно, увеличивалось $\bar{M} E(D_{i\alpha}^2)$. Если $\rho_{\bar{M}}$ равно нулю или отрицательно, то многие из величин $D_{i\alpha}^2$ будут отрицательными, и приведенное ранее решение не годится. В этом случае двухступенчатый отбор менее результативен, чем одноступенчатый.

Перейдем теперь к определению оптимальных z_i . Согласно (11.46)

$$z_i \propto \frac{M_i D_{i\alpha}}{\sqrt{c_\alpha + c_i M_i}}.$$

Поскольку значения отдельных $D_{i\alpha}$ не известны, заменим $D_{i\alpha}$ их средним значением для единиц размера M_i , т. е. значением $\bar{D}_{\alpha, \bar{M}_i} = \sqrt{E(D_{i\alpha}^2 | M_i)}$. Теперь можно сделать следующие выводы.

1. Предположим, что $c_i M_i$, издержки на составление перечня в расчете на исходную единицу, малы по сравнению с c_α — неизменными издержками на исходную единицу. Если $\bar{D}_{\alpha, \bar{M}_i}$ — постоянная величина, то z_i пропорциональны M_i , так что отбор с вероятностями, пропорциональными размеру, будет наилучшим. Если $\bar{D}_{\alpha, \bar{M}_i}$ уменьшается при увеличении M_i , то оптимальные вероятности заключены между z_i , пропорциональными M_i , и z_i , пропорциональными $\sqrt{M_i}$.

2. Если издержки на составление перечня — основная часть затрат, то оптимальные вероятности заключены между z_i , пропорциональными $\sqrt{M_i}$, и $z_i = \text{постоянной}$ (равные вероятности извлечения).

3. Если издержки на составление перечня и неизменные издержки представляют собой величины одного порядка, то хорошее компромиссное решение дают z_i , пропорциональные $\sqrt{M_i}$.

Оптимальное значение k находим, дифференцируя (11.45) по k' . Получаем

$$k = \frac{V \sum M_i S_{d2i}^2}{\sqrt{M_0 c_2} \sum M_i D_{i\alpha} / \sqrt{c_\alpha + c_2 M_i}}. \quad (11.50)$$

Этот результат близок к результату, полученному в параграфе 11.15 для случая отбора единиц с равными вероятностями. Для того чтобы

показать это, заметим, что, как следует из (11.46) и (11.50), оптимальные $m_i = kM_i/z_i$ равны:

$$m_i = \frac{V \sum (M_i/M_0) S_{d2i}^2}{\bar{D}_u \bar{M}_i} \sqrt{(c_u + c_i M_i)/c_z} = \\ = \frac{S_{d2}^2}{S_{du}^2} \sqrt{(c_u + c_i M_i)/c_z}.$$

В таком виде полученное выражение совпадает с выражением для \bar{m}_{opt} в (11.43), если иметь в виду, что здесь роль c_1 из (11.43) играет $c_u + c_i \bar{M}$.

Наконец, оптимальное значение n находим, разрешая относительно n выражение для средних издержек.

11.17. РАССЛОЕННЫЙ ОТБОР. НЕСМЕЩЕННЫЕ ОЦЕНКИ

Для несмещенных оценок обобщение на случай расслоенного отбора производится непосредственно. Индекс h обозначает номер слоя:

M_{oh} — общее число подъединиц в слое;

$M_0 = \sum_h M_{oh}$ — общее число подъединиц в совокупности.

Оценка среднего для совокупности в расчете на подъединицу есть

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h; \quad W_h = \frac{M_{oh}}{M_0},$$

где \bar{y}_h обозначает оценку \bar{Y}_h , среднего значения для слоя в расчете на подъединицу. Далее,

$$V(\bar{y}_{st}) = \sum_h W_h^2 V(\bar{y}_h); \quad v(\bar{y}_{st}) = \sum_h W_h^2 v(\bar{y}_h).$$

Эти формулы легко получаются из приведенных ранее.

Интересно найти условия, при которых оценки становятся равновзвешенными. Для случая отбора с вероятностями, пропорциональными оценке размера (параграф 11.8), оценка среднего для совокупности согласно формуле (11.18) есть

$$\hat{Y}_{pres} = \frac{1}{M_0} \sum_h \frac{1}{n_h} \sum_i \frac{M_{hi} y_{hi}}{m_{hi} z_{hi}},$$

где y_{hi} — суммарное значение по m_{hi} подъединицам, извлеченным из i -й единицы в слое h . В параграфе 11.8 было показано, что оценка ста-

новится равновзвешенной внутри слоя при $M_{hi}/m_{hi}z_{hi} = n_h/f_{oh}$. Если это условие выполняется, то оценка принимает вид

$$\hat{Y}_{pres} = \frac{1}{M_0} \sum_h \frac{1}{f_{oh}} \sum_i y_{hi}.$$

Таким образом, оценка становится равновзвешенной полностью, если ожидаемая общая доля отбора, f_{oh} , одинакова во всех слоях.

Если внутри каждого слоя единицы имеют одинаковый размер (т. е. если $M_{hi} = M_h$), то, как было показано в параграфе 10.10, размещение выборки, приводящее к полностью равновзвешенной оценке, оказывается близким к оптимальному размещению при условии, что $S_{2h}/\sqrt{c_{2h}}$ в допустимых пределах постоянно. Аналогичное утверждение справедливо и здесь. Из (11.23) имеем

$$V(\hat{Y}_{pres}) = \frac{1}{M_0^2} \left[\sum_h \frac{1}{n_h} \sum_i \frac{M_{hi}}{z_{hi}} \left(\frac{Y_{hi}}{z_{hi}} - Y_h \right)^2 + \right. \\ \left. + \sum_h \frac{1}{f_{oh}} \sum_i M_{hi} S_{2hi}^2 \left(1 - \frac{m_{hi}}{M_{hi}} \right) \right],$$

где $M_{hi}/z_{hi}m_{hi} = n_h/f_{oh}$, если мы хотим, чтобы оценка внутри каждого слоя была равновзвешенной. Величина f_{oh} входит в последнее выражение вместе с членом

$$\frac{1}{M_0^2} \sum_h \frac{1}{f_{oh}} \sum_i M_{hi} S_{2hi}^2. \quad (11.51)$$

Множитель m_{hi}/f_{oh} , возникающий из-за пкс на второй ступени отбора, можно переписать в виде $M_{hi}/z_{hi}n_h$ и, таким образом, он представляет собой член порядка $1/n_h$, а не порядка $1/f_{oh}$.

Для функции издержек простого вида ожидаемые издержки можно выразить в виде

$$C = \sum_h c_{1h} n_h + \sum_h c_{2h} f_{oh} M_{oh}, \quad (11.52)$$

так как $f_{oh}M_{oh}$ — ожидаемое число единиц второй ступени, подлежащих извлечению в слое h . Для этой функции издержек затраты на составление перечней уже включены в c_{1h} .

Из (11.51) и (11.52) нетрудно вывести, что при неизменных издержках дисперсия минимальна, если

$$f_{oh} \propto \frac{V \sum M_{hi} S_{2hi}^2}{\sqrt{c_{2h} M_{oh}}} = \frac{1}{\sqrt{c_{2h}}} V \sum (M_{hi}/M_{oh}) S_{2hi}^2.$$

Сумма в этом выражении представляет собой взвешенное среднее величин S_{2hi}^2 . Тем самым утверждение доказано.

Оценка дисперсии получается из формулы (11.25) параграфа 11.8. Для полностью равновзвешенной оценки из (11.25') следует, что v принимает вид

$$v(\hat{Y}_{PPS}) = \frac{1}{(f_h M_h)^2} \sum_h \frac{n_h}{n_h - 1} \sum_i (y_{hi} - \bar{y}_h)^2.$$

11.18. РАССЛОЕННЫЙ ОТБОР. ОЦЕНКИ ПО ОТНОШЕНИЮ

При рассмотрении оценок по отношению возникает старый вопрос: применять ли отдельную или совместную оценку? Раздельная оценка предпочтительнее, если n_h велики в каждом слое, а истинное отношение, вероятно, от слоя к слою меняется. Формулы дисперсии для нее сразу вытекают из соответствующих формул для одного слоя.

В случае совместной оценки и отбора с вероятностями, пропорциональными оценке размера, положим

$$\hat{Y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{z_{hi}} \bar{y}_{hi}; \quad \hat{X}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{z_{hi}} \bar{x}_{hi}.$$

Величины \hat{Y}_h и \hat{X}_h представляют собой несмещенные оценки суммарных значений для слоя, Y_h и X_h . Совместная оценка по отношению имеет вид

$$\hat{R}_c = \frac{\sum_h \hat{Y}_h}{\sum_h \hat{X}_h}.$$

Для того чтобы найти приближенное значение СКО для \hat{R}_c , запишем, как обычно,

$$\hat{R} - R \approx \frac{1}{X} \sum_h (\hat{Y}_h - R \hat{X}_h).$$

Величина $\hat{Y}_h - R \hat{X}_h$ — несмещенная оценка суммарного значения для слоя, $Y_h - R X_h$, переменной $d_{hi} = y_{hi} - R x_{hi}$ при отборе с вероятностями, пропорциональными оценке размера. Следовательно, по формуле (11.23) из параграфа 11.8, подставляя d_{hi} вместо y_{hi} , получаем

$$\text{СКО}(\hat{R}_c) \approx \frac{1}{X^2} \sum_h \frac{1}{n_h} \sum_i \left[z_{hi} \left(\frac{d_{hi}}{z_{hi}} - D_h \right)^2 + \frac{M_{hi}^2 (1 - f_{2hi}) S_{2hi}^2}{z_{hi} m_{hi}} \right],$$

где

$$D_{hi} = Y_{hi} - R X_{hi};$$

$$S_{2hi}^2 = \frac{1}{M_{hi} - 1} \sum_j [(y_{hij} - R x_{hij}) - (\bar{y}_{hi} - R \bar{x}_{hi})]^2.$$

Аналогично из (11.25) получается оценка дисперсии величины R_c

$$v(\hat{R}_c) \approx \frac{1}{X^2} \sum_h \frac{1}{n_h (n_h - 1)} \sum_i (d'_{hi} - \bar{d}'_h)^2,$$

где

$$d'_{hi} = \frac{M_{hi} \bar{d}_{hi}}{z_{hi}}; \quad \bar{d}'_h = \frac{1}{n_h} \sum_i d'_{hi};$$

$$\bar{d}_{hi} = \bar{y}_{hi} - \hat{R}_c \bar{x}_{hi}.$$

Если X не известно, то вместо него подставляется оценка по выборке $\Sigma \hat{X}_h$.

Если общая доля отбора одинакова во всех слоях, то оценка \hat{R}_c сводится к отношению суммарных значений y_{hi} и x_{hi} для выборки.

11.19. ОТБОР С НЕРАВНЫМИ ВЕРОЯТНОСТЯМИ БЕЗ ВОЗВРАЩЕНИЯ

В некоторых обследованиях слои содержат по сравнительно небольшому числу исходных единиц, скажем, от 5 до 15, из которых в выборку извлекаются 2 или 3. Таким образом, исходная доля отбора может составить от 10 до 50%. В таком положении естественна попытка найти удовлетворительный метод отбора исходных единиц без возвращения, который мог бы уменьшить дисперсию с помощью некоторой поправки на конечность совокупности. Для одноступенчатого отбора некоторые основные методы такого типа были описаны в параграфах 9.14 и 9.15. В числовом примере в табл. 9.10 (с 286), где из пяти исходных единиц отбирались две, оказалось весьма примечательным, что в совокупности A , для которой \bar{Y}_i были некоррелированы с z_i , три оценки \hat{Y}_N , \hat{Y}_{SYS} и \hat{Y}_{GI} дали уменьшение дисперсии приблизительно на 40% по сравнению с \hat{Y}_{PPS} , для вычисления которой отбор производился с возвращением. (Это были оценки, полученные с помощью схем оценивания, не искажавших вероятностей извлечения.)

При двухступенчатом отборе выигрыш в точности от применения отбора без возвращения будет меньше. Формулы дисперсий показывают, что этот выигрыш затрагивает только слагаемое дисперсии, обусловленное вариацией между исходными единицами. Хотя дисперсии внутри исходных единиц для отбора с возвращением и без возвращения неодинаковы, они представляют собой величины одного порядка. Более того, если исходная доля отбора велика, то доля отбора на второй ступени будет, вероятно, небольшой, так что главную часть всей дисперсии будет составлять дисперсия на второй ступени. Эти соображения показывают, что для большинства целей необходимость в методах отбора без возвращения не слишком настоятельна. Далее опишем некоторые наиболее простые из соответствующих методов, хотя в настоящее время ни один из них не получил широкого распространения.

При $n = 2$ Йейтс и Гранди (Yates and Grundy, 1953) предложили извлекать первую единицу с вероятностью, пропорциональной неко-

торой характеристике размера, z_i , а вторую — с вероятностью, пропорциональной размеру оставшихся единиц. Оценка суммарного значения для слоя, Y , есть

$$\frac{1}{2} \left(\frac{M_i \bar{y}_i}{z_i} + \frac{M_j \bar{y}_j}{z_j} \right) = \frac{1}{2} (y'_i + y'_j).$$

Эта оценка будет смещенной, поскольку при данном способе отбора вероятности извлечения искажаются, но смещение, по-видимому, несущественно. Величина $(y'_i - y'_j)^2/4$ будет преувеличенной оценкой дисперсии.

Три остальных метода уже были описаны. Один из них состоит в том, чтобы, расположив исходные единицы в случайном порядке, извлечь систематическую выборку «каждого k -го» из соответствующих накопленных значений z_i . В выборку включается при этом каждая исходная единица, которой соответствует некоторое число из систематической выборки (Hartley and Rao, 1962). Несмещенная оценка Y есть $\sum y'_i/n$. Как и для метода Йейтса и Гранди, оценка становится равновзвешенной при $m_i/M_i = f_0/nz_i$, где f_0 — ожидаемая общая доля отбора.

При третьем и четвертом методах слой делится на n групп. Из каждой группы извлекается по одной исходной единице с вероятностью, пропорциональной ее относительному размеру внутри группы, т. е. z_i/Z_g , где $Z_g = \sum z_i$, взятой по группе (скажем, g -й), к которой относится i -я единица. Несмещенная оценка Y есть

$$\hat{Y}_G = \sum_g \frac{Z_g M_g \bar{y}_g}{z_i}. \quad (11.53)$$

Один из возможных способов формирования групп состоит в том, чтобы сделать Z_g , насколько это возможно, одинаковыми, для того чтобы сохранить вероятности извлечения пропорциональными первоначальным z_i . Полезно также, чтобы были приблизительно равны средние для групп, \bar{Y}_g , поскольку оценка дисперсии может быть получена по методу совмещенных слоев.

При четвертом методе единицы распределяются по группам случайным образом и так, чтобы число единиц в каждой группе было по возможности одинаковым (Rao, Hartley and Cochran, 1962). Оценка \hat{Y}_{G2} такая же, как (11.53). Оценка дисперсии, несмещенная при любом n , есть

$$v(\hat{Y}_{G2}) = \frac{(\sum N_g^2 - N)}{(N^2 - \sum N_g^2)} \left[\sum Z_g y_i'^2 - \hat{Y}_{G2}^2 \right] + \sum \frac{Z_g M_g^2}{z_i m_i} (1 - f_{gi}) s_{gi}^2,$$

где N_g — число единиц в g -й группе, $y'_i = M_i \bar{y}_i / z_i$ и $N = \sum N_g$. Как обычно, $s_{gi}^2 = \sum (y_{gi} - \bar{y}_g)^2 / (m_i - 1)$. Если N кратно n , так что $N_g = N/n$, то множитель, стоящий при квадратной скобке, становится равным $(1 - f_1) / (n - 1)$.

11.20. ОБЩИЕ ВЫВОДЫ

Для того чтобы разработать эффективный план многоступенчатой выборки с исходными единицами неодинакового размера, нужно проделать большую предварительную работу. Отбор исходных единиц с вероятностями, пропорциональными некоторой характеристике размера, z_i , дает наилучшие по сравнению с отбором с равными вероятностями результаты в том случае, когда отношения Y_i/z_i некоррелированы с размером, z_i , для основных признаков в обследовании, и размер сильно варьирует. Эти условия часто выполняются при выборочном исследовании, связанном с документами, когда размер исходных единиц (групп документов) определяется организационными или экономическими соображениями, а значения данных в отдельных документах имеют в разных единицах приблизительно один и тот же порядок. Прежде всего нужно сделать следующее.

1. Установите, известны ли размеры единиц точно или приближенно или неизвестны совсем. В последнем случае выясните, можно ли сравнительно легко получить какие-либо сведения о них. Например, Джессен и др. (Jessen et al., 1947) проводили двухступенчатое обследование кварталов в нескольких городах Греции, для которых не было пригодных данных о числе домохозяйств в каждом квартале. Существовало три возможности: а) отобрать кварталы с равными вероятностями; б) быстро объехать город на автомашине для того, чтобы объединить маленькие кварталы в искусственные кварталы, которые «на глаз» содержали бы примерно одинаковое число домохозяйств. При этом были бы исключены кварталы, заведомо не содержащие домохозяйств. После этого отобрать кварталы с равными вероятностями; в) объехать город достаточно медленно, для того чтобы иметь возможность оценить число домохозяйств в каждом квартале. После этого отобрать кварталы с вероятностями, пропорциональными оценкам размера.

2. Выясните, можно ли воспользоваться размером единицы как одной из переменных для расслоения: этот прием рекомендуется, если только он не мешает применить для расслоения некоторую другую переменную, которая могла бы дать значительный выигрыш в точности.

3. Решите, как должны отбираться единицы внутри слоев. Если размеры единиц хотя бы приближенно известны, то наилучшим часто будет отбор с вероятностями, пропорциональными размеру, или квадратному корню из него, хотя это зависит от характера затрат на собственно обследование.

4. Выберите некоторый метод оценивания. При оценивании суммарного или среднего значения для совокупности иногда весьма результативной оказывается оценка по отношению, в которой применяется значение соответствующего признака по последней переписи, если такие сведения имеются. Оценки, основанные на выборочном среднем или взвешенном выборочном среднем, часто бывают более точными, чем несмещенные оценки.

5. Решите, какими должны быть доли отбора и подотбора внутри слоев. Мы рекомендовали выбирать доли подотбора такими, чтобы

оценки были равнозначными. Рекомендуется также дальнейшая корректировка для того, чтобы выборка стала равнозначенной полностью, если только это не будет сопровождаться существенной потерей точности.

Различные материалы по планированию и проведению обследований, предусматривающих двухступенчатый отбор с исходными единицами неодинакового размера, содержатся в следующих работах.

Выборочное исследование по документам

Patton R. (1952). The sampling of records. Public Health Reports, 67, No. 10. (Отбор по карточкам в картотеке.)

Trueblood R. M. and Cyert R. M. (1957). Sampling techniques in accounting. Prentice-Hall. (Применение выборочного метода к изучению дебиторской задолженности).

Выборочные обследования городского населения

Bureau of the Census (1950). A chapter in population sampling. U. S. Government Printing Office.

Kish I. (1952). A two-stage sample of a city. Amer. Sociological Review, 17, 761—769.

Совокупности более общего характера

Gray P. G. and Corlett T. (1950). Sampling for the social survey. Jour. Roy. Stat. Soc., A113, 150—206.

Hemphill F. M. (1952). A sample survey of home injuries. Public Health Reports, 67.

Peaker G. F. (1953). A sampling design used by the Ministry of Education. Jour. Roy. Stat. Soc., A116, 140—165. (Обследование способностей к чтению у подростков 15 лет.)

Кроме того, см. работы Хансена, Хервица и Мэдоу, а также Йейтса*.

Упражнения

11.1. Для всех возможных выборок, которые можно извлечь из искусственной совокупности, приведенной в табл. 11.1, получите оценки с помощью методов Ia, Ib, II и III и проверьте значения общих СКО из табл. 11.2.

11.2. С помощью методов II (отбор с равными вероятностями, несмещенная оценка) и III (отбор с вероятностями, пропорциональными размеру) вычислите дисперсии \hat{y} для примера из табл. 11.1 при $m_1 = 1$. Покажите, что точность метода III по сравнению с методом II при $m_1 = 1$ меньше, чем при $m_1 = 2$. Какой общий вывод можно сделать из этого факта?

11.3. Для совокупности из табл. 11.1, если оценки размера z_i есть 0,1; 0,3 и 0,6, а $m_1 = 2$, покажите, что несмещенная оценка (метод IV) дает меньшую дисперсию, чем отбор с вероятностями, пропорциональными размеру. Чем объяснить этот результат?

11.4. Элементы совокупности с тремя исходными единицами разделены на два класса. Размер единиц M_i и доля элементов, принадлежащих первому классу, P_i равны:

$M_1 = 100$; $M_2 = 200$; $M_3 = 300$; $P_1 = 0,40$; $P_2 = 0,45$; $P_3 = 0,35$.

Для выборки, содержащей 50 элементов из одной исходной единицы, сравните

* Есть русский перевод: Йейтс Ф. Выборочный метод в переписях и обследованиях. М., «Статистика», 1965. — Примеч. ред.

СКО для методов Ia, II и III при оценивании доли элементов совокупности, принадлежащих первому классу. (В формулах дисперсии из параграфа 11.2 S_i^2 приближенно равно $P_i Q_i$.)

11.5. В выборку отобрано с равными вероятностями n исходных единиц. В каждой отобранной единице извлекается одинаковая доля f_2 подъединиц. Пусть a_i из m_i подъединиц, извлеченных в i -й единице, принадлежат классу C. Покажите, что оценка доли элементов в совокупности из класса C, полученная как оценка по отношению с размером в знаменателе (параграф 11.6), имеет вид $\bar{p} = \sum a_i / \sum m_i$. Из формулы (11.12) выведите, что оценка СКО (\bar{p}) имеет вид

$$v(\bar{p}) = \frac{1-f_1}{n\bar{M}^2} \frac{\sum M_i^2 (p_i - \bar{p})^2}{n-1} + \frac{f_1(1-f_2)}{n^2 \bar{m} \bar{M}} \sum \frac{M_i m_i}{m_i - 1} p_i q_i,$$

где $p_i = a_i / m_i$.

11.6. Фирма, имеющая 36 предприятий, хочет проверить состояние некоторого вида оборудования. $M_0 = 25\ 012$ единиц которого находятся в работе. Извлекается случайная выборка, состоящая из 12 предприятий, и на каждом из них проверяется 10%-ная подвыборка. Число проверенных единиц оборудования (m_i) и число единиц с обнаруженными следами повреждения (a_i) приводятся в таблице.

Пред- приятие	m_i	a_i	$p_i = \frac{a_i}{m_i}$	Пред- приятие	m_i	a_i	$p_i = \frac{a_i}{m_i}$
1	65	8	0,123	7	85	18	0,212
2	82	21	0,256	8	73	11	0,151
3	52	4	0,077	9	50	7	0,140
4	91	12	0,132	10	76	9	0,118
5	62	1	0,016	11	64	20	0,312
6	69	3	0,043	12	50	2	0,040

Оцените процент и общее число поврежденных единиц оборудования и найдите оценки их стандартных ошибок.

Замечание. Поскольку $M_i / \bar{M} \approx m_i / \bar{m}$, междуединичное слагаемое дисперсии $v(\bar{p})$ можно вычислять по формуле

$$\frac{1-f_1}{n\bar{m}^2 (n-1)} (\sum a_i^2 - 2\bar{p} \sum a_i m_i + \bar{p}^2 \sum m_i^2),$$

и так как m_i довольно велики, то внутриединичное слагаемое дисперсии находят по формуле

$$\frac{f_1(1-f_2)}{(n\bar{m})^2} \sum a_i q_i.$$

11.7. Покажите, что если исходные единицы отбираются с равными вероятностями и f_2 — постоянная величина, то в обозначениях упражнения 11.5 несмещенная оценка доли для совокупности имеет вид $\bar{p} = N \sum a_i / n M_i f_2$, и что если пренебречь членами порядка $1/m_i$, то дисперсию этой оценки можно найти по формуле

$$v(\bar{p}) = \frac{1-f_1}{n(n-1)\bar{m}^2} \sum (a_i - \bar{a})^2 + \frac{f_1(1-f_2)}{(n\bar{m})^2} \sum a_i q_i.$$

Вычислите \bar{p} и ее стандартную ошибку для данных из упражнения 11.6.

11.8. Выборка объемом n из исходных единиц извлекается с вероятностями, пропорциональными оценке размера x_i (с возвращением), и принята постоянная ожидаемая общая доля отбора f_0 . Покажите, что при оценивании суммарного значения для совокупности несмещенная оценка и оценка по отношению к размеру в знаменателе имеют вид, соответственно, T/f_0 и $TM_0/\sum m_i$, где T — суммарное значение для выборки. (Отсюда следует, что если M_0 не известно, то можно применять несмещенную оценку, но не оценку по отношению к размеру в знаменателе. При оценивании среднего значения для совокупности на подыединицу положение обратное.)

11.9. При исследовании жилищных условий в большом городе один из слоев содержал 100 кварталов, 10 из которых были извлечены с вероятностями, пропорциональными оценке размера (с возвращением). Была принята ожидаемая общая доля отбора $f_0 = 2\%$. Получены следующие суммарные значения количества комнат и числа людей для выборки в каждом квартале:

Квартал	1	2	3	4	5	6	7	8	9	10
Количество комнат	60	52	58	56	62	51	72	48	71	58
Число людей	115	80	82	93	105	109	130	93	109	95

(а) Оцените общее количество комнат, общее число людей в слое и среднее число людей на одну комнату. (б) Вычислите стандартные ошибки суммарного числа людей и среднего числа людей, приходящегося на одну комнату. Воспользуйтесь формулами (11.25') и (11.39), считая, что $m_{2i}/m_1 = f_0/n$.

ЛИТЕРАТУРА

- Gray P. G. and Corlett T. (1950). Sampling for the social survey. *Jour. Roy. Stat. Soc.*, A113, 150—206.
- Hansen M. H. and Hurwitz W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Stat.*, 14, 333—362.
- Hansen M. H. and Hurwitz W. N. (1949). On the determination of the optimum probabilities in sampling. *Ann. Math. Stat.*, 20, 426—432.
- Hartley H. O. and Rao J. N. K. (1962). Sampling with unequal probabilities without replacement. *Ann. Math. Stat.*, 33, 350—374.
- Jebe E. H. (1952). Estimation for sub-sampling designs employing the county as a primary sampling unit. *Jour. Amer. Stat. Assoc.*, 47, 49—70.
- Jessen R. J. et al. (1947). On a population sample for Greece. *Jour. Amer. Stat. Assoc.*, 42, 357—384.
- Madow L. H. (1950). On the use of the county as a primary sampling unit for state estimates. *Jour. Amer. Stat. Assoc.*, 45, 30—47.
- Rao J. N. K., Hartley H. O. and Cochran W. G. (1962). A simple procedure of unequal probability sampling without replacement. *Jour. Roy. Stat. Soc.*, B24.
- Sukhatme P. V. (1954). *Sampling theory of surveys, with applications*. Iowa State College Press, Ames, Iowa.
- Yates F. and Grundy P. M. (1953). Selection without replacement from strata with probability proportional to size. *Jour. Roy. Stat. Soc.*, B15, 253—261.

ГЛАВА 12

ДВОЙНОЙ ОТБОР

12.1. ОПИСАНИЕ МЕТОДА

Как мы уже видели, целый ряд способов отбора опирается на предварительные сведения о некоторой вспомогательной переменной x_i . Для того чтобы применять оценки по отношению и оценки по регрессии, нужно знать среднее значение для совокупности, \bar{X} . Если мы хотим произвести расслоение совокупности по значениям x_i , то должно быть известно распределение частот этой переменной.

Если такие сведения отсутствуют, то иногда можно с относительно малыми затратами получить предварительную выборку большого объема, в которой наблюдается только x_i . Задача этой выборки — обеспечить хорошую оценку \bar{X} или распределения частот x_i . В обследовании, цель которого состоит в получении оценок для некоторой другой переменной y_i , может оказаться целесообразным выделить часть ресурсов на эту предварительную выборку, хотя это и означает, что объем выборки в основном обследовании должен быть уменьшен. Этот метод известен под названием *двойного отбора* или *двухфазного отбора*. Как вытекает из сказанного, этот метод выгоден лишь в том случае, когда выигрыш в точности от применения оценки по отношению или оценки по регрессии или от расслоения значительно перевешивает потерю в точности, вызванную уменьшением объема основной выборки.

Двойной отбор может оказаться полезным в том случае, когда данные об x_i записаны на карточках, но не сведены в таблицу. Например, при обследованиях гражданского населения Германии в 1945 г. выборка для того или иного города обычно извлекалась из регистрационных списков на получение продовольствия. Кроме географического расслоения внутри города, для которого данные обычно уже имелись, требовалось расслоение также по полу и возрасту. Поскольку выборку нужно было извлечь срочно, а списки постоянно находились в употреблении, получить полное распределение населения по полу и возрасту оказалось невозможно. Однако можно было быстро извлечь систематическую выборку довольно большого объема. Все включенные в нее лица были распределены по соответствующим возрастно-половым группам. По этим данным была получена уже гораздо меньшая выборка лиц, подлежащих опросу.

Теоретическое обоснование метода впервые было дано Нейманом (Neuman, 1938).

Совокупность должна быть расслоена на некоторое число классов в соответствии со значениями x_i . Первая выборка представляет собой простую случайную выборку объема n' . Пусть

$W_h = N_h/N$ — доля единиц совокупности, относящихся к слою h ;
 $w_h = n'_h/n'$ — доля единиц первой выборки, относящихся к слою h .
 Тогда w_h служит оценкой W_h .

Вторая выборка представляет собой расслоенную случайную выборку объема n , в которой наблюдается y_i : из слоя h извлекается n_h единиц. Второй выборкой часто служит некоторая подвыборка из первой выборки, но если это удобнее, ее можно извлечь и независимо.

Предполагается, что издержки на получение обеих выборок есть

$$C = nc_n + n'c_{n'}, \quad (12.1)$$

где c_n обычно велико по сравнению с $c_{n'}$.

Задача заключается в том, чтобы выбрать n' и n_h (а следовательно, и n) так, чтобы они минимизировали дисперсию оценки при данных издержках. Затем мы должны проверить, будет ли эта минимальная дисперсия меньше той, которую может дать одинарная простая случайная выборка, где наблюдаются только y_i .

Первый шаг состоит в том, чтобы построить оценку и определить ее дисперсию. Среднее для совокупности есть

$$\bar{Y} = \sum_{h=1}^L W_h \bar{Y}_h.$$

В качестве оценки возьмем

$$\bar{y}_{st} = \sum_{h=1}^L w_h \bar{y}_h.$$

Если извлекается новая выборка, то в данном случае это означает новое извлечение как первой, так и второй выборки. Поэтому как w_h , так и выборочные средние \bar{y}_h будут случайными величинами, подверженными ошибке выборки. Таким образом, задача сводится к случаю расслоения, при котором суммарные значения для слоев точно не известны. При многократном отборе границы слоев предполагаются неизменными.

Теоремы о среднем и о дисперсии \bar{y}_{st} доказываются при небольшом допущении. Предполагается, что каждое $w_h > 0$, т. е. n' достаточно велико, чтобы вероятностью того, что какой-либо слой не будет представлен в большой выборке, можно было пренебречь.

Теорема 12.1. Оценка \bar{y}_{st} есть несмещенная оценка.

Доказательство. Сначала возьмем среднее по всем выборкам при условии, что w_h неизменны. Поскольку \bar{y}_h представляет собой среднее для простой случайной выборки из слоя, $E(\bar{y}_h) = \bar{Y}_h$. Однако, по-

скольку первая выборка также представляет собой простую случайную выборку, среднее по всем возможным извлечениям первой выборки дает $E(w_h) = W_h$. Следовательно,

$$E(\bar{y}_{st}) = E[E(\sum w_h \bar{y}_h | w_h)] = E(\sum w_h \bar{Y}_h) = \sum W_h \bar{Y}_h = \bar{Y}.$$

Теорема 12.2. Если значения n_h не зависят от w_h , то

$$V(\bar{y}_{st}) = \sum_{h=1}^L \left\{ \left[W_h^2 + \frac{g' W_h (1 - W_h)}{n'} \right] \frac{(1 - f_h) S_h^2}{n_h} + \frac{g' W_h (\bar{Y}_h - \bar{Y})^2}{n'} \right\}, \quad (12.2)$$

где $g' = (N - n')/(N - 1)$ и $f_h = n_h/N_h$.

Доказательство. Сначала возьмем среднее по всем выборкам при условии, что w_h неизменны. Среднее значение \bar{y}_{st} по таким выборкам есть $\sum w_h \bar{Y}_h$, так что имеется смещение, равное $\sum (w_h - W_h) \bar{Y}_h$. Значение условной дисперсии \bar{y}_{st} дает теорема 5.3. Следовательно, средний квадрат ошибки равен:

$$E[(\bar{y}_{st} - \bar{Y})^2 | w_h] = \sum_{h=1}^L \frac{w_h^2 (1 - f_h) S_h^2}{n_h} + \left[\sum_{h=1}^L (w_h - W_h) \bar{Y}_h \right]^2. \quad (12.3)$$

Теперь возьмем среднее по всем первым выборкам, когда w_h меняются. (Здесь мы пользуемся предположением о том, что когда w_h меняются, n_h остаются постоянными.) По теореме 3.2

$$V(w_h) = \frac{g' W_h (1 - W_h)}{n'},$$

так что

$$E(w_h^2) = [E(w_h)]^2 + V(w_h) = W_h^2 + \frac{g' W_h (1 - W_h)}{n'}. \quad (12.4)$$

Легко показать также, что

$$E(w_h - W_h)(w_j - W_j) = -\frac{g'}{n'} W_h W_j \quad (h \neq j).$$

Из этих равенств для последнего члена в (12.3) следует, что

$$\begin{aligned} E \left[\sum_{h=1}^L (w_h - W_h) \bar{Y}_h \right]^2 &= \frac{g'}{n'} \left[\sum_{h=1}^L W_h (1 - W_h) \bar{Y}_h^2 - \right. \\ &\quad \left. - 2 \sum_{h=1}^L \sum_{j>h}^L W_h W_j \bar{Y}_h \bar{Y}_j \right] = \frac{g'}{n'} \left(\sum_{h=1}^L W_h \bar{Y}_h^2 - \bar{Y}^2 \right) = \\ &= \frac{g'}{n'} \sum W_h (\bar{Y}_h - \bar{Y})^2. \end{aligned} \quad (12.5)$$

Окончательно, пользуясь (12.4) и (12.5), получаем из (12.3)

$$V(\bar{y}_{st}) = \sum_h^L \left\{ \left[W_h^2 + \frac{g' W_h (1 - W_h)}{n'} \right] \frac{(1 - f_h) S_h^2}{n_h} + \frac{g' W_h (\bar{Y}_h - \bar{Y})^2}{n'} \right\}.$$

В этом выражении члены, не содержащие n' , представляют собой знакомое нам выражение для дисперсии в случае, когда объемы слоев известны точно. Эффект ошибок первой выборки проявляется, следовательно, в небольшом увеличении слагаемого, обусловленного вариацией внутри слоев, и в появлении слагаемого, обусловленного вариацией между слоями.

В большинстве приложений $f_h = n_h/N_h$ пренебрежимо малы. Часто n'/N также мало, так что вместо g' можно подставить 1. При этих условиях

$$V(\bar{y}_{st}) = \sum_h^L \left\{ \left[W_h^2 + \frac{W_h (1 - W_h)}{n'} \right] \frac{S_h^2}{n_h} + \frac{W_h (\bar{Y}_h - \bar{Y})^2}{n'} \right\}. \quad (12.6)$$

Следствие 1. Если n_h зависят от w_h , утверждение теоремы 12.2 немного изменится. Например, если мы хотим получить пропорциональное расслоение, то нужно положить для малой выборки $n_h = n w_h$, поскольку w_h — наилучшие имеющиеся оценки W_h . В более общем виде мы можем взять $n_h = n \lambda_h w_h / \sum \lambda_h w_h$. Подставим эти значения в (12.3) и затем возьмем среднее по всем возможным извлечениям w_h . После некоторых алгебраических преобразований, пренебрегая n_h/N_h , получаем, что дисперсия равна:

$$V(\bar{y}_{st}) = \sum_h^L \left\{ \frac{W_h S_h^2}{n} \left[\frac{Q}{\lambda_h} + \frac{g'}{n'} \left(1 - \frac{Q}{\lambda_h} \right) \right] + \frac{g' W_h (\bar{Y}_h - \bar{Y})^2}{n'} \right\},$$

где $Q = \sum \lambda_h W_h$. Если $\lambda_h = 1$ (пропорциональное расслоение), то $Q = 1$, и мы имеем

$$V(\bar{y}_{st}) = \sum_h^L \left\{ \frac{W_h S_h^2}{n} + \frac{g' W_h (\bar{Y}_h - \bar{Y})^2}{n'} \right\}.$$

Следствие 2. Если по второй выборке должна быть получена оценка доли, то

$$S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h \approx P_h Q_h;$$

$$(\bar{Y}_h - \bar{Y})^2 = (P_h - P)^2.$$

Пренебрегая n_h/N_h , получаем из теоремы 12.2

$$V(p_{st}) \approx \sum_h^L \left\{ \left[W_h^2 + \frac{g' W_h (1 - W_h)}{n'} \right] \frac{P_h Q_h}{n_h} + \frac{g' W_h (P_h - P)^2}{n'} \right\}, \quad (12.7)$$

где P_h — доля в слое h .

В работах Робсона (Robson, 1952) и Робсона и Кинга (Robson and King, 1953) изложенные выше теоретические положения обобщены на случай двухступенчатого отбора и применяются для оценивания читательской аудитории журналов.

12.3. ОПТИМАЛЬНОЕ РАЗМЕЩЕНИЕ

Получение значений n_h и n' , при которых дисперсия минимальна, довольно сложно. Из формулы (12.2) и вида функции издержек следует, что если n' и n заданы, то n_h должны быть пропорциональны

$$S_h \sqrt{W_h^2 + [g' W_h (1 - W_h)]/n'}.$$

Поскольку второй член под знаком корня обычно мал по сравнению с первым, Нейман (Neuman, 1938) предложил брать n_h пропорциональными $W_h S_h$. Таким образом,

$$n_h = \frac{n W_h S_h}{\sum W_h S_h}.$$

Если эти значения подставить в дисперсию (12.2), опуская член при $W_h (1 - W_h)$, и предположить, что n_h/N_h и n'/N пренебрежимо малы, то получим

$$V_{opt} \approx \frac{(\sum W_h S_h)^2}{n} + \frac{\sum W_h (\bar{Y}_h - \bar{Y})^2}{n'}. \quad (12.8)$$

Или, вводя обозначения V_n и $V_{n'}$,

$$V_{opt} \approx \frac{V_n}{n} + \frac{V_{n'}}{n'}. \quad (12.8')$$

Минимизируем теперь это приближенное значение дисперсии, выбирая n и n' при данных издержках, выраженных функцией вида

$$C = n c_n + n' c_{n'}. \quad (12.1)$$

Нетрудно проверить, что минимум достигается при

$$\frac{n}{\sqrt{V_n c_n}} = \frac{n'}{\sqrt{V_{n'} c_{n'}}}. \quad (12.9)$$

Из этого равенства и (12.1) определяются n и n' .

Для дальнейших приложений двойного отбора нам понадобится выражение для минимума дисперсии. Из (12.9)

$$\begin{aligned} \frac{n}{\sqrt{V_n c_n}} &= \frac{n'}{\sqrt{V_{n'} c_{n'}}} = \frac{n c_n + n' c_{n'}}{\sqrt{c_n c_{n'}} (\sqrt{V_n c_n} + \sqrt{V_{n'} c_{n'}})} = \\ &= \frac{C}{\sqrt{c_n c_{n'}} (\sqrt{V_n c_n} + \sqrt{V_{n'} c_{n'}})}. \end{aligned} \quad (12.9')$$

Подставим эти результаты в формулу (12.8') для V_{opt} . Это дает

$$V_{opt} = \frac{(\sqrt{V_n c_n} + \sqrt{V_{n'} c_{n'}})^2}{C}. \quad (12.10)$$

Если первая выборка обходится очень дешево, то предположение, что n'/N пренебрежимо мало, т. е. что $g' = 1$, может не соответствовать действительности. В этом случае можно по-прежнему пользоваться значениями n и n' из (12.9), но нужно только вычесть из (12.10) член $V_{n'}/N$.

Пример. Этот пример искусственный, но он иллюстрирует необходимые вычисления. Рассмотрим приведенные ранее данные по графству Джефферсон (с. 190). Переменная x_i , размер фермы, применяется для разбиения совокупности на два слоя: фермы размером до 160 акров и более 160 акров. Предположим, что определение числа акров под пшеницей на одной ферме (y_i) обходится в 10 раз дороже, чем определение размера фермы (x_i) и пусть издержки составляют

$$C = 100 = n + 0,1 n'. \quad (12.11)$$

Это означает, что если не применять двойного отбора ($n' = 0$), то для оценивания площади под пшеницей мы могли бы взять выборку объемом в 100 ферм.

Необходимые данные о совокупности имеют вид

Слой	W_h	S_h^2	S_h	\bar{Y}_h
1	0,786	312	17,7	19,404
2	0,214	922	30,4	51,626
Совокупность		620		26,297

Согласно (12.10) мы можем сразу начать вычислять V_{opt} . Приведем, однако, и промежуточные вычисления. Находим

$$V_n = (\sum W_h S_h)^2 = 417;$$

$$V_{n'} = \sum W_h (\bar{Y}_h - \bar{Y})^2 = 175,$$

так что согласно (12.9)

$$\frac{n}{n'} = \sqrt{\frac{417}{175} \cdot \frac{1}{10}} = 0,488.$$

Из уравнения издержек (12.11) получаем

$$n' = \frac{100}{0,588} = 170; \quad n = 170 \cdot 0,488 = 83.$$

Здесь читатель может проверить по данным нашего примера, что опущенный в формуле дисперсии (12.2) член, содержащий $W_h (1 - W_h)$, действительно пренебрежимо мал. Тогда из (12.8) получаем

$$V_{opt} = \frac{417}{83} + \frac{175}{170} = 5,02 + 1,03 = 6,05.$$

Для случайной выборки объема 100 без двойного отбора мы имели бы

$$V = \frac{620}{100} = 6,20.$$

Очевидно, что применение двойного отбора дало бы лишь пустячный выигрыш.

Замечание. Возвращаясь к следствию 1 теоремы 12.2, укажем другой возможный подход к выбору n_h . Он состоит в том, чтобы принять $n_h = n \lambda_h w_h / \sum \lambda_h w_h$ и искать λ_h , минимизирующие дисперсию $V(\bar{y}_{st})$, представленную формулой из следствия 1. Можно показать, что оптимальные λ_h пропорциональны S_h в предположении, что $1/n'$ можно пренебречь. Для практических целей этот результат совпадает с уже полученным.

12.4. ОЦЕНКА ДИСПЕРСИИ ПРИ ДВОЙНОМ ОТБОРЕ ДЛЯ РАССЛОЕНИЯ

Несмещенную оценку дисперсии $V(\bar{y}_{st})$ из (12.2) можно построить без особого труда. Мы предполагаем, что n_h/N_h и $1/N$ можно пренебречь.

Теорема 12.3. Несмещенная оценка $V(\bar{y}_{st})$ есть

$$v(\bar{y}_{st}) = \frac{n'}{n' - 1} \sum_h \left\{ \left[w_h^2 - \frac{g' w_h}{n'} \right] \frac{s_h^2}{n_h} + \frac{g' w_h (\bar{y}_h - \bar{y}_{st})^2}{n'} \right\}, \quad (12.12)$$

где $g' = (N - n')/(N - 1)$.

Доказательство. Беря сначала средние по выборкам с неизменными w_h и затем по всем возможным извлечениям w_h , получаем следующие значения математических ожиданий членов внутри скобок:

$$E \sum_h w_h^2 \frac{s_h^2}{n_h} = \sum_h \left[W_h^2 + \frac{g' W_h (1 - W_h)}{n'} \right] \frac{S_h^2}{n_h}; \quad (12.13)$$

$$- E \sum_h \frac{g' w_h}{n'} \frac{s_h^2}{n_h} = - \sum_h \frac{g' W_h}{n'} \frac{S_h^2}{n_h}; \quad (12.14)$$

$$E \sum_h \frac{g' w_h (\bar{y}_h - \bar{y}_{st})^2}{n'} = E \left(\sum_h \frac{g' w_h \bar{y}_h^2}{n'} - \frac{g' \bar{y}_{st}^2}{n'} \right) = \\ = \sum_h \frac{g' W_h \bar{Y}_h^2}{n'} + \sum_h \frac{g' W_h}{n'} \frac{S_h^2}{n_h} - \frac{g' \bar{Y}^2}{n'} - \frac{g' V(\bar{y}_{st})}{n'}. \quad (12.15)$$

Складывая эти три равенства, получаем, сравнивая с (12.2) и полагая, что $1/N$ можно пренебречь,

$$\frac{(n' - 1) E v(\bar{y}_{st})}{n'} = V(\bar{y}_{st}) \left(1 - \frac{g'}{n'} \right) = V(\bar{y}_{st}) \frac{N(n' - 1)}{(N - 1)n'} = \\ = V(\bar{y}_{st}) \frac{(n' - 1)}{n'}.$$

Теорема доказана.

Если n' велико по сравнению с n_h , то $v(\bar{y}_{st})$ сводится к

$$v(\bar{y}_{st}) \approx \sum_h w_h^2 \frac{s_h^2}{n_h} \quad (12.16)$$

Применение этого выражения эквивалентно предположению о том, что ошибками весов w_h для слоев можно пренебречь.

Следствие. Пусть p_h — наблюдаемая доля единиц в слое h , принадлежащих некоторому определенному классу, и $p_{st} = \sum w_h p_h / \sum w_h$ есть оценка соответствующей доли для совокупности. Тогда оценка $V(p_{st})$ есть

$$V(p_{st}) = \frac{n'}{n'-1} \sum_h \left[\left(w_h^2 - \frac{g' w_h}{n'} \right) \frac{p_h q_h}{n_h - 1} + \frac{g' w_h (p_h - p_{st})^2}{n'} \right]$$

Почти во всех случаях это выражение можно заменить более простым:

$$v(p_{st}) \approx \sum_h \left[\frac{w_h^2 p_h q_h}{n_h - 1} + \frac{w_h (p_h - p_{st})^2}{n'} \right]$$

Часто член, содержащий $1/n'$, также можно опустить.

Пример. В простой случайной выборке, содержащей 374 дома, 292 дома заняты белыми семьями и 82 — небелыми. Подвыборка приблизительно каждого четвертого дома дала следующие сведения о числе собственных и сдаваемых внаем домов:

	Собственные дома	Дома, сдаваемые внаем	Всего
Белые семьи	31	43	74
Небелые семьи	4	14	18

Требуется оценить долю домов, сдаваемых внаем, в районе, где отбиралась выборка, и найти стандартную ошибку этой оценки.

Пусть первый слой состоит из домов, в которых живут белые семьи. Тогда:

$$w_1 = \frac{292}{374} = 0,78; \quad w_2 = \frac{82}{374} = 0,22;$$

$$p_1 = \frac{43}{74} = 0,60; \quad p_2 = \frac{14}{18} = 0,78;$$

$$p_{st} = w_1 p_1 + w_2 p_2 = 0,64;$$

$$n' = 374; \quad n_1 = 74; \quad n_2 = 18.$$

Нетрудно проверить, что существенное значение имеет только главный член $v(p_{st})$. Следовательно,

$$v(p_{st}) = \sum_h \frac{w_h^2 p_h q_h}{n_h - 1} = \frac{(0,78)^2 \cdot 0,60 \cdot 0,40}{73} + \frac{(0,22)^2 \cdot 0,78 \cdot 0,22}{17} = 0,00248;$$

$$s(p_{st}) = 0,049.$$

Оценка доли сдаваемых внаем домов равна $0,64 \pm 0,049$. Читатель может проверить, что по сравнению с одноступенчатой простой случайной выборкой объема 92 получается лишь пустячный выигрыш в точности. Ввиду относительно небольшой величины слоя, состоящего из домов небелых семей, для того чтобы двойной отбор был выгодным, разница в доле снимающих жилище между белыми и небелыми семьями должна быть больше.

12.5. ОЦЕНКИ ПО РЕГРЕССИИ

В некоторых случаях применения двойного отбора вспомогательной переменной x_i можно воспользоваться для получения оценки по регрессии величины \bar{Y} . Мы предположим, что совокупность бесконечна и что зависимость между y_i и x_i линейна. Рассмотрим модель

$$y_{i\alpha} = \bar{Y} + B(x_i - \bar{X}) + e_{i\alpha}, \quad (12.17)$$

где второй индекс α вводится для того, чтобы напомнить, что при неизменном x_i случайная переменная $e_{i\alpha}$ подчиняется некоторому распределению частот со средним 0 и дисперсией $S_e^2 = S_y^2(1 - \rho^2)$.

В первой (большой) выборке объема n' мы наблюдаем только значения x_i ; во второй (малой) выборке объема n , мы наблюдаем значения как x_i , так и $y_{i\alpha}$. Оценка \bar{Y} есть

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x}),$$

где \bar{x}' , \bar{x} — средние значения x_i соответственно для первой и второй выборок и b — оценка коэффициента регрессии $y_{i\alpha}$ по x_i , полученная по второй выборке методом наименьших квадратов.

Исследуем теперь ошибку оценки $(\bar{y}_{lr} - \bar{Y})$. Из (12.17) находим

$$\bar{y} = \bar{Y} + B(\bar{x} - \bar{X}) + \bar{e}; \quad (12.18)$$

$$b = \frac{\sum_{i=1}^n (y_{i\alpha} - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$

$$= B + \frac{\sum_{i=1}^n e_{i\alpha}(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (12.19)$$

Подставим полученные выражения для \bar{y} и b в выражение для ошибки оценки. Это дает

$$\bar{y}_{lr} - \bar{Y} = (\bar{y} - \bar{Y}) + b(\bar{x}' - \bar{x}) = B(\bar{x} - \bar{X}) + \bar{e} + B(\bar{x}' - \bar{x}) +$$

$$+ (\bar{x}' - \bar{x}) \frac{\sum e_{i\alpha}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \bar{e} + (\bar{x}' - \bar{x}) \frac{\sum e_{i\alpha}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} + B(\bar{x}' - \bar{X}). \quad (12.20)$$

В обычном регрессионном анализе, когда $\bar{x}' = \bar{X}$, как правило, рассматривается условное распределение частот ошибки оценки $(\bar{y}_{lr} - \bar{Y})$ в многократных выборках при условии, что значения x_i неизмен-

ны. В данном случае при таком подходе, если считать неизменными значения x_i как в первой, так и во второй выборках, оценка окажется смещенной относительно среднего значения условного распределения, поскольку [с — от английского «conditional» — условный]

$$E(\bar{y}_{lr} - \bar{y}) = B(\bar{x}' - \bar{X}).$$

Следовательно, условный средний квадрат ошибки (СКО) оценки \bar{y}_{lr} равен:

$$\text{СКО}(\bar{y}_{lr}) = S_y^2(1 - \rho^2) \left[\frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] + B^2(\bar{x}' - \bar{X})^2. \quad (12.21)$$

Для сравнения с другими методами отбора это выражение неудобно, поскольку СКО зависит от конкретного набора x_i , которые были получены в двух выборках. Вместо этого нам нужно среднее значение СКО по всем возможным извлечениям первой и второй выборок.

Простой результат получается, если предположить, что (а) первая выборка извлекается случайным образом, (б) вторая выборка представляет собой случайную подвыборку из первой и (в) x_i распределены нормально. В этом случае среднее значение СКО оказывается равным:

$$V(\bar{y}_{lr}) = S_y^2(1 - \rho^2) \left[\frac{1}{n} + \left(\frac{1}{n} - \frac{1}{n'} \right) \frac{1}{(n-3)} \right] + \frac{B^2 S_x^2}{n'} = \quad (12.22)$$

$$= \frac{S_y^2(1 - \rho^2)}{n} \left[1 + \frac{(n' - n)}{n'} \frac{1}{(n-3)} \right] + \frac{\rho^2 S_y^2}{n'}, \quad (12.23)$$

поскольку $B^2 S_x^2 = \rho^2 S_y^2$.

Если распределение x_i отличается от нормального, то изменится только член, содержащий $1/(n-3)$. Что касается предположения (б), то малая выборка не обязательно должна извлекаться из большой случайным образом: малую выборку желательно извлекать так для того, чтобы получить больший диапазон значений x_i и уменьшить ошибку выборки для b . Эффект этого состоит в уменьшении, возможно значительном, члена, содержащего $1/(n-3)$.

В некоторых приложениях вторая выборка извлекается независимо от первой. В этом случае рассуждения данного параграфа остаются без изменений вплоть до формулы (12.21). В формуле (12.22) член

$$\frac{1}{n} - \frac{1}{n'}$$

заменяется членом

$$\frac{1}{n} + \frac{1}{n'}.$$

Впервые случай двух независимых выборок был рассмотрен Чемели Боз (Chameli Bose, 1943).

Итак, точное значение члена, содержащего $1/(n-3)$ в выражении для средней дисперсии, остается несколько неопределенным. Однако если $1/n$ пренебрежимо мало, то этим членом также можно пренебречь. Это дает основание сформулировать следующую теорему.

Теорема 12.4. Если первая выборка имеет объем n' , вторая — объем n и $1/n$ пренебрежимо мало, то дисперсия \bar{y}_{lr} , оценки по регрессии при двойном отборе, приближенно равна:

$$V(\bar{y}_{lr}) \approx \frac{S_y^2(1 - \rho^2)}{n} + \frac{\rho^2 S_y^2}{n'}. \quad (12.24)$$

12.6. СРАВНЕНИЕ ДВОЙНОГО ОТБОРА ДЛЯ ОЦЕНКИ ПО РЕГРЕССИИ С ОДИНАРНЫМ ОТБОРОМ

Пользуясь формулой дисперсии (12.24), двойной отбор для оценки по регрессии можно сравнить с одинарным простым случайным отбором в предположении, что (а) первая выборка представляет собой простую случайную выборку, (б) $1/n$ пренебрежимо мало и (в) вторая выборка также представляет собой простую случайную выборку. Результаты, относящиеся к этому случаю, могут служить общим ориентиром и в других случаях.

Запишем

$$V(\bar{y}_{lr}) = \frac{V_n}{n} + \frac{V_{n'}}{n'},$$

где

$$V_n = S_y^2(1 - \rho^2); \quad V_{n'} = \rho^2 S_y^2;$$

$$\text{издержки} = C = nc_n + n'c_{n'}.$$

Задача нахождения оптимальных n и n' и минимальной дисперсии в точности та же, что и при двойном отборе для расслоения (параграф 12.3). Равенство (12.10) дает

$$V_{opt} = \frac{(V_n c_n + V_{n'} c_{n'})^2}{C} = \frac{S_y^2 [V(1 - \rho^2) c_n + \rho^2 c_{n'}]^2}{C}, \quad (12.25)$$

где ρ считается положительным.

Если все средства расходуются на одинарную выборку без оценки по регрессии, то эта выборка имеет объем $n_s = C/c_n$ [с — от английского «single» — одинарный] и дисперсия ее среднего равна:

$$V(\bar{y}) = \frac{S_y^2}{n_s} = \frac{c_n S_y^2}{C}. \quad (12.26)$$

Следовательно, двойной отбор дает меньшую дисперсию, если

$$c_n > [V(1 - \rho^2) c_n + \rho^2 c_{n'}]^2$$

Это неравенство можно записать двумя способами:

$$\frac{c_n}{c_{n'}} > \frac{(1 + \sqrt{1 - \rho^2})^2}{\rho^2} = \frac{\rho^2}{(1 - \sqrt{1 - \rho^2})^2} \quad (12.27)$$

$$\rho^2 > \frac{4c_n c_{n'}}{(c_n + c_{n'})^2} \quad (12.28)$$

Неравенство (12.27) показывает, что при некотором данном значении ρ двойной отбор приведет к увеличению точности лишь тогда, когда отношение издержек в расчете на единицу во второй выборке к издержкам в расчете на единицу в первой выборке превзойдет некоторое

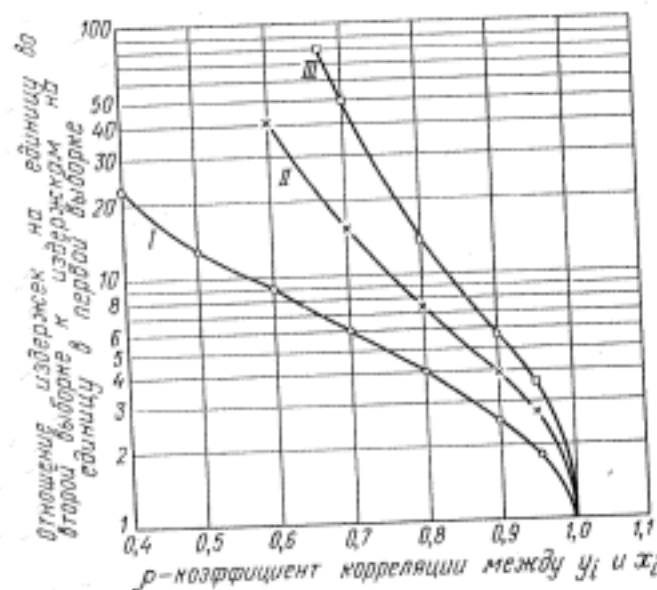


Рис. 12.1. Зависимость между $c_n/c_{n'}$ и ρ для трех заданных значений относительной точности двойного и одинарного отбора:

Кривая I: двойной и одинарный отбор одинаково точны.
Кривая II: двойной отбор дает 25%-ное увеличение точности.

Кривая III: двойной отбор дает 50%-ное увеличение точности.

критическое значение. Если даны c_n и $c_{n'}$, то для того чтобы двойной отбор стал выгодным, (12.28) указывает критическое значение, которое должно превзойти ρ^2 .

На рис. 12.1 по горизонтальной оси отложены значения ρ , а по вертикальной (в логарифмическом масштабе) — значения отношения $c_n/c_{n'}$. Кривая I показывает зависимость между этими величинами, когда двойной отбор и одинарный отбор одинаково точны; кривая II — когда $V_{\text{дв}} = 0,8 V(y)$, т. е. когда двойной отбор дает 25%-ное увеличение точности; кривая III соответствует 50%-ному увеличению точности при двойном отборе. Например, при $\rho = 0,8$ двойной отбор столь же точен, сколь и одинарный, если $c_n/c_{n'}$ равно 4;

он дает 25%-ное увеличение точности, если $c_n/c_{n'}$ приблизительно равно 7 1/2, и 50%-ное увеличение точности, если $c_n/c_{n'}$ приблизительно равно 13.

При практическом применении эти кривые будут преувеличивать выигрыш в точности, достигаемый при двойном отборе, потому что наилучшие значения n и n' приходится либо оценивать по прежним данным, либо определять ориентировочно. Поэтому, прежде чем решить, применять ли двойной отбор, нужно сделать некоторые поправки на ошибки в оценках этих величин.

При любом ρ выигрыш в точности от применения двойного отбора имеет некоторую верхнюю границу. Она достигается, когда получение сведений относительно \bar{x}' не требует затрат ($c_{n'} = 0$). Верхняя граница относительной точности равна $1/(1 - \rho^2)$.

12.7. ОЦЕНКА ДИСПЕРСИИ ПРИ ДВОЙНОМ ОТБОРЕ ДЛЯ ОЦЕНКИ ПО РЕГРЕССИИ

Если членами, содержащими $1/n$, можно пренебречь, то $V(\bar{y}_{lr})$ выражается формулой (12.24):

$$V(\bar{y}_{lr}) \approx \frac{S_y^2(1 - \rho^2)}{n} + \frac{\rho^2 S_x^2}{n'}$$

Для модели с линейной регрессией величина

$$s_{y,x}^2 = \frac{1}{n-2} \left[\sum_{i=1}^n (y_i - \bar{y})^2 - b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

есть несмещенная оценка $S_y^2(1 - \rho^2)$, где индекс y теперь опущен. Поскольку

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

есть несмещенная оценка S_y^2 , разность

$$s_y^2 - s_{y,x}^2$$

есть несмещенная оценка $\rho^2 S_y^2$.

Таким образом, оценка $V(\bar{y}_{lr})$ по выборке есть

$$v(\bar{y}_{lr}) = \frac{s_{y,x}^2}{n} + \frac{s_y^2 - s_{y,x}^2}{n'} \quad (12.29)$$

Если вторая выборка очень мала и членами, содержащими $1/n$, пренебречь нельзя, то можно предложить оценку дисперсии вида

$$v(\bar{y}_{lr}) = s_{y,x}^2 \left[\frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] + \frac{s_y^2 - s_{y,x}^2}{n'}$$

Это — некоторого рода гибрид условной дисперсии и средней дисперсии.

Если первая выборка применяется для того, чтобы получить \bar{x}' для оценки по отношению величины \bar{Y} , то оценкой будет

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} \bar{x}' \quad (12.30)$$

Для того чтобы найти приближенное значение дисперсии, запишем

$$\begin{aligned} \bar{y}_R - \bar{Y} &= \frac{\bar{y}}{\bar{x}} \bar{x}' - \bar{Y} = \left(\frac{\bar{y}}{\bar{x}} \bar{X} - \bar{Y} \right) + \frac{\bar{y}}{\bar{x}} (\bar{x}' - \bar{X}) = \\ &= \frac{\bar{X}}{\bar{x}} (\bar{y} - R\bar{x}) + \frac{\bar{y}}{\bar{x}} (\bar{x}' - \bar{X}). \end{aligned}$$

Первое слагаемое представляет собой ошибку обычной оценки по отношению (параграф 2.9). В параграфе 2.9, для того чтобы получить приближенную дисперсию оценки, мы заменяли в этом члене множитель \bar{X}/\bar{x} единицей. Во втором слагаемом при том же порядке приближения мы заменяем множитель \bar{y}/\bar{x} отношением для совокупности $R = \bar{Y}/\bar{X}$. Тогда

$$\bar{y}_R - \bar{Y} \approx (\bar{y} - R\bar{x}) + R(\bar{x}' - \bar{X}). \quad (12.31)$$

Если первая и вторая выборки извлекаются *независимо*, то, предполагая, что членами пкс можно пренебречь, имеем

$$V(\bar{y}_R) = \frac{S_y^2 - 2RS_{yx} + R^2 S_x^2}{n} + \frac{R^2 S_x^2}{n'}. \quad (12.32)$$

Если вторая выборка представляет собой случайную подвыборку из первой, то перепишем (12.31) в виде

$$\bar{y}_R - \bar{Y} = (\bar{y} - R\bar{X}) + R(\bar{x}' - \bar{x}) = (\bar{y} - \bar{Y}) + R(\bar{x}' - \bar{x}).$$

Можно проверить, что если пренебречь пкс, то

$$\begin{aligned} V(\bar{y} - \bar{Y}) &= \frac{S_y^2}{n}; \\ \text{cov}[(\bar{y} - \bar{Y}) R(\bar{x}' - \bar{x})] &= -RS_{yx} \left(\frac{1}{n} - \frac{1}{n'} \right); \\ V[R(\bar{x}' - \bar{x})] &= R^2 S_x^2 \left(\frac{1}{n} - \frac{1}{n'} \right). \end{aligned}$$

Следовательно, $V(\bar{y}_R)$ принимает вид

$$V(\bar{y}_R) = \frac{S_y^2 - 2RS_{yx} + R^2 S_x^2}{n} + \frac{2RS_{yx} - R^2 S_x^2}{n'}. \quad (12.33)$$

Заметим, что как формула (12.32), так и формула (12.33) имеют общий вид

$$V(\bar{y}_R) = \frac{V_n}{n} + \frac{V_{n'}}{n'}.$$

Поэтому оптимальные n и n' и минимальная дисперсия для сравнения с одинарным отбором отыскиваются так же, как это делалось при отборе для расслоения и для получения оценок по регрессии.

При оценивании дисперсии по выборке в (12.32) и (12.33) можно подставить величины S_y^2 , S_{yx} , S_x^2 и R . Получающиеся при этом оценки $v(\bar{y}_R)$ не будут несмещенными, но, с точностью до принятого в нашем анализе порядка приближения, они, по-видимому, приемлемы.

12.9. ПОВТОРНОЕ ВЫБОРОЧНОЕ ИССЛЕДОВАНИЕ ОДНОЙ И ТОЙ ЖЕ СОВОКУПНОСТИ

По мере того как росло доверие к выборочному методу, стала обычной практика применения выборочных обследований для периодического сбора важных данных, предназначенных для регулярной публикации. Это обусловлено отчасти осознанием того, что при быстро меняющемся населении данные переписей, проводимых через большие интервалы времени, имеют ограниченную ценность. Даже очень точные сведения о характеристиках населения, скажем, в июле 1950 г. и в июле 1960 г., мало могут помочь планированию, для которого нужны данные об этом населении в 1964 г. Больше пользы может принести серия небольших выборок, получаемых ежегодно или через еще более короткие интервалы времени.

Когда одна и та же совокупность (если не считать ее изменений, вызываемых самим ходом времени) подвергается выборочному исследованию неоднократно, исследователь находится в идеальном положении для того, чтобы получить реалистичные оценки как издержек, так и дисперсий, и воспользоваться методами, приводящими к оптимальной результативности отбора. Один из важных вопросов заключается в том, как часто и каким образом должна меняться выборка со временем. На решение этого вопроса влияют многие соображения. Люди могут не хотеть раз за разом предоставлять сведения одного и того же типа. На опрашиваемых может влиять информация, которую они получают в ходе опроса и это может со временем постепенно уменьшать представительность сообщаемых ими сведений. Однако иногда при второй беседе опрашиваемые отвечают более охотно, чем при первой, и если собираются сведения специального или конфиденциального характера, то второе посещение может обеспечить более достоверные данные, чем первое.

В оставшейся части этой главы рассматриваются вопросы о размещении выборки и связанные с этим проблемы вычисления оценок по ряду чередующихся выборок. Эти темы отнесены к содержанию настоящей главы, поскольку здесь может быть применена методика двойного отбора.

Если имеются данные ряда последовательных выборок, нас могут интересовать оценки величин трех видов:

1. Изменение \bar{Y} от одного момента отбора к другому.
2. Среднее значение \bar{Y} по всем моментам отбора.
3. Среднее значение \bar{Y} для самого последнего момента отбора.

В большинстве обследований интерес сосредоточивается на текущем среднем (3), особенно если можно предполагать, что характеристики совокупности быстро меняются с течением времени. С другой стороны, для совокупности, в которой такие изменения происходят медленно, может оказаться в основном приемлемым среднее за год (2), взятое по 12 ежемесячным или по четырем ежеквартальным выборкам. Так, например, обстоит дело при изучении распространенности хронических заболеваний большой продолжительности. Относительно заболеваний, распространение которых имеет выраженный сезонный характер, основной интерес представляют текущие данные, но для сравнения различных районов и разных лет полезны также средние за год. Оценки изменения (1) требуются преимущественно в тех случаях, когда хотят изучить влияние на совокупность различных факторов, которые, как известно, должны были на нее воздействовать. Например, если принят законопроект, который, как предполагается, должен стимулировать жилищное строительство, то интересно выяснить, увеличились ли темпы нового строительства в следующем году (имея в виду, что их увеличение может и не быть полностью обусловлено этим законопроектом).

Предположим, что мы можем изменять или оставлять без изменения состав выборки и что общий объем выборки для всех моментов отбора должен остаться одним и тем же. Если мы хотим максимизировать точность, то относительно тактики замещения выборки можно рекомендовать следующее:

1. Для получения оценки изменения лучше всего оставить одну и ту же выборку для всех моментов отбора.
2. Для получения оценки среднего по всем моментам отбора лучше каждый раз извлекать новую выборку.
3. При получении текущих оценок как сохранение одной и той же выборки, так и ее полная замена при каждом моменте отбора дают одинаковую точность. Однако лучше, чем эти альтернативы, может оказаться замена части выборки при каждом моменте отбора.

Утверждения 1 и 2 справедливы, потому что почти всегда существует положительная корреляция между значениями признака у одной и той же единицы для двух последовательных моментов отбора. Оценка изменения по некоторой единице имеет дисперсию, равную $S_1^2 + S_2^2 - 2\rho S_1 S_2$, где индекс относится к номеру момента отбора. Если изменение оценивается по двум различным единицам, то дисперсия равна $S_1^2 + S_2^2$. При оценивании среднего по двум моментам отбора дисперсия равна $(S_1^2 + S_2^2 + 2\rho S_1 S_2)/4$, если единица сохраняется в выборке, и $(S_1^2 + S_2^2)/4$, если отобрана новая единица.

Утверждение 3, которое менее очевидно, будет исследовано в следующих параграфах.

12.10. ОТБОР В ДВА МОМЕНТА

Предположим, что в оба момента отбора выборки имеют один и тот же объем n и что основной интерес представляют текущие оценки. Тактика замещения была исследована Джессеном (Jessen, 1942). Для простоты мы предположим, что применяется простой случайный отбор и что дисперсия значений y_i для совокупности, S^2 , в оба момента одна и та же.

Среднее значение для первой выборки имеет дисперсию S^2/n , причем никакими прежними сведениями мы не пользуемся. При извлечении второй выборки в ней остается m единиц из первой выборки [m означает число совпадающих (matched) единиц]. Остальные $n - m$ единиц [$n - m$ — число несовпадающих (unmatched)] исключаются и на их место извлекаются новые.

Обозначения:

\bar{y}_{1n} — среднее по несовпадающей части для момента h ;

\bar{y}_{1m} — среднее по совпадающей части для момента h ;

\bar{y}_h — среднее по всей выборке для момента h .

Совпадающие и несовпадающие части второй выборки дают независимые оценки \bar{y}'_{2n} , \bar{y}'_{2m} величины \bar{Y}_2 , как указано в табл. 12.1.

Таблица 12.1
ОЦЕНКИ ПО НЕСОВПАДАЮЩЕЙ И СОВПАДАЮЩЕЙ ЧАСТЯМ ВЫБОРКИ

	Оценка	Дисперсия
Несовпадающая часть	$\bar{y}'_{2n} = \bar{y}_{2n}$	$\frac{S^2}{n} = \frac{1}{W_{2n}}$
Совпадающая часть	$\bar{y}'_{2m} = \bar{y}_{2m} + b(\bar{y}_1 - \bar{y}_{1m})$	$\frac{S^2(1-\rho^2)}{m} + \rho^2 \frac{S^2}{n} = \frac{1}{W'_{2m}}$

Для совпадающей части мы применяем оценку по регрессии для двойного отбора, где «большой» выборкой служит первая выборка, а вспомогательной переменной x_i — значение y_i для первого момента. Дисперсию \bar{y}'_{2m} получаем по формуле (12.24), с. 361; заметим, что n и n' в (12.24) отвечают соответственно нашим m и n .

Наилучшая совместная оценка \bar{Y}_2 получается путем взвешивания двух независимых оценок весами, обратными их дисперсиям. Если W_{2n} , W'_{2m} — величины, обратные дисперсиям, то эта оценка имеет вид

$$\bar{y}_2 = \phi_2 \bar{y}'_{2n} + (1 - \phi_2) \bar{y}'_{2m}, \quad (12.34)$$

где

$$\phi_2 = \frac{W_{2n}}{W_{2n} + W'_{2m}}.$$

Из теории метода наименьших квадратов известно, что дисперсия \bar{y}_2 есть

$$V(\bar{y}_2) = \frac{1}{W_{2u} + W_{2m}}.$$

Подставив значения дисперсий из табл. 12.1, после упрощения получаем

$$V(\bar{y}_2) = \frac{S^2(n - u\rho^2)}{n^2 - u^2\rho^2}. \quad (12.35)$$

Заметим, что при $u = 0$ (полное совпадение выборок) или при $u = n$ (полное замещение) эта дисперсия имеет одно и то же значение, S^2/n .

Оптимальное значение u находим, минимизируя (12.35) по переменной u . Это дает

$$\frac{u}{n} = \frac{1}{1 + \sqrt{1 - \rho^2}}; \quad \frac{m}{n} = \frac{\sqrt{1 - \rho^2}}{1 + \sqrt{1 - \rho^2}}. \quad (12.36)$$

Подставив оптимальное значение u в (12.35), получаем минимальное значение дисперсии

$$V_{opt}(\bar{y}_2) = \frac{S^2}{2n} [1 + \sqrt{1 - \rho^2}]. \quad (12.37)$$

В табл. 12.2 для ряда значений ρ указан оптимальный процент совпадающих единиц и относительный выигрыш в точности по сравнению со случаем, когда выборка замещается полностью.

Таблица 12.2
ОПТИМАЛЬНЫЙ ПРОЦЕНТ СОВПАДАЮЩИХ ЕДИНИЦ

ρ	Оптимальный процент совпадающих единиц	Выигрыш в точности (в %)	Выигрыш (в %) при	
			$\frac{m}{n} = \frac{1}{3}$	$\frac{m}{n} = \frac{1}{4}$
0,5	46	7	7	6
0,6	44	11	11	9
0,7	42	17	17	15
0,8	38	25	25	23
0,9	30	39	39	39
0,95	24	52	50	52
1,0	0	100	67	75

Оптимальный процент совпадающих единиц ни разу не превышает 50 и постепенно убывает по мере увеличения ρ . При $\rho = 1$ формула указывает значение $m = 0$, которое не соответствует нашим предположениям, поскольку предполагалось, что m достаточно велико. В этом случае правильным решением было бы взять $m = 2$. Для того чтобы точно определить линию регрессии, достаточно оставить во второй выборке две единицы, наблюдавшиеся в первой.

Наибольший возможный выигрыш в точности равен 100%; он достигается при $\rho = 1$. Пока ρ не велико, выигрыш оказывается умеренным.

Хотя оптимальный процент совпадающих единиц меняется с изменением ρ , на практике для всех признаков, наблюдаемых в обследовании, можно принять только одно значение процента. В двух правых столбцах табл. 12.2 указан процентный выигрыш в точности, когда в выборках совпадает одна треть и одна четверть всех единиц. Оба эти значения вполне приемлемы в отношении признаков, для которых ρ не превосходит 0,95.

12.11. ОТБОР БОЛЕЕ ЧЕМ В ДВА МОМЕНТА

Общая задача замещения применительно как к текущим оценкам, так и к оценкам изменения была исследована Йейтсом (Yates, 1960) и Паттерсоном (Patterson, 1950). Если число моментов отбора больше двух, то появляется возможность применить более разнообразные варианты замещения. В момент h мы можем получить выборку, часть которой совпадает с выборкой в момент $h - 1$, часть — с выборкой в моменты $h - 1$ и $h - 2$ и т. д. Для того чтобы улучшить текущую оценку, мы можем применить множественную регрессию, опираясь на все части выборки, совпадающие с предыдущими моментами. Мы можем также уточнить текущую оценку для момента $h - 1$ после того, как стали известны данные для момента h . В уточненной оценке можно воспользоваться регрессией момента $h - 1$ как по моменту $h - 2$, так и по моменту h , если предположить, что имеются надлежащие совпадающие части выборки.

Данный параграф содержит введение в этот раздел теории. Мы ограничимся рассмотрением текущих оценок, в которых применяется регрессия по данным лишь последней из предшествующих выборок. Подобное ограничение приводит к некоторой потере в точности, однако поскольку при увеличении промежутка времени между моментами отбора корреляция ρ обычно уменьшается, такой проигрыш в точности редко бывает большим. Далее везде предполагается, что дисперсия S^2 и коэффициент корреляции ρ между значениями признака у одной и той же единицы в два последовательных момента постоянны.

Обозначим для h -го момента через m_h и u_h число единиц, соответственно совпадающих и несовпадающих с единицами, отобранными в $(h - 1)$ -й момент. Две возможные оценки \bar{y}_h указаны в табл. 12.3. Единственное изменение в методе оценивания по сравнению со случаем двух моментов отбора (табл. 12.1) состоит в том, что для регрессионной поправки в оценке, получаемой по совпадающей части выборки, мы принимаем вместо выборочного среднего \bar{y}_{h-1} улучшенную оценку \bar{y}'_{h-1} .

Дисперсия оценки \bar{y}_{hm} по совпадающей части из табл. 12.3 получена на основе формулы (12.24), приведенной в конце параграфа

Оценки \bar{V}_h для h -го момента отбора

Таблица 12.3

	Оценка	Дисперсия
Несовпадающая часть	$\bar{y}_{hu} = \bar{y}_{hu}$	$\frac{S^2}{n} - \frac{1}{W_{hu}}$
Совпадающая часть	$\bar{y}_{hm} = \bar{y}_{hm} + b(\bar{y}_{h-1} - \bar{y}_{h-1,m})$	$\frac{S^2(1-\rho^2)}{m} + \rho^2 V(\bar{y}_{h-1}) = \frac{1}{W_{hm}}$

12.5. Заметим, что (а) наше m соответствует n из (12.24) и (б) член $\rho^2 S^2/n$, соответствующий члену $B^2 E(\bar{x}' - \bar{X})^2$ в (12.21), заменяется выражением $\rho^2 V(\bar{y}_{h-1})$, поскольку $B = \rho$ и \bar{y}_{h-1} соответствует применявшемуся ранее \bar{x}' .

Исследуем теперь точность, получаемую в том случае, если в каждый момент отбора применяются оптимальные m_h и u_h и оптимальные веса. Будет показано, что оптимальное m_h/n_h от одного момента отбора к другому постепенно увеличивается, быстро приближаясь к предельному значению $1/2$.

Наилучшей оценкой \bar{y}_h будет оценка с весами, обратно пропорциональными дисперсиям,

$$\bar{y}_h = \phi_h \bar{y}_{hu} + (1 - \phi_h) \bar{y}_{hm}, \quad (12.38)$$

где $\phi_h = W_{hu}/(W_{hu} + W_{hm})$. Это дает

$$V(\bar{y}_h) = \frac{1}{W_{hu} + W_{hm}} = \frac{g_h S^2}{n},$$

где g_h обозначает отношение дисперсии в момент h к дисперсии в первый момент отбора. Подставляя W_{hu} и W_{hm} , взятые из табл. 12.3, получаем

$$\frac{S^2}{V(\bar{y}_h)} = \frac{n}{g_h} = S^2(W_{hu} + W_{hm}) = u_h + \frac{1}{\frac{(1-\rho^2)}{m_h} + \frac{\rho^2 g_{h-1}}{n}}. \quad (12.39)$$

Найдем теперь m_h и u_h , максимизирующие эту величину, и, следовательно, минимизирующие $V(\bar{y}_h)$. Записав $u_h = n - m_h$ и дифференцируя правую часть формулы (12.39) по m_h , получим

$$\frac{1-\rho^2}{m_h^2} = \left(\frac{1-\rho^2}{m_h} + \frac{\rho^2 g_{h-1}}{n} \right)^2.$$

Отсюда, разрешая это выражение относительно оптимального \hat{m}_h , имеем

$$\frac{\hat{m}_h}{n} = \frac{\sqrt{1-\rho^2}}{g_{h-1}(1+\sqrt{1-\rho^2})}. \quad (12.40)$$

Если найденную величину подставить в формулу (12.39), то она, после некоторых алгебраических преобразований, примет вид

$$\frac{1}{g_h} = 1 + \frac{1-\sqrt{1-\rho^2}}{g_{h-1}(1+\sqrt{1-\rho^2})}. \quad (12.41)$$

Это соотношение можно записать в виде

$$r_h = 1 + b r_{h-1},$$

где $r_h = 1/g_h$ и $r_1 = 1/g_1 = 1$. Повторное применение этого рекуррентного соотношения приводит к равенству

$$\frac{1}{g_h} = r_h = 1 + b + b^2 + \dots + b^{h-1} = \frac{1-b^h}{1-b},$$

где согласно (12.41) $b = (1 - \sqrt{1-\rho^2}) / (1 + \sqrt{1-\rho^2})$. Поскольку $0 < b < 1$, предел отношения дисперсий, g_∞ , равен

$$g_\infty = 1 - b = \frac{2\sqrt{1-\rho^2}}{1 + \sqrt{1-\rho^2}}. \quad (12.42)$$

Следовательно, дисперсия \bar{y}_h стремится к

$$V(\bar{y}_\infty) = \frac{S^2}{n} \left(\frac{2\sqrt{1-\rho^2}}{1 + \sqrt{1-\rho^2}} \right). \quad (12.43)$$

Окончательно предельное значение \hat{m}_h получается из (12.40) в виде

$$\frac{\hat{m}_\infty}{n} = \frac{\sqrt{1-\rho^2}}{g_\infty(1+\sqrt{1-\rho^2})} = \frac{1}{2}$$

независимо от значения ρ .

В табл. 12.4 для ряда значений h приведены полученные по формуле (12.40) значения оптимального процента единиц, которые должны совпадать, $100 \hat{m}_h/n$, и соответствующие этим значениям дисперсии при $\rho = 0,7; 0,8; 0,9$ и $0,95$.

ОПТИМАЛЬНЫЙ ПРОЦЕНТ СОВПАДАЮЩИХ ЕДИНИЦ И ЗНАЧЕНИЯ ДИСПЕРСИЙ

Таблица 12.4

h	Процент совпадающих единиц $100 \hat{m}_h/n$				$g_h = n V(\bar{y}_h)/S^2$			
	при $\rho =$				при $\rho =$			
	0,7	0,8	0,9	0,95	0,7	0,8	0,9	0,95
2	42	38	30	24				
3	49	47	42	36	0,857	0,800	0,718	0,656
4	50	49	47	43	0,837	0,762	0,646	0,566
5	50	50	49	46	0,834	0,753	0,622	0,515
∞	50	50	50	50	0,833	0,751	0,613	0,495
					0,833	0,750	0,607	0,476

Для четвертого момента отбора оптимальный процент совпадающих единиц близок к 50 для всех указанных значений ρ , в то время как для второго и третьего моментов отбора процент совпадающих единиц должен быть меньше. Уменьшение дисперсии, т. е. $(1 - g_h)$, при ρ , меньших 0,8, оказывается умеренным.

12.12. УПРОЩЕНИЕ И ДАЛЬНЕЙШЕЕ РАЗВИТИЕ ИЗЛОЖЕННЫХ МЕТОДОВ

При практическом применении произведенный в предыдущем параграфе анализ может потребовать некоторого видоизменения. Мы предполагали, что все варианты тактики замещения требовали одинаковых издержек и были в равной степени осуществимы. Если мы имеем дело с совокупностями людей, то расходы на опрос будут, вероятно, ниже, если одни и те же единицы остаются в выборке на ряд моментов отбора. Если интерес представляют оценки изменения среднего и суммарного значений для совокупности, то это обстоятельство также указывает на необходимость сохранить от одного момента отбора к другому более половины единиц.

Удобно также сохранять веса и доли совпадающих единиц постоянными, а не менять их в каждый момент отбора. Поэтому мы исследуем дисперсии \bar{y}_h и оценки изменения $(\bar{y}_h - \bar{y}_{h-1})$ в предположении, что m , u и ϕ остаются постоянными. Мы продолжаем пользоваться равенством $V(\bar{y}_h) = g_h S^2/n$, хотя в действительности значение g_h будет отличаться от его значения в предыдущем параграфе.

Теперь оценка есть

$$\bar{y}_h' = \phi \bar{y}_{hu}' + (1 - \phi) \bar{y}_{hm}'.$$

Подставляя выражения для двух дисперсий (из табл. 12.3), получаем

$$\begin{aligned} V(\bar{y}_h') &= \frac{g_h S^2}{n} = \phi^2 V(\bar{y}_{hu}') + (1 - \phi)^2 V(\bar{y}_{hm}') = \\ &= S^2 \left[\frac{\phi^2}{u} + \frac{(1 - \phi)^2 (1 - \rho^2)}{m} \right] + \frac{S^2 \rho^2 (1 - \phi)^2 g_{h-1}}{n}. \end{aligned}$$

Следовательно,

$$g_h = \left[\frac{\phi^2}{\mu} + \frac{(1 - \phi)^2 (1 - \rho^2)}{\lambda} \right] + \rho^2 (1 - \phi)^2 g_{h-1}, \quad (12.44)$$

где $\mu = u/n$ и $\lambda = m/n$. Запишем это равенство в виде

$$g_h = a + b g_{h-1}.$$

Применяя его последовательно для разных h , получаем, пользуясь тем, что $g_1 = 1$,

$$g_h = \frac{a(1 - b^{h-1})}{1 - b} + b^{h-1}.$$

Поскольку $b = \rho^2 (1 - \phi)^2$ меньше 1, предельное значение равно:

$$g_\infty = \frac{a}{1 - b} = \frac{\lambda \phi^2 + \mu (1 - \phi)^2 (1 - \rho^2)}{\lambda \mu [1 - \rho^2 (1 - \phi)^2]}. \quad (12.45)$$

Значение веса ϕ , при котором предельная дисперсия минимальна можно найти, дифференцируя (12.45). Это приведет к квадратному уравнению, нужный корень которого равен:

$$\phi_{opt} = \frac{V(1 - \rho^2) [V(1 - \rho^2 + 4\lambda\mu\rho^2 - V(1 - \rho^2))]}{2\lambda\rho^2}.$$

На практике значение ρ точно не известно и для разных признаков будет разным. Обычно можно выбрать некоторое простое, приемлемое для всех признаков значение. Очевидно, что ϕ_{opt} будет меньше, чем $\mu = u/n$, поскольку совпадающая часть выборки дает большую точность на единицу, чем несовпадающая часть. Например, при $\mu = 0,25$, т. е. когда не совпадает 1/4 часть выборки, ϕ_{opt} при $\rho = 0,7; 0,8; 0,9$ оказывается равным соответственно 0,216; 0,198 и 0,164. При значениях ρ в этом интервале подошло бы значение $\phi = 0,2$.

Для оценки изменения имеем

$$V(\bar{y}_h' - \bar{y}_{h-1}') = V(\bar{y}_h') + V(\bar{y}_{h-1}') - 2\text{Cov}(\bar{y}_h' \bar{y}_{h-1}'). \quad (12.46)$$

Для того чтобы найти ковариацию, заметим, что если y_{hi} , $y_{h-1,i}$ — значения признака у i -й единицы среди единиц, совпадающих в моменты h и $(h - 1)$, то наша модель примет вид

$$y_{hi} = \bar{Y}_h + \rho(y_{h-1,i} - \bar{Y}_{h-1}) + e_{hi},$$

где e_{hi} независимы от значений y . По этой модели находим подстановкой

$$\bar{y}_{hm}' = \bar{y}_{hm} + \rho(\bar{y}_{h-1} - \bar{y}_{h-1,m}) = \bar{Y}_h + \rho(\bar{y}_{h-1} - \bar{Y}_{h-1}) + \bar{e}_{hm}.$$

Следовательно, ковариация \bar{y}_{hm}' и \bar{y}_{h-1}' есть $\rho V(\bar{y}_{h-1})$. Но

$$\begin{aligned} \text{Cov}(\bar{y}_h' \bar{y}_{h-1}') &= \text{Cov}(\phi \bar{y}_{hu}' + (1 - \phi) \bar{y}_{hm}' \bar{y}_{h-1}') = \\ &= \rho(1 - \phi) V(\bar{y}_{h-1}'), \end{aligned}$$

поскольку \bar{y}_{hu}' независимы от \bar{y}_{h-1}' . Отсюда и из (12.46) получаем

$$V(\bar{y}_h' - \bar{y}_{h-1}') = \frac{S^2}{n} [g_h + g_{h-1} (1 - 2\rho(1 - \phi))]. \quad (12.47)$$

По формулам (12.44) и (12.47) дисперсии \bar{y}_h' и $(\bar{y}_h' - \bar{y}_{h-1}')$ можно вычислить для любых значений m , ϕ и ρ . В табл. 12.5 приведены эти дисперсии для $\lambda = m/n = 1/2$ и $3/4$. Вес ϕ был взят равным 0,35 при $\lambda = 1/2$ и равным 0,2 при $\lambda = 3/4$.

Полученные результаты показывают, что увеличение с 1/2 до 3/4 доли единиц, оставляемых в выборке в следующий момент отбора, приводит к очень небольшому увеличению дисперсии текущей оценки и к существенно большему уменьшению дисперсии оценки изменения. Например, при $\rho = 0,8$ увеличение $V(\bar{y}_h)$ составляет приблизительно 5%, в то время как $V(\bar{y}_h - \bar{y}_{h-1})$ уменьшается более чем на 20%. Поэтому, если мы хотим получить как текущие оценки, так и оценки изменения, то приемлемой практически тактикой будет сохранение в последующий момент отбора 2/3, 3/4 или 4/5 объема выборки, полученной в предыдущий момент.

Сравнение $V(\bar{y}_h)$ при $\lambda = 1/2$ в табл. 12.5 с оптимальными дисперсиями из табл. 12.4 показывает, что применение постоянного веса и неизменного значения $\lambda = 1/2$ дает лишь небольшую потерю в точности.

Если ρ больше 0,8, то коэффициент регрессии $b = \rho$ лишь с небольшой дополнительной потерей в точности можно заменить единицей. Это приводит к оценке \bar{y}_h вида

$$\bar{y}_h^* = \phi \bar{y}_{hm} + (1 - \phi)(\bar{y}_{h-1} + \bar{y}_{hm} - \bar{y}_{h-1, m}). \quad (12.48)$$

В проводимом ежемесячно Бюро переписи США важным Текущем обследовании населения (Current Population Survey) каждый месяц замещается четверть единиц второй ступени, так что отдельное домохозяйство остается в выборке на протяжении четырех последовательных месяцев. Домохозяйство не подвергается обследованию в течение следующих восьми месяцев, но затем снова возвращается в выборку еще на четыре месяца, что несколько увеличивает точность сравнения по годам.

Составная оценка, применяемая в этом обследовании, напоминает оценку (12.48), но несколько отличается от нее, и имеет вид

$$\bar{y}_h^* = (1 - K) \bar{y}_h + K(\bar{y}_{h-1} + \bar{y}_{hm} - \bar{y}_{h-1, m}), \quad (12.49)$$

где K — постоянный взвешивающий множитель. Отличие состоит в том, что вместо \bar{y}_{hm} в (12.48) здесь стоит \bar{y}_h , текущая оценка по всей выборке. Величины \bar{y}_{hm} , $\bar{y}_{h-1, m}$, \bar{y}_h в (12.49) представляют собой оценки по отношению довольно сложного вида. Дисперсия \bar{y}_h [найденная Бершо (Bershad)], приведена у Хансена, Хервица и Мэдоу (Hansen, Hurwitz and Madow, 1953). Поскольку исходные единицы не меняются, описанная тактика замещения влияет только на внутри-единичное слагаемое дисперсии $V(\bar{y}_h)$.

При другой тактике чередования в каждый момент отбора извлекается новая выборка и никакие единицы в двух выборках не совпадают. При ежемесячном обследовании такой план вполне пригоден, если наибольший интерес представляют годовые и в меньшей степени полугодовые или квартальные оценки, например при исследовании заболеваемости, где особое значение придается хроническим болезням. Если для каждой единицы опрос дает сведения как для предыду-

Таблица 12.5

ВЛИЯНИЕ ДОЛИ СОВПАДАЮЩИХ ЕДИНИЦ НА $V(\bar{y}_h)$ И $V(\bar{y}_h - \bar{y}_{h-1})$

λ	ρ						ρ					
	ρ			ρ			ρ			ρ		
	0,7	0,8	0,9	0,95	0,9	0,8	0,7	0,8	0,9	0,95	0,9	0,8
	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
2	0,88 0,91	0,82 0,88	0,75 0,84	0,71 0,82	0,97 0,79	0,78 0,60	0,97 0,79	0,78 0,60	0,58 0,40	0,47 0,30	0,58 0,40	0,47 0,30
3	0,86 0,88	0,77 0,83	0,66 0,76	0,60 0,72	0,94 0,77	0,74 0,58	0,94 0,77	0,74 0,58	0,53 0,39	0,43 0,29	0,53 0,39	0,43 0,29
4	0,85 0,87	0,76 0,81	0,63 0,72	0,56 0,66	0,93 0,77	0,73 0,57	0,93 0,77	0,73 0,57	0,52 0,38	0,41 0,28	0,52 0,38	0,41 0,28
5	0,85 0,87	0,75 0,80	0,62 0,69	0,54 0,62	0,93 0,76	0,72 0,57	0,93 0,76	0,72 0,57	0,51 0,38	0,41 0,28	0,51 0,38	0,41 0,28
∞	0,85 0,87	0,75 0,79	0,62 0,67	0,53 0,58	0,93 0,76	0,72 0,57	0,93 0,76	0,72 0,57	0,51 0,37	0,40 0,28	0,51 0,37	0,40 0,28

* Значение λ — доля совпадающих единиц.

шего месяца, так и для текущего месяца, то мы можем принять составную оценку вида

$$\bar{y}_h^* = \bar{y}_h + \phi_h (\bar{y}_{h-1}^* - \bar{y}_{h-1, h}), \quad (12.50)$$

где \bar{y}_h — оценка, получаемая по текущим данным в текущей выборке;

$\bar{y}_{h-1, h}$ — оценка, получаемая по данным предыдущего месяца в текущей выборке;

\bar{y}_{h-1}^* — составная оценка для предыдущего месяца.

Теоретический анализ этой оценки дан Хансеном, Хервицем и Мэдоу (Hansen, Hurwitz and Madow, 1953) и Вудруфом (Woodruff, 1959), который применял такие оценки при обследовании розничной торговли, а также Эклером (Eckler, 1955). В обследовании розничной торговли (Retail Trade Survey) составная оценка включает оценку по отношению и имеет вид

$$\bar{y}_h^* = (1 - W) \bar{y}_h + W \left(\frac{\bar{y}_h}{\bar{y}_{h-1, h}} \right) \bar{y}_{h-1}^*,$$

где W — взвешивающий множитель. Поскольку между наблюдениями по месяцам существует тесная, в среднем около 0,98, корреляция, достигается существенный выигрыш в точности. В следующем месяце вычисляются уточненная составная оценка для месяца h , учитывающая результаты для месяца h по новой выборке, отобранной в месяце $(h + 1)$.

При таком методе существенно, чтобы данные, получаемые по текущей выборке для предшествующего месяца, были сравнительно точны. Это условие может не выполняться, если опрашиваемый сообщает сведения по памяти, без какой-либо записи. В то же время метод может вполне себя оправдывать, если получаемые данные принадлежат к тому типу данных, которые опрашиваемый тщательно и регулярно записывает.

Упражнения

12.1. На обследование для получения оценки некоторой доли отпущено 3000 долл. Издержки на проведение основного обследования в расчете на единицу отбора составляют 10 долл. Имеются сведения, содержащиеся в картотеке, — издержки на их получение 0,25 долл. на единицу отбора, — которые позволяют разделить единицы на два слоя приблизительно одинаковой величины. Оцените оптимальные n , n' и соответствующее значение $V(\bar{y}_{st})$, если истинная доля составляет 0,2 для слоя 1 и 0,8 для слоя 2. Даст ли двойной отбор выигрыш в точности по сравнению с одинарным отбором? (Отношениями n'/N , n_h/N_h можно пренебречь.)

12.2. Для значений W_h и P_h из упражнения 12.1 найдите значения отношения издержек c_h/c_n , при которых двойной отбор будет экономичнее одинарного отбора.

12.3. Совокупность состоит из L слоев одинаковой величины. Обозначим через V_{ran} дисперсию среднего для простой случайной выборки и через V_{st} , V_{ds} — соответствующие дисперсии при расслоенном случайном отборе с про-

порциональным размещением и двойным отборе для расслоения. Покажите, что приближенно [ds — от английского «double sampling» — двойной отбор]

$$nV_{ran} = \bar{S}_h^2 + \frac{\sum (\bar{y}_h - \bar{Y})^2}{L};$$

$$nV_{st} = \bar{S}_h^2;$$

$$nV_{ds} = \bar{S}_h^2 + \frac{n}{n'} \frac{\sum (\bar{y}_h - \bar{Y})^2}{L},$$

где \bar{S}_h^2 — средняя дисперсия внутри слоев (как N , так и n' можно считать большими по сравнению с L , а n_h при двойном отборе можно считать равными n/L).

Отсюда, обозначив через $(RP)_{st}$ относительную точность (relative precision) расслоенной выборки по сравнению с простой случайной выборкой и вводя аналогично $(RP)_{ds}$, покажите, что

$$(RP)_{ds} = \frac{(RP)_{st}}{1 + (n/n') [(RP)_{st} - 1]}.$$

При $(RP)_{st} = 2$ рассмотрите зависимость $(RP)_{ds}$ от n/n' . Насколько малым должно быть это отношение, чтобы $(RP)_{ds} = 1,9$?

12.4. Пусть при двойном отборе для оценки по регрессии $\rho = 0,8$. Во сколько раз n' должно быть больше n , если потери в точности, вызванная выборочной ошибкой среднего значения большой выборки, должна быть меньше 10%?

12.5. В одном из случаев применения двойного отбора для регрессии малая выборка имела объем, равный 87, а большая — 300 единицам. Следующие данные относятся к малой выборке:

$$\Sigma (y_i - \bar{y})^2 = 17283; \quad \Sigma (y_i - \bar{y})(x_i - \bar{x}) = 5114;$$

$$\Sigma (x_i - \bar{x})^2 = 3248.$$

Вычислите стандартную ошибку оценки по регрессии величины \bar{y} .

12.6. Для $\rho = 0,95$ проверьте данные из табл. 12.4 относительно оптимального процента совпадающих единиц и выигрыша в точности по сравнению со случаем полного замещения выборки. Вычислите соответствующие значения процентов выигрыша в точности, когда во второй момент отбора сохраняется треть единиц, отобранных в первый момент, а в каждый из последующих моментов сохраняется половина единиц.

12.7. При простом случайном отборе в каждом из двух моментов предположим, что во второй момент применяется следующая оценка (в обозначениях параграфа 12.10):

$$\bar{y}_2^* = (1 - \phi) (\bar{y}_1 + \bar{y}_{2m} - \bar{y}_{1m}) + \phi \bar{y}_{2n}.$$

(а) Покажите, что если пренебречь п.к.с., то

$$V(\bar{y}_2^*) = \frac{S^2}{n} \left\{ (1 - \phi)^2 \frac{[1 + \mu(1 - 2\rho)]}{\lambda} + \frac{\phi^2}{\mu} \right\},$$

где $\lambda = m/n$, $\mu = n/n'$. (б) При данных значениях ρ , λ , μ найдите значение ϕ , минимизирующее $V(\bar{y}_2^*)$. Покажите, что при ρ , большем 1/2, наилучшее значение веса ϕ заключено между μ и $\mu/(1 + \mu)$.

12.8. При $\mu = 1/4$, $\lambda = 1/2$, $\rho = 0,8$ и $\rho = 0,9$ сравните $V(\bar{y}_2^*)$ из предыдущего упражнения с дисперсией оптимальной составной оценки по регрессии \bar{y}_2 , выраженной формулой (12.35). (В оценке \bar{y}_2^* положите $\phi = 0,2$ при $\mu = 1/4$ и $\phi = 0,4$ при $\mu = 1/2$.) Проверьте, что при этих значениях ρ оценка \bar{y}_2 почти так же точна, как и \bar{y}_2^* , как при $\mu = 1/4$, так и при $\mu = 1/2$.

12.9. Каждый месяц извлекается независимая выборка объема n . По каждой такой выборке получают сведения за текущий и за предыдущий месяцы. Применяется та же составная оценка \bar{y}'_h , что и в (12.50), параграф 12.12.

$$\bar{y}'_h = \bar{y}_h + \phi_h (\bar{y}'_{h-1} - \bar{y}_{h-1}, n).$$

Модель имеет вид

$$y_{hi} = \bar{Y}_h + \rho (y_{h-1, i} - \bar{Y}_{h-1}) + e_{hi},$$

где e_{hi} независимы от значений y и имеют дисперсию, равную $(1 - \rho^2)$. Покажите, что

$$(a) \quad \bar{y}'_h - \bar{Y}_h = \bar{e}_h + \phi_h (\bar{y}'_{h-1} - \bar{Y}_{h-1}) + (\rho - \phi_h) (\bar{y}_{h-1, n} - \bar{Y}_{h-1}).$$

$$(b) \quad \text{Если } V(\bar{y}'_h) = g_h S^2/n, \text{ где } S^2 \text{ постоянно для всех моментов отбора, то}$$

$$g_h = (1 - \rho^2) + \phi_h^2 g_{h-1} + (\rho - \phi_h)^2.$$

(в) Оптимальное $\phi_h = \rho/(1 + g_{h-1})$ и соответствующее оптимальное g_h равно:

$$g_h = 1 - \frac{\rho^2}{1 + g_{h-1}}.$$

(г) Предельное значение g_h равно: $g_\infty = \sqrt{1 - \rho^2}$. Эти результаты были получены Эклером (Eckler, 1955).

ЛИТЕРАТУРА

- Bose Chameli (1943). Note on the sampling error in the method of double sampling. *Sankhya*, 6, 330.
- Eckler A. R. (1955). Rotation sampling. *Ann. Math. Stat.*, 26, 664—685.
- Hansen M. H., Hurwitz W. N. and Madow W. G. (1953). *Sample survey methods and theory*. John Wiley and Sons, New York.
- Jessen R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agr. Exp. Sta. Res. Bull.* 304.
- Neyman J. (1938). Contribution to the theory of sampling human populations. *Jour. Amer. Stat. Assoc.*, 33, 101—116.
- Patterson H. D. (1950). Sampling on successive occasions with partial replacement of units. *Jour. Roy. Stat. Soc.*, B12, 241—255.
- Robson D. S. (1952). Multiple sampling of attributes. *Jour. Amer. Stat. Assoc.*, 47, 203—215.
- Robson D. S. and King A. J. (1953). Double sampling and the Curtis impact survey. *Cornell Univ. Agr. Exp. Sta. Mem.* 231.
- Woodruff R. S. (1959). The use of rotating samples in the Census Bureau's Monthly Surveys. *Proc. Social Statistics Section. Amer. Stat. Assoc.*, 130—138.
- Yates F. (1960). *Sampling methods for censuses and surveys*. Charles Griffin and Co., London, third edition. Есть русский перевод: Яйтс Ф. Выборочный метод в переписях и обследованиях. М., «Статистика», 1965.

ГЛАВА 13

ИСТОЧНИКИ ОШИБОК ПРИ ОБСЛЕДОВАНИЯХ

13.1. ВВЕДЕНИЕ

Излагая теорию выборочного метода в предыдущих главах, мы везде предполагали, что применяется один из способов вероятностного отбора и что значение наблюдения y_i у i -й единицы представляет собой точное значение для этой единицы. Ошибка оценки возникала исключительно из-за случайной вариации выборочных значений, вызываемой тем, что мы наблюдаем только n единиц, а не всю совокупность N единиц.

Эти предположения достаточно хорошо оправдываются при более простых видах обследований, в которых применяются точные методы наблюдения, а качество работы сравнительно высоко. В сложных обследованиях, особенно если процесс наблюдения связан с серьезными трудностями, эти предположения могут быть далеки от реальности.

Можно назвать три следующих дополнительных источника ошибок.

1. Отсутствие данных по некоторым единицам в извлеченной выборке. Это может произойти из-за недосмотра или — при обследовании совокупности людей — из-за того, что с некоторыми лицами не удалось встретиться или же они, встретившись с исследователем, отказались отвечать на вопросы.

2. Ошибки наблюдения по некоторым единицам. Метод наблюдения может быть неточным или приводить к смещению. При обследовании совокупности людей опрашиваемые могут не располагать верными сведениями или давать неправильные ответы.

3. Ошибки, возникающие при проверке записей, кодировании и сведении результатов в таблицы.

Перечисленные источники ошибок вызывают необходимость в некотором видоизменении обычной теории выборочного исследования. Основная цель такого видоизменения состоит в том, чтобы выработать правила распределения ресурсов между задачами уменьшения ошибок случайного отбора и задачами уменьшения других ошибок и разработать такие методы вычисления стандартных ошибок и доверительных границ, которые сохраняют силу и при наличии других ошибок.

13.2. ЭФФЕКТ НЕПОЛУЧЕНИЯ ОТВЕТА

Термином *неполучение ответа* мы будем обозначать отсутствие данных по некоторым единицам в извлеченной выборке. Изучая неполучение ответа, удобно считать всю совокупность разделенной на два

«слоя»: первый из них включает все единицы, по которым, если они попадают в выборку, можно получить данные, второй — единицы, по которым данные получить нельзя. Состав этих двух слоев сильно зависит от того, какие методы применяются для нахождения единиц и для получения данных. Так, например, обследование, в котором каждый дом посещается, если это необходимо, по крайней мере, три раза и в котором всех, кто отказывается отвечать, опрашивает инструктор-контролер, обладающий умением убеждать, будет иметь гораздо меньший слой «неотвечивших», чем обследование, в котором каждый дом посещается только один раз.

Такое разделение на два различных слоя представляет собой, конечно, сильное упрощение. Будет ли та или иная единица найдена и обследована при данном числе попыток, отчасти зависит от случая. При более детальном исследовании этой проблемы следовало бы приписать каждой единице некоторую вероятность, соответствующую ее шансам быть обследованной при данном методе наблюдения, если она попала в выборку.

Выборка не дает никаких сведений относительно единиц из слоя 2 — неотвечивших. Это не имело бы значения, если бы можно было предположить, что у слоя 2 те же характеристики, что и у слоя 1. Однако при проверке этого обстоятельства часто оказывалось, что единицы из слоя «неотвечивших» отличаются от единиц, доступных для наблюдения. Иллюстрацией могут служить данные табл. 13.1. Они получены по результатам экспериментального выборочного исследования фруктовых садов штата Северная Каролина в 1946 г. Садоводам трижды рассылали по почте один и тот же опросный лист. По одному из вопросов — о числе фруктовых деревьев — имелись данные для всей совокупности (Finkner, 1950).

Таблица 13.1
РЕЗУЛЬТАТЫ ТРЕХ ОБРАЩЕНИЙ В ОБСЛЕДОВАНИЕ, ПРОВОДИВШЕМСЯ ПО ПОЧТЕ

	Число садоводов	Процент совокупности	Среднее число фруктовых деревьев на одного садовода
Отвечившие при первом обращении	300	10	456
Отвечившие при втором обращении	543	17	382
Отвечившие при третьем обращении	434	14	340
Неотвечившие после трех обращений	1 839	59	290
По всей совокупности	3 116	100	329

Отчетливо видно, что с каждым обращением число фруктовых деревьев на одного садовода постепенно уменьшается; оно составляет 456 для ответивших на первое обращение по почте, 382 — при втором обращении, 340 — при третьем и 290 — для неотвечивших на все три письма. Общий процент ответивших оказался невысоким: более половины садоводов не предоставили данных даже после трех обращений.

Рассмотрим теперь влияние неполучения ответа на оценку по данным выборки. Пусть N_1 и N_2 — числа единиц в двух слоях и пусть $W_1 = N_1/N$ и $W_2 = N_2/N$, так что W_2 — доля случаев неполучения ответа для совокупности. Предположим, что из совокупности извлекается простая случайная выборка. После завершения опроса мы имеем данные для простой случайной выборки из слоя 1, но не имеем данных по слою 2. Следовательно, величина смещения выборочного среднего есть

$$E(\bar{y}) - \bar{Y} = \bar{Y}_1 - \bar{Y} = \bar{Y}_1 - (W_1 \bar{Y}_1 + W_2 \bar{Y}_2) = W_2 (\bar{Y}_1 - \bar{Y}_2). \quad (13.1)$$

Величина смещения представляет собой произведение доли неотвечивших и разности средних по двум слоям. Поскольку выборка не дает никаких сведений о значении \bar{Y}_2 , то величина смещения остается неизвестной, если только мы не сможем указать границы для \bar{Y}_2 по иным, чем выборка, источникам. Для непрерывной переменной те границы, которые можно указать с уверенностью, часто столь широки, что практически бесполезны.

Поэтому в случае непрерывных переменных при любой значимой доле неотвечивших приемлемые доверительные границы для \bar{Y} по результатам выборки обычно указать невозможно. Нам остается лишь принять некоторое предположение относительно величины смещения, не подкрепленное никакими фактическими данными.

При отборе для оценивания долей положение несколько облегчается, поскольку неизвестная доля P_2 для слоя 2 должна находиться между 0 и 1. Если W_2 известно, то эти границы для P_2 позволяют нам построить доверительные границы для доли совокупности P . Предположим, что извлечена простая случайная выборка объемом в n единиц и получены данные по n_1 единицам выборки. Тогда в предположении, что n_1 достаточно велико, 95%-ные доверительные границы для P_1 имеют вид

$$p_1 \pm 2\sqrt{p_1 q_1 / n_1},$$

где p_1 — доля в выборке, причем пкс не учитывается.

Если мы хотим получить доверительное утверждение относительно P , то для полной уверенности можно считать $P_2 = 0$ при нахождении нижней границы, \hat{P}_L , и $P_2 = 1$ при нахождении верхней границы, \hat{P}_U . Таким образом, в качестве 95%-ных доверительных границ можно взять

$$\hat{P}_L = W_1 (p_1 - 2\sqrt{p_1 q_1 / n_1}) + W_2 (0); \quad (13.2)$$

$$\hat{P}_U = W_1 (p_1 + 2\sqrt{p_1 q_1 / n_1}) + W_2 (1). \quad (13.3)$$

Нетрудно проверить, что эти границы взяты «с запасом», т. е. что

$$Pr(\hat{P}_L \leq P \leq \hat{P}_U) > 0,95.$$

Эти границы можно несколько сузить с помощью более тщательного доказательства (Cochran, Mosteller and Tukey, 1954), так как P_2 не может быть равным 0 и 1 одновременно, как предполагалось ранее.

К сожалению, если W_2 не очень мало, то эти границы чрезмерно широки. В табл. 13.2 указаны средние значения границ для выборки объема $n = 1000$ и ряда значений W_2 и p_1 . Поскольку границы в (13.2) и (13.3) зависят от значения n_1 (числа ответивших в выборке), то при составлении табл. 13.2 мы полагали $n_1 = n W_1$, т. е. среднему числу ответивших.

Таблица 13.2
95%-НЫЕ ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ ДЛЯ P (В %) ПРИ $n=1000$

Процент неответив- ших, 100 W_2	Выборочный процент, 100 p_1							
	5		10		20		50	
	нижняя	верхняя	нижняя	верхняя	нижняя	верхняя	нижняя	верхняя
0	3,6	6,4	8,1	11,9	17,5	22,5	46,7	53,2
5	3,4	11,1	7,6	16,3	16,5	26,5	44,4	55,6
10	3,2	15,8	7,2	20,8	15,6	30,4	42,0	58,0
15	3,0	20,5	6,8	25,2	14,7	34,3	39,6	60,4
20	2,8	25,2	6,3	29,7	13,7	38,3	37,2	62,8

Быстрое увеличение ширины доверительного интервала при возрастании W_2 очевидно. Интересно выяснить, какими должны быть значения n для того, чтобы обеспечить ту же ширину доверительного интервала при W_2 , равном нулю. Их легко найти при $p_1 = 50\%$. Для $W_2 = 5\%$ табл. 13.2 дает значение половины ширины доверительного интервала, равное 5,6. Эквивалентный объем выборки n_e (е — от английского «equivalent» — эквивалентный) в предположении, что случаев неполучения ответа нет, определяется из уравнения

$$5,6 = 2\sqrt{50 \cdot 50/n_e};$$

$$n_e = 320.$$

Для $W_2 = 10, 15$ и 20% значения n_e составляют соответственно 155, 90 и 60. Очевидно, имеет смысл потратить существенную часть ресурсов на сокращение числа случаев неполучения ответа.

Интересный метод определения объема выборки с учетом возможного неполучения ответа был предложен Бернбаумом и Серкенем (Bernbaum and Sirken, 1950a, 1950b). Доля случаев неполучения ответа, W_2 , предполагается известной по опыту прежних обследований данного типа. О величинах P_1 , P_2 или P предварительных предположений не делается. Тогда, если бы неответивших не было и мы бы хотели, чтобы абсолютная ошибка выборочной доли была меньше d , нужно было бы взять (в соответствии с параграфом 4.4)

$$n = \frac{t_{\alpha}^2 PQ}{d^2},$$

где t_{α} — квантиль нормального распределения, соответствующий вероятности α того, что ошибка превысит d . Не имея предварительных

сведений относительно P , следует, как в наименее благоприятном случае, считать $P = 0,5$. Отсюда

$$n = \frac{t_{\alpha}^2}{4d^2}. \quad (13.4)$$

Принимая наименее благоприятное сочетание смещения $W_2 \times (P_1 - P_2)$ и значения P_1 , Бернбаум и Серкен показали, что значение n , для которого ошибка, с вероятностью α , еще не превышает d , равно:

$$n \approx \frac{t_{\alpha}^2}{4d(d - W_2)W_1} - 1. \quad (13.5)$$

Заметим, что если $W_2 > d$, то этим условиям не удовлетворяет ни одно значение n . При $W_2 = 0$ это уравнение сводится к (13.4), если не считать члена -1 , который обусловлен приближением, допущенным при выводе формулы. Некоторые значения n , полученные с помощью метода Бернбаума и Серкена, приведены в табл. 13.3.

Таблица 13.3
НАИМЕНЬШИЕ ЗНАЧЕНИЯ n ПРИ ДАННОМ ПРЕДЕЛЕ ОШИБКИ
 d И ВЕРОЯТНОСТИ $\alpha = 0,05$

Процент неответивших, 100 W_2	d (в %)			
	20	15	10	5
0	24	43	96	384
2	27	50	122	653
4	31	60	166	2000
6	36	75	255
8	43	99	521
10	52	142
15	112

Из этой таблицы можно сделать те же печальные выводы, что и из табл. 13.2. Если мы довольствуемся грубой оценкой ($d = 20$), то при доле случаев неполучения ответа до 10% еще можно обойтись удвоением объема выборки. Однако при любом значительном проценте неответивших достигнуть высоко гарантированной точности, увеличивая объем выборки среди отвечающих, невозможно или слишком дорого стоит.

13.3. ВИДЫ НЕПОЛУЧЕНИЯ ОТВЕТА

В последующих параграфах описываются некоторые способы решения проблемы неполучения ответа. Дадим примерную классификацию видов неполучения ответа.

1. *Необнаруженные* — те единицы выборки, которых оказалось невозможным найти или посетить.

Такая проблема возникает, например, при территориальных единицах отбора, когда исследователь должен отыскать в городском квартале все жилища (согласно некоторому определению жилища) и сос-

тавить их перечень. Ее может породить также применение неполных перечней единиц. Иногда погода или плохие средства сообщения не дают возможности посетить некоторые единицы в течение всего периода обследования.

2. *Не оказавшиеся дома.* Эта группа включает лиц, живущих в определенном месте, но временно отсутствующих дома. В семьях, где родители работают или в бездетных семьях, труднее заставить кого-нибудь дома, чем в семьях с очень маленькими детьми или с очень старыми людьми, прикованными к дому.

3. *Неспособные дать ответ.* Опрашиваемый может не располагать сведениями по некоторым вопросам или не хотеть их сообщить. Мерой предосторожности против этого служит тщательная формулировка вопросов в опросном листе и его предварительное испытание.

4. *«Крепкие орешки».* К этой группе относятся лица, которые категорически отказываются отвечать, не в состоянии ответить на вопросы, или те, кто находится далеко от дома в течение всего времени, отведенного на обследование соответствующего участка. Они представляют собой источник смещения, которое нельзя устранить, сколько бы усилий мы ни прилагали для увеличения полноты записей.

Отыскание необнаруженных и сбор данных о них — сложная задача. При территориальном отборе один из методов состоит в том, чтобы повторно посетить исходные единицы и тщательно составить для проверки перечень подлежащих обследованию. Иногда на то, что некоторые люди или жилища пропущены, указывает сравнение подсчетов их числа с аналогичными подсчетами при другом обследовании. Если главной основой выборки служит адресный справочник, то выборку из него можно дополнить выборкой территорий с целью охватить обследованием участки города (новые здания), не полностью представленные в справочнике, а в тех участках города, где справочник кажется точным, проверить, нет ли адресов, пропущенных в справочнике. Обследования, в которых применялись эти методы, описаны в работах Киша и Хесса (Kish and Hess, 1958), а также Вулси (Woolsey, 1956), где рассматривается проблема неполноты охвата.

Что касается не оказавшихся дома, то в тех обследованиях, где ответы на вопросы может дать любой взрослый, которого удалось застать, дело обстоит проще, чем в тех, где нужно опросить лишь одного, случайным образом отобранного взрослого человека. Опрос одного взрослого человека предпочтительнее в тех случаях, когда обследуются отдельные лица и человек не может правильно ответить за другого, или когда внутри домохозяйства наблюдается высокая корреляция, так что опрос более чем одного человека на семью становится неэкономичным. В связи с этим полезный метод отбора одного лица из домохозяйства был разработан Кишем (Kish, 1949). Исходная схема была предназначена для домохозяйств, включающих не более чем шесть подлежащих опросу лиц, но эта же методика применима как для больших, так и для меньших домохозяйств.

Обследователь записывает на бланке всех лиц в семье, которых нужно опросить, и затем нумерует их: сначала мужчин в порядке убывания возраста, затем в том же порядке женщин. На каждом бланке

напечатан один из восьми возможных вариантов отбора, указанных в табл. 13.4.

Таблица 13.
ПРАВИЛА ОТБОРА ОДНОГО ИЗ ЧЛЕНОВ ДОМОХОЗЯЙСТВА ДЛЯ ОПРОСА

Относительная частота применения	Код варианта	Если число взрослых в домохозяйстве равно					
		1	2	3	4	5	6
		отбирайте человека под номером					
1/6	A	1	1	1	1	1	1
1/12	B1	1	1	1	1	2	2
1/12	B2	1	1	1	2	2	2
1/6	C	1	1	2	2	3	3
1/6	D	1	2	2	3	4	4
1/12	E1	1	2	3	3	3	5
1/12	E2	1	2	3	4	5	5
1/6	F	1	2	3	4	5	6

Каждый подлежащий опросу член домохозяйства данной величины имеет одинаковую вероятность быть отобранным, за исключением домохозяйств с 5 взрослыми, в которых эта вероятность для лиц, получающих номера 3 и 5, несколько больше. Поскольку по вариантам A, B и C будут преимущественно опрашиваться мужчины, обследование домохозяйств, попадающих под эти варианты, лучше производить в вечерние часы.

13.4. ПОВТОРНЫЕ ОБРАЩЕНИЯ

Обычно заранее определяется число повторных обращений или минимальное число обращений к какой-либо единице, которые необходимы, прежде чем исключить эту единицу ввиду «невозможности войти в контакт» с ней. Стивен и Маккарти (Stephan and McCarthy, 1958) приводят основанные на ряде обследований данные о том, какой процент всей выборки опрашивается при каждом обращении. Средние данные приведены в табл. 13.5.

Таблица 13.5
ЧИСЛО ОБРАЩЕНИЙ, ПОТРЕБОВАВШИХСЯ ДЛЯ ПОЛНОГО ОПРОСА*

Кто опрашивается	Процент опрошенных при			Процент неответивших	Итого
	первом обращении	втором обращении	третьем и последующих обращениях		
Любой взрослый**	70	17	8	5	100
Случайно отобранный взрослый	37	32	23	8	100

* Под полным опросом (completed interview) подразумевается, что получены ответы на все вопросы опросного листа и повторного обращения к данному лицу не требуется. — Примеч. ред.

** Два обследования, в которых опрашивались соответственно домашние хозяйки и работники ферм, были включены в группу «любой взрослый».

В обследованиях, где на вопросы мог отвечать любой взрослый, при первом обращении было опрошено 70%, а после двух обращений — 87% выборки. Для случаев, когда должен был опрашиваться случайно отобранный взрослый, очевидно увеличение издержек на выборочное исследование: здесь первое обращение дало только 37% требуемое число опросов. Заметный успех второго обращения отражает усилия, приложенные исследователями для того, чтобы заранее установить, когда намеченные для опроса лица окажутся дома.

Данных об относительных издержках на последующие обращения по сравнению с первым обращением опубликовано мало. Можно ожидать, что последующие обращения будут более дорогостоящими в расчете на один полный опрос, поскольку для них дома, находящиеся на отведенном исследователю участке, расположены более редко, а их обитатели, как можно предположить, проводят вне дома в среднем больше времени, чем остальные. По опыту обследований в Великобритании Дербин (Durbin, 1954) считает, что последующие обращения могут быть менее дорогостоящими, чем можно предположить. Далее приводятся опубликованные Дербином и Стюартом (Durbin and Stuart, 1954) данные специального исследования об оценках относительных издержек в расчете на один полный опрос (т. е. суммы, затраченной на i -е обращение, деленной на число вновь опрошенных при этом обращении) для каждого обращения, вплоть до пятого.

Определяя такие затраты, следует соблюдать осторожность. Если нужного лица не оказалось дома при первом обращении, то исследователь может затратить время, чтобы узнать, когда это лицо будет дома, и условиться о встрече. При расчетах издержек это время нужно отнести ко второму обращению, а не к неудавшемуся первому.

Таблица 13.6

ОТНОСИТЕЛЬНЫЕ ИЗДЕРЖКИ В РАСЧЕТЕ НА ОДИН НОВЫЙ ПОЛНЫЙ
ОПРОС ПРИ i -М ОБРАЩЕНИИ

Обращение	1	2	3	4	5
Относительные издержки	100	112	127	151	250

Более полезной характеристикой служат средние издержки в расчете на один полный опрос по всем опросам, проведенным в результате i обращений. По этому показателю определяются относительные издержки на проведение n полных опросов при условии, что отказ от опроса какого-либо лица допускается лишь после i попыток. Для того чтобы вычислить эти средние издержки, мы должны знать, сколько лиц опрашивается при каждом обращении. В табл. 13.7 соответствующие вычисления проделаны для двух видов обследований. К первому виду относятся такие обследования, в которых отвечать на вопросы может любой взрослый, ко второму — те, где опрашивается взрослый, отобранный случайным образом. Данные о числе проведенных опросов взяты из табл. 13.5.

Таблица 13.7

ОТНОСИТЕЛЬНЫЕ ИЗДЕРЖКИ НА ОДИН ПОЛНЫЙ ОПРОС ПРИ i ОБРАЩЕНИЯХ

Обращение	Относительные издержки	Опрашиваемый — любой взрослый					Опрашиваемый — «случайно отобранный» взрослый	
		при i -м обращении		при i обращениях			число опросов	издержки на один опрос
		число опросов	издержки на опросы	общее число опросов	общие издержки	издержки на один опрос		
1	100	0,70 n_0	70 n_0	0,70 n_0	70 n_0	100	0,37 n_0	100
2	112	0,17 n_0	19,04 n_0	0,87 n_0	89,04 n_0	102	0,32 n_0	106
3	127	0,07 n_0	8,89 n_0	0,94 n_0	97,93 n_0	104	0,16 n_0	110
4	151	0,04 n_0	6,04 n_0	0,98 n_0	103,97 n_0	106	0,09 n_0	114
5	250	0,02 n_0	5,00 n_0	1,00 n_0	108,97 n_0	109	0,06 n_0	122

Подробности вычислений указаны только для обследований первого вида; для обследований второго вида они делались точно так же. Символом n_0 обозначен исходный объем выборки.

Требование трехкратного обращения увеличивает издержки в расчете на один полный опрос по сравнению с однократным обращением лишь на 4%, когда опрашивается любой взрослый, и лишь на 10%, когда опрашивается случайно отобранный взрослый. Насколько типичны эти данные, сказать трудно, но если необходимые сведения о расходах и об объеме выборки уже собраны, то при этом способе расчета мы получаем реалистичные оценки издержек на требуемые повторные обращения. Имеет значение также и фактор времени: повторные обращения отодвигают срок получения окончательных результатов.

13.5. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ЭФФЕКТА ПОВТОРНЫХ ОБРАЩЕНИЙ

Деминг (Deming, 1953) разработал полезную и гибкую математическую модель, позволяющую более подробно изучить результаты различной тактики повторного обращения. Совокупность разделена на r классов в соответствии с вероятностью того, что опрашиваемый окажется дома. Пусть

w_{ij} — вероятность того, что опрашиваемый из j -го класса будет опрошен при i -м обращении или до него;

p_j — доля лиц совокупности, принадлежащих к j -му классу;

μ_j — среднее значение признака для j -го класса;

σ_j^2 — дисперсия признака для j -го класса.

Предположим для простоты, что для всех классов $w_{ij} > 0$, хотя модель нетрудно обобщить, так чтобы она включала и лиц, которых невозможно опросить. Если \bar{y}_{ij} — среднее значение признака у лиц из j -го класса, опрошенных при i -м обращении или до него, то предполагаем также, что $E(\bar{y}_{ij}) = \mu_j$.

Истинное среднее значение признака для совокупности равно:

$$\bar{\mu} = \sum_j p_j \mu_j. \quad (13.6)$$

Рассмотрим состав выборки после i обращений. Всех лиц, попавших в выборку, можно разбить на $(r+1)$ классов следующим образом: опрошенные из первого класса; опрошенные из второго класса; и т. д. $(r+1)$ -й класс охватывает всех неопрошенных после i обращений. Если пренебречь пкс, то числа попадающих в эти $(r+1)$ классов подчиняются полиномиальному распределению

$$[w_{i1} p_1 + w_{i2} p_2 + \dots + w_{ir} p_r + (1 - \sum w_{ij} p_j)]^{n_i},$$

где n_i — первоначальный объем выборки.

Отсюда следует, что число опрошенных за i обращений n_i распределено по биномиальному закону с числом испытаний, равным n_0 , и вероятностью успеха $\sum w_{ij} p_j$. Следовательно,

$$E(n_i) = \text{ожидаемому числу опрошенных за } i \text{ обращений} = n_0 \sum w_{ij} p_j. \quad (13.7)$$

При неизменном n_i число проведенных опросов n_{ij} ($j = 1, 2, \dots, r$) подчиняется полиномиальному распределению с вероятностями $w_{ij} p_j / \sum w_{ij} p_j$. Отсюда вытекает, что

$$E(n_{ij} | n_i) = \frac{n_i w_{ij} p_j}{\sum w_{ij} p_j}.$$

Следовательно, если \bar{y}_i — выборочное среднее, получаемое после i обращений, то

$$E(\bar{y}_i | n_i) = E\left(\frac{\sum n_{ij} \bar{y}_{ij}}{n_i}\right) = \frac{\sum n_i w_{ij} p_j \mu_j}{n_i \sum w_{ij} p_j} = \frac{\sum w_{ij} p_j \mu_j}{\sum w_{ij} p_j} = \bar{\mu}_i. \quad (13.8)$$

Поскольку результат не зависит от n_i , безусловное среднее значение \bar{y}_i также есть $\bar{\mu}_i$. Следовательно, смещение оценки \bar{y} равно $(\bar{\mu}_i - \bar{\mu})$.

Условная дисперсия \bar{y}_i при данном n_i находится аналогично:

$$V(\bar{y}_i | n_i) = \frac{\sum w_{ij} p_j [\sigma_j^2 + (\mu_j - \bar{\mu}_i)^2]}{n_i \sum w_{ij} p_j}. \quad (13.9)$$

Для того чтобы получить, пренебрегая членами порядка $1/n_i^2$, приближенное значение безусловной дисперсии, нужно заменить величину n_i в (13.9) ее математическим ожиданием из (13.7).

Окончательно получаем, что средний квадрат ошибки оценки, получаемой после i обращений, есть

$$\text{СКО}(\bar{y}_i | i) = V(\bar{y}_i | i) + (\bar{\mu}_i - \bar{\mu})^2. \quad (13.10)$$

Следует также учесть издержки на i обращений. Математическое ожидание числа новых опросов, проводимых при k -м обращении, есть $n_0 \sum (w_{kj} - w_{k-1,j}) p_j$. Следовательно, если c_k — издержки на проведение одного полного опроса при k -м обращении, то математическое ожидание общих издержек на i обращений есть $n_0 C(i)$, где

$$C(i) = c_1 \sum w_{1j} p_j + c_2 \sum (w_{2j} - w_{1j}) p_j + \dots + c_i \sum (w_{ij} - w_{i-1,j}) p_j.$$

Пример. В табл. 13.8 представлена совокупность, состоящая из трех классов. Значения p_j и w_{ij} подобраны таким образом, чтобы имитировать обследования, в которых опрашивается случайно отобранный взрослый. Вероятности проведения опроса в этих трех классах при первом обращении, w_{1j} , взяты равными 0,6; 0,3 и 0,1. При втором и последующих обращениях условные вероятности проведения опроса ранее пропущенного лица составляют 0,9; 0,5 и 0,2. Эти числа взяты более высокими, чем соответствующие вероятности при первом обращении для того, чтобы отразить влияние дополнительных усилий обследователей.

Таблица 13.8
ХАРАКТЕРИСТИКИ ТРЕХ КЛАССОВ

	Класс		
	1	2	3
p_j	0,45	0,50	0,05
w_{1j}	0,6+0,4 [1-(0,1) $^{i-1}$]	0,3+0,7 [1-(0,5) $^{i-1}$]	0,1+0,9 [1-(0,8) $^{i-1}$]
$I \mu_j$	55	50	45
$II \mu_j$	60	50	40

Необходимо оценить некоторый процент, близкий к 50, пользуясь биномиальным распределением. Рассматриваются два набора μ_j (I и II). Для простоты все дисперсии внутри классов, $\sigma_j^2 = \mu_j (100 - \mu_j)$, считаются равными 2500. Относительные издержки на проведение одного полного опроса при последовательных обращениях взяты из табл. 13.6.

В табл. 13.9 указаны (а) ожидаемое общее число опросов, проведенных после i обращений, (б) средние издержки на эти обращения в расчете на один опрос и (в) смещение $(\bar{\mu}_i - \bar{\mu})$ оценки \bar{y} при предположениях I и II относительно μ_j .

При предположении II, например, истинное среднее для совокупности, $\bar{\mu}$, равно 54%. Среднее $\bar{\mu}_i$ значений, полученных при первом обращении, равно 56,235%, что дает смещение, равное + 2,235%, как указано в таблице. Тактика, предусматривающая три обращения, уменьшает это смещение до + 0,842%.

Таблица 13.9

ЧИСЛО ОПРОСОВ, ИЗДЕРЖКИ НА ОДИН ОПРОС И СМЕЩЕНИЯ

Число требуемых обращений	Число проведенных опросов	Средние издержки на один опрос	Смещение I	Смещение II
1	0,425 n_0	100	+1,118	+2,235
2	0,771 n_0	105	+0,711	+1,421
3	0,882 n_0	108	+0,421	+0,842
4	0,933 n_0	110	+0,266	+0,532
5	0,960 n_0	114	+0,180	+0,360

Значения СКО (\bar{y}), получаемые при различной тактике повторных обращений и при заданной сумме расходов, сравниваются между собой. Сначала рассматривается случай, когда имеющихся средств достаточно для того, чтобы взять $n_0 = 500$ только при одном обращении. Согласно табл. 13.9 ожидаемое число опросов, проводимых при первом обращении, равно $E(n_1) = 500 \cdot 0,425 = 212,5$. Если делаются два обращения, то (при сохранении той же суммы расходов) это ожидаемое число опросов должно быть уменьшено до $E(n_2) = 212,5/1,05 = 202,4$. Аналогично определяется ожидаемое число опросов для 3, 4 и 5 обращений. Эти значения $E(n_i)$ подставляются в формулу (13.9) для того, чтобы получить $V(\bar{y})$ и, следовательно, СКО (\bar{y}).

В табл. 13.10 приведены значения СКО для трех величин затрат, соответствующих $n_0 = 500, 1000, 2000$ при единственном обращении. При $n_0 = 500$ приведены также значения СКО (\bar{y}) для случая «смещение отсутствует», при котором каждое $\mu_i = 50$. Этот столбец показывает эффект повторных обращений в том случае, когда они не нужны, поскольку, если ограничиться единственным обращением, никакого смещения не возникнет.

Таблица 13.10

ЗНАЧЕНИЯ СКО (\bar{y}) ПРИ РАЗЛИЧНОЙ ТАКТИКЕ ОБРАЩЕНИЙ И НЕИЗМЕННЫХ ЗАТРАТАХ

Требуемое число обращений	$n_0 = 500$ (только при одном обращении)			$n_0 = 1000$		$n_0 = 2000$	
	смещение отсутствует	I*	II*	I	II	I	II
1	11,8	13,0	16,9	7,1	10,9	4,2	8,0
2	12,4	12,9	14,6	6,7	8,3	3,6	5,2
3	12,7	12,9	13,6	6,5	7,1	3,4	3,9
4	13,0	13,1	13,4	6,6	6,9	3,3	3,6
5	13,5	13,5	13,8	6,8	6,9	3,4	3,5

* Соответствующие столбцы отвечают совокупностям с меньшими (I) и большими (II) величинами смещения, указанными в табл. 13.8.

Наименьшие значения СКО при соответствующей тактике повторных обращений выделены жирным шрифтом. Рассмотрим сначала случай наименьшего объема выборки, $n_0 = 500$. Когда повторные обра-

щения не нужны (при отсутствии смещения), тактика, предусматривающая до четырех обращений, ведет лишь к умеренному увеличению СКО. При предположении I, с меньшим смещением, различная тактика обеспечивает приблизительно одну и ту же достоверность, хотя оптимальная тактика требует трех обращений. При предположении II удовлетворительные результаты дают три — пять обращений, а если ограничиться одним обращением, то СКО будет приблизительно на 25% больше минимального.

При больших объемах выборки оптимальное число обращений увеличивается до четырех или пяти, а тактика единственного обращения ведет к более значительной потере достоверности.

Рассмотренный пример служит, конечно, лишь иллюстрацией. Значение же этого метода состоит в том, что коль скоро накоплены сведения об издержках и об относительных смещениях, он дает возможность разработать экономичную тактику для любого конкретного вида обследований.

13.6. ОПТИМАЛЬНАЯ ДОЛЯ ОТБОРА СРЕДИ НЕОТВЕТИВШИХ

Другой подход, разработанный Хансеном и Хервицем (Hansen and Hurwitz, 1946), заключается в том, чтобы, после того как сделана первая попытка опросить попавших в выборку лиц, извлечь случайную подвыборку из числа оставшихся неопрошенными и приложить максимальные усилия к тому, чтобы опросить каждого в этой подвыборке. Первоначально этот метод был разработан для обследований, в которых первый раз опрос производили по почте, а затем подвыборка лиц, не вернувших опросные листы, обследовалась более дорогостоящим методом устного опроса.

Сперва, с помощью обычных практических приемов берется простая случайная выборка объемом в n единиц. Пусть n_1 — число единиц выборки, по которым получены интересовавшие нас сведения, и n_2 — число единиц выборки в слое неответивших. В дальнейшем, с помощью более интенсивных усилий, получают сведения по случайной выборке объемом в r_2 единиц из числа n_2 единиц. Пусть

$$n_2 = k r_2 \quad (k > 1). \quad (13.11)$$

Тогда средняя доля отбора в первом слое в k раз больше, чем во втором. Это следует из того, что если k определено заранее, то

$$E\left(\frac{n_1}{N_1}\right) = E\left(\frac{n_2}{N_2}\right) = k E\left(\frac{r_2}{N_2}\right).$$

Значения n (первоначальный объем выборки) и k выбираются так, чтобы обеспечить определенную точность при наименьших издержках. Издержки на получение выборки равны:

$$C = c_0 n + c_1 n_1 + c_2 r_2,$$

где c_0 — издержки в расчете на одну единицу; c_1 — издержки на проведение первой попытки, c_2 — издержки на обработку данных

первой попытки и c_2 — издержки на получение и обработку данных во втором слое. Поскольку значения n_1 и n_2 до проведения первой попытки неизвестны, при планировании выборки пользуются *ожидаемыми* издержками. Ожидаемые значения n_1 и n_2 равны соответственно $W_1 n$ и $W_2 n/k$, где W_1, W_2 — истинные доли единиц в двух слоях. Тогда ожидаемые издержки равны

$$c_0 n + c_1 W_1 n + \frac{c_2 W_2 n}{k}. \quad (13.12)$$

Пусть \bar{y}_1, \bar{y}_{2r} — средние значения для выборки в двух слоях. Индекс r вводится для того, чтобы напомнить, что объем выборки во втором слое равен r_2 . В качестве оценки среднего для совокупности возьмем

$$\bar{y}' = \frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}_{2r}). \quad (13.13)$$

Заметим, что второму слою придается вес n_2 , хотя выборка из него имеет объем, равный только r_2 . Это делается для того, чтобы оценка была несмещенной.

Описанная методика представляет собой частный случай двойного отбора с расслоением. Первая, или «большая» выборка объема n дает оценку отношения объемов слоев, n_1/n_2 . Вторая, или «малая» выборка имеет объем n_1 в первом слое и r_2 во втором.

Для того чтобы найти $V(\bar{y}')$, запишем

$$\bar{y}' = \frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}_{2n}) + \frac{n_2}{n} (\bar{y}_{2r} - \bar{y}_{2n}), \quad (13.14)$$

где \bar{y}_{2n} — среднее для всей выборки объема n_2 в слое 2.

Первый член в правой части равенства есть среднее для случайной выборки объема n во всей совокупности. Следовательно, его дисперсия равна

$$\frac{(N-n)}{N} \frac{S^2}{n},$$

где S^2 — дисперсия для всей совокупности. Заметим далее, что при нахождении дисперсии \bar{y}' удвоенное произведение первого и второго членов дает нуль. Это происходит потому, что

$$E[\bar{y}_{2n}(\bar{y}_{2r} - \bar{y}_{2n})] = 0$$

по всем случайным выборкам объема r_2 , которые могут быть извлечены из неизменной выборки объема n_2 .

Рассмотрим второй член в правой части (13.14). Пусть \bar{Y}_2 — среднее значение для слоя неответивших. Тогда имеем

$$(\bar{y}_{2r} - \bar{Y}_2) = (\bar{y}_{2r} - \bar{y}_{2n}) + (\bar{y}_{2n} - \bar{Y}_2),$$

так что

$$E(\bar{y}_{2r} - \bar{Y}_2)^2 = E(\bar{y}_{2r} - \bar{y}_{2n})^2 + E(\bar{y}_{2n} - \bar{Y}_2)^2.$$

Удвоенное произведение дает нуль по тем же причинам, что и раньше. Далее, \bar{y}_{2r} есть среднее для простой случайной выборки объема r_2 во втором слое, а \bar{y}_{2n} — среднее для простой случайной выборки объема n_2 в том же слое. Следовательно, при *неизменных* n_2 и r_2

$$\frac{(N_2 - r_2)}{N_2} \frac{S_2^2}{r_2} = E(\bar{y}_{2r} - \bar{y}_{2n})^2 + \frac{(N_2 - n_2)}{N_2} \frac{S_2^2}{n_2},$$

где S_2^2 — дисперсия в слое неответивших. Это дает

$$E(\bar{y}_{2r} - \bar{y}_{2n})^2 = S_2^2 \left(\frac{1}{r_2} - \frac{1}{n_2} \right) = S_2^2 \frac{n_2 - r_2}{n_2 r_2} = S_2^2 \frac{k-1}{n_2},$$

так как $n_2 = k r_2$.

Следовательно, складывая дисперсии двух членов из (13.14) при неизменном n_2 , находим

$$\begin{aligned} V(\bar{y}') &= \frac{(N-n)}{N} \frac{S^2}{n} + \left(\frac{n_2}{n} \right)^2 \frac{(k-1)}{n_2} S_2^2 = \\ &= \frac{(N-n)}{N} \frac{S^2}{n} + \frac{(k-1)}{n^2} n_2 S_2^2. \end{aligned} \quad (13.15)$$

Поскольку $E(n_2) = n W_2$, ожидаемая дисперсия есть

$$\bar{V}(\bar{y}') = \frac{(N-n)}{N} \frac{S^2}{n} + \frac{(k-1) W_2}{n} S_2^2. \quad (13.16)$$

Первый член представляет собой дисперсию, которая была бы получена, если бы в слое неответивших были отобраны все n_2 единиц. Второй член есть приращение дисперсии, вызванное тем, что отбирается только r_2 единиц из n_2 .

Значения n и k подбираются теперь такими, чтобы при заданном значении ожидаемой дисперсии (13.16) были минимальны средние издержки (13.12).

Решением служат

$$k_{opt} = \sqrt{\frac{c_2 (S^2 - W_2 S_2^2)}{S_2^2 (c_0 + c_1 W_1)}}; \quad (13.17)$$

$$n_{opt} = \frac{N [S^2 + (k-1) W_2 S_2^2]}{NV + S^2}, \quad (13.18)$$

где V — заданная величина дисперсии оценки среднего для совокупности.

Для того чтобы получить решение, нужно знать W_2 ; часто его можно оценить по опыту предыдущих обследований. Кроме S^2 , значение которой следует оценить заранее при любой задаче определения объема выборки, в решении участвует также величина S_2^2 , дисперсия в слое неответивших. Естественно, что значение S_2^2 предсказать труднее; оно, вероятно, не совпадает с S^2 . Например, при проводимых по почте обследованиях большинства видов экономической деятельности ответы чаще поступают от более крупных единиц, дисперсия между которыми больше, чем дисперсия между неответившими.

Если W_2 известно неточно, то для получения удовлетворительного приближения нужно найти n_{opt} для интервала предполагаемых значений W_2 между 0 и надежной верхней границей. Максимальное из этих n_{opt} принимается в качестве первоначального объема выборки n . После того как вернутся заполненными отправленные по почте опросные листы, становится известным значение n_2 . Теперь выражение для дисперсии (13.15) можно разрешить относительно k для того, чтобы найти его значение, обеспечивающее желательную дисперсию V . При таком методе издержки обычно оказываются лишь незначительно больше оптимальных издержек, которые получились бы, если W_2 было известно.

Пример. Этот пример взят из статьи Хансена и Хервица (Hansen and Hurwitz, 1946). Первая выборка опрашивается по почте, причем ожидается, что процент ответивших составит около 50%. Желательно иметь такую точность, которая была бы получена для простой случайной выборки объемом в 1000 единиц, если бы неответивших не было. Издержки на отправку по почте одного опросного листа составляют 10 центов, а издержки на обработку заполненного опросного листа — 40 центов. Устный опрос одного человека стоит 4,1 долл.

Сколько опросных листов нужно разослать и какой процент неответивших нужно опросить устно?

В обозначениях функции издержек (13.12) издержки в расчете на одну единицу (в долларах) равны:

c_0 (издержки на первую попытку) = 0,1;
 c_1 (издержки на обработку данных в расчете на одного ответившего) = 0,4;
 c_2 (издержки на получение и обработку данных в расчете на одного неответившего) = 4,5.

Оптимальные n и k можно вычислить по формулам (13.17) и (13.18). Если дисперсии S^2 и S_2^2 считать равными и предположить, что N велико, то

$$k_{opt} = \sqrt{\frac{c_2(1-W_2)}{c_0+c_1W_1}} = \sqrt{\frac{4,5 \cdot 0,5}{0,1+0,4 \cdot 0,5}} = \sqrt{7,5} = 2,739;$$

$$n_{opt} = \frac{S^2 [1 + (k-1)W_2]}{V} = 1000 \{1 + 1,739 \cdot 0,5\} \doteq 1870.$$

Заметим, что мы должны положить $S^2/V = 1000$, или $V = S^2/1000$, так как последнее выражение есть дисперсия, которую имело бы среднее для выборки объемом в 1000 единиц при отсутствии неответивших.

Таким образом, должно быть разослано 1870 опросных листов. Приблизительно 935 из них не будут возвращены и мы должны провести устный опрос по случайной подвыборке объемом в 935/2,739, или опросить 341 человека. Издержки составят 2095 долл.

Как указал Дербин (Durbin, 1954), маловероятно, что применение подбора приведет к заметному выигрышу, если только c_2 не велико по сравнению с $(c_0 + c_1W_1)$. Эти две величины сравниваются потому,

что $(c_0 + c_1W_1)$ — ожидаемые издержки на проведение первой попытки и обработку результатов в расчете на одну единицу, а c_2 имеет тот же смысл для второй попытки. Из равенств, приведенных ранее, можно вывести, что отношение издержек на получение заданного V при $k = 1$ (отсутствие подбора) к минимальным издержкам при оптимальном k составляет

$$\frac{S^2(c_0+c_1W_1+c_2W_2)}{[V(S^2-W_2S_2^2)(c_0+c_1W_1)+\sqrt{c_2W_2S_2^2}]^2} =$$

$$= \frac{c_0+c_1W_1+c_2W_2}{[VW_1(c_0+c_1W_1)+\sqrt{c_2W_2}]^2}$$

при условии, что S^2 и S_2^2 приблизительно равны. Если отношение c_2 к $(c_0 + c_1W_1)$ обозначить через r , то указанное отношение издержек примет вид

$$\frac{1+rW_2}{(VW_1+\sqrt{rW_2})^2}.$$

Например, для $r = 4$, отношение издержек равно 1,029 при $W_1 = 0,5$; 1,074 — при $W_1 = 0,8$ и 1,061 — при $W_1 = 0,9$. Если, однако, S^2 существенно больше, чем S_2^2 , то подбор должен дать больший выигрыш.

При расслоенном отборе оптимальные значения n_h и k_h для отдельных слоев имеют весьма сложный вид. Для того чтобы получить хорошее приближение, нужно сначала с помощью методов из параграфов 5.5 и 5.6 оценить объемы выборки по слоям, n_{hA} , которые были бы необходимы, если бы неответивших не было. Далее, из (13.18) при $W_2 = 0$ имеем

$$n_0 = \frac{NS^2}{NV+S^2}.$$

Следовательно, (13.18) можно переписать в виде

$$n_{opt} = n_0 \left[1 + \frac{(k-1)W_2S_2^2}{S^2} \right].$$

Это равенство, примененное отдельно к каждому слою, дает приближение к оптимальным n_h . Оптимальные значения k_h находим, применяя в каждом слое формулу (13.17).

Описанные приемы можно применять при получении оценок по отношению и по регрессии. В случае оценки по отношению величины S^2 и S_2^2 нужно заменить величинами S_d^2 и S_{d2}^2 , где $d_i = y_i - Rx_i$. В случае оценки по регрессии вместо S^2 берется $S^2(1-\rho^2)$ и вместо S_2^2 берется $S_2^2(1-\rho^2)$.

13.7. ПОПРАВКИ НА СМЕЩЕНИЕ БЕЗ ПОВТОРНЫХ ОБРАЩЕНИЙ

Остроумный метод уменьшения смещений, присутствующих в результатах первого обращения, был предложен Хартли (Hartley, 1946) и развит в дальнейшем Полицем и Симмонсом (Politz and Simmons, 1949, 1950; Simmons, 1954). Будем считать, что все обращения пропе-

ходят в течение какого-либо из шести будних вечеров недели. Опрашиваемый должен ответить, был ли он дома в то время, когда проводится опрос, в каждый из пяти предшествующих будних вечеров. Если опрашиваемый сообщает, что он был дома t вечеров из пяти, то в качестве оценки частоты π , с которой он находится дома во время опроса, принимается отношение $(t+1)/6$.

Результаты первого обращения разбиваются на шесть групп в соответствии со значениями t (0, 1, 2, 3, 4, 5). Обозначим через n_t число опросов, проведенных в t -й группе, и через \bar{y}_t — среднее значение в ней наблюдаемого признака. Оценка Полица — Симмонса среднего значения для совокупности μ имеет вид

$$\bar{y}_{PS} = \frac{\sum_{t=0}^5 6n_t \bar{y}_t / (t+1)}{\sum_{t=0}^5 6n_t / (t+1)}.$$

Этот подход подчеркивает, что результаты первого обращения состоят в основном из наблюдений для лиц, проводящих дома большую часть времени. Поскольку лицо, проводящее дома, в среднем, долю π всего времени, имеет вероятность быть включенным в выборку π , его ответ должен получить вес, равный $1/\pi$. В качестве оценки $1/\pi$ служит величина $6/(t+1)$. Таким образом, оценка \bar{y}_{PS} смещена меньше, чем выборочное среднее по данным первого обращения, \bar{y} , но ее дисперсия больше, потому что невзвешенное среднее здесь заменяется взвешенным с весами, оцениваемыми по выборке.

При изучении среднего и дисперсии \bar{y}_{PS} мы воспользуемся обозначениями параграфа 13.5. Разделим совокупность на классы так, что лица из j -го класса проводят дома долю времени π_j . Заметим, что t -я группа (т. е. лица, находившиеся дома t вечеров из пяти предшествующих) содержит лиц из разных классов. Пусть n_{jt} , \bar{y}_{jt} — число лиц и среднее значение признака для лиц, принадлежащих одновременно классу j и группе t . Тогда \bar{y}_{PS} можно записать в следующем виде:

$$\bar{y}_{PS} = \frac{\sum \sum 6n_{jt} \bar{y}_{jt} / (t+1)}{\sum \sum 6n_{jt} / (t+1)}.$$

Это оценка типа оценки по отношению. Обозначим ее через N/D . Для больших выборок ее среднее значение приблизительно равно $E(N)/E(D)$.

Пусть n_0 — первоначальный объем выборки (число опрошенных плюс число не оказавшихся дома) и n_j — число опрошенных из класса j . Сделаем следующие предположения:

(а) $\frac{n_j}{n_0}$ есть биномиальная оценка $p_j \pi_j$;

(б) $E(n_{jt} | n_j) = n_j \frac{5!}{t!(5-t)!} \pi_j^t (1-\pi_j)^{5-t}$;

(в) $E(\bar{y}_{jt}) = \mu_j$ для всех j и t .

Предположение (б) может вызвать возражения. Не вдаваясь в подробное обсуждение, заметим, что, приняв его, мы считаем, что люди правильно сообщают о том, сколько раз они были дома.

При данном j , пользуясь предположением (б), имеем

$$E \sum_{t=0}^5 n_{jt} \left(\frac{6}{t+1} \right) = n_j \sum_{t=0}^5 \left(\frac{6}{t+1} \right) \frac{5!}{t!(5-t)!} \pi_j^t (1-\pi_j)^{5-t} = \frac{n_j}{\pi_j} [1 - (1-\pi_j)^6].$$

Отсюда, пользуясь предположением (а), получаем

$$E(D) = \sum_{j=1}^r \frac{E(n_j)}{\pi_j} [1 - (1-\pi_j)^6] = n_0 \sum_{j=1}^r p_j [1 - (1-\pi_j)^6].$$

Далее, поскольку $E(\bar{y}_{jt}) = \mu_j$ для любых j и t , приходим к приближенному равенству

$$E(\bar{y}_{PS}) = \bar{\mu}_{PS} \approx \frac{\sum_{j=1}^r p_j \mu_j [1 - (1-\pi_j)^6]}{\sum_{j=1}^r p_j [1 - (1-\pi_j)^6]}.$$

Поскольку истинное среднее $\bar{\mu} = \sum p_j \mu_j$, величина \bar{y}_{PS} по-прежнему сохраняет некоторое смещение. В некотором смысле эта оценка имеет то же смещение, что и \bar{y}_e , выборочное среднее, получаемое при применении метода повторных обращений в случае, когда делается, при необходимости, до шести обращений. В параграфе 13.5 из уравнения (13.8) следовало, что метод повторных обращений с i обращениями дает несмещенную оценку $\bar{\mu}_i = \sum w_{ij} p_j \mu_j / \sum w_{ij} p_j$, где w_{ij} — вероятность того, что лицо из класса j будет опрошено, если оно попадает в выборку. Пусть $w_{ij} = \pi_j$. Если при дальнейших обращениях вероятность застать дома ранее опрошенного человека остается равной π_j , то

$$w_{ij} = [1 - (1-\pi_j)^i],$$

так что $\bar{\mu}_{PS} = \bar{\mu}_6$. Однако (при применении метода повторных обращений) может оказаться, что вероятность быть опрошенным при последующем обращении будет больше π_j из-за того, что при первом или вообще при предшествующих обращениях обследователь получает некоторые сведения об отсутствующем. В этом случае метод повторных обращений с шестью обращениями даст меньшее смещение.

Дисперсия \bar{y}_{PS} имеет весьма сложный вид. С помощью обычного приближения для оценки по отношению, следуя Демингу (Deming, 1953), ее можно записать в виде

$$V(\bar{y}_{PS}) \approx \frac{1}{n_0 U} \{ \sum p_j p_j B_j [\sigma_j^2 + (\mu_j - \bar{\mu}_{PS})^2] + (n_0 - 1) \sum (\pi_j p_j)^2 (B_j - A_j^2) (\mu_j - \bar{\mu}_{PS})^2 \},$$

$$U = 1 - \sum p_j (1 - \pi_j)^6;$$

$$A_j = \frac{1}{\pi_j} [1 - (1 - \pi_j)^6];$$

$$B_j = \sum_{t=0}^5 \left[\frac{6}{(1+t)} \right]^2 \frac{5!}{t!(5-t)!} \pi_j^t (1 - \pi_j)^{5-t}.$$

Хотя реальную значимость каждого члена этой формулы трудно представить без применения ее к конкретным совокупностям, можно все же сделать два общих замечания. Если μ_j мало различаются, т. е. если смещение для первых обращений умеренное, то главную роль в этой формуле играет первый член

$$\frac{1}{n_0 U} \sum \pi_j p_j B_j \sigma_j^2.$$

Это выражение обычно бывает на 25—30% больше, чем дисперсия невзвешенного среднего для первых обращений. Кроме того, $V(\bar{y}_{PS})$ содержит член, который не убывает при увеличении n_0 и для очень больших выборок становится существенным.

В заключение заметим, что, как показал анализ имитированных совокупностей, проведенный Демингом (Deming, 1953), Дербином (Durbin, 1954) и автором этой книги, описанный только что метод имеет большие преимущества по сравнению с методом повторных обращений в том случае, когда смещения для первых обращений существенны и объем выборки велик. Уменьшение СКО при тех же издержках, однако, невелико, если только повторные обращения не обходятся значительно дороже, чем предполагалось ранее. Метод Полица — Симмонса имеет то преимущество, что он экономит время. К его недостаткам относится возможность неточного и неполного определения значений t , в анализе она не учитывалась. Описанный метод можно, как предложил Симмонс (Simmons, 1954), применять в сочетании с несколькими повторными обращениями.

Были предложены и некоторые другие методы исключения смещения, вызываемого «не оказавшимися дома». Бартоломью (Bartholomew, 1961) рассматривал обследование с двумя обращениями. Он предположил, что исследователь, путем тщательных расспросов, может добиться того, что вероятность встретить при втором обращении лиц, не оказавшихся дома при первом обращении, будет приблизительно одинаковой для всех этих лиц. Если это условие выполняется, то n_2 лиц, опрашиваемых при втором обращении, представляют собой случайную подвыборку из числа $(n_0 - n_1)$ лиц, пропущенных при первом обращении. Следовательно, выражение $[n_1 \bar{y}_1 + (n_0 - n_1) \bar{y}_2] / n_0$ есть несмещенная оценка среднего для выборки, которую мы первоначально намеревались получить. Этот метод хорошо проявил себя в некоторых обследованиях в Великобритании, в которых Бартоломью его применял. Киш и Хесс (Kish and Hess, 1959) указали на возможность применения при повторных обследованиях данных о не-

ответивших в прошлых обследованиях в качестве данных о неответивших в текущем обследовании. Для случая, когда смещение при первых обращениях обнаруживает определенную закономерность, как это было в табл. 13.1, некоторые экстраполяционные методы для оценивания средних результатов, которые могли бы показать неответившие, были предложены Хендриком (Hendricks, 1949).

13.8. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ОШИБОК НАБЛЮДЕНИЯ

Мысленно можно себе представить, что возможно произвести большое число независимых наблюдений i -й единицы. Пусть y_{ia} — значение, получаемое при α -м наблюдении. Тогда

$$y_{ia} = \mu_i + e_{ia},$$

где μ_i — истинное значение, e_{ia} — ошибка наблюдения.

Понятие «истинное значение» требует комментариев. Для некоторых признаков это понятие просто и конкретно. Например, при инвентаризации, проводимой путем выборочной проверки, истинным числом каких-то деталей можно считать число этих деталей, лежащих на полке склада в 12 часов дня в определенный день. В некоторых случаях истинное значение можно определить, указывая процесс, с помощью которого оно устанавливается. Истинное диастолическое давление крови у человека в указанное время можно определить как значение, получаемое в результате его измерения стандартным прибором согласно тщательно разработанным правилам. Мы, однако, отдаем себе отчет в том, что наш стандартный прибор производит измерения с ошибками и что можно рассчитывать на создание со временем более точного прибора. Что касается других признаков, таких, как, например, некоторые аспекты отношения служащего к его нанимателю или ощущение человеком своей способности справляться с повседневными проблемами, то никто не может утверждать, что он владеет удовлетворительным методом измерения их «истинного значения». Тем не менее это понятие остается полезным даже в таких случаях.

При повторных наблюдениях одной и той же единицы ошибки наблюдения e_{ia} будут следовать некоторому распределению частот. Обозначим через β_i и σ_i^2 — среднее и дисперсию e_{ia} для i -й единицы. Величина β_i представляет собой смещение, возникающее при наблюдении. Значения β_i и σ_i^2 зависят, конечно, от характера изучаемого признака и от средств измерения. Они могут зависеть также от многих других факторов. В отношении совокупностей людей на ответы опрашиваемых могут влиять преобладающие в данный момент экономические условия и политические настроения, а также масштаб и характер предварительных сведений об обследовании, которые получает население.

Следующий шаг состоит в том, чтобы рассмотреть, как меняются ошибки наблюдения при переходе от одной единицы к другой. Здесь также могут возникнуть различные осложнения.

Что касается слагаемого, представляющего собой смещение, β_i , то может существовать постоянное смещение, скажем, $E(\beta_i) = \beta$, присутствующее всем единицам совокупности. Кроме него будет присутствовать

слагаемое ($\beta_i - \beta$), следующее некоторому распределению частот во всей совокупности. Это слагаемое может быть коррелировано с истинным значением μ_i , например измерительное устройство может систематически преуменьшать большие и преувеличивать малые значения μ_i .

Может также существовать корреляция между значениями e_{ia} для различных единиц в одной и той же выборке. Простейшим примером служит «смещение обследователя». Иногда в средних значениях y_{ia} , полученных разными обследователями для сравнимых частей одной и той же совокупности, обнаруживались просто поразительные различия (см. (Lienau, 1941), (Mahalanobis, 1946) и (Bart, 1957)).

Подобный эффект возникал также, когда выборки для оценивания будущего урожая отбирали разные бригады обследователей и когда химические или биологические анализы производились в разных лабораториях. Участие человека — не единственный фактор, вызывающий корреляцию между единицами, наблюдаемыми приблизительно в одно и то же время. На многие процессы наблюдения влияет погода; иногда применяется сырье, качество которого меняется от партии к партии. Как указывают Хансен, Хервиц и Бершо (Hansen, Hurwitz and Bershad, 1961), при оценивании нынешней продажной цены домов, построенных несколько лет назад, обнаружилось, что обследователи и домовладельцы, определяя стоимость домов, не продававшихся в течение многих лет, руководствовались ценами тех домов в выборке, которые были проданы недавно. Таким образом, средняя цена, полученная по выборке, может зависеть от порядка, в каком появляются в выборке недавно проданные дома.

Для того чтобы исследовать такие внутривыборочные корреляции в наиболее общем виде, требуется более сложная модель, чем представленная здесь. В частности, обозначения для e_{ia} и β_i должны были бы указывать на то обстоятельство, что их величина может зависеть от того, какие еще единицы оказались в выборке. Однако те виды корреляции, которые, по-видимому, наиболее типичны для практики, вполне могут быть отражены представленной моделью или ее простыми обобщениями.

Слагаемые ошибки наблюдения сведены в табл. 13.11.

Таблица 13.11
СЛАГАЕМЫЕ ОШИБКИ НАБЛЮДЕНИЯ ПО i -Й ЕДИНИЦЕ

Символ	Характер слагаемого ошибки
β	Смещение, постоянное для всех единиц
$\beta_i - \beta$	Переменная составляющая смещения, подчиняющаяся при изменении i некоторому распределению частот со средним значением нуль, и, возможно, коррелированная с истинным значением μ_i .
$d_{ia} = e_{ia} - \beta_i$	Флуктуирующая составляющая ошибки, при неизменном i и меняющемся α , подчиняющаяся некоторому распределению частот со средним значением нуль и дисперсией σ_i^2 .

Мы уже отмечали ранее, что величины β_i и d_{ia} для различных единиц в одной и той же выборке могут быть коррелированы друг с другом.

Модели, схожие в общих чертах с только что описанной, исследовались в работах Хансена и др. (Hansen et al., 1951), Сукхатма и Сетха (Sukhatme and Seth, 1952), а также Хансена, Хервица и Бершо (Hansen, Hurwitz and Bershad, 1961).

13.9. ЭФФЕКТ ПОСТОЯННОГО СМЕЩЕНИЯ

Предположим, что значения наблюдений y_i у всех единиц подвержены некоторому постоянному смещению β , величина которого неизвестна. Тогда среднее значение \bar{y} для простой случайной выборки также подвержено смещению β . В оценке дисперсии выборочного среднего это смещение сокращается, поскольку оценка вычисляется как сумма квадратов вида $(y_i - \bar{y})^2$. Следовательно, доверительные границы для \bar{Y} , вычисленные по данным выборки обычным путем, не будут учитывать этого смещения. Аналогичные утверждения справедливы и для расслоенного случайного отбора.

В сущности, то же положение возникает и в случае оценок по отношению и по регрессии. Рассмотрим оценку по регрессии

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}),$$

где как y_i , так и x_i могут быть подвержены постоянным смещениям соответственно β_y и β_x . Поскольку при этом значение оценки b , полученной методом наименьших квадратов, не меняется, и смещение β_x сокращается в члене $(\bar{X} - \bar{x})$, величина \bar{y}_{lr} имеет смещение β_y . Нетрудно проверить, что выборочная оценка дисперсии $V(\bar{y}_{lr})$ не содержит слагаемого, обусловленного этими смещениями.

В случае оценки по отношению

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} \bar{X}$$

смещение в первом приближении также равно β_y , поскольку для больших выборок $E(\bar{X}/\bar{x})$ приблизительно равно 1, даже если x_i подвержены постоянному смещению. Для больших выборок оценка дисперсии по выборке

$$v(\bar{y}_R) = \frac{(N-n)}{Nn} \frac{\sum (y_i - \hat{R}x_i)^2}{n-1},$$

рассматриваемая в качестве оценки

$$E(\bar{y}_R - \bar{Y})^2,$$

т. е. в качестве оценки дисперсии относительно смещенного среднего, \bar{Y} , почти не имеет смещения.

Из всего сказанного следует, что постоянного смещения по данным выборки обнаружить нельзя. Как мы уже видели (параграф 1.7), 95%-ные доверительные границы почти не изменяются, если отношение β_i к стандартной ошибке оценки среднего меньше 0,1, но если это отношение превышает указанное значение, то вычисленные доверительные границы могут ввести в заблуждение. Оценки изменения от одного периода времени к другому или от одного слоя к другому остаются несмещенными при условии, что смещение везде одинаково.

13.10. ЭФФЕКТ ОШИБОК, НЕКОРРЕЛИРОВАННЫХ ВНУТРИ ВЫБОРКИ

Если постоянным смещением можно пренебречь и ошибки наблюдения некоррелированы внутри выборки, то остаются справедливыми обычные формулы, предназначенные для оценивания стандартных ошибок выборочных оценок при условии, что члены пкс пренебрежимо малы. Это утверждение будет доказано для случая простого случайного отбора.

Нашу модель можно записать в виде

$$y_{i\alpha} = \mu_i + \beta_i + d_{i\alpha} = \mu'_i + d_{i\alpha}, \quad (13.19)$$

где $\mu'_i = \mu_i + \beta_i$ есть среднее значение наблюдений i -й единицы, получаемых с помощью определенного процесса. Поскольку постоянное смещение не учитывается,

$E(\beta_i) = 0$; $E(\mu'_i) = \mu$ — истинному среднему для совокупности.

Мы предполагаем, и для большинства обследований это справедливо, что по каждой единице производится только одно наблюдение. Выборочные средние величины $y_{i\alpha}$, μ'_i и $d_{i\alpha}$ обозначим через \bar{y}_α , $\bar{\mu}'$ и \bar{d}_α . Согласно (13.19)

$$\bar{y}_\alpha - \mu = (\bar{\mu}' - \mu) + \bar{d}_\alpha.$$

Поскольку, по определению $d_{i\alpha}$, величина $E(d_{i\alpha}|i) = 0$, среднее \bar{y}_α есть несмещенная оценка μ при простом случайном отборе. Далее,

$$(\bar{y}_\alpha - \mu)^2 = (\bar{\mu}' - \mu)^2 + \bar{d}_\alpha^2 + 2\bar{d}_\alpha(\bar{\mu}' - \mu). \quad (13.20)$$

Следовательно,

$$V(\bar{y}_\alpha) = E(\bar{\mu}' - \mu)^2 + E(\bar{d}_\alpha^2); \quad (13.21)$$

удвоенное произведение обращается в нуль, так как среднее значение $d_{i\alpha}$ равно нулю для любого i .

Сначала возьмем среднее по многократным наблюдениям одного и того же набора n единиц. Поскольку $d_{i\alpha}$ для различных единиц независимы,

$$E(\bar{d}_\alpha^2) = \frac{1}{n^2} \sum_i \sigma_i^2.$$

Если теперь взять среднее по всем простым случайным выборкам, то мы получим

$$V(\bar{y}_\alpha) = \frac{1-f}{n} \frac{\sum_i (\mu'_i - \mu)^2}{N-1} + \frac{1}{nN} \sum_i \sigma_i^2 = \frac{1-f}{n} S_{\mu'}^2 + \frac{1}{n} \sigma_d^2, \quad (13.22)$$

где σ_d^2 обозначает среднее значений дисперсий ошибок наблюдения.

Заметим, что $S_{\mu'}^2$ — дисперсия величин $(\mu_i + \beta_i)$ для совокупности. Она обычно больше, чем S_μ^2 , дисперсия истинных значений μ_i для совокупности, хотя она может быть и меньше, если корреляция μ_i и β_i отрицательна. При $n = N$, т. е. когда выборка охватывает всю совокупность, $V(\bar{y}_\alpha)$ не обращается в нуль, потому что ошибки наблюдения дают слагаемое σ_d^2/N .

Из уравнения (13.22), как заметили Хансен, Хервиз и Бершо (Hansen, Hurwitz and Bershaw, 1961), вытекает интересное следствие для случая, когда оценивается доля P для совокупности. Пусть для любой единицы истинное значение μ_i равно 1, если единица принадлежит классу C , и 0 в противном случае. Тогда

$$E(\mu_i) = P; \quad S_\mu^2 = \frac{\sum_i (\mu_i - \mu)^2}{N-1} = \frac{N}{N-1} PQ. \quad (13.23)$$

Если наблюдение производится с ошибками, то это значит, что некоторые единицы классифицируются неправильно. Для каждой такой единицы при многократных наблюдениях регистрируемое значение $y_{i\alpha}$ иногда будет равно 1, а иногда 0. Пусть P_i — доля наблюдений i -й единицы, при которых $y_{i\alpha} = 1$. Тогда $y_{i\alpha}$ будет случайной переменной, распределенной по биномиальному закону со средним значением $\mu'_i = P_i$ и с дисперсией $d_{i\alpha}$, равной $P_i Q_i$. Следовательно, если выборочная оценка есть $p_\alpha = \bar{y}_\alpha$, то (13.22) принимает вид

$$V(p_\alpha) = \frac{1-f}{n} \frac{\sum_i (P_i - P)^2}{N-1} + \frac{1}{nN} \sum_i P_i Q_i < \frac{1}{n(N-1)} \left[\sum_i (P_i - P)^2 + \sum_i P_i Q_i \right] = \frac{1}{n(N-1)} \left(\sum_i P_i^2 - NP^2 + \sum_i P_i - \sum_i P_i^2 \right) = \frac{N}{n(N-1)} PQ = \frac{S_\mu^2}{n} \quad (13.24)$$

на основании (13.23).

Величина S_μ^2/n есть то значение, которое приняла бы дисперсия $V(p_\alpha)$, если бы распределение по классам производилось без ошибок и n/N было пренебрежимо мало. Если распределение по классам происходит с ошибками, то результат в формуле (13.24), согласно которой $V(p_\alpha) \leq S_\mu^2/n$, обнадеживает. Он может показаться, однако, парадоксальным, поскольку можно было ожидать, что эти ошибки должны оказывать большее влияние на $V(p_\alpha)$. Объяснить его следует тем, что при

оценивании доли корреляция между μ_i и β_i всегда отрицательна. Если $\mu_i = 1$, то $\beta_i \leq 0$, поскольку $P_i = (\mu_i + \beta_i) \leq 1$, а если $\mu_i = 0$, то $\beta_i \geq 0$, поскольку $P_i \geq 0$.

Возвращаясь к случаю непрерывной переменной, запишем обычную формулу оценки дисперсии $V(\bar{y}_\alpha)$ по выборке

$$v(\bar{y}_\alpha) = \frac{1-f}{n} s^2 = \frac{1-f}{n} \frac{\sum (y_{i\alpha} - \bar{y}_\alpha)^2}{n-1}.$$

Согласно (13.19)

$$y_{i\alpha} - \bar{y}_\alpha = (\mu'_i - \bar{\mu}') + (d_{i\alpha} - \bar{d}_\alpha).$$

Возводя в квадрат и беря среднее сначала по многократным наблюдениям, а затем по всем возможным извлечениям выборки, получаем

$$E(s^2) = E\left[\frac{\sum (y_{i\alpha} - \bar{y}_\alpha)^2}{n-1}\right] = S_{\mu'}^2 + \sigma_d^2.$$

Следовательно,

$$Ev(\bar{y}_\alpha) = \frac{1-f}{n} S_{\mu'}^2 + \frac{1-f}{n} \sigma_d^2. \quad (13.25)$$

Сравнивая с формулой (13.22), мы видим, что $v(\bar{y}_\alpha)$ имеет отрицательное смещение, равное σ_d^2/N . С другой стороны, если в выражении для $v(\bar{y}_\alpha)$ опустить член пкс $(1-f)$, то значение дисперсии будет преувеличено на величину $S_{\mu'}^2/N$.

Тем же путем можно показать, что остаются справедливыми приведенные в предыдущих главах формулы выборочных оценок дисперсий для расслоенного отбора и для многоступенчатого отбора и что при больших выборках остаются справедливыми приближенные формулы для оценок по отношению и оценок по регрессии при условии, что ошибки наблюдения величины $y_{i\alpha}$ и $x_{i\alpha}$ некоррелированы внутри выборки и что поправками на конечность совокупности можно пренебречь. (Ошибка $y_{i\alpha}$ может быть коррелирована с ошибкой для соответствующего $x_{i\alpha}$.)

В связи с тем что сказанным возникает вопрос: в какого рода обследованиях ошибки наблюдения внутри выборки некоррелированы? Поскольку корреляция внутри выборки может возникать в процессе наблюдения, при дублировании наблюдений, при исправлении и кодировании записей, особенно если по ходу дела принимаются субъективные решения, и при переносе данных для машинной обработки, то с предположением о том, что корреляция отсутствует, спешить не следует. Однако в обследованиях, производимых на основании документов, при саморегистрации (как, например, при обследованиях по почте), когда лица из одной выборки не могут консультироваться друг с другом, и в обследованиях совокупностей, состоящих из неодушевленных предметов, когда наблюдения объективны, возможность такой корреляции должна быть минимальной.

13.11. ЭФФЕКТ КОРРЕЛЯЦИИ МЕЖДУ ОШИБКАМИ ВНУТРИ ВЫБОРКИ

Модель

$$y_{i\alpha} = \mu_i + \beta_i + d_{i\alpha} = \mu'_i + d_{i\alpha}$$

дает возможность рассмотреть некоторые из наиболее типичных видов корреляции внутри выборки в предположении, что величины $d_{i\alpha}$ для единиц в одной и той же выборке коррелированы. При нахождении $V(\bar{y}_\alpha)$ все рассуждения, проведенные в параграфе 13.8, полностью сохраняются вплоть до равенства (13.21):

$$V(\bar{y}_\alpha) = E(\bar{\mu}' - \mu)^2 + E(\bar{d}_\alpha^2). \quad (13.21)$$

Теперь

$$\bar{d}_\alpha^2 = \frac{1}{n^2} \left(\sum_i d_{i\alpha}^2 + 2 \sum_{i>j} d_{i\alpha} d_{j\alpha} \right).$$

Следовательно,

$$E(\bar{d}_\alpha^2) = \frac{1}{n} \sigma_d^2 + \frac{2n(n-1)}{2n^2} E(d_{i\alpha} d_{j\alpha}), \quad (13.26)$$

где произведения во втором члене берутся по всем парам единиц в одной и той же выборке. По аналогии с гнездовым отбором средний коэффициент корреляции внутри выборки, ρ_w , можно определить с помощью равенства:

$$E(d_{i\alpha} d_{j\alpha}) = \rho_w \sigma_d^2.$$

Отсюда, пользуясь (13.22) и (13.26), получаем

$$V(\bar{y}_\alpha) = \frac{1-f}{n} S_{\mu'}^2 + \frac{\sigma_d^2}{n} [1 + (n-1) \rho_w]. \quad (13.27)$$

Среднее значение $v(\bar{y}_\alpha)$ определяется тем же путем и оказывается равным:

$$Ev(\bar{y}_\alpha) = \frac{1-f}{n} S_{\mu'}^2 + \frac{1-f}{n} \sigma_d^2 (1 - \rho_w). \quad (13.28)$$

Поскольку для большинства видов ошибок наблюдения ρ_w , по-видимому, положительно, обычная формула для $v(\bar{y}_\alpha)$, как правило, преуменьшает значение дисперсии. Значительно ли такое преуменьшение, зависит от соотношения между величинами $S_{\mu'}^2$, σ_d^2 и ρ_w .

Эта модель отражает только простейший вид корреляции внутри выборки. При расслоенном отборе, например, работник, кодирующий записи, может обрабатывать материалы по нескольким слоям и из-за неправильного понимания инструкций может внести в эти данные ошибки, которые будут коррелированы между слоями. Рассмотренную математическую модель можно приспособить к ситуациям такого типа,

13.12. ЭФФЕКТ ОШИБОК НАБЛЮДЕНИЯ. РЕЗЮМЕ

Согласно описанной модели среднее для простой случайной выборки, \bar{y} , было бы несмещенным и имело бы дисперсию S_y^2/n (если пренебречь пкс), если бы все наблюдения были вполне достоверными. Из-за ошибок наблюдения рассмотренных видов это среднее может иметь смещение величины β и средний квадрат его ошибки есть

$$\text{СКО}(\bar{y}_\alpha) = \frac{1}{n} \{S_y^2 + \sigma_\alpha^2 [1 + (n-1)\rho_\alpha]\} + \beta^2, \quad (13.29)$$

где $\mu'_i = \mu_i + \beta$.

Формула (13.29) содержит два члена: S_y^2/n и $\sigma_\alpha^2(1 - \rho_\alpha)/n$, которые с увеличением n уменьшаются пропорционально $1/n$. Два оставшихся члена ($\rho_\alpha\sigma_\alpha^2$ и β^2) на первый взгляд кажутся не зависящими от n . Такое предположение было бы, вероятно, слишком упрощенным. Всякое существенное изменение объема выборки может потребовать некоторого изменения в методах непосредственного наблюдения, а это может повлиять на ρ_α и на β^2 . Однако при изменении n эти два члена должны меняться (если они вообще меняются) сравнительно медленно. Поэтому для больших выборок величина СКО будет, по-видимому, определяться этими двумя членами, а роль обычной выборочной дисперсии становится незначительной, что мешает судить по ней о действительной достоверности выборочных данных.

13.13. ИЗУЧЕНИЕ ОШИБОК НАБЛЮДЕНИЯ

В последние годы при анализе практики выборочного исследования много усилий было посвящено изучению ошибок наблюдения. Целью этого было установить, какие факторы оказывают наибольшее влияние на слагаемые СКО, и найти пути уменьшения этого влияния. Некоторые из основных методов такого анализа описаны в этом и следующем параграфах. Сейчас уже ясно, что прогресс в этом направлении будет медленным и обойдется недешево. Одна из причин этого заключается в том, что, как уже упоминалось, ошибки наблюдения неразрывно связаны с характером изучаемых признаков и с процессом наблюдения. Данные об ошибках наблюдения в одном обследовании редко можно считать пригодными для других обследований.

В идеале наилучший способ изучения ошибок наблюдения предполагает получение истинных значений μ_i . На практике такой подход ограничен признаками, для которых существует приемлемый способ нахождения μ_i , а также издержками, связанными с необходимым для этого обследованием и возможностью его провести. Примеры можно найти у Беллока (Belloc, 1954), который сравнивал данные о том, лечился ли человек в больнице, полученные при опросе на дому, с данными, основанными на больничной регистрации, и у Грея (Gray, 1955), который сравнивал сведения об отпусках по болезни, сообщаемые служащими, с записями в их личных делах. Проверки такого типа — называемые иногда «проверками по документам» — можно

производить по таким признакам, как возраст, занятие, число лет обучения и плата за автомобиль. Одна из трудностей состоит в том, что иногда в документах нет сведений, позволяющих точно установить, то ли же самое лицо было опрошено.

Не располагая подходящим способом определения истинного значения, можно поступить иначе, а именно, вновь провести наблюдение некоторым независимым методом, который считается более достоверным. Киш и Лансинг (Kish and Lansing, 1954), например, для того чтобы оценить продажную цену домов, которая ранее уже была сообщена самими владельцами, приглашали профессиональных оценщиков. При обследованиях заболеваемости ответы опрашиваемых сравнивались либо с записями врача о состоянии здоровья опрашиваемого, либо со сведениями, полученными после полного медицинского обследования [(Sagen, Dunham and Simmons, 1959) и (Trussell and Elinson, 1959)]. Результаты такого рода сравнений не всегда легко поддаются интерпретации в терминах описанной ранее модели, поскольку и более совершенные средства наблюдения могут давать ошибки, однако такие сравнения, по крайней мере, обнаруживают, по каким признакам результаты, полученные с помощью обычных средств и более совершенных средств, согласуются, а по каким нет.

При обследованиях домохозяйств еще одна возможность состоит в том, чтобы повторно опросить некоторую подвыборку из числа уже опрошенных. В одном варианте этого метода роль более достоверного средства наблюдения играет бригада исследователей, проводящих повторный опрос: в нее включаются лучшие работники и опросный лист делается более подробным и тщательно испытывается. При таком подходе основное значение придается отысканию признаков, относительно которых первый опрос оказался недостоверным, и причин этой недостоверности. В другом варианте бригаду для повторного опроса стремятся подобрать из работников того же уровня, что и для первого. Если можно считать, что две эти бригады исследователей дают два независимых набора наблюдений: $y_{i\alpha} = \mu'_i + d_{i\alpha}$ и $y_{i\alpha'} = \mu'_i + d_{i\alpha'}$, то половина квадрата разности, $(y_{i\alpha} - y_{i\alpha'})^2/2$, могла бы служить оценкой дисперсии ответов, σ_i^2 , для соответствующего признака. При обоих подходах возникает затруднение, связанное с тем, что второй ответ может не быть независимым от первого. Опрашиваемый может просто повторить по памяти ответ, который он дал при первом опросе, так что расхождения в ответах преуменьшат σ_i^2 . Если ради сокращения этого влияния увеличить промежуток времени между двумя посещениями, то при втором опросе опрашиваемый может уже плохо помнить то, о чем его спрашивали раньше, так что расхождения в ответах преувеличат σ_i^2 .

Иногда удается произвести общее сравнение результатов двух различных обследований. Для целого ряда признаков данные переписи населения США можно сравнить с аналогичными данными Текущего обследования населения (Current Population Survey), проводимого в то же самое время. Поскольку Текущее обследование населения считается более достоверным, особенно в отношении тех признаков, наблюдение которых затруднительно, то можно приближенно оценить величину

смещения β , вызванного ошибками наблюдения в данных переписи (Hansen, Hurwitz and Bershady, 1961). Стивен и Маккарти (Stephan and McCarthy, 1958) анализируют результаты сопоставления данных, полученных путем отбора квотами и путем вероятностного отбора.

13.14. ВЗАИМОПРОНИКАЮЩИЕ ПОДВЫБОРКИ

Этот метод, предложенный Махаланобисом (Mahalanobis, 1946), особенно полезен при изучении коррелированных ошибок. Для того чтобы представить его в наиболее простом виде, предположим, что случайная выборка объемом в n единиц случайным образом разбивается на k подвыборок, каждая из которых содержит $m = n/k$ единиц. Собственно обследование и разработка материалов выборки планируются так, чтобы не возникало корреляции между ошибками наблюдения любых двух единиц в различных подвыборках. Предположим, например, что корреляция, которую мы должны исследовать, возникает исключительно из-за смещений, обусловленных работой исследователей (interviewer bias — в дальнейшем — смещение у исследователей. — Примеч. перев.). Если каждому из k исследователей выделена отдельная подвыборка и нет корреляции между ошибками наблюдения у разных исследователей, то это и будет примером применения упомянутого метода.

Применяя уже рассмотренную математическую модель, удобно снабдить единицы двойным индексом. Пусть

$$y_{ija} = \mu'_{ij} + d_{ija},$$

где i — номер подвыборки (исследователя) и j — член внутри подвыборки. Пкс не учитывается.

Поскольку i -я подвыборка представляет собой случайную подвыборку, она сама есть простая случайная выборка объема m . Следовательно, согласно (13.27), дисперсия ее среднего есть

$$V(\bar{y}_{ia}) = \frac{1}{m} \{S_{\mu'}^2 + \sigma_d^2 [1 + (m-1) \rho_w]\},$$

где ρ_w — коэффициент корреляции между значениями d_{ija} , полученными одним и тем же исследователем. Так как ошибки в разных подвыборках независимы,

$$V(\bar{y}_a) = \frac{1}{k} V(\bar{y}_{ia}) = \frac{1}{n} \{S_{\mu'}^2 + \sigma_d^2 [1 + (m-1) \rho_w]\}. \quad (13.30)$$

На основе данных выборки мы можем провести дисперсионный анализ, разложив дисперсию на слагаемые, имеющие своим источником вариацию «между исследователями (подвыборками)» и «у исследователей (внутри подвыборок)». Нетрудно проверить, что математические ожидания значений средних квадратов имеют вид, указанный в табл. 13.12.

Таблица 13.12
МАТЕМАТИЧЕСКИЕ ОЖИДАНИЯ СРЕДНИХ КВАДРАТОВ
(НА ОСНОВАНИИ ОТДЕЛЬНОЙ ЕДИНИЦЫ)

	Степени свободы	Средние квадраты	E (средних квадратов)
Между обследователями (подвыборками)	$k-1$	$s_b^2 = \frac{m \sum (\bar{y}_{ia} - \bar{y}_a)^2}{k-1}$	$s_{\mu'}^2 + \sigma_d^2 [1 + (m-1) \rho_w]$
У обследователей	$k(m-1)$	$s_w^2 = \frac{\sum \sum (y_{ija} - \bar{y}_{ia})^2}{k(m-1)}$	$s_{\mu'}^2 + \sigma_d^2 (1 - \rho_w)$

Из этого анализа вытекают два важных утверждения. Сравнивая с (13.30), мы видим, что s_b^2/n есть несмещенная оценка $V(\bar{y}_a)$. Таким образом, описываемый метод дает оценку ошибки, которая учитывает смещения у исследователей. Кроме того, этот анализ дает сведения о значении ρ_w . Величина F -отношения s_b^2/s_w^2 дает критерий для проверки нулевой гипотезы $\rho_w = 0$. Величина $(s_b^2 - s_w^2)/m$ есть несмещенная оценка $\rho_w \sigma_d^2$. Этот факт позволяет нам ответить на вопрос: какую часть $V(\bar{y}_a)$ составляет слагаемое, обусловленное смещением у исследователей? Для того чтобы представить ситуацию, когда смещение у исследователей отсутствует, нужно положить $\rho_w = 0$ и допустить, что σ_d^2 остается без изменения. При этих предположениях $V(\bar{y}_a)$ принимает вид

$$V'(\bar{y}_a) = \frac{1}{n} (S_{\mu'}^2 + \sigma_d^2).$$

Несмещенная оценка этой величины есть

$$v'(\bar{y}_a) = \frac{1}{n} \left[s_w^2 + \frac{(s_b^2 - s_w^2)}{m} \right] = \frac{1}{n} \frac{[(m-1) s_w^2 + s_b^2]}{m}.$$

Это выражение нужно сравнить с s_b^2/n , оценкой фактической дисперсии, $V(\bar{y}_a)$.

В другой модели, которая дает те же результаты, смещение у i -го исследователя выражается членом g_i , среднее значение которого по всем возможным i равно нулю, а дисперсия σ_g^2 [I — от английского «interviewer» — исследователь]. Таким образом,

$$y_{ija} = \mu'_{ij} + g_i + d_{ija}, \quad (13.31)$$

где d_{ija} теперь некоррелированы одно с другим и с g_i . Для этой модели

$$V(\bar{y}_a) = \frac{1}{n} (S_{\mu'}^2 + m \sigma_g^2 + \sigma_d^2) \quad (13.32)$$

и из дисперсионного анализа следует, что s_b^2 , как и ранее, оказывается несмещенной оценкой $nV(\bar{y}_a)$. Ситуация, когда смещения у исследова-

телей отсутствуют, воспроизводится, если положить $g_i = 0$ и заменить σ_a^2 членом $(\sigma_i^2 + \sigma_j^2)$ для того, чтобы дисперсия ошибки, с которой исследователь оценивает значения признака у отдельной единицы, осталась без изменения.

Как подразумевает само название *взаимопроницающие* подвыборки, очень важно, чтобы подвыборки выбирались случайным образом. Обычно каждому обследователю отводят для наблюдения единицы, расположенные на небольшом участке территории вблизи его дома с тем, чтобы уменьшить путевые расходы. В этом случае любые реальные различия между средними \bar{y}_i для различных участков выступают при дисперсионном анализе как смещения у обследователей. Таким образом, s_b^2/n будет преувеличивать значение $V(\bar{y}_a)$.

Описанный метод применим при расслоенном и при многоступенчатом отборе. Если мы заинтересованы лишь в том, чтобы получить несмещенную оценку $V(\bar{y}_a)$, то для этого нужно лишь, чтобы выборка состояла из нескольких подвыборок одинаковой структуры и чтобы мы могли быть уверены поэтому в независимости ошибок наблюдения в разных подвыборках. Строго говоря, для этого необходимо, чтобы для разных подвыборок были отведены разные бригады обследователей, разные контролеры и разные средства обработки полученных данных. Если \bar{y}_{ia} — среднее значение для i -й подвыборки, то величина $\sum (\bar{y}_{ia} - \bar{y}_a)^2 / k(k-1)$ есть несмещенная оценка дисперсии $V(\bar{y}_a)$ с $(k-1)$ степенями свободы. Это утверждение справедливо, потому что подвыборку можно рассматривать как отдельную составную единицу отбора, а всю выборку — как некоторую простую случайную выборку, в действительности состоящую из этих составных единиц, причем ошибки наблюдения в разных составных единицах некоррелированы. Следовательно, здесь можно применить результаты параграфа 13.10.

Многочисленные приложения этого метода, иногда называемого *дублированным отбором*, описаны Демингом (Deming, 1960), который широко им пользовался. Его преимущества рассмотрены также в работах Джонса (Jones, 1955) и Кула (Коор, 1960). Применение взаимопроницающих выборок обычно увеличивает путевые расходы обследователей, но это увеличение можно умерить, расслоив выборку на компактные участки. Каждый слой может содержать, например, две случайные подвыборки, порученные разным обследователям. Каждый обследователь должен вместо половины слоя объезжать весь слой. Каждый слой вносит в оценку дисперсии одну степень свободы.

13.15. ОБОБЩЕНИЕ НА БОЛЕЕ СЛОЖНЫЕ СХЕМЫ ОТБОРА

Вообще интерпретация результатов дисперсионного анализа зависит от характера применяемой схемы отбора и обычно должна составлять отдельный этап работы. В качестве иллюстрации рассмотрим двухступенчатую расслоенную выборку из L слоев. В каждом слое извлекается n' исходных единиц и каждому из k обследователей выделяется случайная подвыборка n' исходных единиц. Предполагается, что исходные единицы имеют одинаковый размер. Поскольку сово-

купность компактна, каждый обследователь работает в каждом слое. Примерно по такой схеме производилось обследование работающих в Бенгалии (Bengal Labour Enquiry) 1941—1942 гг., описанное Махаланобисом (Mahalanobis, 1946), и обследование здоровья населения района Арсенал в Питтсбурге, описанное Хорвицем (Horvitz, 1952).

В обозначениях второй модели (13.31) среднее значение для j -й исходной единицы, наблюдаемой i -м обследователем в слое h , можно представить как

$$\bar{y}_{hij} = \bar{\mu}_h + g_i + w_{hi} + (\bar{\mu}'_{hij} - \bar{\mu}_h) + \bar{d}_{hij},$$

где $\bar{\mu}_h$ — истинное среднее для слоя. Как и ранее, $\bar{\mu}'_{hij} = \bar{\mu}_{hij} + \bar{b}_{hij}$. Предполагается, что среднее значение в слое величины $(\bar{\mu}'_{hij} - \bar{\mu}_h)$ равно нулю, а ее дисперсия есть $\sigma_{\mu'}^2$. Величины \bar{d}_{hij} некоррелированы и имеют средние значения нуль при любых h, i, j и дисперсию σ_d^2 .

В последней формуле содержится новый член w_{hi} , обозначающий взаимодействие обследователь \times слой. Особых доказательств необходимости этого члена для большего соответствия модели реальности нет, однако, если слои заметно различаются по экономическому уровню, то обследователь может допустить разное смещение в разных слоях. Предполагается, что w_{hi} распределены со средним значением нуль и дисперсией σ_{IS}^2 [IS — от английского «interviewer \times stratum» — обследователь \times слой].

Математические ожидания средних квадратов наиболее важных в дисперсионном анализе членов выглядят следующим образом. Анализ производится на основании среднего значения на поединицу для отдельной исходной единицы.

	Степени свободы	Средние квадраты	E (средних квадратов)
Между обследователями	$k-1$	s_b^2	$\sigma_{\mu'}^2 + \sigma_d^2 + n'\sigma_{IS}^2 + n'Lo^2$
Взаимодействие обследователь \times слой	$(k-1)(L-1)$	s_{IS}^2	$\sigma_{\mu'}^2 + \sigma_d^2 + n'\sigma_{IS}^2$
Между исходными единицами у одного и того же обследователя	$kL(n'-1)$		$\sigma_{\mu'}^2 + \sigma_d^2$

Этот анализ дает нам возможность оценить оба слагаемых, σ_{IS}^2 и σ_d^2 . Если n/N пренебрежимо мало, то несмещенную оценку $kn'LV(\bar{y}_a)$ дает средний квадрат, обусловленный вариацией «между обследователями».

Пример. Пример взят из работы Хорвица (Horvitz, 1952). Имеем $L=6$, $k=18$, $n'=1$. Для числа болевших в течение предыдущего месяца в расчете на одно домохозяйство средние квадраты (в наших

обозначениях) принимают значения $s_y^2 = 0,0759$, $s_{ys}^2 = 0,0147$. Требуется оценить слагаемое $V(\bar{y}_a)$, обусловленное смещением у обследователей.

В обозначениях модели дисперсионный анализ дает

$$s_y^2 = 0,0759 \sim \sigma_\mu^2 + \sigma_a^2 + \sigma_{ys}^2 + 6\sigma_j^2;$$

$$s_{ys}^2 = 0,0147 \sim \sigma_\mu^2 + \sigma_a^2 + \sigma_{ys}^2.$$

Третий член (часть дисперсии, обусловленная вариацией между исходными единицами у одного и того же обследователя) вычислить нельзя, поскольку $n' = 1$.

Оценка σ_j^2 равна $(0,0759 - 0,0147)/6 = 0,0102$. При $n' = 1$ оценить σ_{ys}^2 невозможно. Для того чтобы имитировать ситуацию, при которой смещения у обследователей отсутствуют, мы предположим, что слагаемое ошибки наблюдения по отдельной поединице (домохозяйству), обусловленное обследователем, имеет дисперсию $\sigma_{ys}^2 + \sigma_j^2$, но что эти слагаемые для разных домохозяйств независимы. Поскольку каждому обследователю в отдельном слое выделено приблизительно 26 домохозяйств, общая дисперсия среднего на исходную единицу должна быть равной

$$\sigma_\mu^2 + \sigma_a^2 + \frac{1}{26} (\sigma_{ys}^2 + \sigma_j^2).$$

Мы не можем оценить эту величину, за исключением случая, когда $\sigma_{ys}^2 = 0$. Величина

$$0,0147 + \frac{1}{26} (0,0102) = 0,0151 \sim \sigma_\mu^2 + \sigma_a^2 + \sigma_{ys}^2 + \frac{1}{26} \sigma_j^2$$

дает преувеличенную оценку, но если σ_{ys}^2 мало, то преувеличение не должно быть значительным. Эту величину нужно сравнить с 0,0759, фактической дисперсией, вычисленной на основании среднего значения на одну исходную единицу. Слагаемое общей дисперсии, обусловленное смещением у обследователей, равно приблизительно 80%. Оценка, полученная Хорвицем, который принимал для s_{ys}^2 другое значение, равна 72%.

13.16. КОНТРОЛИРУЕМЫЕ ЭКСПЕРИМЕНТЫ, ВКЛЮЧЕННЫЕ В ОБСЛЕДОВАНИЕ

Дальнейшее углубление идеи взаимопроникающих подвыборок состоит в том, чтобы включить экспериментальное сравнение некоторых элементов процесса наблюдения или собственно обследования в само обследование. Вероятно, старейшим примером такого рода служит применение «расщепленного» опросного листа. Опросный лист подготавливается в двух вариантах q и q' , отличающихся формулировкой некоторых вопросов или их порядком. При одной из схем эксперимента каждый вариант опросного листа получает отбираемая случайным образом половина единиц в выборке или в какой-либо ее части. Для

того или иного признака можно вычислить среднее значение разности $\bar{y}_q - \bar{y}_{q'}$ и ее стандартную ошибку и применить некоторый критерий, чтобы проверить, существует ли относительное смещение. Если есть возможность, то для выяснения, не приводит ли один из вариантов опросного листа к большей вариации ответов, чем другой, можно воспользоваться в качестве критерия отношением дисперсий $s_y^2/s_{y'}^2$. При другой схеме единицы распределяются по парам таким образом, чтобы от каждого члена пары можно было ожидать сходных ответов. Члены пары получают разные варианты опросного листа. Если разбиение по парам было действенным, то эта схема приводит к более точной оценке $\bar{y}_q - \bar{y}_{q'}$, хотя и к менее точной оценке отношения дисперсий.

Идея, таким образом, заключается в том, чтобы, соблюдая необходимые правила контроля, включая рандомизацию, типичные для правильно поставленного эксперимента, превратить часть обследования в контролируемый эксперимент. При планировании более сложных обследований, для того чтобы получить несмещенные оценки эффекта тех или иных особенностей обследования, представляющих наибольший интерес, нужно затратить значительные усилия. Приведем два примера.

При переписи населения США 1950 г. в 24 графствах двух штатов был поставлен эксперимент, позволивший оценить, какую роль играет смещение у обследователей. В нем приняло участие более 700 обследователей. Графства были разделены на 125 участков, в среднем по 6500 жителей. Хотя участки различались по размеру, в среднем объем работы в каждом из них был рассчитан на шестерых обследователей. Участок, на котором должны были работать k обследователей, разделялся на $2k$ подучастков и каждому обследователю случайным образом выделялось два подучастка. Внутри каждого участка оценка дисперсии обследователя могла быть получена путем дисперсионного анализа подобного представленному в табл. 13.12. Описание эксперимента и его результаты содержатся в работах Хансона и Маркса (Hanson and Marks, 1958) и Хансена, Хервица и Бершо (Hansen, Hurwitz and Bershad, 1961).

В примере, описанном Дербинном и Стьюартом (Durbin and Stuart, 1954), эксперимент производился не попутно, а был главной целью обследования. При обследовании населения ряда районов шести городов требовалось оценить: (а) эффект двух типов объединения опросов в гнезда в отношении процента отвечающих, затрат на опрос и точности и (б) эффект смещений у обследователей. В каждом городе работало по два обследователя от каждого из трех агентств, участвовавших в обследовании. Тем самым в основной части эксперимента участвовало $3 \times 2 \times 6 = 36$ обследователей. Основой выборки служили списки избирателей, причем в каждом городе отбор производился в одном избирательном округе.

При негнездовой выборке (тип 1) каждому обследователю поручалась систематическая выборка объемом в 30 фамилий, извлеченная из общего списка. Для выборки типа 2 извлекалось по систематической выборке 15 фамилий в каждом из двух избирательных участков внутри данного округа. Для выборки типа 3 выборка внутри каждого из-

бирательного участка извлекалась по отдельной улице или по группе небольших улиц. Таким образом, наиболее плотно размещалась выборка типа 3, а тип 2 занимал промежуточное положение. Общая схема распределения типов выборки, сделанная по принципу латинского квадрата, изображена в табл. 13.13.

Таблица 13.13

СХЕМА ОБСЛЕДОВАНИЯ (РАСПРЕДЕЛЕНИЕ ТИПОВ ВЫБОРКИ) для ИЗУЧЕНИЯ ЭФФЕКТА ОБЪЕДИНЕНИЯ в ГНЕЗДА и СМЕЩЕНИЯ у ОБСЛЕДОВАТЕЛЕЙ

Агентство	Город					
	I	II	III	IV	V	VI
A	1	2	1	3	3	2
B	2	3	2	1	1	3
C	3	1	3	2	2	1

Например, в городе II два исследователя из агентства A работали с двумя различными выборками типа 2 и т. д. Эта клетка таблицы дает 1 степень свободы для среднего квадрата «между исследователями» и 2 степени свободы для среднего квадрата «между гнездами для одного и того же исследователя». Анализ эффекта объединения в гнезда, имевшего несколько интересных особенностей, здесь не приводится.

Описанные в этом параграфе приемы обладают двумя преимуществами. Поскольку специально на изучение методов наблюдения и приемов отбора средства отпускаются довольно редко, то, возможно, единственный способ предпринять такое изучение состоит в том, чтобы включить его в какое-то текущее исследование. Кроме того, результаты, полученные таким путем, вероятно, будут больше соответствовать практическим условиям. Надо, однако, заметить, что когда часть обследования представляет собой контролируемый эксперимент, трудно точно воспроизвести рядовые условия проведения обследования. Даже если приняты специальные меры, чтобы скрыть такой контроль, контролеры и исследователи, вероятно, почувствуют, что часть их работы имеет какой-то особый характер: поэтому контролируемый эксперимент всегда нарушает тем или иным образом обычный ход дела. Если мы ожидаем возникновения этой проблемы и заранее исследуем ее, то можно многое сделать, чтобы обеспечить получение нужных сведений.

13.17. ВЫВОДЫ

Ошибки, обусловленные не отбором, можно классифицировать по их влиянию на результаты, полученные в предыдущих главах, следующим образом:

1. Наиболее важное следствие неполного охвата и неполучения ответа заключается в том, что оценки могут оказаться смещенными, потому что часть совокупности, не затронутая обследованием, может отличаться от той, которая подвергалась выборочному исследованию. Теперь существует достаточно свидетельств того, что такие смещения значительно разнятся от признака к признаку и от обследования к об-

следованию, будучи иногда большими, а иногда пренебрежимо малыми. Второе следствие состоит, конечно, в том, что дисперсии оценок оказываются увеличенными, потому что фактически наблюдаемая выборка меньше намечавшейся. Это обстоятельство можно учесть, хотя бы приближенно, при определении объема намечаемой выборки.

2. Ошибки наблюдения, независимые для различных единиц внутри выборки и равные в среднем нулю по всей совокупности, должным образом учитываются и обычными формулами, по которым вычисляются стандартные ошибки оценок при условии, что поправки на конечность совокупности можно пренебречь. Такие ошибки уменьшают точность оценок, а поэтому имеет смысл выяснять, насколько это уменьшение существенно.

3. Если ошибки наблюдения по различным единицам в выборке коррелированы, то обычные формулы стандартных ошибок оказываются смещенными. Стандартные ошибки будут, по-видимому, слишком малыми, поскольку на практике корреляции в большинстве случаев положительны. Воздействие такого рода искажений легко просмотреть и часто оно может остаться незамеченным.

4. Труднее всего обнаружить постоянное смещение, действующее одинаково на все единицы. Этого смещения не может устранить никакая обработка данных выборки.

Как показывает содержание этой главы, изучение этих проблем — дело не легкое и не быстрое. Тем не менее основы для него уже заложены. Было проявлено много изобретательности в разработке методов обнаружения и контроля ошибок, обусловленных не отбором. Хотя в этой области трудно делать широкие обобщения, следовало бы постепенно накапливать информацию о характере и величинах ошибок наблюдения в обследованиях разных типов. Необходимо также более глубоко исследовать, чего можно добиться с помощью правильного обучения исследователей и контроля их работы, пробных обследований, разработки систем контроля качества собственно обследования, а также с помощью анализа достоинств и недостатков проведенной работы после окончания обследования.

Упражнения

13.1. Предположим, что с помощью методов обследования разной интенсивности можно добиться того, что слой «ответивших» будет составлять 60, 80, 90 или 95% всей совокупности. Для некоторого процента, подлежащего оценке, истинные средние для слоя «ответивших» равны: в 60%-ном слое — 40,7; в 80%-ном слое — 43,5; в 90%-ном слое — 44,8; в 95%-ном слое — 45,4. Для последних 5% — 59,0. (а) Покажите, что для метода, при котором слой «ответивших» составляет лишь 60%, квадратный корень из среднего квадрата ошибки оценки процента для всей совокупности равен

$$\sqrt{2414/n + 28,94},$$

где n — число полученных заполненных опросных листов. (б) Покажите, что значение квадратного корня из среднего квадрата ошибки, равное 5%, не может быть получено с помощью метода, для которого слой ответивших составляет 60%, но его можно получить с помощью методов, для которых слой ответивших составляет 80% и более, если число заполненных опросных листов несколько

превышает 100. (в) С помощью каких методов и при каких объемах выборки может быть получено значение квадратного корня из среднего квадрата ошибки, равное 2%?

13.2. Для случая 13.1 (в) предположим, что получение одного заполненного опросного листа при применении метода, дающего 90% ответивших, обходится в среднем в 5 долл. Получение одного заполненного опросного листа для следующих 5% совокупности обходится в среднем в 20 долл. Какой из методов — с 90% или с 95% ответивших — обойдется дешевле, если нужно получить значение квадратного корня из среднего квадрата ошибки, равное 2%?

13.3. Совокупность состоит из двух слоев одинаковой величины. Вероятность того, что опрашиваемый окажется дома и согласится ответить на вопросы, равна при каждом обращении 0,9 для лиц из слоя 1 и 0,4 для лиц из слоя 2. (а) Покажите, что в обозначениях параграфа 13.5

$$w_{11} = 1 - (0,1)^i; w_{12} = 1 - (0,6)^i.$$

(б) Пусть исходный объем выборки равен n_0 . Вычислите ожидаемое значение общего числа опросов при 1, 2, 3, 4 и 5 обращениях. (в) Относительные издержки на один полный опрос при i -м обращении составляют для $i = 1, 2, 3, 4, 5$ соответственно 100, 120, 150, 200 и 300. Вычислите средние издержки на один опрос по всем опросам, проводимым вплоть до i -го обращения. (г) Средства, отпущенные на обследование, достаточно для проведения 300 полных опросов при первом обращении. Если обследование должно продолжаться вплоть до i -го обращения, то каковы (при тех же средствах) будут ожидаемое значение общего числа полных опросов для $i = 1, 2, 3, 4, 5$?

13.4. Пусть в условиях упражнения 13.3 оценивается процент лиц, обладающих некоторым признаком. Этот процент подчиняется биномиальному распределению и имеет среднее значение, равное 40% для лиц из слоя 1 и 60% для лиц из слоя 2. (а) Вычислите смещение выборочного среднего при $i = 1, 2, 3, 4, 5$ обращениях. (б) Вычислите дисперсии выборочных средних в тех же условиях финансирования, что и в пункте (г) упражнения 13.3. (Для упрощения вычислений можно считать, что дисперсии равны $2600/p_i$, где p_i — ожидаемое значение общего числа опросов.) (в) Какая тактика дает наименьший СКО?

13.5. Для метода, описанного в параграфе 13.6 (подбор среди непрошеных), покажите, что произведение VC , где V — заданная дисперсия $V(\bar{y})$ и C — ожидаемые издержки, равно

$$S'^2 C' + c_2 W_2^2 S_2^2 + k W_2^2 S_2^2 C' + \frac{1}{k} c_2 W_2^2 S'^2,$$

где

$$S'^2 = S^2 - W_2^2 S_2^2, \quad C' = c_0 + c_1 W_1$$

и что минимальное значение VC равно

$$(S' \sqrt{C'} + V c_2 W_2 S_2)^2.$$

13.6. При обследовании домашней птицы и свиней на приусадебных участках и в некоторых небольших хозяйствах (Gray, 1957) после рассылки анкеты по почте с несколькими напоминаниями был проведен опрос некоторой подвыборки ответивших. Заранее было решено взять $k = 2$ (т. е. 50%-ную подвыборку). В результате обследования по одному из важных признаков были получены следующие данные (в обозначениях параграфа 13.6):

$$\frac{c_1}{c_0} \approx 0,15; \quad \frac{c_2}{c_0} \approx 9,5; \quad W_1 \approx 0,8; \quad S^2 \approx S_1^2.$$

Найдя VC для $k = 2$ и для оптимального k , определите, удачным ли был выбор $k = 2$.

13.7. В обследовании по методу Поляца — Симмонса из намеченной выборки объемом в 660 человек при первом обращении дома оказалось 390 опроши-

ваемых. Далее приведены данные о числе лиц, указавших, что они были дома в 0, 1, 2, 3, 4, и 5 из пяти предшествующих вечеров, и о числе лиц, ответивших «да» на один из вопросов обследования.

	0/5	1/5	2/5	3/5	4/5	5/5
Число бывших дома	14	35	55	74	91	118
Число ответивших «да»	4	13	20	30	42	56

Вычислите оценку Поляца — Симмонса для доли лиц в исследуемой совокупности, которые ответили бы «да», и сравните ее с простой биномиальной оценкой.

13.8. Совокупность объемом $N = 6$ содержит три единицы, для которых правильный ответ на некоторый вопрос — ответ «да», и три единицы, для которых такой ответ — «нет». Из-за ошибок наблюдения вероятность получить ответ «да» по единице, для которой правилен ответ «да», равна 0,9, и вероятность получить ответ «нет» по единице, для которой правилен ответ «нет», также равна 0,9. (а) Найдите распределение всех возможных ответов для выборки объемом в 2 единицы, покажите, что вероятности того, что выборка содержит 0, 1 и 2 ответа «да», равны соответственно 0,218; 0,564 и 0,218. (б) Покажите, что дисперсия оценки доли ответов «да» равна 0,1090. Проверьте формулы (13.22) и (13.24) из параграфа 13.10. (в) Какой была бы дисперсия оценки доли ответов «да», если бы не было ошибок наблюдения?

13.9. В части обследования трудовых ресурсов в Бенгалии, проводившегося в 1942 г. (Mahalanobis, 1946) в каждом из трех слоев извлекалась выборка объемом приблизительно в 175 семей. Выборка в каждом слое подразделялась на пять случайных подвыборок, которые отводились разным исследователям. Во всех трех слоях работало пять исследователей. Для признака «расходы на питание» соответствующая часть дисперсионного анализа (на основании отдельной семьи) выглядит следующим образом:

	Степени свободы	Средние квадраты	Математические ожидания средних квадратов
Между исследователями	4	22,3	$\sigma_{\mu}^2 + \sigma_d^2 + 35\sigma_{JS}^2 + 105\sigma_I^2$
Взаимодействие исследователя × слой	8	9,6	$\sigma_{\mu}^2 + \sigma_d^2 + 35\sigma_{JS}^2$
Внутри подвыборок	510	9,9	$\sigma_{\mu}^2 + \sigma_d^2$

В обозначениях параграфа 13.15 соответствующая модель слагаемых значения наблюдения для отдельной семьи имеет вид

$$y_{hij} = \bar{\mu}_h + g_i + w_{hi} + (\mu_{hij} - \bar{\mu}_h) + d_{hij}\alpha.$$

Дисперсии:

$$\sigma_I^2 \quad \sigma_{JS}^2 \quad \sigma_{\mu}^2 \quad \sigma_d^2.$$

Проверьте выражения, приведенные для математических ожиданий средних квадратов, и вычислите оценку части общей дисперсии среднего, которую можно приписать смещению у исследователей.

ЛИТЕРАТУРА

- Barr A. (1957). Differences between experienced interviewers. *App. Stat.*, 6, 180—188.
- Bartholomew D. J. (1961). A method of allowing for «Not-at-homes» bias in sample surveys. *App. Stat.*, 10, 52—59.
- Belloc N. B. (1954). Validation of morbidity survey data by comparison with hospital records. *Jour. Amer. Stat. Assoc.*, 49, 832—846.
- Birnbaum Z. W. and Sirken M. G. (1950a). Bias due to nonavailability in sampling surveys. *Jour. Amer. Stat. Assoc.*, 45, 98—111.
- Birnbaum Z. W. and Sirken M. G. (1950b). On the total error due to non-interview and to random sampling. *Int. Jour. Opinion and Attitude Res.*, 4, 179—191.
- Cochran W. G., Mosteller F. and Tukey J. W. (1954). *Statistical problems of the Kinsey Report*. American Statistical Association, Washington, D. C., p. 280.
- Deming W. E. (1953). On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias of non-response. *Jour. Amer. Stat. Assoc.*, 48, 743—772.
- Deming W. E. (1960). *Sample design in business research*. John Wiley and Sons, New York.
- Durbin J. (1954). Non-response and call-backs in surveys. *Bull. Int. Stat. Inst.*, 34, 2, 72—86.
- Durbin J. and Stuart A. (1954). Callbacks and clustering in sample surveys: an experimental study. *Jour. Roy. Stat. Soc.*, A, 117, 387—428.
- Finkner A. L. (1950). Methods of sampling for estimating commercial peach production in North Carolina. *North Carolina Agr. Exp. Stat. Tech. Bull.* 91.
- Gray P. G. (1955). The memory factor in social surveys. *Journ. Amer. Stat. Assoc.*, 50, 344—363.
- Gray P. G. (1957). A sample survey with both a postal and an interview stage. *App. Stat.*, 6, 139—153.
- Hansen M. H. et al. (1951). Response errors in surveys. *Jour. Amer. Stat. Assoc.*, 46, 147—190.
- Hansen M. H. and Hurwitz W. N. (1946). The problem of nonresponse in sample surveys. *Jour. Amer. Stat. Assoc.*, 41, 517—529.
- Hansen M. H., Hurwitz W. N. and Bershad M. (1961). Measurement errors in censuses and surveys. *Bull. Int. Stat. Inst.*, 38, 2, 359—374.
- Hanson R. H. and Marks E. S. (1958). Influence of the interviewer on the accuracy of survey results. *Jour. Amer. Stat. Assoc.*, 53, 635—655.
- Hartley H. O. (1946). Discussion of paper by F. Yates. *Jour. Roy. Stat. Soc.* 109, 37.
- Hendricks W. A. (1949). Adjustment for bias by non-response in mailed surveys. *Agr. Econ. Res.*, 1, 52—56.
- Horvitz D. G. (1952). Sampling and field procedures of the Pittsburgh morbidity survey. *Pub. Health Reports*, 67, 1003—1012.
- Jones H. W. (1955). Investigating the properties of a sample mean by employing random subsample means. *Jour. Amer. Stat. Assoc.*, 51, 54—83.
- Kish L. (1949). A procedure for objective respondent selection within the household. *Jour. Amer. Stat. Assoc.*, 44, 380—387.
- Kish L. and Hess I. (1958). On noncoverage of sample dwellings. *Jour. Amer. Stat. Assoc.*, 53, 509—524.
- Kish L. and Hess I. (1959). A «replacement» procedure for reducing the bias of nonresponse. *Amer. Statistician*, 13, 4, 17—19.
- Kish L. and Lansing J. B. (1954). Response errors in estimating the value of homes. *Jour. Amer. Stat. Assoc.*, 49, 520—538.
- Koop J. C. (1960). On theoretical questions underlying the technique of replicated or interpenetrating samples. *Proc. Soc. Statistics Sect. Amer. Stat. Assoc.*, 196—205.
- Lienau C. C. (1941). Selection, training and performance of the National Health Survey field staff. *Amer. Jour. Hygiene*, 34, 110—132.
- Mahalanobis P. C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Jour. Roy. Stat. Soc.*, 109, 325—370.
- Politz A. N. and Simmons W. R. (1949, 1950). An attempt to get the «not at homes» into the sample without callbacks. *Jour. Amer. Stat. Assoc.*, 44, 9—31 and 45, 136—137.
- Sagen O. K., Dunham R. E. and Simmons W. R. (1959). Health statistics from record sources and household interviews compared. *Proc. Social Statistics Sect. Amer. Stat. Assoc.*, 6—15.
- Simmons W. R. (1954). A plan to account for «not-at-homes» by combining weighting and callbacks. *Jour. of Marketing*, 11, 42—53.
- Stephan F. and McCarthy P. J. (1958). *Sampling Opinions*. John Wiley and Sons, New York, p. 243.
- Sukhatme P. V. and Seth G. R. (1952). Non-sampling errors in surveys. *Jour. Ind. Soc. Agr. Stat.*, 4, 5—41.
- Trussell R. E. and Elinson J. (1959). *Chronic illness in a large city*. Harvard Univ. Press, pp. 339—370.
- Woolsey T. D. (1956). Sampling methods for a small household survey. *Pub. Health Monographs* No. 40.

ОТВЕТЫ К УПРАЖНЕНИЯМ

- 1.5. 80400 долл. и 82960 долл.
 1.6. Доверительная вероятность равна приблизительно 0,054 (соответствует $t = -1,67$ с 25 степенями свободы). При этом предполагается, что поступления в будущем подчиняются тому же распределению частот, что и выборка объемом в 26 поступлений.
 1.7. Если СКО целиком складывается из смещения, то оценка всегда дает ошибку величиной $1/\sqrt{СКО}$. В этом случае, следовательно, вероятность того, что ошибка $> 1/\sqrt{СКО}$, равна 1, а вероятности того, что ошибка $> 1,96/\sqrt{СКО}$ или $> 2,576/\sqrt{СКО}$ равны нулю. В табл. 1.2 $Pr(> 1/\sqrt{СКО})$ стремится к 1/3.
 2.4. $\hat{Y} = 51\,473$. Вероятность равна приблизительно 0,9.
 2.5. Да. $\sigma(\hat{Y})$ равна 98,4.
 2.6. $\hat{Y} = 20\,238$; $s(\hat{Y}) = 849$.
 2.7. (а) Для государственных $\hat{R} = 15,46$; для частных $\hat{R} = 12,75$. (б) Для государственных $s(\hat{R}) = 0,761$; для частных $s(\hat{R}) = 0,727$. Пикс принимается равной 100/468. (в) $14,2 < R < 16,7$.
 2.8. Отношение разности к стандартной ошибке разности = $2,71/1,186 = 2,28$, P равно приблизительно 0,023. Заметьте, что при вычислении стандартной ошибки разности пикс не учитывается.
 2.9. (а) 9408; стандартная ошибка = 780; (б) 9472; стандартная ошибка = 1104.
 2.10. Стандартная ошибка (в тыс.) = (а) 14 800; (б) 3900; (в) 3140.
 2.11. 9,2. (а) 2,7; (б) 2,4.
 2.12. (а) $n = 60$, по 30 единиц из каждой области изучения; (б) $n = 80$ будет достаточным, если число домовладельцев в выборке заключено между 20 и 60. При $n = 80$ вероятность этого события равна 0,54 (по биномиальным таблицам). При $n = 100$ эта вероятность равна 0,94.
 2.14. (а) 420; (б) 490; (в) обе несмещенные; (г) от оценки (б).
 3.2. 1066, 1334 по формуле (3.17) при аппроксимации нормальным распределением.
 3.3. Обеспечивает почти полную уверенность.
 3.6. (а) $76,2 \pm 3,6\%$; (б) 1738 ± 280 семей.
 3.7. 1789 ± 268 семей.
 3.8. Справедливо точное равенство:

$$\frac{V(\hat{A}_1)}{V(\hat{A}_1')} = \frac{N_1^2 n Q_1}{N^2 n_1 (1-\pi) (Q_1 + P_1 \pi)}$$

Далее, $N_1 = N(1-\pi)$ и для больших выборок $n_1 \approx n(1-\pi)$. Из этих равенств следует нужное утверждение. Для того чтобы $V(\hat{A}_1)/V(\hat{A}_1')$ было малым, нужно чтобы $\pi(1-Q_1)/Q_1$ было большим. Это значит, что Q_1 должно быть малым; другими словами, доля единиц в области 1, принадлежащих классу С, должна быть большой. При заданном Q_1 π должно быть большим.

3.9. Все методы дают $A_U = 13$. В случае применения гипергеометрического распределения вероятность того, что в выборке не будет единиц из класса С, равна 0,0601 при $A_U = 12$ и 0,0434 при $A_U = 13$. В случае применения биномиального распределения $P_U = 0,4507$ и $\sqrt{1-P_U} = 0,4114$, что дает $A_U = 12,3$. Метод из примера 3 параграфа 3.6 дает 0,061 и 0,044.

- 3.11. По-видимому, оценка (б) более точна.
 3.12. Наибольшее значение равно PQ/n по сравнению с $PQ/мл$, получаемым по биномиальной формуле. Оно образуется, когда каждое гнездо состоит целиком из 1 или целиком из 0. Наименьшее значение может быть равно нулю, если каждое гнездо дает одну и ту же долю P . (Это возможно лишь при некоторых значениях P и $мл$.)
 3.13. Дисперсия равна 0,00184 для метода отношения и 0,00160 по биномиальной формуле.
 3.14. Средний объем выборки = m/P .

- 4.1. 735 домов. Такой объем выборки нужен, чтобы оценить процент домохозяйств, имеющих два автомобиля при $P = 10\%$.
 4.2. Приблизительно 260 листов.
 4.3. (а) 2475; (б) 4950.
 4.4. $n = 21$ (принимая $t = 2$).
 4.5. $n = 484$. Для числа безработных коэффициент вариации был бы приблизительно 15%.
 4.6. Еще 62.
 4.7. (а) $n = 278$; (б) $n = 2315$; (в) $n = 3046$.
 4.8. Если предполагается, что внутри каждой группы размер колледжей распределен равномерно, то берем $S^2 = 0,0834$ или $S = 0,289$. Тогда оценки в четырех группах равны 230, 580, 2030 и 11 600. Если для четвертой группы воспользоваться распределением с плотностью, изображаемой правильным треугольником, то $S = 0,24h$, что дает для этой группы значение 9600.

$$4.11. \quad n_{opt} = \left(\frac{UNS}{2c\sqrt{2\pi}} \right)^{2/3}$$

- 5.1. (а) Нейманово размещение дает $n_1 = 0,87$; $n_2 = 3,13$. (б) Оценка может принимать три возможных значения при неймановом размещении и девять — при пропорциональном размещении. $V_{opt}(\bar{y}_{st}) = 1/6 = 0,167$; $V_{prop}(\bar{y}_{st}) = 7/12 = 0,583$. (г) По формуле (5.21) $V_{opt}(\bar{y}_{st}) = 0,159$.
 5.2. (а) $n_1 = 375$; $n_2 = 625$; (б) $n_1 = 750$; $n_2 = 250$.
 5.3. Относительная точность = 181% при пропорциональном размещении и 214% при оптимальном размещении.
 5.5. Максимальное относительное увеличение наблюдается, когда $W_1 = W_2$ и равно 0,030 при $c_2/c_1 = 2$ и 0,111 при $c_2/c_1 = 4$.
 5.6. (а) $n_1/n = 1/3$; $n_2/n = 2/3$; (б) $n = 264$, $n_1 = 88$, $n_2 = 176$; (в) 1936 долл.
 5.7. (а) 2288 долл. по сравнению с 1936 долл. (б) Нет. Минимальные издержки на собственно обследование, необходимые для того, чтобы получить $V = 1$, равны 2230 долл.
 5.8. (а) $n_1 = 384$; $n_2 = 192$; (б) $n_1 = 400$; $n_2 = 1600$; (в) $n_1 = 1200$; $n_2 = 2400$.
 5.9. Относительный прирост = 1/9.
 5.10. $n_1 = 541$; $n_2 = 313$; $n_3 = 146$.
 5.12. Для совокупности 1 $V_{prop} = 0,143/n$; $V_{opt} = 0,134/n$. Для совокупности 2 $V_{prop} = 0,0491/n$; $V_{opt} = 0,0423/n$. Уменьшение дисперсии, вызванное применением оптимального размещения, составляет приблизительно 6% для совокупности 1 по сравнению с 14% для совокупности 2.
 5.14. (а) Если в качестве компромиссного предположения взять $P_1 = 45\%$; $P_2 = 25\%$ и $P_3 = 7,5\%$, то приходим к $n_1 = 268$; $n_2 = 116$; $n_3 = 16$; (б) стандартная ошибка = 0,0225; (в) стандартная ошибка = 0,0241.

5А.4. Нет. В каждом из наиболее неблагоприятных случаев $[\Sigma (w_h - W_h) \bar{Y}_h]^2 = (0,105)^2 = 0,011$. Таким образом, при расслоении СКО (\bar{y}_{st}), выраженной формулой (5А.6), равен $0,0108 + 0,0110 = 0,0218$. При простом случайном отборе $V(y) = 0,0177$.

5А.5. (а) $n = 1024$. Оптимальное размещение для второй переменной (средняя величина вклада) удовлетворяет обоим условиям. (б) Размещение, выражаемое формулой (5А.14), с. 142, при $\lambda = 0,09$ удовлетворяет обоим условиям при $n = 1031$.

5А.6. $W_1 = 0,728$; $W_2 = 0,272$. $S_1 = 1,806$; $S_2 = 4,698$ (условные величины). (а) Оптимальные объемы выборки по слоям следующие: $n_1 = 0,507 n$; $n_2 = 0,493 n$; (б) $V(\bar{y}) = 31,95/n$; $V_{opt}(\bar{y}_{st}) = 6,72/n$.

$$5А.7. (б) \int_0^1 V(\bar{y}) dy = \int_0^1 \sqrt{2(1-y)} dy = 2\sqrt{2} [1 - (1-a)^{3/2}]/3.$$

Следовательно, необходимо чтобы $[1 - (1-a)^{3/2}] = 1/2$.

5А.8. Оптимальными будут значения $L = 7$ при $\rho = 0,95$; $L = 5$ при $\rho = 0,9$ и $L = 4$ при $\rho = 0,8$. В качестве хорошего компромиссного значения можно взять либо $L = 5$, либо $L = 6$.

5А.9 (а) Выигрыш в точности составляет приблизительно 110%. (б) Выигрыш от пропорционального расслоения по сравнению с простым случайным отбором составляет приблизительно 90%.

5А.10. (а) 3,733; (б) 1,111; (в) 8,222.

6.1. Для оценки по отношению $V(\bar{Y}_R) = N^2(1-f) S_d^2/n$ и для простого распространения $V(\bar{Y}) = N^2(1-f) S_y^2/n$, где $d = (y - Rx)$. Для выборки объемом в 21 домохозяйство оценки S_d^2 и S_y^2 таковы: для числа детей $s_d^2 = 0,49$; $s_y^2 = 1,61$; для количества автомобилей $s_d^2 = 0,41$; $s_y^2 = 0,39$; для количества телевизоров $s_d^2 = 0,51$; $s_y^2 = 0,45$. Оценка по отношению представляется лучшей при оценивании числа детей.

6.2. Выигрыш = 66%. По крайней мере, 11 участков (при оценке по отношению).

6.3. Квадратичные границы (27 100; 29 870); границы при аппроксимации нормальным распределением (27 030; 29 700).

6.5. Воспользуйтесь для оценивания $R = \bar{Y}/\bar{X}$ теоремой 6.3. При больших выборках применяйте \bar{y}/\bar{x} , если $\rho < (\text{коэффициент вариации } x)/2$ (коэффициент вариации y) и \bar{y}/\bar{x} в противном случае.

6.6. СКО равны 46,5 для раздельной оценки по отношению и 40,6 для совместной оценки по отношению. В обоих случаях слагаемым СКО, обусловленным смещением, можно пренебречь.

6.7. Для метода Лахири $V(\bar{Y}) = 40,1$.

6.8. Оценка общей численности населения = 116,21 миллиона. Относительная дисперсия равна 0,00111, так что стандартная ошибка равна $0,0333 \times 116,21 = 3,87$ миллиона. Оценка отклоняется от истинного значения не более чем на однократную стандартную ошибку.

6.9. Значения оценок равны: (а) 1896, (б) 1660, (в) 1689. Для случая (в) находим $w_1 = 2,38$; $w_2 = -1,38$. Выборочные оценки стандартных ошибок равны: (а) 256, (б) 36,9, (в) 18,6. Для нахождения стандартной ошибки в случае (б) применялась формула: стандартная ошибка = $\hat{Y}_R \sqrt{(1-f)(c_{yy} + c_{11} - 2c_{y1})/n}$, где \hat{Y}_R — оценка по отношению величины \bar{Y} , т. е. 1660. Для нахождения стандартной ошибки в случае (в) применялась $\bar{Y}_{MR} = 1689$.

7.1. Значение оценки = 11 080; стандартная ошибка = 152 (учитывая пкс).

7.2. Не будет, поскольку b очень близко к 1.

7.3. $\bar{Y}_{tr} = 28 177 \pm 570$. Относительная точность составляет 113%.

7.4. $27 751 \pm 694$.

7.6. Для «разностной» оценки $V(\bar{y}) = S_d^2/n$, для линейной оценки по регрессии $V(\bar{y}_{tr}) = S_d^2 S_y^2/n (S_d^2 + S_y^2)$. Оценка по регрессии имеет меньшую дисперсию, однако если S_d^2/S_y^2 мало, то преимущество этой оценки незначительно.

7.7. $V(\bar{Y}_{tr}) = 34,5$; $V(\bar{Y}_{trc}) = 10,3$.

8.1. Дисперсии равны: 8,19 (систематическая); 11,27 (простая случайная); 8,25 (расслоенная, 2); 7,46 (расслоенная, 1).

8.2. $V_{opt} = 0,00141$; $V_{opt} = 0,00340$.

8.3. Систематическая выборка должна быть лучше при оценивании доли лиц польского происхождения, так как по этой переменной существует географическое расслоение. При оценивании доли детей, она будет, по-видимому, хуже, потому, что интервал отбора, 5, совпадает со средним размером домохозяйства. Это же соображение справедливо, хотя и в меньшей степени, для оценивания доли мужчин.

8.4. Дисперсии имеют следующие значения: для доли лиц мужского пола $V_{opt} = 0,0204$; $V_{opt} = 0,0216$; для доли детей $V_{opt} = 0,0204$; $V_{opt} = 0,0776$; для доли лиц в семьях специалистов $V_{opt} = 0,0192$; $V_{opt} = 0,0016$.

8.5. Фактическая дисперсия = 8,19. Метод (а) дает 11,29. Оценка дисперсии по методу (б) для отдельной выборки равна $(1-f)(\bar{y}_1 - \bar{y}_2)^2/4$, где \bar{y}_1 , \bar{y}_2 — средние значения для каждой из двух половин выборки. Среднее значение оценки равно 3,24. Столь сильное преуменьшение дисперсии неожиданно.

8.7. Дисперсия для обоих методов равна $(k^2 - 1)/6$.

8.8. Простой случайный отбор лучше за исключением случаев, когда $n = 1$ или $k = 1$.

9.1. Относительные издержки при применении четырех типов единиц равны 100; 90,1; 79,7 и 77,8 (за основу сравнения принимается единица первого типа).

9.2. Относительная точность домохозяйства как единицы отбора составляет 211% при оценивании отношения числа мужчин к числу женщин и 38% при оценивании доли людей, посетивших врача.

9.3. Относительная точность большой единицы составляет 0,566 при простом случайном отборе и 0,625 при расслоенном случайном отборе.

9.5. (а) $M = 5$; (б) $M = 1$.

9.6. Оптимальное значение M должно уменьшиться, потому что путевые расходы, которые меняются как \sqrt{n} , при увеличении n становятся относительно менее значительными.

9.7. (а) 34 242; (б) 5534; (в) 6493.

9.9. (а) Если среднее квадратичное отклонение для больших единиц из класса h пропорционально M_h . (б) Если вероятности пропорциональны $\sqrt{M_h}$.

9.10. $V(\bar{Y}_w) = 1,75$; $V(\bar{Y}_{opt}) = 0,50$; $V(\bar{Y}_{G2}) = 0,33$.

10.1. (а) 2,00; (б) 2,13.

10.3. (а) 165/n; (б) 148,5/n; (в) 132/n.

10.4. (а) $n = 660$ полей; (б) $n = 530$ полей. Оценивание процента белка требует меньшего числа полей, чем оценивание урожайности.

10.5. $c_1/c_2 = 8$.

10.7. (а) 0,93%; (б) 0,51%; (в) 0,36%.

10.8. (а) Подходят значения $m_0 = 7$ или $m_0 = 8$; (б) 89% для $m_0 = 7$ и 93% для $m_0 = 8$; (в) 86% для $m_0 = 7$ и 89% для $m_0 = 8$.

11.2. Относительная точность метода III по отношению к методу II падает с 3,02 до 2,75. Если две схемы отбора отличаются в основном слагаемыми дисперсии, обусловленными вариацией между единицами, относительная точность лучшей схемы будет, вообще говоря, уменьшаться при увеличении отношения внутрисредней дисперсии к общей дисперсии.

11.3. Грубо говоря, дело в том, что при наших данных значения Y_1/M_1 более устойчивы, чем значения Y_2/M_2 . Если бы применялись значения $z_1 = 1/33$, $8/33$ и $24/33$, то для метода IV междуединичное слагаемое дисперсии исчезло бы.

- 11.4. Общая дисперсия: 0,00504 (I а); 0,02358 (II); 0,00554 (III).
- 11.6. Оценка процента поврежденных единиц оборудования $14,2 \pm 2,16$.
- 11.7. Оценка процента поврежденных единиц оборудования $13,9 \pm 2,49$.
- 11.9. (а) Общее количество комнат 29 400, общее число людей 50 550, среднее число людей на одну комнату 1,72; (б) значения стандартных ошибок: для общего числа людей 2440, для среднего числа людей на одну комнату 0,066.
- 12.1. $n = 267$; $n' = 1320$ или $n = 268$; $n' = 1280$. $V(p_{st})$, при оптимальном размещении равна 6,67, если p_{st} выражена в процентах. В случае одинарного отбора $V(p) = 8,33$.
- 12.2. $c_n/c_1 > 9$.
- 12.4. $n' > 16n$.
- 12.5. По формуле (12.29) стандартная ошибка = 1,25.
- 12.6. Процентный выигрыш в точности для моментов отбора от второго до шестого составляет соответственно 50, 75, 91, 100 и 105.
- 12.8. $nV(\bar{y}_2)/S^2$ и $nV(\bar{y}_2)/S^2$ имеют следующие значения. При $\mu = 1/4$, $\rho = 0,8; 0,885; 0,875$; при $\mu = 1/4$, $\rho = 0,9; 0,843; 0,840$; при $\mu = 1/2$, $\rho = 0,8; 0,824; 0,810$; при $\mu = 1/2$, $\rho = 0,9; 0,752; 0,746$.
- 13.1. (а) При 90%-ном слое ответивших с 1047 полными опросами или при 95%-ном слое ответивших с 701 полным опросом.
- 13.2. Применение метода с 90%-ным слоем ответивших обойдется в 5235 долл. В случае 95%-ного слоя ответивших средние издержки на один полный опрос равны 5,7895 долл., а общие издержки составят 4058 долл.
- 13.3. (б) $0,65 \mu_a; 0,815 \mu_a; 0,8915 \mu_a; 0,9351 \mu_a; 0,9611 \mu_a$. (в) 100, 104, 108, 112, 117; (г) 300, 288, 277, 267, 256.
- 13.4. (а) Смещение (в %) = $-3,85; -2,15; -1,21; -0,69; -0,40$; (б) дисперсия равны: 8,67; 9,03; 9,39; 9,74; 10,16; (в) четыре обращения.
- 13.6. Выбор был удачным. VC для $k = 2$ всего на 2% превышает минимальное.
- 13.7. Оценка Поляца — Симмонса дает 39,7%; биномиальная оценка дает 42,3%.
- 13.8. (в) Дисперсия была бы равна 0,1.
- 13.9. Если бы ошибки наблюдения каждого обследователя были независимы от семьи к семье, то дисперсия выборочного среднего была бы равна $(\sigma_m^2 + \sigma_a^2 + \sigma_s^2 + \sigma_f^2)/525$, а не $(\sigma_m^2 + \sigma_a^2 + 35\sigma_s^2 + 105\sigma_f^2)/525$. Сглаженное дисперсия, обусловленное смещением у обследователей, составляет приблизительно 55% общей дисперсии.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Автокоррелированные совокупности (autocorrelated populations), 238

Аналитические обследования (analytical surveys), 18

Асимметричная совокупность (skewed population)

выборки из конкретной совокупности, 55

Асимметрия (skewness)

коэффициент асимметрии, 58

влияние расслоения, 59

влияние на доверительные границы, 56

Биномиальное распределение (binomial distribution), 69

в качестве приближения к гипергеометрическому, 73

для оценивания долей, 69

доверительные границы, 73

неправильное применение при гнездовом отборе, 79—82

таблицы, 69

Взаимопроницающие подвыборки (interpenetrating subsamples), 408

дисперсия оценки, 408

при расслоенном отборе, 410

оценка дисперсии, 409

Вероятностный отбор (probability sampling), определение и свойства, 24

Вероятность, пропорциональная оценке размера (probability proportional to estimated size)

двухступенчатый отбор без возвращения, 345—346

отбор без возвращения, 280

при двухступенчатом отборе, 292, 296, 320—322, 328—342

при одноступенчатом отборе, 272—275

оптимальная характеристика размера, 275

Вероятность, извлечения, пропорциональная размеру (probability proportional to size), при одноступенчатом отборе, 271

в сравнении с равными вероятностями отбора, 276

в сравнении с оценкой по отношению, 276

метод получения выборки, 271

отбор без возвращения, 280—287

при расслоенном отборе, 280

Вероятность, извлечения, пропорциональная размеру, при двухступенчатом отборе, 317

дисперсия несмещенной оценки, 331

дисперсия оценки по отношению, 336

в сравнении с оценкой по отношению, 332

отбор без возвращения, 345—346

Вес слоя (stratum weight), 104

Внутригнездовая корреляция (intra-cluster correlation), 230, 262

Географическое расслоение (geographic stratification), 118

формирование слоев, 151

выигрыш в точности от применения, 118

Гипергеометрическое распределение (hypergeometric distribution), 70

в качестве условного распределения, 75

пример, 70

доверительные границы, 71

таблицы (ссылка), 71

Гнездовой отбор (cluster sampling), одноступенчатый

влияние размера гнезда на дисперсию, 263—265

дисперсия среднего, 262

оптимальный размер и тип гнезда, 263, 266

оценивание в случае единиц неодинакового размера, 269
оценивание долей, 78, 267
применение отбора с вероятностями пропорциональными размеру, 269
сравнение с простым случайным отбором элементов, 263
Гнездовой отбор (с подотбором), см. двухступенчатый отбор
Границы слоев (boundaries of strata), правило для нахождения, 146
Двойной отбор (double sampling), применение в обследованиях, описание, 135, 351
неответивших, 391—395
Двойной отбор для оценок по отношению, 364
дисперсия, 364
оптимальный объем выборки, 364, 365
оценка дисперсии, 365
Двойной отбор для оценок по регрессии, 359
дисперсия, 360—362
оптимальные объемы выборки, 361
оценка дисперсии, 363
сравнение с простым случайным отбором, 361
Двойной отбор для расслоения, 352
дисперсия, 354
оптимальные объемы выборки, 355
оценивание долей, 354
оценка дисперсии, 357
сравнение с простым случайным отбором, 356
Двухступенчатый отбор (two-stage sampling), единицы одинакового размера, преимущества, 291
дисперсия оценки среднего, 296
дисперсия оценки доли, 299
применение пробного обследования, 306
обозначения, 296
определение, 291
оптимальные доли отбора и подотбора, 301
оценка дисперсии, 298
расслоенный отбор исходных единиц, 310
таблица для определения оптимального объема подвыборки (subsampling), 304
Двухступенчатый отбор (единицы неодинакового размера)
сравнение отбора с равными вероятностями и отбора с вероятностями, пропорциональными размеру, 318, 332

оптимальные доли отбора и подотбора, 337—342
отбор исходных единиц без возвращения, 345
отбор единиц с равными вероятностями, 316, 319, 323, 326
отбор единиц с вероятностями, пропорциональными характеристике размера, 319—322, 328—332
оценивание долей, 334
оценка по отношению с размером в знаменателе, 323
оценка по отношению с другой переменной в знаменателе, 334—336
при расслоении, 342—345
Двухфазный отбор (two-phase sampling), см. Двойной отбор
Дисперсия (variance)
определение S^2 и σ^2 , 37
как функция размера единицы, 263
для совокупности, предварительные оценки, 92
Дисперсия выборочных оценок, см. Стандартная ошибка
Доверительные границы (confidence limits), 27
влияние неполучения ответа, 381—383
влияние смещения, 28
для гипергеометрического распределения, 71
для долей или процентов, 71, 75
для оценок по отношению, 182
обоснованность аппроксимации нормальным распределением, 54
при простом случайном отборе, 41
при расслоенном случайном отборе, 111
Доли (proportions)
оценивание, 64
влияние неполучения ответа, 386—383
влияние доли для совокупности, P , на точность, 67
в случае более чем двух классов, 74, 75
объем выборки, 86—90
при гнездовом отборе, 78—82, 267—269
при двойном отборе, 354
при двухступенчатом отборе (единицы неодинакового размера), 334
при двухступенчатом отборе (единицы одинакового размера), 229
при простом случайном отборе, 64—82
при расслоенном случайном отборе, 122—126

Доля отбора (sampling fraction), 35
первой ступени, 297
общая, 329
второй ступени, 297
Достоверность (accuracy), 30
Дублированный отбор (replicated sampling), 27, 410
Единица отбора (sampling unit), определение, 21
дисперсия, 44
метод нахождения оптимальной единицы, 253—261, 263—267
оптимальная характеристика размера, 275
Замещение (matching) при повторном обследовании одной и той же совокупности, 365, 367
Издержки на собственно обследование (field costs)
влияние на оптимальный размер единицы, 57, 267
Изменение (change)
оценки изменения, 365—366, 372—376
Изуемая совокупность (target population), 20
Исходная единица (отбора) (primary (sampling) unit), 292
«Истинное значение» («correct value»), 399
Качественные признаки (attributes), выборочное исследование, см. Доли
Квадратичные доверительные границы для оценки по отношению, 164
Квота (quota), 155
Квотный отбор (quota sampling), 155
Ковариация выборочных средних (covariance of sample means), 39
Компромиссное размещение выборки по слоям, 137
Контролируемые эксперименты (controlled experiments), при изучении ошибок наблюдения, 412
Контролируемый отбор (controlled selection), 146
Концевые поправки (end corrections), 237
Коррелограмма (correlogram), 239
Коши—Буняковского неравенство, 113
Коэффициент вариации (coefficient of variation) оценки, 68
Коэффициент корреляции (correlation coefficient)
внутригнездовой, 229, 262
внутри систематической выборки, 229
для конечной совокупности, 176

Кумулянт (cumulant), κ_4 , 60
Латинские квадраты (latin squares), использование при систематическом отборе, 249
Линейная оценка по регрессии, см. оценка по регрессии
Многомерная оценка по отношению (multivariate ratio estimate), 203
Множитель распространения (inflation factor), 35
Надсовершенство (superpopulation), 234
Накопленного χ^2 правило (cumulative χ^2 rule), 149
Наилучшая линейная несмещенная оценка (best linear unbiased estimate), 184
Направленный отбор (purposive selection), 25
Нейманово размещение (Neyman allocation), 113, 147
наилучшие границы слоев, 147
Необнаруженные (noncoverage), 383
Неполучение ответа (nonresponse), 379
влияние повторных обращений, 385—391
влияние на доверительные границы, 381—383
влияние на дисперсию при расслоенном отборе, 167
вызванное им смещение, 380—383
метод Полица—Симмонса, 395—398
необходимый объем выборки, 380—383
оптимальная доля отбора среди неответивших, 391—395
причины неполучения ответа, 383—385
Несмещенная оценка (unbiased estimate) (метод оценивания), определение, 26, 36
Нормальное распределение (normal distribution), 24
как аппроксимация биномиального, 72
как аппроксимация гипергеометрического, 72
как предельное распределение среднего для выборки, 54
обоснованность аппроксимации, 54—61
применение в обследованиях, 26
Область изучения (domain of study), 48

Обозначения

- в модели влияния повторных обращений, 387
- для двухступенчатого отбора, 296
- для долей, 64, 77
- для оценок дисперсий, 41
- для оценок по отношению, 173
- для простого случайного отбора, 34
- для расслоенного отбора, 104
- Обследуемая совокупность (sampled population), 20
- Общая доля отбора (over-all sampling fraction), 329
- Объем выборки необходимый (size of sample needed)
 - при вычислении доверительных границ с помощью нормальной аппроксимации в случае непрерывных данных, 57, 58
 - при вычислении доверительных границ для долей с помощью нормальной аппроксимации, 72
 - при оценивании оптимальной доли подотбора, 304, 305
- Описательные обследования (descriptive surveys), 18
- Оптимальное размещение при расслоенном двухступенчатом отборе, 310
- Оптимальное размещение при расслоенном отборе (optimum allocation in stratified sampling)
 - в аналитических обследованиях, 163
 - влияние отклонений от оптимального размещения, 131
 - влияние ошибок при определении S_y , 133
 - влияние ошибок при определении объемов слоев, 133
 - в случае более чем 100%-ного отбора, 119
 - в случае оценок по отношению, 192
 - определение по предварительным данным, 117, 149
 - при двойном отборе, 355
 - при изучении нескольких признаков, 135, 137
 - при заданном объеме выборки, 113, 124
 - при заданных общих издержках, 111, 124
 - при оценивании долей, 123
 - сравнение с пропорциональным размещением, 115, 125
 - сравнение с простым случайным отбором, 115, 116
- Оптимальный объем подвыборки
 - исходные единицы одинакового размера, 301
 - исходные единицы неодинакового размера, 338—342
 - Оптимальный процент замещения при отборе более чем в два момента, 370
 - Основа выборки (frame), 21
 - оценивание в случае, когда основа выборки содержит единицы, не принадлежащие к совокупности, 51
 - Отбор без возвращения (sampling without replacement), 34
 - с одинаковыми вероятностями, 280—287, 292—296, 345
 - Отбор более чем в два момента, 369
 - Отбор в два момента (sampling on two occasions), 367
 - Отбор по решетке (lattice sampling), 249
 - Отбор с возвращением (sampling with replacement), 34
 - Отклонение от нормальности (non-normality)
 - влияние на выборочную дисперсию, 59
 - влияние на выборочные средние, 56
 - влияние на доверительные границы, 56
 - влияние расслоения, 59
 - встречающиеся в практике выборочного метода, 55
 - Относительная дисперсия (relative variance), 177
 - Относительная чистая точность (relative net precision), 235
 - Относительное смещение (relative bias), 179
 - Относительная точность (relative precision)
 - метод вычисления, 118
 - оптимального и общего размещения, 131
 - простого и расслоенного случайного отбора, 114, 124, 154
 - вид совокупности, обеспечивающей максимальный выигрыш в точности, 116
 - Оценка по отношению (ratio estimate), 44
 - верхняя граница относительного смещения, 188
 - дисперсия, 45, 176
 - доверительные границы, 182
 - достоверность приближенной дисперсии, 177
 - как частный случай оценки по регрессии, 209
 - метод Лахири, 197

- поправки для уменьшения смещения, 195
- смещение, 199
- состоятельность, 175
- сравнение с отбором с вероятностями, пропорциональными размеру, 276
- сравнение со средним на единицу, 183
- сравнение с оценкой по регрессии, 218
- сравнение с расслоенным отбором, 167
- оптимальное размещение, 192
- оптимальные условия для применения, 185
- оценка дисперсии, 46, 181
- оценка Хартли—Росса, 195
- при гнездовом отборе, 80—82, 323, 331, 334—336
- при двойном отборе, 364
- при расслоенном двухступенчатом отборе, 344
- при расслоенном случайном отборе, 186
- стандартная ошибка при сравнении двух отношений, 200
- условия, при которых оценка не смещена, 179, см. также Совместная оценка по отношению и Раздельная оценка по отношению
- Оценка по регрессии (regression estimate), 228—223
 - дисперсия, 213
 - достоверность дисперсии для больших выборок, 215
 - оценка дисперсии, 214
 - при двойном отборе, 359
 - применения, 210
 - при повторном отборе из одной и той же совокупности, 367—376
 - при расслоенном случайном отборе, 219
 - смещение, 216, 217
 - сравнение со средним на единицу, 218
 - сравнение с оценкой по отношению, 218
 - условия, при которых оценка не смещена, 217, см. также Совместная оценка по регрессии и Раздельная оценка по регрессии
- Оценки дисперсий для совокупности
 - влияние ошибок при определении S_y на точность расслоенного отбора, 131
 - при определении объема выборки, 92
- Ошибка (error), границы (limits of), 87—89

- Ошибки наблюдения (errors of measurement), 399
 - влияние корреляции между ошибками внутри выборки, 406
 - влияние ошибок наблюдения, выводы, 406
 - влияние ошибок некоррелированных внутри выборки, 402
 - влияние постоянного смещения, 401
 - использование взаимопроницающих подвыборок, 408
 - математическая модель изучения, 399
 - методы изучения, 406
- Ошибки при обследованиях (errors in surveys), виды, 379
- Перепись (census), сравнение с выборочным обследованием, 16
- Периодическая вариация, влияние на систематический отбор, 237
- Повторное обследование одной и той же совокупности (repeated sampling), 365
- Повторные обращения (call-backs), 385
 - влияние на долю ответивших, 385—391
 - математическая модель влияния, 387—391
 - оптимальная тактика, 389—391
 - относительные издержки, 385—387
 - попытки избежать, 391—395
- Повторный опрос (reinterview) при изучении ошибок наблюдения, 407
- Подотбор (subsampling), единицы одинакового размера, 291
- Подотбор среди неответивших (subsampling of nonrespondent), 391—395, см. также двухступенчатый отбор, трехступенчатый отбор
- Полномасштабное распределение (multinomial distribution), 74
- Поправка на конечность совокупности (пке) (finite population correction)
 - для оценок по отношению, 46, 176
 - для оценок по регрессии, 211
 - когда можно не учитывать, 39
 - при двухступенчатом отборе, 276
 - при расслоенном случайном отборе, 107
- Поправка на непрерывность (correction for continuity), 72, 78
- Потери (loss), связанные с ошибками при оценивании, 98
- Потерь функция (loss function), 98, 138
- Пределы ошибки желательные (limits of error tolerable), 88—91

Приемочный контроль (acceptance sampling), применение выборочного метода, 18, 99

Признак (item) определение, 34

Пробное обследование (pilot survey) применение при оценивании дисперсий для совокупности, 94 применение для оценивания оптимальных долей отбора и подотбора, 305

Пропорциональное размещение при расслоенном отборе дисперсия, 108, 123 получение равновзвешенной выборки, 105 правило применения, 118, 125 при отборе для оценивания долей, 124 сравнение с оптимальным размещением, 105, 118, 123 сравнение с простым случайным отбором, 105 сравнение с расслоением после извлечения выборки, 153

Простой случайный отбор (simple random sampling), 33 дисперсия доли для выборки, 64 дисперсия среднего для выборки, 37, 43 доверительные границы доли для выборки, 71 доверительные границы среднего для выборки, 41 объем выборки при оценивании долей, 91 объем выборки для оценивания среднего, 89 оценка дисперсии доли для выборки, 66 оценка дисперсии среднего для выборки, 40 распределение доли для выборки, 69, 70 способ получения выборки, 33 точность в сравнении с расслоенным случайным отбором, 114, 124, 155

Проценты, оценивание, см. Доли, оценивание

Путевые расходы (travel costs), 112

Равновзвешенная оценка (self-weighting estimate), 105 при двухступенчатом отборе, 324, 327, 329, 331, 333

Равномерное распределение (rectangular distribution) дисперсия, 95

Раздельная (separate) оценка по отношению, 186 возможность смещения, 187 в сравнении с совместной оценкой по отношению, 189

дисперсия, 186 оптимальное размещение, 192 оценка дисперсии, 198 при расслоенном двухступенчатом отборе, 344

Раздельная (separate) оценка по регрессии возможность смещения, 221 в сравнении с совместной оценкой по регрессии, 223 дисперсия, 220, 221 оценка дисперсии, 221

Размещение выборки при расслоенном отборе, см. оптимальное размещение при расслоенном отборе

Распределение Стьюдента (Student t-distribution), 27, 58

Расслоение (stratification) 103 влияние числа слоев на точность, 151 наилучшая переменная, 116 необходимость применения, 103 по двум признакам, 144 после получения выборки, 153 при двойном отборе, 352 при двухступенчатом отборе, 342, 344

Расслоенный случайный отбор (stratified random sampling), 103 вид совокупности, для которой выигрыш в точности значителен, 116 дисперсия p_{st} , 123 дисперсия y_{st} , 107 для оценок по отношению, 186 для оценок по регрессии, 219 доверительные границы в случае непрерывных данных, 110 объем выборки, 120, 126 оптимальное размещение, 111 оценивание средних и суммарных значений для областей изучения, 164 оценка p_{st} , 123 оценка y_{st} , 105 оценка дисперсии y_{st} , 110; p_{st} , 123 при двухступенчатом отборе, 310 сравнение с отбором с вероятностями, пропорциональными размеру, 276 сравнение с оценкой по отношению, 190, 276 сравнение с простым случайным отбором, 109 сравнение с систематическим отбором, 233—247 формирование слоев, 146 при одной единице в слое, 159

Расслоенный отбор, общая формула дисперсии оценки, 106

Регрессионный коэффициент (regression coefficient), для конечной совокупности, 210, 211

Редкие признаки (rare items), необходимость большой выборки

Систематический отбор (systematic sampling) в автокоррелированных совокупностях, 238 в двумерных совокупностях, 247 влияние периодической вариации, 237 в реальных совокупностях, 241 в совокупностях, единицы которых расположены в «случайном» порядке, 233 в совокупностях с линейным трендом, 235 дисперсия оценки, 227 конечные поправки, 237 метод извлечения, 225, 227 оценивание дисперсии, 243—246 преимущества, 225 при двухступенчатом отборе, 299 при одноступенчатом гнездовом отборе, 284 расслоенный систематический отбор, 246 рекомендации относительно применения, 250 связь с гнездовым отбором, 227 сравнение с простым случайным отбором, 228, 233, 247 сравнение с расслоенным отбором, 233—246

Слой (stratum) границы слоев, правило нахождения, 146

Слагаемое дисперсии (component of variance), вариация между средними для исходных единиц, 301

Случайный отбор, см. Простой случайный отбор

«Случайная точка» («random point»), метод отбора, 260

Смещение (bias), 27 влияние на доверительные границы, 28, 29 в оценках по отношению, 188 в оценках по регрессии, 216, 217 допустимая величина, 30 применение смещенных оценок, 27—29 связанное с неполучением ответа, 381 связанное с ошибками при определении веса слоя, 133 у обследователей, 400

Смещение (y) обследователей (interviewer bias), 400

математическая модель, 406—412

Совместная оценка по отношению (combined ratio estimate), 187 верхняя граница относительного смещения, 188 в сравнении с раздельной оценкой по отношению, 189 дисперсия, 188 оптимальное размещение, 192 оценка дисперсии, 191 при расслоенном двухступенчатом отборе, 344 при оценивании средних для областей изучения, 166

Совместная оценка по регрессии (combined regression estimate), 220 в сравнении с раздельной оценкой по регрессии, 223 дисперсия, 222 оценка дисперсии, 222 смещение, 220

Совокупность (population) автокоррелированная, 238 двумерная, 247, 248 изучаемая, 20 обследуемая, 20 реальная (данные), 241 с линейным трендом, 235 с периодической вариацией, 237 со «случайным» порядком расположения единиц, 233

Совмещенные слои (collapsed strata), 160, 247

Составление перечня подъединиц в исходных единицах (listing of primary units) влияние издержек составления перечня на оптимальную вероятность отбора, 338—342

Составная оценка (composite estimate) при повторном обследовании одной и той же совокупности, 374

Среднее квадратичное отклонение (standard deviation), 26

Средний квадрат ошибки (mean square error), СКО, определение, 30 необходимость применения, 30 связь с дисперсией и смещением, 30

Стандартная ошибка (standard error) оценки доли для области изучения, 77, 78 оценки доли при гнездовом отборе, 78—82, 267—269 оценки доли при двухступенчатом отборе, 299—300 оценки доли при простом случайном отборе, 66—68 оценки доли при расслоенном отборе, 123—124

оценки по отношению, 45—47, 178, 179

оценки по отношению при рас-
слоенном отборе, 186—188
оценки по отношению при двух-
ступенчатом отборе, 335, 336, 344,
345

оценки по регрессии, 210, 213, 215
оценки по регрессии при рас-
слоенном отборе, 220, 222
оценки среднего для области изу-
чения, 49

оценки среднего для области изу-
чения при расслоенном отборе,
165—167

оценки среднего на элемент при
двухступенчатом отборе, 297—
299, 308, 315—332, 342, 343

оценки среднего на элемент при
отборе с неодинаковыми вероят-
ностями, 272—275, 282

оценки среднего при гнездовом
отборе, 79, 80

оценки среднего при отборе с не-
одинаковыми вероятностями без
возвращения, 282, 346

оценки среднего при простом слу-
чайном отборе, 38

оценки суммарного значения для
области изучения, 51

оценки суммарного значения для
области изучения при расслоен-
ном отборе, 50—53

оценки суммарного значения для
совокупности при простом слу-
чайном отборе, 38

разности двух отношений, 200,
203

разности оценок средних для об-
ластей изучения, 53, 54

Территориальный отбор (area sam-
pling), 383

Текущие оценки (current estimates)
при повторном обследовании одной
и той же совокупности, 365—376

Точность (precision) 30

Трехступенчатый отбор (threestage
sampling), 307

дисперсия среднего на единицу
третьей степени, 308

оптимальные доли отбора и под-
отбора, 309

Функция издержек (cost function),
в аналитических исследованиях,
163

при двойном отборе для оценки
по регрессии, 361

при двойном отборе для рассло-
ения, 352

при нахождении объема выбор-
ки, 98

при нахождении оптимальных ве-
роятностей отбора исходных еди-
ниц 337—343

при нахождении оптимальной до-
ли отбора среди неотвечавших,
391

при нахождении оптимального
размера единицы, 265

при нахождении числа слоев, 153

при расслоенном случайном отбо-
ре, 112

Характеристика (characteristic) см.
Признак, 34

Характеристика размера (measures
of size), 272

оптимальная для одноступенчато-
го гнездового отбора, 275

Центрально расположенная система-
тическая выборка (centrally located
systematic sample), 226

Чередующиеся выборки (rotation sam-
pling) см. Повторное обследование
одной и той же совокупности

Экцесс (kurtosis), 60

влияние на дисперсию для выбор-
ки, 59

Элемент (element), 67, 253

УКАЗАТЕЛЬ ИМЕН

Айзенхарт (Eisenhart), 301

Армитедж (Armitage P.), 114

Бартлетт (Bartlett M. S.), 217

Бартоломью (Bartholomew D. J.), 398

Беллок (Belloc N. B.), 406

Бершо (Berstad M.), 374, 400, 401,
403, 413

Бернбаум (Bernbaum Z. W.), 382, 383

Бернерт (Bernert E. H.), 151

Блис (Blythe R. H.), 98, 101

Блэк (Black C. A.), 249, 254

Боз Чемелли (Bose Chameli), 360

Большев Л. Н., 69

Боярский А. Я., 8

Брайант (Bryant E. C.), 144

Брукс (Brooks S.), 303, 306

Буныковский В. Я., 113

Вудруф (Woodruff R. S.), 376

Вулси (Woolsey T. D.), 384

Гаек (Hájek J.), 54, 183

Гаусс (Gauss K.), 185

Герни (Gurney M.), 149, 185, 193

Гранди (Grundy P. M.), 95, 281, 282,
283, 285, 345, 346

Грей (Gray P. G.), 314, 406

Гудмен (Goodman L. A.), 196

Гудмен (Goodman R.), 146

Далениус (Dalenius T.), 14, 138, 140,
146, 148, 149, 164, 169

Дас (Das A. C.), 248

Дейвид (David F. N.), 185

Делюри (De Lury D. B.), 71, 73, 246

Деминг (Deming W. E.), 14, 18, 95,
387, 397, 398, 410

Дербин (Durbin J.), 165, 167, 199, 292,
386, 394, 398, 413

Дес Радж (Des Raj), 277, 283, 284

Джебе (Jebe E. H.), 14, 319

Джексон (Jessen R. J.), 22, 53, 100,
118, 136, 144, 185, 257, 265, 347, 367

Джонс (Jones H. W.), 410

Джонсон (Johnson F. A.), 91, 255

Доула (Dough J. A.), 14

Дружинин Н. К., 11

Жаркович (Žarković S. S.), 163, 277

Йейтс (Yates F.), 14, 73, 98, 138, 162,
163, 165, 209, 237, 241, 246, 249

Кайерт (Cyert R. M.), 14, 250, 277,
281, 282, 283, 285, 345, 346, 348, 369

Кейфитц (Keyfitz N.), 163, 191, 206

Кендэл М., 8

Кенуи (Quenouille M. H.), 199, 240,
248

Кинг (King A. J.), 151, 355

Клиш (Kish L.), 14, 82, 146, 203, 384,
398, 407

Ковалевский А. Г., 8

Кокрен (Cochran W. G.), 5, 6, 8, 9, 10,
11, 14, 150, 151, 153, 240, 246, 277,
285, 398

Кокс (Cox D. R.), 92

Кокс Г., 5

Корлетт (Corlett T.), 314

Корнелл (Cornell F. G.), 14, 121, 122

Корнфилд (Cornfield J.), 42, 94, 268

Коши (Cauchy), 113, 128

Куп (Koop J. C.), 410

Кэлвин (Calvin L. D.), 14

Лахри (Lahiri D. B.), 196, 197, 198,
199, 206, 421

Лансинг (Lansing J. B.), 407

Либман (Lieberman G. J.), 71

Линдберг (Lindeberg), 54

Макней (McVay F. E.), 264, 269

Маккарти (McCarthy P. J.), 155, 385,
408

Маккарти (McCarthy D. E.), 151, 312

Маккензи (Mackenzie W. A.), 241

Маркс (Marks E. S.), 413

Матерн (Matern B.), 238, 241, 246, 247, 248
 Махаланобис (Mahalanobis P. C.), 291, 408, 411
 Мурти (Murthy M. N.), 283
 Мидзуно (Mizuno H.), 197, 207
 Микки (Mickey M. R.), 223
 Милн (Milne A.), 238, 248
 Мостеллер, 69
 Мэдоу (Madow L. H.), 230, 234, 238, 267, 305
 Мэдоу (Madow W. G.), 14, 30, 36, 54, 230, 234, 263, 374, 376
 Нарсин (Narsin R. D.), 283, 284, 286
 Нейман (Neyman J.), 8, 113, 116, 169, 185, 352, 355
 Нордин (Nordin J. A.), 99
 Олкин (Olkin I.), 203, 206
 Осборн (Osborne J. G.), 241, 246
 Оуэн (Owen D. B.), 71
 Паттерсон (Patterson H. D.), 249, 369
 Полиц (Politz A. N.), 395, 396, 398, 416, 417, 423
 Райфа (Raiffa H.), 99
 Рао (Rao J. N. K.), 284, 285
 Рао (Rao S.), 14
 Ренни (Rényi A.), 54
 Робсон (Robson D. S.), 196, 355
 Роджерс (Rogers S.), 14
 Ромиг (Romig H. G.), 69, 73
 Росс (Ross A.), 180, 188, 195, 198
 Рурке, 69
 Санделиус (Sandelius M.), 85
 Серкен (Sirken M. G.), 382, 383
 Сетх (Seth G. R.), 401
 Симмонс (Simmons W. R.), 395, 396, 398, 416, 417, 423
 Ситтин (Sittig J.), 99
 Слоним (Slonim M. J.), 18
 Смирнов Н. В., 69
 Старовский В. Н., 8
 Стивен (Stephan F. F.), 14, 134, 151, 154, 155, 385, 408
 Стюарт (Stuart A.), 386, 413
 Студент (Student), 27, 41, 58, 59, 182
 Сукхатм (Sukhatme P. V.), 133, 177, 254, 271, 325, 329, 401
 Тейлор, 177, 199
 Томас, 69

Томпсон (Thompson D. J.), 282, 288
 Тьюки (Tukey J. W.), 44
 Уишарт (Wishart J.), 44
 Уолд (Wold H.), 241
 Уотсон (Watson D. J.), 208
 Уэст (West Q. M.), 58, 60
 Феллер (Feller W.), 54
 Финкнер (Finkner A. L.), 14
 Финни (Finney D. J.), 85, 238, 241
 Фишер (Fisher R. A.), 58, 60, 73
 Хамис (Khamis S. H.), 63
 Хансен (Hansen M. H.), 14, 30, 36, 82, 163, 177, 193, 263, 267, 271, 305, 317, 319, 338, 341, 348, 374, 376, 391, 394, 400, 401, 403, 413
 Хансон (Hanson R. H.), 413
 Хартли (Hartley H. O.), 14, 144, 165, 180, 188, 195, 196, 198, 284, 285, 395
 Хауземан (Houseman E. E.), 53, 118
 Хегуд (Hagood M. J.), 151
 Хейнс (Haynes J. D.), 248
 Хендрикс (Hendricks W. A.), 265, 399
 Хервиц (Hurwitz W. N.), 14, 30, 36, 82, 193, 263, 267, 271, 305, 317, 319, 338, 341, 348, 374, 376, 391, 394, 400, 401, 403, 413
 Хесс (Hess I.), 203, 384, 398
 Ходжес (Hodges J. L.), 148, 169
 Хомейер (Hornmeyer P. G.), 249, 254
 Хорвиц (Horvitz D. G.), 282, 288, 411, 412
 Хотинский В. И., 8
 Шлайфер (Shlaifer R.), 93, 99
 Штейн (Stein C.), 92
 Шварц (Schwarz), 113, 128
 Чанг (Chung J. H.), 71, 73
 Чупров А. А., 8, 113
 Эванс (Evans W. D.), 114, 133
 Эклер (Eckler A. R.), 376
 Эрдош (Erdős P.), 54
 Юл, 8
 Ястребский Б. С., 8
 Ятц (Youtz C.), 14

УКАЗАТЕЛЬ ГЕОГРАФИЧЕСКИХ НАЗВАНИЙ

Австрия, 17
 Айова, 100, 118, 136, 189, 190, 228
 Арсенал, район, 411
 Балтимор, 81, 250
 Бенгалия, 411, 417
 Великобритания, 17, 398
 Голландия, 17
 Германия, 352
 Греция, 185, 347
 Джефферсон, графство, 189, 190, 356
 Детройт, 55
 Индианаполис, 19
 Индия, 307
 Италия, 17
 Калифорния, 118
 Канзас, 312

Нью-Йорк, 55
 Питтсбург, 411
 Северная Каролина, 193, 260, 319, 334, 380
 Сенека, графство, 58
 США, 17, 55, 101, 109, 117, 121, 149
 Филадельфия, 55
 Флорида, 118
 Цейлон, 17
 Чехословакия, 17
 Чикаго, 55
 Швеция, 17
 Эймс, 288

Составитель указателей Н. М. Сонин

ОГЛАВЛЕНИЕ

Предисловие к русскому переводу	5
Предисловие	12
ГЛАВА 1. ВВЕДЕНИЕ	15
1.1. Преимущества выборочного метода	15
1.2. Примеры применения выборочного метода	17
1.3. Основные проблемы выборочного обследования	19
1.4. Роль теории выборочного метода	23
1.5. Вероятностный отбор	24
1.6. Применение нормального распределения	26
1.7. Смещение и его роль	27
1.8. Средний квадрат ошибки	30
Упражнения	31
Литература	32
ГЛАВА 2. ПРОСТОЙ СЛУЧАЙНЫЙ ОТБОР	33
2.1. Простой случайный отбор	33
2.2. Определения и обозначения	34
2.3. Свойства оценок	35
2.4. Дисперсии оценок	37
2.5. Поправка на конечность совокупности	38
2.6. Оценивание стандартной ошибки по выборке	40
2.7. Доверительные границы	41
2.8. Другой метод доказательства	43
2.9. Оценивание отношения	44
2.10. Оценки средних значений для подсовкупностей	48
2.11. Оценки суммарных значений для подсовкупностей	50
2.12. Сравнение средних значений для областей изучения	53
2.13. Обоснованность аппроксимации нормальным распределением	54
2.14. Влияние отклонения распределения от нормального на выборочную дисперсию	59
Упражнения	61
Литература	63
ГЛАВА 3. ОТБОР ДЛЯ ОЦЕНИВАНИЯ ДОЛЕЙ И ПРОЦЕНТОВ	64
3.1. Качественные признаки	64
3.2. Дисперсии выборочных оценок	64
3.3. Влияние P на стандартные ошибки	67
3.4. Биномиальное распределение	69
3.5. Гипергеометрическое распределение	70
3.6. Доверительные границы	71
3.7. Классификация по нескольким признакам	74
3.8. Доверительные границы при классификации по нескольким признакам	74
3.9. Условное распределение p	75
3.10. Доли и суммарные значения для подсовкупностей	77
3.11. Сравнение между различными областями	78
3.12. Оценивание долей при гнездовом отборе	78
Упражнения	83
Литература	85
ГЛАВА 4. ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ	86
4.1. Гипотетический пример	86
4.2. Анализ проблемы	87
4.3. Задание уровня точности	88
4.4. Формула для n при отборе для оценивания долей	89
4.5. Формула для n в случае непрерывных переменных	91

4.6. Предварительные оценки дисперсий для совокупности	92
4.7. Объем выборки при изучении нескольких признаков	96
4.8. Объем выборки при необходимости получить оценки для подразделений совокупности	96
4.9. Объем выборки с точки зрения теории решений	97
Упражнения	99
Литература	101
ГЛАВА 5. РАССЛОЕННЫЙ СЛУЧАЙНЫЙ ОТБОР	103
5.1. Описание	103
5.2. Обозначения	104
5.3. Свойства оценок	105
5.4. Оценка дисперсии и доверительные границы	110
5.5. Оптимальное размещение	111
5.6. Сравнительная точность расслоенного случайного отбора и простого случайного отбора	114
5.7. При каких условиях расслоение обеспечивает большой выигрыш в точности?	116
5.8. Размещение, требующее более чем 100%-ного отбора	119
5.9. Определение объема выборки в случае непрерывных переменных	120
5.10. Расслоенный отбор для оценивания долей	122
5.11. Выигрыш в точности при расслоенном отборе для оценивания долей	124
5.12. Определение объема выборки при оценивании долей	126
Упражнения	126
Литература	130
ГЛАВА 5А. ДРУГИЕ ПРОБЛЕМЫ РАССЛОЕННОГО ОТБОРА	131
5А.1. Эффект отклонений от оптимального размещения	131
5А.2. Эффект ошибок в значениях величины слоев	133
5А.3. Проблема размещения при изучении нескольких признаков	135
5А.4. Другие способы размещения при изучении нескольких признаков	137
5А.5. Расслоение по двум признакам для небольших выборок	144
5А.6. Формирование слоев	146
5А.7. Число слоев	151
5А.8. Расслоение после извлечения выборки	153
5А.9. Отбор квотами	155
5А.10. Оценивание по выборке выигрыша, получаемого от расслоения	155
5А.11. Оценивание дисперсии при одной единице на слой	159
5А.12. Упрощенное вычисление стандартных ошибок	160
5А.13. Слои как области изучения	163
5А.14. Оценивание суммарных и средних значений для подсовкупностей	164
Упражнения	168
Литература	170
ГЛАВА 6. ОЦЕНКИ ПО ОТНОШЕНИЮ	172
6.1. Методы оценивания	172
6.2. Оценка по отношению	173
6.3. Приближенная дисперсия оценки по отношению	175
6.4. Достоверность приближенного значения дисперсии	177
6.5. Смещение оценки по отношению	178
6.6. Оценивание дисперсии по выборке	180
6.7. Доверительные границы	182
6.8. Сравнение оценки по отношению с оценкой по среднему на единицу	183
6.9. Условия, при которых оценка по отношению оптимальна	184

6.10. Оценка по отношению при расслоенном случайном отборе	186
6.11. Совместная оценка по отношению	187
6.12. Сравнение совместной и раздельной оценок	189
6.13. Упрощенное вычисление дисперсии	191
6.14. Оптимальное размещение для оценки по отношению	192
6.15. Несмещенные оценки типа оценок по отношению	194
6.16. Сравнение двух отношений	200
6.17. Многомерные оценки по отношению	203
Упражнения	205
Литература	207
ГЛАВА 7. ОЦЕНКИ ПО РЕГРЕССИИ	208
7.1. Линейные оценки по регрессии	208
7.2. Оценки по регрессии при заданном b	209
7.3. Оценка по регрессии, когда b вычисляется по выборке	212
7.4. Достоверность формулы для $V(\bar{y}_h)$ при больших выборках	215
7.5. Дополнительные замечания о смещении	217
7.6. Сравнение оценки по регрессии с оценкой по отношению и по среднему на единицу	118
7.7. Оценки по регрессии при расслоенном отборе	219
7.8. Выборочные оценки коэффициентов регрессии	221
7.9. Сравнение двух видов оценок по регрессии	223
Упражнения	223
Литература	224
ГЛАВА 8. СИСТЕМАТИЧЕСКИЙ ОТБОР	225
8.1. Описание	225
8.2. Связь систематического отбора с гнездовым	227
8.3. Дисперсия оценки среднего	227
8.4. Сравнение систематического отбора со случайным расслоенным отбором	233
8.5. Совокупности со «случайным» порядком расположения единиц	233
8.6. Совокупности с линейным трендом	235
8.7. Совокупности с периодической вариацией	237
8.8. Автокоррелированные совокупности	238
8.9. Реальные совокупности	241
8.10. Оценивание дисперсии по отдельной выборке	243
8.11. Расслоенный систематический отбор	246
8.12. Двумерный систематический отбор	247
8.13. Резюме	249
Упражнения	250
Литература	252
ГЛАВА 9. ОДНОСТУПЕНЧАТЫЙ ГНЕЗДОВОЙ ОТБОР	253
9.1. Почему необходим гнездовой отбор	253
9.2. Одно простое правило	254
9.3. Сравнение точности по данным выборочного обследования	258
9.4. Дисперсия, выраженная через внутригнездовую корреляцию	262
9.5. Дисперсия как функция размера единицы отбора	263
9.6. Функция издержек	265
9.7. Гнездовой отбор для оценивания долей	267
9.8. Гнездовые единицы неодинакового размера	269
9.9. Отбор с вероятностями, пропорциональными размеру единицы	271
9.10. Теория для отбора с произвольными вероятностями	272
9.11. Оптимальная характеристика размера единицы	275

9.12. Сравнительная точность способов отбора и оценивания	276
9.13. Обобщение на случай расслоенного отбора	280
9.14. Отбор с неравными вероятностями без возвращения	280
9.15. Другие подходы	283
9.16. Некоторые сравнения при $n = 2$	285
Упражнения	287
Литература	289

ГЛАВА 10. ПОДОТБОР ПРИ ЕДИНИЦАХ ОДИНАКОВОГО РАЗМЕРА

10.1. Двухступенчатый отбор	291
10.2. Две полезные теоремы	292
10.3. Дисперсия оценки среднего при двухступенчатом отборе	296
10.4. Оценивание дисперсии	297
10.5. Оценивание долей	299
10.6. Оптимальные доли отбора и подотбора	301
10.7. Оценивание σ^2 по данным пробного обследования	305
10.8. Трехступенчатый отбор	307
10.9. Расслоенный отбор единиц	310
10.10. Оптимальное размещение при расслоенном отборе	310
Упражнения	312
Литература	313

ГЛАВА 11. ПОДОТБОР ПРИ ЕДИНИЦАХ НЕОДИНАКОВОГО РАЗМЕРА

11.1. Введение	314
11.2. Методы отбора при $n = 1$	315
11.3. Отбор с вероятностями, пропорциональными оценке размера	320
11.4. Сводка методов при $n = 1$	322
11.5. Методы отбора при $n > 1$	322
11.6. Отбор единиц с равными вероятностями. Оценка по отношению с размером в знаменателе	323
11.7. Отбор единиц с равными вероятностями. Несмещенная оценка	326
11.8. Отбор единиц с вероятностями, пропорциональными характеристике размера. Несмещенная оценка	328
11.9. Отбор единиц с вероятностями, пропорциональными размеру. Несмещенная оценка	331
11.10. Отбор единиц с вероятностями, пропорциональными характеристике размера. Оценка по отношению с размером в знаменателе	331
11.11. Сравнение схем отбора и оценивания	332
11.12. Отношение с другой переменной в знаменателе	334
11.13. Дисперсия оценки по отношению при отборе с равными вероятностями	335
11.14. Дисперсия оценки по отношению при отборе с вероятностями, пропорциональными оценке размера	336
11.15. Определение долей отбора и подотбора. Отбор с равными вероятностями	337
11.16. Доли отбора и подотбора при отборе с вероятностями, пропорциональными оценке размера	338
11.17. Расслоенный отбор. Несмещенные оценки	342
11.18. Расслоенный отбор. Оценки по отношению	344
11.19. Отбор с неравными вероятностями без возвращения	345
11.20. Общие выводы	347
Упражнения	348
Литература	350

ГЛАВА 12. ДВОЙНОЙ ОТБОР

12.1. Описание метода	351
12.2. Двойной отбор для расслоения	352
12.3. Оптимальное размещение	355

12.4. Оценка дисперсии при двойном отборе для расслоения	357
12.5. Оценки по регрессии	359
12.6. Сравнение двойного отбора для оценки по регрессии с однократным отбором	361
12.7. Оценка дисперсии при двойном отборе для оценки по регрессии	363
12.8. Оценки по отношению	364
12.9. Повторное выборочное исследование одной и той же совокупности	365
12.10. Отбор в два момента	367
12.11. Отбор более чем в два момента	369
12.12. Упрощение и дальнейшее развитие изложенных методов	372
Упражнения	376
Литература	378
ГЛАВА 13. ИСТОЧНИКИ ОШИБОК ПРИ ОБСЛЕДОВАНИЯХ	379
13.1. Введение	379
13.2. Эффект неполучения ответа	379
13.3. Виды неполучения ответа	383
13.4. Повторные обращения	385
13.5. Математическая модель эффекта повторных обращений	387
13.6. Оптимальная доля отбора среди неответивших	391
13.7. Поправки на смещение без повторных обращений	395
13.8. Математическая модель ошибок наблюдения	399
13.9. Эффект постоянного смещения	401
13.10. Эффект ошибок, некоррелированных внутри выборки	402
13.11. Эффект корреляции между ошибками внутри выборки	405
13.12. Эффект ошибок наблюдения. Резюме	406
13.13. Изучение ошибок наблюдения	406
13.14. Взаимопроницающие подвыборки	408
13.15. Обобщение на более сложные схемы отбора	410
13.16. Контролируемые эксперименты, включенные в обследование	412
13.17. Выводы	414
Упражнения	415
Литература	418
Ответы к упражнениям	420
Предметный указатель	425
Указатель имен	433
Указатель географических названий	435

Уильям Кокрен

МЕТОДЫ ВЫБОРОЧНОГО ИССЛЕДОВАНИЯ

Редактор Е. В. Крестянинова Техн. редактор Р. И. Феоктистова
 Корректоры А. Т. Сидорова, С. С. Нисаревская
 Худ. редактор Т. В. Стихно. Переплет художника В. С. Сергеевой

Сдано в набор 13/XI 1975 г. Подп. к печати 21/V 1976 г.
 Формат бумаги 60×90/16. Бумага № 2. Объем 27,5 печ. л. Уч.-изд. л. 29,32.
 Усл. п. л. 27,5 Тираж 10500 экз. (Тематич. план 1976 г. № 114).
 Заказ № 581 Цена 1 р. 92 к.

Издательство «Статистика», Москва, ул. Кирова, 30.

Московская типография № 4 Союзполиграфпрома при Государственном комитете
 Ситета Министров СССР по делам издательства, полиграфии и книжной торговли,
 Москва, Б-41, Б. Перевская ул., 46