Методы выборочных наблюдений

2021 год

Оглавление

I. Простая случайная выборка без возвращения (ПСВ без ВО) 4

II.Простая случайная выборка с возвращением 10

III.Стратифицированная выборка 13

IV.Кластерная выборка, двухэтапная выборка и систематический отбор 25

V. Систематическая выборка 41

VI. Выборки с равными и неравными вероятностями включения 45

VII. Список использованной литературы 57

Введение

Я проработал в органах статистики разных стран мира более тридцати лет и моя деятельность была тесно связана с проведением выборочных статистических наблюдений предприятий, организаций и населения. Ежегодно я участвовал в проведении тысячи обследований в отдельных областях статистики - строительстве, промышленности, сельском хозяйстве, торговле и сфере услуг, а также исследованиях социально-демографических характеристик населения.

Таким образом, выборочные обследования охватывали максимально широкий спектр: обследования конъюнктуры и деловой активности в сфере услуг и торговли, деятельности малых предприятий и бюджетных организаций, сельскохозяйственной деятельности крестьянских хозяйств и индивидуальных предпринимателей, бюджетов домашних хозяйств, уровня доходов и расходов населения, занятости и безработицы, потребительских цен и цен производителей и т.п.

Отнюдь не всегда на практике выборочные обследования представляют собой вероятностные выборки, позволяющие оценить точность получаемых оценок по выборке и сделать обоснованный вывод о репрезентативности выборки. В силу разных причин, в том числе организационных, достаточно широко используются методы основного массива, ценза, направленного отбора и др. Точность и достоверность получаемых таким образом оценок показателей - вопрос дискуссионный. Возможно, в некоторых случаях в них присутствует элемент случайности и они могут расцениваться как вероятностый метод отбора. Однако, эта ситуация требует отдельного изучения, исследования и анализа.

Выборочная методология, несмотря на очевидные премущества (например, экономия финансовых средств) сложна с нескольких точек зрения: с математической, так как основана на достаточно сложной и не всегда очевидной и проработанной для каждого случая математической базе, так и организационной, в силу того, что порождает значительное количество практических проблем.

Следует отметить, что повсеместная автоматизация в органах статистики привела к тому, что поколение молодых специалистов не особо осведомлены о деталях и ограничениях выборочной методики, в соответствии с которой проводится руководимое ими обследование, так как компьютерная программа, созданная специалистами предыдущего поколения совместно с IT-разработчиками, выполняет расчеты самостоятельно и не требует их участия в вычислениях. При этом она из года в год выдает приемлимые для организатора обследований результаты.

Между тем выборочные исследования – это теория и противоречивая практика, компромиссы и тяжелые сомнения, парадоксы и ограничения, правовые коллизии и преимущества, очевидные плюсы и провалы…

В данном материале представлены наиболее распространенные методы выборочных исследований, которые используются многими странами мира для формирования выборок, получения оценок по выборке, а также характеристик точности полученных оценок.

Алексей Хохловский, Санкт-Петербург

I. Простая случайная выборка без возвращения (ПСВ без ВО)

Краткое описание

Простая случайная выборка без возвращения является наиболее востребованным выборочным планом. Этот выборочный метод считается наиболее простым, так как отбор производится из целой совокупности. Существует также простая случайная выборка с возвращением, но она используется не часто, в том числе статистическими службами. Однако, в связи с тем, что относительно простая оценка дисперсии по выборке с возвращением может применяться при определенных условиях в случае отбора с возвращением, то соответствующие формулы дисперсии кратко описаны далее.

Определение простой случайной выборки без возвращения (ПСВ без ВО)

Формальное определение ПСВ без ВО следующее. Рассмотрим совокупность U, состоящую из N элементов, т.е. U=(1,2,...N). ПСВ без ВО – это метод отбора n-элементов из совокупности U, при котором все возможные подмножества U, состоящие из n-элементов, имеют одинаковую вероятность включения в выборку. Таким образом, существует $\begin{bmatrix} N \\ n \end{bmatrix}$ возможных подмножеств совокупности U объема n.

Схема отбора ПСВ без ВО

В практической деятельности ПСВ без ВО производится путем последовательного отбора случайных чисел между 1 и N, а затем отбора соответствующих им единиц элементов совокупности, до тех пор пока не будет сформирована выборка объема n. Если число уже было извлечено, то выбирается случайным образом новое число.

Хорошей альтернативой случайному отбору является систематический отбор, который описан далее в данном.

Применимость

ПСВ без ВО часто используется при проведении обследований предприятий и организаций (бизнес-обследований).

Обычно совокупность предприятий расслаивается (стратифицируется) и ПСВ без ВО применяется по отношению к каждой страте (слою) (см. следующую главу о стратифицированной выборке). ПСВ без ВО часто

используется в качестве бенчмаркинга для оценки качества других планов выборки путем сравнения дисперсий оценок, полученных в соответствии с различными планами выборки.

Описание

Начнем с краткого описания параметров совокупности. Для некоторого целевого показателя суммарное значение показателя обозначается Y, среднее значение - $\bar{Y}$, дисперсия показателя в совокупности - σ_y^2, скорректированная дисперсия - S_y^2, коэффициент вариации - CV_y. Скорректированная дисперсия часто используется для упращения формул ПСВ без ВО. Более того Y_k обозначено значение целевого показателя k-го элемента, k=1,…N. Описанные выше параметры определяются по следующим формулам:

$$Y = Y_1 + Y \ldots + Y_N = \sum_{k=1}^{N} Y_k \qquad \textbf{2.1.}$$

$$\bar{Y} = \frac{1}{N} Y = \frac{1}{N} \sum_{k=1}^{N} Y_k$$

$$\sigma_y^2 = \frac{1}{N} \sum_{k=1}^{N} (Y_k - \bar{Y})^2$$

$$S_y^2 = \frac{N}{N-1} \sigma_y^2 = \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - \bar{Y})^2$$

$$CV_y = \frac{S_y}{\bar{Y}}$$

n-наблюдений (т.е. объема выборки) целевого показателя в выборке обозначаются маленькими буквами $y_{1,\ldots,} y_n$.Среднее выборочное значение $\overline{y_s}$, выборочная дисперсия s_y^2, коэффициент вариации cv_y являются наиболее важными параметрами выборки. Эти параметры определяются следующим образом:

$$\overline{y_S} = \frac{1}{n} \sum_{k=1}^{n} y_k \qquad \textbf{2.2.}$$

$$s_y^2 = \frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y_S})^2$$

$$cv_y = \frac{s_y}{\overline{y_s}}$$

Оценка среднего значения, суммарного значения, скорректированной дисперсии

Для трех из пяти параметров, описанные в формулах 2.1., точные оценки приведены ниже. Эти параметры: суммарное значение совокупности, среднее значение, скорректированная дисперсия. Интуитивно среднее значение $\overline{y_s}$ воспринимается как допустимая оценка среднего значения и рассчитывается следующим образом:

$$\hat{\bar{Y}} = \bar{y}_s = \frac{1}{n}\sum_{k=1}^{n} y_k \quad \textbf{2.3.}$$

Шапочка ˆ означает оценку соответствующих параметров. Оценку суммарного значения можно вычислить по следующей формуле:

$$\hat{Y} = N\hat{\bar{Y}} = N\bar{y}_s = \sum_{k=1}^{n}(\frac{N}{n})y_k \quad \textbf{2.4.}$$

Параметр N/n в данном выражении называется *весом включения* выборки так как представляет собой отношение объема совокупности к объему выборки.

Таким образом, выборочная дисперсия – оценка скорректированной дисперсии совокупности:

$$\hat{S}_y^2 = s_y^2 = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \overline{y_S})^2 \quad \textbf{2.5.}$$

Несмещенность оценок

Оценка является несмещенной если математическое ожидание величины эквивалентно оцениваемому показателю. Для того, чтобы определить матожидания (прямых) оценок среднего или суммарного значения для ПСВ без ВО, удобно переписать формулу в иной форме. Формула оценки среднего значения можно представить следующим образом:

$$\hat{\bar{Y}} = \bar{y}_s = \frac{1}{n}\sum_{k=1}^{N} a_k Y_k \quad \textbf{2.6.}$$

где бинарная случайная переменная a_k определяется:

$a_k = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix}$ (**2.7.**), где 1 если k элемент включен в выборку и 0 если k элемент не включен в выборку.

Бинарную случайную переменную a_k также называют индикатором отбора. Этот индикатор обладает двумя свойствами:

$$E(a_k) = P(a_k = 1) = \frac{n}{N} \qquad \text{k=1,…N} \quad \textbf{2.8.}$$

$$E(a_k a_l) = \begin{Bmatrix} n/N \\ \frac{n(n-1)}{N(N-1)} \end{Bmatrix}$$

k=l ∧ k=1, …N

$1 \le k \ne l \le N$ **2.9.**

Используем формулу 2.6. в качестве начальной точки. Затем на основе свойств, приведенных выше, и опираясь на формулу индикатора отбора a_k, убеждаемся в том, что оценка является *несмещенной оценкой* $\bar{Y}$:

$$E(\hat{\bar{Y}}) = E(\frac{1}{n}\sum_{k=1}^{N} a_k Y_k) = \frac{1}{n}\sum_{k=1}^{N} E(a_k) Y_k = \frac{1}{n}\sum_{k=1}^{N} \frac{n}{N} Y_k = \frac{1}{N}\sum_{k=1}^{N} Y_k = \bar{Y}$$

Легко установить, что оценка суммарного значения также является несмещенной:

$$E(\hat{Y}) = E(N\hat{\bar{Y}}) = NE(\hat{\bar{Y}}) = N\bar{Y} = Y$$

Индикатор отбора также полезен при доказательстве того, что s_y^2 в формуле 2.5. является несмещенной оценкой скорректированной дисперсии. Иными словами:

$$E(s_y^2) = E\left[\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y}_s)^2\right] = S_y^2 \qquad \textbf{2.10}$$

Расчет дисперсии

Может быть доказано, что дисперсия оценки среднего значения является функцией скорректированной дисперсии:

$$\mathrm{var}(\hat{\bar{Y}}) = \frac{1}{n}(1-f)S_y^2 = \frac{1}{N}\left(\frac{1-f}{f}\right)S_y^2 \qquad f = \frac{n}{N} \qquad \textbf{2.11}$$

1-f является поправкой на конечную совокупность - пкс, *f* называется выборочной долей или долей выборки.

Основываясь на результате дисперсия оценки суммарного значения для совокупности:

$$\mathrm{var}(\hat{Y}) = N^2\,\mathrm{var}(\hat{\bar{Y}}) = N(\frac{1-f}{f})S_y^2 \qquad \textbf{2.12}$$

Значение параметра f зависит от значения параметра (1-f)/f, как это показана на графике 2.1. Если при фиксированном объема совокупности N объем выборки n увеличивается, то выборочная доля f становится равно 1. В этом случае, как это видно на графике, параметр (1-f)/f приближается к 0. Другими словами дисперсия оценки суммарного значения приближается к нулю по мере того как увеличивается объем выборки (см. формулу 2.12).

Дисперсия оценки среднего значения может быть рассчитана следующим образом:

$$\hat{\text{var}}(\hat{\bar{Y}}) = \frac{1}{N}\left(\frac{1-f}{f}\right)\hat{S}_y^2 = \frac{1}{N}\left(\frac{1-f}{f}\right)s_y^2 \quad \textbf{2.13}$$

где s_y^2 представляет собой выборочную дисперсию. Несмещенность оценки следует из того, что выборочная дисперсия s_y^2 - оценка скорректированной дисперсии S_y^2.

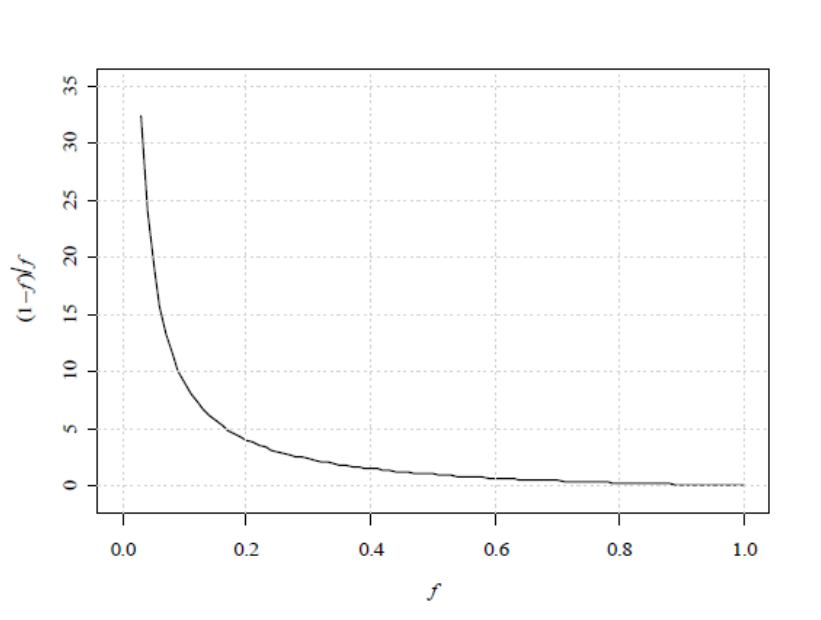
Несмещенная оценка дисперсии оценки $\hat{Y}$ может быть рассчитана:

$$\hat{\text{var}}(\hat{Y}) = N\left(\frac{1-f}{f}\right)s_y^2 \quad 2.14$$

s_y^2 - выборочная дисперсия.

Несмещенность дисперсии оценки является прямым следствием (доказанным) несмещенности s_y^2 в качестве оценки S_y^2.

График 2.1. Зависимость параметра f от параметра (1-f)/f



Коэффициент вариации

Помимо дисперсии оценки расчитывается также коэффициент вариации. Коэффициент вариации оценки среднего значения, обозначаемый $CV(\hat{\bar{Y}})$, и коэффициет вариации оценки суммарного значения $CV(\hat{Y})$ определяют по формулам:

$$CV(\hat{\bar{Y}}) = \frac{\sqrt{\text{var}(\hat{\bar{Y}})}}{\bar{Y}}$$

$$CV(\hat{Y}) = \frac{\sqrt{\text{var}(\hat{Y})}}{Y} \quad \textbf{2.15}$$

Вышеприведенная формула содержит дисперсию оценки среднего значения показателя на основе ПСВ без ВО. Эта формула может быть использована для вывода выражения, приведенного ниже, в соответствии с которым коэфициент вариации оценки рассчитывается:

$$CV^2(\hat{\bar{Y}}) = \frac{1}{N}(\frac{1-f}{f})CV_v^2 \quad \textbf{2.16}$$

Другими словами существует прямая связь между коэффициентом вариации оценки среднего значения показателя и коэффициентом вариации совокупности.

Учитывая, что существует взаимосвязь между суммарным значением совокупности и средним значением совокупности, между оценкой суммарного значения и оценкой среднего значения, можно сделать вывод о том, что коэффициенты вариации двух оценок связаны следующим образом:

$$CV(\hat{Y}) = CV(\hat{\bar{Y}}) \quad \textbf{2.17}$$

Приведенное выше выражение показывает, что прямые оценки суммарного значения и среднего значения одинаково точны.

К сожалению, на практике невозможно рассчитать значения этих коэфициентов и, следовательно, они должны быть оценены. Таким образом, оценкой коэфициента вариации оценки является:

$$\hat{CV}(\hat{Y}) = \frac{s_y}{\bar{y}_s}\sqrt{\frac{1}{N}(\frac{1-f}{f})} \quad \textbf{2.18}$$

так как

$$\hat{CV}_y = cv_y = \frac{s_y}{\bar{y}_s}$$

Важно отметить, что выборочная доля часто очень мала на практике, т.е. можно считать ее равной 0. Это фактически означает, что в данном случае выборка с возвращением эквивалентна выборке без возвращения.

Интуитивно также понятно, что на значение оценки не влияет присутствие возвращения, вследствие того, что объем совокупности намного превосходит объем выборки.

В следующем разделе обсуждается выборка с возвращением.

II.Простая случайная выборка с возвращением

Предположим мы извлекаем простую случайную выборку с возвращением с равными вероятностями включения 1/N (ПСВ с ВО). Тогда $y_1,...y_n$ может рассматриваться как n независимых извлечений из $Y_1,...Y_n$ с математическим ожиданием и дисперсией:

$$E(y_k) = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y} \quad k = 1,...,n$$

$$\text{var}(y_k) = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \bar{Y})^2 = \sigma_y^2 \quad k = 1,...,n \qquad \mathbf{2.19}$$

Отметим, что эти две формулы также применяются к выборке без возвращения. В случае выборки с ВО эти две формулы приводят прямо соответственно к математическому ожиданию и дисперсии оценки $\hat{\bar{Y}} = \bar{y}_s$:

$$E(\bar{y}_S) = \frac{1}{n}\sum_{k=1}^{n} E(y_k) = \bar{Y}$$

$$\text{var}(\bar{y}_s) = \frac{1}{n}\sigma_y^2 \quad \mathbf{2.20}$$

Сравнивнение формул дисперсии, приведенных в этом разделе, с формулами из предыдущего раздела подверждает сделанный ранее вывод о том, что формулы дисперсии приблизительно одни и те же, если f мало.

В заключении необходимо отметить, что в случае выборки с возвращением верно равенство $E(s_y^2) = \sigma_y^2$. Доказательство выражений опускается, так оно почти в значительной степени такое же, как и для выборки без возвращения.

Определение объема выборки

Наиболее часто задаваемый вопрос в практической деятельности - насколько большой должен быть объем выборки, чтобы обеспечить

определенную точность показателя Y - суммарного значения совокупности. Наиболее часто используемый критерий точности - это 95% доверительный интервал $I_{95}(Y).I_{95}(Y)$ - означает, что с 95% вероятностью неизвестное значение параметра находится в пределах интервала $I_{95}(Y)$.

Предположим, что показатель $\hat{Y}$ приблизительно нормально распределен, интервал с 95% вероятностью в случае, если n достаточно велико, определяется по формуле:

$$I_{95}(Y) = (\hat{Y} - 1{,}96\sqrt{\mathrm{var}(\hat{Y})}, \hat{Y} + 1{,}96\sqrt{\mathrm{var}(\hat{Y})}).$$

Вышеуказанное то выражение можно переформулировать следующим образом:

$$I_{95}(Y) = \hat{Y} \pm 1{,}96\frac{\sqrt{\hat{\mathrm{var}}(\hat{Y})}}{\hat{Y}}100\% = \hat{Y} \pm \hat{CV}(\hat{Y}) \times 196\% \quad \mathbf{2.21}$$

Это также можно пояснить на простом примере. Предположим $\hat{Y} = 120$, а $\hat{\mathrm{var}}(\hat{Y}) = 25$. Мы тогда получаем $I_{95}(Y) =$ (120-9,8; 120+9,8) = (110,2;129,8). В соответствии с формулой, приведенной выше, можно представить этот интервал следующим образом: $120 \pm 8{,}2\%$.

Приведенные выше уравнения также позволяют определить объем выборки, который необходим для того, чтобы оставаться в пределах установленного размера неопределенности, скажем r%.

Из формулы

$$196 \times CV_y\sqrt{\frac{1}{N} \times \frac{1-f}{f}} \langle r$$

следует, что

$$\frac{196^2 \times CV_y{}^2}{N \times r^2 + 196^2 \times CV_y^2} \langle f$$

Если 5% неопределенность признается приемлемой, выборочная доля f для совокупности, где N=1000, CV_y=0,7 должна удовлетворять неравенству:

$$\frac{196^2 \times 0.49}{1000 \times 25 + 196^2 \times 0{,}49} \approx 0{,}43 = f_{\min} < f$$

Для совокупности объема N=1000 эта минимальная доля отбора соотвествует минимальному объему выборки $n_{\min} = 430$.

На практике CV_y должно быть каким-то образом оценено. Оценки могут часто основываться на последних данных. Иногда возможно использовать сопоставимые данные с той же самой волативностью.

Оценивание доли

Необходимо рассчитать процент людей с определенными характеристиками, например количество людей с высоким уровнем заработной платы. Y_k принимает следующие значения:

$Y_k = \begin{cases} 1 \\ 0 \end{cases}$, где

1- если работник имеет высокую заработную плату;

0- если у работника невысокая заработная плата;

$k = 1,...N$

Если $P = \overline{Y}$ - доля работников с высоким уровнем заработной платы в совокупности, тогда (в этой ситуации $Y_k = Y_k^2$):

$$\overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k = P$$

$$\sigma_y^2 = \frac{1}{N}\sum_{k=1}^{N}(Y_k - P)^2 = \frac{1}{N}\sum_{k=1}^{N} Y_k^2 - P^2 = P - P^2 = PQ$$

$Q = 1 - P$

$$S_y^2 = \frac{N}{N-1}PQ$$

Доля работников по выборке с высокой заработной платой по выборке $p = \hat{P}$ в соотвествии с формулами 2.3 и 2.11:

$p = \overline{y}_s$

$$\text{var}(p) = \frac{1}{N}\left(\frac{1-f}{f}\right)S_y^2 = \frac{1}{N-1}\left(\frac{1-f}{f}\right)PQ.$$

Дисперсия оценки может быть вычислена следующим образом:

$$\hat{\text{var}}(p) = \frac{1}{N}\left(\frac{1-f}{f}\right)s_y^2 = \frac{1}{N}\left(\frac{1-f}{f}\right)\left(\frac{1}{n-1}\right)\sum_{k=1}^{n}(y_k - p)^2 = (1-f)(\frac{1}{n-1})p(1-p)$$

Коэффициент вариации оценки p может быть оценен:

$$\hat{C}V(p) = \sqrt{(1-f)(\frac{1}{n-1})\left(\frac{1-p}{p}\right)}$$

Из предыдущих формул следует что для интервала p с 95% уровнем доверительности и с сравнительно большим объемом n $I_{95}(p)$ может быть оценен:

$$I_{95}(p) = p \pm 196\% \times \sqrt{(1-f)(\frac{1}{n-1})(\frac{1-p}{p})}$$

Из вышеуказанной формулы следует, что степень неорпределенности сильно возрастает по мере уменьшения p. Например, в случае ПСВ без ВО параметр P оценивается 0,001 при n=50 000 и f≈0, и в этом случае относительный размер неопределенности равен 28%. Этот пример демонстрирует трудности оценивания малых областей с определенной точностью.

Показатели качества

Показателями качества ПСВ без ВО являются:

- пределы неопределенности оценок;

- уровень неответов.

Уровень неответов может серьезно повлиять на качество результатов, если уровень неответов высок и является селективным (избирательным). Смещение, вызванное селективным неответом, иногда можно скорректировать с помощью вспомогательной информации, относящейся как к вероятности ответа, так и к целевому показателю.

Другим важным предположением при реализации случайной выборки является то, что основа, из которой выборка извлекается, близка к целевой совокупности, о которой должны быть сделаны выводы.

III.Стратифицированная выборка

Краткое описание

Определение стратифицированной случайной выборки предполагает разбиение целевой совокупности на подсовокупности или страты. Страты не должны пересекаться и вместе должны охватывать всю совокупность. Случайная выборка после этого отбирается из каждого слоя без возвращения. Различные выборки без возвращения взаимно независимы. Другими словами, вместо того, чтобы осуществлять один большой отбор без возвращения - выполняется отбор в несколько мелких этапов.

Статистические службы часто использует региональные, демографические или социально-экономические данные для стратификации. Стратификация, т. е. рабиение целевой совокупности на страты, требует, чтобы в основе выборки имелась вся необходимая вспомогательная информация.

Например, для стратификации по отраслям промышленности информация о каждой компании должна быть в наличии в основе выборки. Регистр предприятий может содержать информацию о стандартной отраслевой классификации и размере компании. Эти данные являются вспомогательными переменными, используемыми для стратификации. Информация о каждом человеке в Базе личных данных муниципалитета (GBA) содержит много персональных данных, таких как пол, дата рождения, семейное положение и адрес, которые могут быть использованы для стратификации населения.

Стратифицированная случайная выборка может быть использована по целому ряду причин.

Во-первых, стратификация является распространенным способом повышения точности оценок (т. е. уменьшения дисперсии), в частности при оценке характеристик всей совокупности. Некоторые показатели, например выручка компании, могут иметь такую большую дисперсию в совокупности, что для получения надежных выводов необходимы очень большие выборки. Если удается сформировать группы, в пределах которых целевая переменная изменяется мало, то в этом случае, стратифицированная выборка позволит получить более точные результаты, чем простая случайная выборка (при одинаковом размере выборки).Точность повышается в связи с тем, что дисперсия внутри слоев меньше, чем дисперсия совокупности в целом.

Во-вторых, интерес зачастую представляет не только совокупность в целом, но и конкретные подсовокупности, а также проведение сравнений между ними. При простой случайной выборке сколько элементов в конечном итоге содержится в слоях - это вопрос случая. Небольшие подсовокупности, в частности, будут тогда недостаточно представлены в выборке. Стратификация позволяет добиться достаточной представительности всех интересующих подсовокупностей в выборке и это обеспечивает надежность получаемых результатов.

В-третьих, при стратификации можно использовать различные методы сбора данных для разных слоев. Например, может быть принято решение о том, что в обследованиях предприятий малый бизнес обследуется с помощью краткого бумажного вопросника, а большие компании - с помощью подробного телефонного или личного опроса. Методы отбора и оценки могут также отличаться для каждого слоя.

В-четвертых, по административным причинам основа выборки часто подразделяется на "естественно образованные" части, которые могут находится в географически разных местах. В этом случай отдельный отбора может быть более экономичным.

Детализированное описание

Совокупность делится на H страт. Отдельная страта обозначается h, h = 1,…, H и состоит из N_h элементов. Страты не должны пересекаться. Другими словами, каждый элемент должен принадлежать только одной страте. Сумма страт представляет собой совокупность:

$$\sum_{h=1}^{H} N_h = N$$

где N - объем генеральной совокупности.

Предположим, что объем каждой страты N_h известен. Значения целевого показателя k-ого элемента в страте обозначается Y_{hk}, где h, h = 1,…, H и k = 1, …,N_h.

Объем выборки в страте h обозначается n_h, при этом по определению $\sum_h n_h = n$. y_{hk} - значение, полученнное по выбороке в страте h, где h = 1,…, H и k = 1, …,n_h.

Для целевого показателя определяются следующие параметры в страте: суммарное количество Y_h, среднее значение $\bar{Y}_h$, дисперсия σ^2_{yh}, скорректированная дисперсия S^2_{yh} , коэффициент вариации CV_{yh} .

Параметры рассчитываются по следующим формулам:

$$Y_h = \sum_{k=1}^{N_h} Y_{hk}$$

$$\bar{Y}_h = \frac{1}{N_h} Y_h$$

$$\sigma^2_{yh} = \frac{1}{N_h} \sum_{k=1}^{N_h} (Y_{hk} - \bar{Y}_h)^2 \quad \textbf{3.1.}$$

$$S^2_{yh} = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (Y_{hk} - \bar{Y}_h)^2$$

$$CV_{yh} = \frac{S_{yh}}{\bar{Y}_h}$$

Все вышеуказанные параметры относятся к параметрам совокупности в страте h.

План расслоенной выборки предполагает отбор выборки из каждой страты.Значения по выборке рассчитываются в каждой страте: среднее

значение в страте h $\bar{y}_h$,выборочная дисперсия s^2_{yh} , коэффициент вариации cv_{yh} . Параметры по выборке рассчитываются следующим образом:

$$\bar{y}_h = \frac{1}{n_h}\sum_{k=1}^{n_h} y_{hk}$$

$$s^2_{yh} = \frac{1}{n_h - 1}\sum_{k=1}^{n_h}(y_{hk} - \bar{y}_h)^2 \quad \textbf{3.2.}$$

$$cv_{yh} = \frac{s_{yh}}{\bar{y}_h}$$

Взаимосвязь между параметрами совокупности и параметрами в страте

Каждый элемент в совокупности принадлежит только одной страте. Это свойство страты позволяет установить взаимосвязь между параметрами совокупности и параметрами страты. Например, верны следующие выражения:

$$Y = \sum_{h=1}^{H}\sum_{k=1}^{N_h} Y_{hk} = \sum_{h=1}^{H} Y_h$$

$$Y = \frac{1}{N}Y = \frac{1}{N}\sum_{h=1}^{H} Y_h = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{Y}_h = \sum_{h=1}^{H}(\frac{N_h}{N})\bar{Y}_h \quad \textbf{3.3.}$$

Другими словами суммарное значение совокупности может рассматриваться как сумма суммарных значений страт, а среднее значение совокупности – взвешенные средние значения в стратах. Фактор взвешивания среднего значения в страте тесно связан с относительным размером рассматриваемой страты.

Выражение для скорректированной дисперсии для совокупности S^2_{yh} может быть переписано с учетом скорректированных дисперсий в страте, т.е.:

$$S^2_y = \sum_{h=1}^{H}(\frac{N_h - 1}{N - 1})S_{yh}{}^2 + \sum_{h=1}^{H}(\frac{N_h}{N-1})(\bar{Y}_h - \bar{Y})^2 \quad \textbf{3.4.}$$

Первый член справа от знака равенства обозначает дисперсию внутри слоя, так она состоит из дисперсий отдельных дисперсий в стратах. Второй член справа от знака равенства представляет собой дисперсию между слоями и представляет собой отклонение средних значений в стратах от значения показателя совокупности.

С учетом вышеуказанного выражения коэффициента вариации равен:

$$CV_y^2 = \sum_{h=1}^{H}(\frac{N_h - 1}{N-1})(\frac{\overline{Y}_h}{\overline{Y}})CV_{yh}{}^2 + \sum_{h=1}^{H}(\frac{N_h}{N-1})((\frac{\overline{Y}_h}{\overline{Y}}) - 1)^2 \quad \textbf{3.5}.$$

Оценка среднего значения и суммарного значения совокупности

Простой случайный отбор производится отдельно для каждой страты в расслоенной выборке. Т.е. из совокупности N_h из страты h производится простой случайный отбор для формирования выборки объема n_h . Оценки среднего значения совокупности и суммарного значения в страте:

$\hat{\overline{Y}}_h = \overline{y}_h$

$\hat{Y}_h = N_h \overline{y}_h$ **3.6**.

Эти две формулы используются для оценки суммарного и среднего значений совокупности.

Суммарное значение Y соответствует сумме суммарных значений страт Y_h. Поэтому логично оценивать суммарное значение с помощью оценок суммарных значений страт. Оценка, которая предусматривает использование стратификации, называется оценкой стратификации. Оценка стратификации суммарного значения обозначается $\hat{Y}_{ST}$ и определяется по формуле:

$$\hat{Y}_{ST} = \sum_{h=1}^{H}\hat{Y}_h = \sum_{h=1}^{H} N_h \overline{y}_h \quad \textbf{3.7}.$$

По аналогии оценка среднего значения по расслоенной (стратифицированной) выборке обозначается как $\hat{\overline{Y}}_{ST}$ и определяется:

$$\hat{\overline{Y}}_{ST} = \frac{1}{N}\hat{Y}_{ST} = \sum_{h=1}^{H}(\frac{N_h}{N})\overline{y}_h \quad \textbf{3.8}.$$

ПСВ без ВО осуществляется из каждой страты для формирования выборки объема n_h. Этот тип случайной выборки обладает несколькими известными свойствами и, соответственно, оценки среднего и суммарных значений в страте и можно считать несмещенными. Принимая во внимание несмещенность оценок суммарного значения, можно вычислить математическое ожидание суммарной оценки стратифицированной выборки:

$$E(\hat{Y}_{ST}) = \sum_{h=1}^{H} E(\hat{Y}_h) = \sum_{h=1}^{H} Y_h = Y$$

Таким образом, стратифицированная оценка является несмещенной оценкой суммарного значения показателя. В связи с этим матожидание оценки среднего значения:

$$E(\hat{\bar{Y}}_{ST}) = \frac{1}{N} E(\hat{Y}_{ST}) = \frac{1}{N} Y = \bar{Y}.$$

Формула показывает, что $\hat{\bar{Y}}_{ST}$ является несмещенной оценкой среднего значения совокупности.

Доказательство несмещенности оценок, приведенных выше, подтверждают то, что ясно интуитивно. Если оценки суммарных значений страт являются несмещенными, то тогда взвешенная сумма оценок суммарных значений страт также не является несмещенной. Похожее рассуждение относится к оценкам средних значений в слоях и взвешенной сумме оценок средних значений в слоях.

Дисперсия

Дисперсия, например суммарного значения в страте, может быть вычислена (см. 2.12.):

$$\text{var}(\hat{Y}_{yh}) = N_h (\frac{1 - f_h}{f_h}) S_{yh}^2 \quad \textbf{3.9}.$$

В этой формуле f_h выборочная доля в страте h, т.е. $f_h = n_h / N_h$.

ПСВ без ВО выполняется независимо в различных старатах раслоенной выборки. Следовательно, дисперсия стратифицированной оценки суммарного значения равно сумме дисперсий отдельных оценок суммарных значений в стратах. Т.е:

$$\text{var}(\hat{Y}_{ST}) = \sum_{h=1}^{H} \text{var}(\hat{Y}_h) = N \sum_{h=1}^{H} (\frac{1 - f_h}{f_h})(\frac{N_h}{N}) S_{yh}^2 \quad \textbf{3.10.}$$

Используя вышеприведенную формулу можно вычислить дисперсию стратифицированной оценки среднего значения совокупности:

$$\text{var}(\hat{\bar{Y}}_{ST}) = \frac{1}{N^2} \text{var}(\hat{Y}_{ST}) = \frac{1}{N} \sum_{h=1}^{H} (\frac{1 - f_h}{f_h})(\frac{N_h}{N}) S_{yh}^2 \quad \textbf{3.11.}$$

Дальнейшее изучение формул показывает, что обе дисперсии малы, если только S_{yh}^2 дисперсии в страте также малы. Маленький размер дисперсии в страте означает что целевые показатели мало варьируют в пределах страты; другими словами страты внутренне однородны, т.е. гомогенны.

Дисперсия оценки также зависит от используемой схемы размещения, так как дисперсия внутри страты зависит от размера случайной выборки.

Дисперсия стратифицированной оценки суммарного значения – это линейная взвешенная комбинация дисперсий в стратах. Оценка этой дисперсии может таким образом быть получена путем оценки индивидуальных дисперсий в стратах. Выборочная дисперсия ПСВ без ВО представляет собой несмещенную оценку дисперсии совокупности и может быть рассчитана:

$$\hat{\text{var}} = (\hat{Y}_{ST}) = N\sum_{h=1}^{H}(\frac{1-f_h}{f_h})(\frac{N_h}{N})\hat{S}_{yh}^2 = N\sum_{h=1}^{H}(\frac{1-f_h}{f_h})(\frac{N_h}{N})s_{yh}^2 \quad \textbf{3.12.}$$

По аналогии дисперсия оценки среднего значения может быть оценена следующим образом:

$$\hat{\text{var}} = (\hat{\bar{Y}}_{ST}) = \frac{1}{N^2}\hat{\text{var}}(\hat{Y}_{ST}) = \frac{1}{N}\sum_{h=1}^{H}(\frac{1-f_h}{f_h})(\frac{N_h}{N})s_{yh}^2 \quad \textbf{3.13.}$$

Так как выборочная дисперсия является несмещенной оценкой скорректированной дисперсии совокупности при ПСВ без ВО – оценки в **3.12.** и **3.13.** являются несмещенными.

Коэффициент вариации

Коэффициенты вариации стратифицированных оценок суммарного и средних значений определяются следующим образом:

$$CV(\hat{Y}_{ST}) = \frac{\sqrt{\text{var}(\hat{Y}_{ST})}}{Y}$$

$$CV(\hat{\bar{Y}}_{ST}) = \frac{\sqrt{\text{var}(\hat{\bar{Y}}_{ST})}}{\bar{Y}}$$

Можно также вывести, что коэффициенты вариации:

$$CV(\hat{\bar{Y}}_{ST}) = \frac{\sqrt{\text{var}(\hat{\bar{Y}}_{ST})}}{\bar{Y}} = \frac{(\frac{1}{N})\sqrt{\text{var}(\hat{Y}_{ST})}}{(\frac{1}{N})Y} = CV(\hat{Y}_{ST})$$

Следовательно, расслоенные оценки суммарного значения и среднего значения одинаково точны. Можно также выразить коэффициент вариации по расслоенной выборке через коэффициенты вариации в страте:

$$CV^2(\hat{Y}_{ST}) = N\sum_{h=1}^{N}(\frac{1-f_h}{f_h})(\frac{N_h}{N})(\frac{\bar{Y}_h}{\bar{Y}})^2 CV_{yh}^2$$

Оценка плана выборки

Основная идея использования стратифицированных выборок состоит в том, что стратификация позволяет получить более точную оценку, чем простая случайная выборка. Если наблюдения внутри страт в целом более однородны, чем в совокупности, то уменьшение дисперсии в стратах приведет к уменьшению дисперсии оценки.

Выборочный план может быть оценен путем сравнения дисперсии оценки полученной по выборке, выполненный по определенному плану, с соответствующей дисперсией оценки ПВС без ВО. Отношение двух дисперсий называется эффектом плана и обозначается DEFF. Эффект плана для стратифицированной выборки суммарного значения совокупности и среднего значения определяется как (при условии, что $\text{var}(\hat{Y}_{EAZT}) \neq 0$):

$$DEFF(\hat{Y}_{ST}) = \frac{\text{var}(\hat{Y}_{ST})}{\text{var}(\hat{Y}_{ПСВбезВО})}$$

$$DEFF(\hat{\bar{Y}}_{ST}) = \frac{\text{var}(\hat{\bar{Y}}_{ST})}{\text{var}(\hat{\bar{Y}}_{ПСВбезВО})} \quad \textbf{3.14.}$$

Если DEFF равно единице, то это означает что обе оценки равнозначны: точность стратифицированной оценки такая же, как у ПСВ без ВО. Если DEFF меньше единицы, то стратифицированная случайная выборка более эффективна, чем ПСВ без ВО того же объема. При DEFF больше единицы - она менее эффективна, чем ПСВ без ВО.

Эффект плана стратифицированной выборки для среднего значения совокупности может быть также рассчитан следующим образом:

$$DEFF(\hat{\bar{Y}}_{ST}) = \frac{\text{var}(\hat{\bar{Y}}_{ST})}{\text{var}(\hat{\bar{Y}}_{ПСВбезВО})} = \frac{(1/N^2)\,\text{var}(\hat{Y}_{ST})}{(1/N^2)\,\text{var}(\hat{Y}_{ПСВбезВО})} = DEFF(\hat{Y}_{ST})$$

Другими словами эффект плана стратифицированной выборки для среднего и суммарного значения равны. Это неудивительно, так как эффект плана обычно не более чем отношение квадратов коэффициентов вариации стратифицированной оценки и оценки ПСВ без ВО.

Компактное выражение эффекта плана стратифицированной выборки суммарного значения по совокупности может быть получено, если существует простая взаимосвязь между $\text{var}(\hat{Y}_{ST})$ и $\text{var}(\hat{Y}_{SRSWOR})$.К сожалению, такая взаимосвязь отсутсвует в общем случае. Однако, такая взаимосвязь существует при соблюдении (не обязательно строгих) условий.

Предположим, что выборочная доля f_h может быть константой, что приводит к упрощению выражения **3.10**:

$$\text{var}(\hat{Y}_{ST}) = N\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\frac{N_h}{N}S_{yh}^2 = N\left(\frac{1-f}{f}\right)\sum_{h=1}^{H}(\frac{N_h}{N})S_{yh}^2 \quad \textbf{3.15.}$$

Дисперсия оценки суммарного значения ПСВ без ВО того же самого размера зависит от скорректированной дисперсии суммарного значения S_y^2 (2.12).

Предположим также, что выполнено следующее условие:

$$N_h \approx (N_h - 1) \wedge N \approx (N-1) \quad \textbf{3.16.}$$

Основываясь на этом предположении выражение $\text{var}(\hat{Y}_{SRSWOR})$ может быть переписано:

$$\text{var}(\hat{Y}_{SRSWOR}) \approx \text{var}(\hat{Y}_{ST}) + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H} N_h(\bar{Y}_h - \bar{Y})^2 \quad \textbf{3.17.}$$

Резюмируя, дисперсия оценки суммарного значения ПСВ без ВО может быть определена как сумма дисперсий стратифицированных оценок суммарных значений и слагаемое, которое прямо пропорционально значению дисперсии между стратами. Этот результат позволяет аппроксимировать эффект плана стратифицированной оценки:

$$DEFF(\hat{Y}_{ST}) \approx \frac{\text{var}(\hat{Y}_{ST})}{\text{var}(\hat{Y}_{ST}) + (\frac{1-f}{f})\sum_{h=1}^{H} N_h(\bar{Y}_h - \bar{Y})^2} \quad \textbf{3.18.}$$

Согласно этой аппроксимации эффект плана стратифицированных оценок всегда меньше или равен 1.Кроме того, в данном случае верно, что большая межстратовая дисперсия может дать более точные оценки по стратификации.

Именно поэтому целесообразно при планировании стратификации максимизировать разброс средних значений страт $\bar{Y}_h$ как можно больше.

Таким образом, в случае применения стратифицированной выборки с постоянной выборочной долей в страте и достаточно большим объемом совокупности и страт, оценки суммарного и среднего значения совокупности, по крайней мере, так же точны, как соответствующие оценки ПСВ без ВО.

Размещение

В приведенных выше разделах речь шла о стратифицированной выборке с определенным размером N_h и n_h. В подразделах ниже будет кратко рассмотрен вопрос о том, сколько наблюдений должно присутствовать в каждом слое.

Если мы хотим оценить среднее значение в слое или другие параметры страты с определенным заранее уровнем точности, то нам необходимо рассчитать размеры объемов выборок в слоях с использованием формул, приведенных ранее.

Вместе эти объемы образуют общую выборку. Часто, однако, на практике общий объем выборки определен, и вопрос заключается в том, как распределить эти n элементов по слоям.

В этом разделе рассматриваются три различных метода размещения.

Пропорциональное размещение

Пропорциональное размещение основано на идее репрезентативной выборки. Стратифицированная выборка отбирается таким образом, чтобы она отражала совокупность с точки зрения стратификацированной переменной (переменных).

Размер выборки в каждом слое берется пропорционально размеру слоя:

$$n_h = \frac{N_h}{N} \times n \qquad \textbf{3.19.}$$

При необходимости нецелые значения n_h округляются, чтобы получить целочисленный размер выборки. Используя (неокругленный) объем выборки в слое, выборочная доля каждого слоя может быть рассчитана:

$$f_h = n_h / N_h = n / N$$

Другими словами, выборочная доля в каждом слое одинаковая. Условие, лежащее в основе 3.15, соблюдается автоматически. Кроме того, 3.17 и 3.18 при соблюдении дополнительного условия 3.16.

Наконец, вероятность включения элемента k в слой h можно вычислить следующим образом:

$$\pi_{kh} = n_h / N_h = \frac{N_h}{N}\frac{n}{N_h} = n / N$$

Все элементы совокупности, независимо от страты, имеют одинаковую вероятность быть отобранным в выборку и, следовательно, имеют

одинаковый вес при вычислении оценок. В данном случае мы имеем дело с *самовзвешенной выборкой.*

При ПСВ без ВО каждый элемент совокупности имеет вероятность n/N быть отобранным в выборку. Стратифицированная выборка позволяет исключить некоторые «экстремальные» выборки, которые не были бы исключены при ПСВ без ВО (например, выборка только с высокими или низкими доходами).

Оптимальное размещение

Когда различные скорректированные дисперсии в слоях равны,то пропорциональное размещение является лучшим методом для повышения точности. В выборке, где элементы различаются по размеру (например, предприятия, школы, муниципалитеты), более крупные элементы, как правило, демонстрируют большую вариативность целевой переменной, чем мелкие единицы.

Примером может служить компании, занимающиеся международной торговлей; более крупные компании будут демонстрировать большую вариативность экспорта и импорта, чем небольшие компании. В этом случае целесообразно иметь более высокую долю крупных компаний в выборке.

Если скорректированные дисперсии в стратах сильно различаются, то оптимальное размещение выборки приводит к минимальной дисперсии оценок суммарного и среднего значения совокупности. Этот метод предусматривает расчет объем выборки в слое h по следующей формуле:

$$n_h = \frac{N_h S_{hy}}{\sum_{h=1}^{H} N_h S_{hy}} \times n \qquad \textbf{3.20.}$$

где n_h округляется до целого значения.

Оптимальное размещение действительно приводит к равным вероятностям включения всех элементов совокупности; вероятность включения в выборку пропорциональна S_{yh} и, следовательно, варьируется между стратами.

Количество отобранных элементов в страте, больше, в тех случаях, когда страта занимает большую долю в совокупности и когда страта относительно неоднородна, т. е. имеет относительно большой S_{yh}. При применении формулы 3.20 рассчитанный размер выборки может быть больше, чем фактический размер слоя, и в этом случае включается все элементы слоя.

Определение оптимального распределения требует знания (соотношения) величин скорректированных дисперсий в стратах. Эта

информация редко известна на практике, но эти дисперсии в некоторых случаях могут быть аппроксимированы на основе данных предыдущих обследований.

Если можно ожидать, что различия в стратах не будут слишком сильными, то можно предположить, что $S_{yh} = S_y$, и вышеуказанная формула будет сводится к пропорциональному размещению.

Стоимость

Помимо точности оценки, затраты, также важны при проведении выборочных обследований. Следует при размещении учитывать различия в затратах на одно наблюдение между слоями.

Например, из-за различных методов сбора данных, используемых при отборе в этих слоях, требуется меньшее количество наблюдений в относительно дорогих слоях. Предположим, что для выполнения полевых работ выделен определенный бюджет, т.е. есть максимальные общие затраты равны C. Тогда простая функция затрат будет рассчитываться по формуле:

$$C = c_0 + \sum_{h=1}^{H} n_h c_h \quad \textbf{3.21.}$$

где c_0 представляет собой фиксированные расходы и c_h (переменные) затраты на наблюдение в страте h.

Теперь необходимо распределить случайную выборку между слоями, так чтобы, минимизировать дисперсию оценки с учетом общих затрат C.

(3.20) вытесняет выражение, которое предполагает, что общий размер выборки должен быть равен n. Это можно доказать, учитывая, что размер выборки в страте h равен :

$$n_h = \frac{\frac{N_h S_{hy}}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h S_{hy}}{\sqrt{c_h}}} \times n$$

при этом минимизируется дисперсия оценки суммарного значения, полученной по расслоенной выборке.

Доказательство приведено в книге Кокрена (1977, Глава 5). Таким образом, отбирается больше элементов из слоя, если он составляет большую долю в совокупности, дисперсия внутри слоя большая, а наблюдения в слое стоят недорого.

Оптимальное распределение на самом деле является частным случаем, когда затраты в каждом слое одинаковые. Оптимальное размещение также называют размещение по Нейману.

Доли

Если целевая переменная Y является индикаторной переменной, которая может принимать значения 0 и 1, то среднее значение равно дроби. Таким образом, доля элементов с заданным свойством (доля "успехов") в выборке в слое h, таким образом, равна $p_h = \bar{y}_h$. Доля "успехов" в совокупности может быть оценена с помощью стратификационной оценки, т. е.:

$$\hat{P}_{st} = \sum_{h=1}^{H} (\frac{N_h}{N}) p_h$$

Дисперсия этой стратификационной оценки рассчитывается по формуле 3.10. В практических ситуациях дисперсия может быть оценена по формуле:

$$\hat{\text{var}}(\hat{P}_{st}) = \sum_{h=1}^{H} (1 - f_h)(\frac{N_h}{N})^2 \left[\frac{p_h(1-p_h)}{n_h - 1} \right]$$,так как дисперсия в страте может быть рассчитана:

$$s_h^2 = p_h(1-p_h)(\frac{n_h}{n_h - 1})$$

Показатель качества

Важным критерием качества стратифицированной выборки является уменьшение дисперсии по сравнению с ПСВ без ВО. Это находит свое отражение в эффекте плана DEFF. Чем значительнее уменьшение дисперсии, тем меньше значение DEFF. Уменьшение дисперсии особенно велико, когда средние значения целевой переменной Y страт сильно варьируются.

IV.Кластерная выборка, двухэтапная выборка и систематический отбор

Краткое описание

Принцип трех видов планов выборки, рассмотренных в этой главе, заключается в следующем: распределение целевой совокупности на кластеры. Распределение должно быть таким, чтобы кластеры не пересекались, при этом все вместе они должны охватывать всю совокупность.

Кластеры выбираются случайным образом, а затем каждый выбранный кластер наблюдается полностью. Таким образом, кластерная выборка может быть интерпретирована как случайная выборка групп. В этом случае также есть возможность отбора кластеров с возвращением или без возвращения.

Первый этап двухэтапной выборки, как и в случае кластерной выборки, предусматривает отбор кластеров случайным образом. Затем на втором этапе двухэтапного отбора извлекаются случайным образом элементы из каждого выбранного кластера.

Систематический отбор также рассматривается в этой главе, поскольку он может рассматривается как своего рода кластерная выборка.

Применимость

В предыдущей главе стратификация была представлена как один из методов, при котором совокупность сначала делят на подсовокупности (страты), а затем отбирают выборку из каждого отдельного слоя.

Стратифицированная выборка обычно увеличивает точность оценок. Кластерная выборка, на первый взгляд, очень напоминает стратифицированную выборку, при этом ее свойства совершенно иные. Например, кластерная выборка обычно обладает более низкой точностью, чем простая случайная выборка. Таким образом, кластерная выборка применяется только в тех случаях, когда этого требует практическая ситуация или когда потеря точности компенсируется существенным сокращением затрат на сбор данных.

Потеря точности возникает из-за того, что, в отличие от стратифицированной выборки, в случае выборки отбираются не все подсовокупности.

Тем не менее, возможно иногда достичь приемлимую точность на втором этапе путем извлечения выборок из кластеров, которые были отобраны на первом этапе. Поскольку кластеры больше не наблюдаются полностью, можно рассмотреть возможность отбора большего количества кластеров на первом этапе для того, чтобы получить более четкую картину совокупности в целом. Двухступенчатая выборка особенно для кластеров, которые являются однородными с точки зрения целевой переменной.

Примером двухэтапной выборки является кластерная выборка всех домашних хозяйств в регионах или округах. Выборка домашних хозяйств отбирается на втором этапе из каждого района, выбранного на первом этапе. Преимуществом региональной кластерной выборки такого рода является снижение общих расходов на поездки интервьюеров по сравнению со случайной выборкой. Расходы на поездки сокращаются, поскольку интервьюер теперь может опросить относительно большое число людей, живущих близко друг к другу в кластерах.

Все рассмотренные до сих пор методы отбора предполагали наличие основы выборки удовлетворительного качества для целевой совокупности.

Очевидно, что так происходит не всегда. При этом основа выборки может быть доступна для подсовокупностей целевой совокупности.

Предположим, что мы хотим провести опрос учащихся средних школ. У организаторов обследования нет в наличии полного списка учащихся средней школы, но они без особого труда могут найти список средних школ.

Очевидно, что каждая школа будет обладать информацией о своих учениках.

Другим примером "естественных" групп являются муниципалитеты, домохозяйства, предприятия и дома престарелых. Эти группы образуют кластеры.

Подробное описание

Далее описываются три метода отбора выборок.

Во-первых, рассматривается кластерная выборка. В ней отобранные в выборку кластеры наблюдаются полностью.

Во-вторых, представлена двухступенчатая выборка, которая предполагает отбор подвыборки из кластеров, отобранных на первом этапе.

Наконец, рассмотрена систематическая выборка, которую можно считать частным случаем кластерной выборки.

V Кластерная выборка

По отношению к целевой переменной Y можно выделить пять параметров для совокупности.

Как и в случае стратифицированной выборки, в кластерной выборке совокупность делится на подсовокупности. Подсовокупности называются кластерами или первичными выборочными единицами (ПВЕ). Неявное предположение состоит в том, что вся совокупность полностью охвачена кластерами и что отдельные кластеры не содержат общих элементов.

Кластеры в плане выборки первоначально интерпретируются как единицы обследования (поэтому кластеры не могут пересекаться в совокупности).

Отдельный кластер обозначается d, d,=1…N, M_d – число элементов в кластере d. Элементы первичных выборочных единиц называются вторичными единицами выборки (ВЕВ). Объем кластера M_d обычно известен.

Общее количество элементов совокупности обозначается M и вычисляется по следующей формуле:

$$\sum_{d=1}^{N} M_d = M$$

Средний размер кластера, обозначаемый $\overline{M}$:

$\overline{M} = M / N$.

Значение целевой переменной элемента k в кластере d обозначается Y_{dk}, где d =1,.. ,N и k =1, .. ,M_d. Для целевой переменной параметры в кластере вычисляются по формулам:

$$Y_d = \sum_{k=1}^{M_d} Y_{dk} \quad \textbf{4.1.}$$

$$\bar{Y}_d = \frac{1}{M_d} Y_d$$

$$\sigma_{yd}^2 = \frac{1}{M_d} \sum_{k=1}^{M_d} (Y_{dk} - \bar{Y}_d)^2$$

$$S_{yd}^2 = \frac{1}{M_d - 1} \sum_{k=1}^{M_d} (Y_{dk} - \bar{Y}_d)^2$$

К вышеперечисленным четырем параметрам можно добавить среднее значение по кластерам $\bar{Y}_{CT}$ и скорректированную дисперсию по кластерам S_{yCT}^2:

$$\bar{Y}_{CT} = \frac{1}{N} Y = \frac{1}{N} \sum_{d=1}^{N} Y_d$$

$$S_{yCT}^2 = \frac{1}{N-1} \sum_{d=1}^{N} (Y_d - \bar{Y}_{CT})^2 \quad \textbf{4.2.}$$

Простая случайная выборка n кластеров выбирается из совокупности N кластеров. Каждый из выбранных кластеров наблюдается в полном объеме, так что каждый кластер содержит m_d наблюдений, $m_d = M_d$.

Таким образом, можно интерпретировать случайную кластерную выборку как простую случайную выборку n из N первичных выборочных единиц со значениями показателей кластеров в качестве наблюдений. Значения показателей отобранных в n кластерах обозначаются строчными буквами $y_1,...,y_n$.

Выборное среднее значение по кластерам $\bar{y}_{CT}$ и дисперсия s_{yCT}^2:

$$\bar{y}_{CT} = \frac{1}{n} \sum_{d=1}^{n} y_d \quad \textbf{4.3.}$$

$$s_{yCT}^2 = \frac{1}{n-1} \sum_{l=1}^{n} (y_d - \bar{y}_{CT})^2 .$$

Взаимосвязь между параметрами совокупности и параметрами кластеров

Взаимосвязь между параметрами совокупности и кластера определяется:

$$Y = \sum_{d=1}^{N} \sum_{k=1}^{M_d} Y_{dk} = \sum_{d=1}^{N} Y_d$$

$$\bar{Y} = \frac{Y}{N} = \frac{1}{M}\sum_{d=1}^{N} Y_d = \frac{1}{M}\sum_{d=1}^{N} \bar{Y}_d M_d = \sum_{d=1}^{N} \bar{Y}_d \left(\frac{M_d}{M}\right)$$

Приведенные выше выражения показывают, что суммарное значение совокупности равно сумме всех суммарных значений кластеров, а среднее значение совокупности можно интерпретировать как взвешенную сумму средних значений показателей кластеров.

Дисперсия совокупности может быть определена на основе дисперсий подсовокупностей, как и в случае стратификации. Верна следующая формула:

$$\sigma_y^2 = \sum_{d=1}^{N} \left(\frac{M_d}{M}\right)\sigma_{yd}^2 + \sum_{d=1}^{N} \frac{M_d}{M}(\bar{Y}_d - \bar{Y})^2 \quad \textbf{4.4}$$

Теперь можно вывести зависимость между скорректированной дисперсией совокупности S_y^2 и скорректированными дисперсиями кластеров S_{yd}^2:

$$S_y^2 = N\left(\frac{\bar{M}-1}{M-1}\right)S_{\text{int}\,ra}^2 + \bar{M}\left(\frac{N-1}{M-1}\right)S_{\text{int}\,er}^2 \quad \textbf{4.5.}, \text{ где}$$

$$S_{\text{int}\,ra}^2 = \frac{1}{N}\sum_{d=1}^{N}\left(\frac{M_d - 1}{\bar{M}-1}\right)S_{yd}^2$$

$$S_{\text{int}\,er}^2 = \frac{1}{N-1}\sum_{d=1}^{N}\left(\frac{M_d}{\bar{M}}\right)(\bar{Y}_d - \bar{Y})^2$$

Первый член (в правой части от знака «=») прямо пропорционален $S_{\text{int}\,ra}^2$, который называется внутрикластерной дисперсией. Внутрикластерная дисперсия состоит из N скорректированных дисперсий кластера. Второй член (в правой части от знака «=») прямо пропорциональна $S_{\text{int}\,er}^2$ - межкластерной дисперсии. Межкластерная дисперсия зависит от разницы между средним значением в кластере и средним значением совокупности кластеров.

Оценки среднего значения совокупности и суммарного значения совокупности

В сущности, кластерная выборка - это простая случайная выборка, с кластерами в качестве единиц и содержимым кластеров в качестве объектов наблюдения. Кроме того, если предположить, что кластеры отбираются без возвращения, то выборка кластеров является просто результатом отбора без возвращения, примененного к значениям кластера.

Оценки параметров совокупности, основанные на этой специфической интерпретации кластерной выборки как отбора содержимого кластеров, называются кластерными оценками.

Основываясь на методе отбора без возвращения кластеров, оценка среднего значения по кластерной выборке рассчитывается следующим образом:

$$\hat{\bar{Y}} = \bar{y}_{CT} = \frac{1}{n}\sum_{d=1}^{n} y_d$$

Аналогично, оценка скорректированной дисперсии значений кластеров выглядит следующим образом:

$$\hat{S}^2_{yCT} = s_{yCT}{}^2 = \frac{1}{n-1}\sum_{d=1}^{n}(y_d - \bar{y}_{CT})^2$$

Существует зависимость между суммарным значением совокупности Y и средним значением в целом по кластерам $\bar{Y}_{CT}$, которая может быть получена из формул, приведенных выше. Существует также аналогичная связь между средним значением совокупности и средним значением по кластерам. Таким образом:

$$Y = N\bar{Y}_{CT} \quad \wedge \quad Y = \frac{Y}{M} = \frac{1}{M}\bar{Y}_{CT} \quad \textbf{4.6.}$$

Основываясь на этих взаимосвязях и оценке среднего значения по кластерам получаем кластерную оценку суммарного значения совокупности Y, обозначенную как $\hat{Y}_{CL}$ и кластерную оценку среднего значения совокупности $\hat{\bar{Y}}_{CL}$.

$$\hat{Y}_{CL} = N\hat{\bar{Y}}_{CT} = N\bar{y}_{CT} = (\frac{1}{f})\sum_{d=1}^{n} y_d \quad \textbf{4.7.}$$

$$\hat{\bar{Y}}_{CL} = \frac{\hat{Y}_{CL}}{M} = \frac{1}{M}\hat{\bar{Y}}_{CT} = \frac{1}{M}\bar{y}_{CT} = (\frac{1}{f})(\frac{1}{M})\sum_{d=1}^{n} y_d$$

В вышеуказанных формулах f = n/N представляет собой долю кластера в случайной выборке. Следует отметить, что выборочное среднее $\bar{y}_{CT}$ в кластере закономерно возникает в этих оценках.

Несмещенность этих оценок

Поскольку процедура кластерной выборки является ПСВ без ВО значений кластера, оценка среднего значения кластеров, как известно, является несмещенной. На основании этого можно также сделать вывод о том, что выборочная дисперсия по кластерам является несмещенной оценкой скорректированной дисперсии значений по кластерам.

По определению кластерные оценки суммарного и среднего значение являются масштабированными версиями оценок значений кластеров. Этот факт вкупе с несмещенностью оценки среднего значения по кластерам и

зависимых величин 4.6,означает,что кластерные оценки суммарного и среднего значений 4.7 являются несмещенными.

Расчет вариации

Было продемонстрировано, что несмещенность оценки среднего значения кластера является основой несмещенности двух кластерных оценок. Аналогично, вычисление дисперсии оценки среднего значения кластеров является основой для расчета дисперсии кластерных оценок.

Дисперсия оценки среднего значения кластеров может быть рассчитана следующим образом:

$$\mathrm{var}(\hat{\bar{Y}}_{CT}) = \mathrm{var}(\bar{y}_{CT}) = \frac{1}{N}\left(\frac{1-f}{f}\right)S^2_{yCT} \quad \mathbf{4.8}$$

На основе этого результата можно сделать вывод о том, что дисперсия кластерной оценки суммарного значения равна:

$$\mathrm{var}(\hat{Y}_{CL}) = N^2\ \mathrm{var}(\hat{\bar{Y}}_{CT}) = N\left(\frac{1-f}{f}\right)S^2_{yCT} \quad \mathbf{4.9}$$

и дисперсия кластерной оценки среднего значения:

$$\mathrm{var}(\hat{\bar{Y}}_{CL}) = \mathrm{var}\left(\frac{\hat{Y}_{CL}}{M}\right) = \frac{1}{M^2}\mathrm{var}(\hat{Y}_{CL}) = \frac{1}{M}\times\frac{1}{\overline{M}}\left(\frac{1-f}{f}\right)S^2_{yCT} \quad \mathbf{4.10}$$

Оценки для показателей 4.8 – 4.10 могут быть получены путем замены S^2_{yCT} в соответствующих формулах на s^2_{yCT}. Так как выборочная дисперсия значений кластеров является несмещенной оценкой скорректированной дисперсии значений кластеров, оценки дисперсий (кластерных оценок), полученные таким образом, являются несмещенными.

Оценка плана (дизайна) выборки

Для оценки кластерной выборки необходимо рассчитать эффект плана при использовании кластерной выборки для получения оценок суммарного и среднего значений. Это предусматривает сравнение дисперсии кластерной оценки с дисперсией оценки ПСВ без ВО того же размера.

К сожалению, размер кластерной выборки заранее не известен. При этом, в среднем кластерная выборка состоит из $(n\times\overline{M})$ наблюдений. Сравнение с оценкой по ПСВ без ВО также базируется на выборке $(n\times\overline{M})$ из M элементов.

Эффект плана кластерных оценок суммарного и среднего значения определяется по формулам (при этом $\mathrm{var}(\hat{Y}_{SRSWOR}) \neq 0$):

$$DEFF(\hat{Y}_{CL}) = \frac{\text{var}(\hat{Y}_{CL})}{\text{var}(\hat{Y}_{SRSWOR})}$$

$$DEFF(\hat{\bar{Y}}_{CL}) = \frac{\text{var}(\hat{\bar{Y}}_{CL})}{\text{var}(\hat{\bar{Y}}_{SRSWOR})} \quad \textbf{4.11}$$

При этом верно следующее соотношение:

$$DEFF(\hat{\bar{Y}}_{CL}) = \frac{\text{var}(\hat{\bar{Y}}_{CL})}{\text{var}(\hat{\bar{Y}}_{SRSWOR})} = \frac{(\frac{1}{M^2})\,\text{var}(\hat{Y}_{CL})}{(\frac{1}{M^2})\,\text{var}(\hat{Y}_{SRSWOR})} = DEFF(\hat{Y}_{CL})$$

Дисперсия оценки суммарного значения ПСВ без ВО определяется:

$\text{var}(\hat{Y}_{SRSWOR}) = M\frac{(1-f)}{f}S_y^2$, где

$f = (n \times \overline{M})/M = n/N.$

С учетом вышеизложенного эффект плана кластерной оценки суммарного значения может быть рассчитан:

$$DEFF(\hat{\bar{Y}}_{CL}) = \frac{\text{var}(\hat{Y}_{CL})}{\text{var}(\hat{Y}_{SRSWOR})} = \frac{N(\frac{1-f}{f})S_{yCT}^2}{M(\frac{1-f}{f})S_y^2} = \frac{1}{\overline{M}}\frac{S_{yCT}^2}{S_y^2} \quad \textbf{4.12}$$

Таким образом, эффект плана рассчитывается как отношение скорректированной дисперсии кластерной оценки показателя и скорректированной дисперсии суммарного значения совокупности. Скорректированная дисперсия суммарного значения совокупности является фиксированной и не зависит от плана выборки. Иначе обстоит дело с скорректированной дисперсией по кластерам, которая полностью определяется планом кластерной выборки.

Детальный анализ эффекта плана возможен при дополнительном предположение о том, что существует простая связь между S_{yCT}^2 и S_y^2 (см далее).

Кластеры одинакового размера

Анализ дизайна плана основан на предположении о том, что все кластеры содержат одинаковое количество элементов:

$$M_d = \overline{M} = M / N$$

Результаты простой случайной выборки значений кластеров существенно упрощаются исходя из этого предположения. Например, выражение 4.5. может быть переписано следующим образом:

$$S_y^2 = (\frac{M-N}{M-N})S_{\text{int}\,ra}^2 + \frac{1}{\overline{M}}(\frac{N-1}{M-1})S_{yCT}^2 \quad \textbf{4.13}$$

Этот результат демонстрирует требуемое соотношение между скорректированной дисперсией по кластерам и скорректированной дисперсией совокупности. Теперь можно продолжить разработку формулы 4.12. Анализируя предыдущие формулы, можно сделать вывод о том, что в соответствии с вышеприведенным предположением межкластерная дисперсия прямо пропорциональна скорректированной дисперсии S_{yCT}^2.

4.13 облегчает расчет эффекта плана кластерной оценки суммарного значения:

$$DEFF(\hat{Y}_{CL}) = -(\frac{M-N}{N-1})(\frac{S_{\text{int}\,ra}^2}{S_y^2}) + (\frac{M-N}{N-1}) \quad \textbf{4.14}$$

Первый член в правой части от знака " = " прямо пропорционален отношению внутрикластерной дисперсии и дисперсии совокупности. Другими словами, эффект плана (кластерной оценки суммарного значения) представляет собой убывающую функцию линейной внутрикластерной дисперсии $S_{\text{int}\,ra}^2$

Формула, приведенная выше, показывает непрямую связь между способом кластеризации и эффектом плана оценки, поскольку внутрикластерная дисперсия зависит от дизайна кластера. Это означает, что эффект плана максимален в случае минимальной внутрикластерной дисперсии и минимален в случае максимальной внутрикластерной дисперсии. Таким образом, предел эффекта плана может быть определен следующим образом:

$$0 \le DEFF(\hat{Y}_{CL}) \le \frac{M-1}{N-1} \quad \textbf{4.15}$$

Нижний предел эффекта плана (кластерной оценки суммарного значения совокупности) достигается тогда и только тогда, когда межкластерная дисперсия равна нулю (см. также 4.5).

Кластерная выборка часто оказывается менее эффективной на практике, чем ПСВ без ВО. Таким образом, эта потеря эффективности должна быть компенсирована более низкими затратами кластерной выборки.

Как правило, большая внутрикластерная дисперсия приводит к кластерной оценке (суммарного значения совокупности) с большей точностью. В свою очередь, большее значение внутрикластерная дисперсия означает меньшую внутреннюю однородность кластеров. Поэтому целесообразно сделать кластеры как можно более разнородными.

Альтернативный анализ эффекта плана кластерной оценки суммарного значения совокупности предполагает внутриклассовый коэффициент корреляции. Внутриклассовый коэффициент корреляции ρ_c определяется как коэффициент корреляции Пирсона для $N \times \overline{M}(\overline{M}-1)$ парных наблюдений (Y_{dk} ,Y_{dl}) внутри кластеров, где k ≠ l и d = 1,…, N. Формальное определение таково:

$$\rho_c = \frac{\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l\neq l}^{\overline{M}}(Y_{dk}-\overline{Y})(Y_{dl}-\overline{Y})}{(M-1)(\overline{M}-1)S_y^2} \quad \textbf{4.16}$$

Коэффициент внутриклассовой корреляции или коэффициент внутрикластерной корреляции является мерой однородности внутри кластеров. Он показывает, насколько схожи или различны элементы в кластере. Максимальная однородность кластеров соответствует $\rho_c = 1$. Таким образом, минимальная однородность кластера соответствует максимальной внутрикластерной дисперсии.

При ближайшем рассмотрении обнаруживается возможность выражения коэффициента внутрикластерной через отношение скорректированной дисперсии кластеров и скорректированной дисперсии:

$$\rho_c = \left[\frac{N-1}{(M-1)(\overline{M}-1)}\right]\frac{S_{yCT}^2}{S_y^2} - \frac{1}{\overline{M}-1} \quad \textbf{4.17}$$

Этот результат показывает, что коэффициент внутрикластерной корреляции линейно возрастающая функция скорректированной дисперсии кластеров S_{yCT}^2.

Другими словами, ρ_c возрастающая функция межкластерной дисперсии $S_{\text{int}er}^2$. Коэффициент p_c минимален для наименьших значений межкластерной дисперсии и максимален для наибольшей межкластерной дисперсии.

При этом верно неравенство:

$$-\left(\frac{1}{\overline{M}-1}\right) \le p_c \le 1 \quad \textbf{4.18}$$

Наконец, уравнение 4.16 позволяет эффект плана кластерной оценки суммарного значения переформулировать как функцию от p_c. Анализ показывает, что для N>>1:

$$DEFF\left[\hat{Y}_{CL}\right] = \frac{M-1}{M-\overline{M}}\left(1+(\overline{M}-1)p_c\right) = (1+(\overline{M}-1)p_c) \quad \textbf{4.19}$$

Таким образом, эффект плана является линейной функцией p_c; аппроксимация 4.19 имеет место быть, когда число кластеров N велико. Линейно возрастающий характер отношения означает, что эффект плана достигает минимального значения при наименьшем значении коэффициента внутрикластерной корреляции.

При прочих равных условиях максимальный эффект плана достигается при максимально большом внутрикластерном коэффициенте корреляции. На практике коэффициенты внутриклассовой корреляции часто положительны. В частности, когда кластеры образуют "естественные" группы в совокупности, элементы внутри кластера будут больше похожи друг на друга, чем элементы, выбранные случайным образом из совокупности.

Это сходство может быть объяснено общим происхождением или окружающей обстановкой. Когда элементы внутри кластера похожи друг на друга, внутрикластерная дисперсия S^2_{intra} будет относительно мала по сравнению со скорректированной дисперсией S^2_y, так что значение p_c будет положительным.

Кроме того, p_c будет отрицательным только в том случае, если элементы в кластере демонстрируют большую дисперсию, чем случайно отобранные элементы. Это случается редко, но может произойти с искусственно сформированными кластерами, в которых элементы распределяются случайным образом.

Двухступенчатая выборка

Рассмотренная выше кластерная выборка всегда предусматривала обследование всех элементов отобранного кластера. Было показано, что кластерная выборка не очень эффективна, когда кластеры достаточно однородны. Если элементы внутри кластера похожи друг на друга, то обследование всех элементов может оказаться пустой тратой времени и денег;ту же самую информацию можно получить, исследовав только несколько элементов отобранного кластера.

В таких случаях кластерная выборка будет менее эффективна, и, возможно, более эффективно в этом случае извлечь выборку из отобранных кластеров. Этот метод называется двухэтапной выборкой.

Первый этап двухэтапной выборки заключается в случайном отборе кластеров или первичных единиц выборки. Второй этап двухэтапной выборки предусматривает случайный отбор элементов (вторичных единиц выборки), из каждой отобранной первичной единицы выборки.

Предполагая, что кластеры однородны, двухэтапная выборка обычно позволяет отобрать большее количество кластеров на первом этапе, и это может повысить точность оценивания. Этот метод позволяет лучше контролировать размер выборки, в которой первичные единицы сильно варьируются. Следует отметить, что и в этом случае основа выборки требуется только для отобранных первичных единиц.

Точность и затраты на двухэтапную выборку находятся примерно между затратами на кластерную выборку и стратифицированную выборку. Двухэтапная выборка обычно обходится дороже, чем кластерная выборка того же размера выборки, но дешевле, чем стратифицированная выборка. С другой стороны, двухэтапная выборка, как правило, более точна, чем кластерная выборка, и менее точна, чем стратифицированная выборка.

Двухступенчатую выборку легко расширить до многоступенчатой выборки (с тремя, четырьмя или более этапами). Например, отбор учащихся средней школы может включать в себя отбор школ, затем числа классов в отобранной школе и, наконец, отбор числа учащихся в выбранном классе.

Другой пример - последовательный отбор по регионам, городам, почтовым индексам и адресам. Единицы на заключительном этапе представляют собой элементы выборки. Здесь обсуждение ограничивается двухэтапной выборкой, не только в силу ее простоты, но также потому, что это более распространенный план выборки на практике, чем многоступенчатая выборка.

Планирование двухэтапной выборки включает выбор методов отбора как для первичных, так и для вторичных единиц выборки. Эти два метода не обязательно должны быть одинаковыми. На каждом этапе можно выбирать единицы различными способами: с помощью случайного или систематического отбора, отбора единиц пропорционально размеру (эти методы рассмотрены далее). При этом единицы могут быть сначала стратифицированы (расслоены).

Помимо методов отбора на двух этапах, необходимо также решить вопрос о распределении выборки на первом и втором этапах. Что лучше выбрать относительно большое число первичных единиц выборки и небольшое количество вторичных единиц выборки или относительно небольшое число первичных единиц выборки и большое количество вторичных единиц выборки? Первый способ, как правило, дает более точные оценки, но стоит дороже.

Текущий анализ подразумевает ПСВ без ВО в ходе реализации обоих этапов. Используемые обозначения почти аналогичны тем, что используется для построения кластерной выборки. Совокупность снова

делится на N непересекающихся кластеров или первичных выборочных единиц.

На первом этапе выбираются n первичных единиц выборки. Затем на втором этапе из каждой первичной единицы выборки выбираются вторичные единицы выборки m_d, где в общем случае $m_d \neq M_d$. Предполагается, что размер выборки m_d, d =1,...,n определяется заранее.

Значение целевой переменной для элемента k в выбранном кластере d обозначается через y_{dk}, k = 1,..., m_d. Кроме того, определяют для каждой (под)выборки две величины $\bar{y}_{sd}$ и s^2_{yd} :

$$\bar{y}_{sd} = \frac{1}{m_d}\sum_{k=1}^{m_d} y_{dk}$$

$$s^2_{yd} = \frac{1}{m_d - 1}\sum_{k=1}^{m_d}(y_{dk} - \bar{y}_{sd})^2 \quad \textbf{4.20.}$$

Оценки при двухэтапной выборке

Двухэтапная выборка не предполагает наблюдения всех вторичных единиц отбора в отобранных первичных единицах. Поэтому необходимо оценить кластерные параметры (выбранных) первичных единиц. Суммарное значение по выборке кластера y_d в первичной выборочной единице d может быть оценено:

$$\hat{y}_d = M_d \bar{y}_{sd} = \frac{M_d}{m_d}\sum_{k=1}^{m_d} y_{dk} \qquad (d=1,\ldots n) \quad \textbf{4.21.}$$

Другими словами, оценка суммарного значения в кластере равна взвешенной сумме наблюдаемых вторичных единиц (в упомянутой ранее первичной единице выборки). Весовой коэффициент является обратной величиной доли отбора первичных выборочных единиц d.

Среднее кластерное значение $\bar{Y}_{CT}$ может быть вычислено:

$$\hat{\bar{Y}}_{CT} = \bar{\hat{y}}_{CT} = \frac{1}{n}\sum_{d=1}^{n}\hat{y}_d$$

Следовательно, кластерные оценки суммарного и среднего значений:

$$\hat{Y}_{CL2} = N\hat{\bar{Y}}_{CT} = \frac{N}{n}\sum_{d=1}^{n}\hat{y}_d$$

$$\hat{\bar{Y}}_{CL2} = \frac{1}{\bar{M}}\hat{\bar{Y}}_{CT} = \frac{N}{nM}\sum_{d=1}^{n}\hat{y}_d$$

Несмещенность оценок

Тот факт, что отобранный кластер (первичная выборочная единица) наблюдается уже не полностью, а посредством отбора ПСВ без ВО m_d единиц из M_d, не оказывает никакого влияния на анализ смещения оценок. То есть все оценки, описанные ранее в этом подразделе, являются несмещенными.

Расчет дисперсии

Дисперсии оценок $\hat{\bar{Y}}_{CL}$ и $\hat{Y}_{CL}$ в обычной кластерной выборке определяются только различиями между кластерами. Однако кластерная оценка $\hat{y}_d$ в двухэтапной выборке является стохастической переменной. Следовательно, вариация внутри кластеров также вносит свой вклад в дисперсию оценок. Таким образом, дисперсия кластерной оценки суммарного значения $\hat{Y}_{CL2}$ содержит дополнительный член:

$$\mathrm{var}(\hat{Y}_{CL2}) = N(\frac{1-f_1}{f_1})S^2_{yCT} + \sum_{d=1}^{N} M_d(\frac{1-f_{2,d}}{f_1 f_{2,d}})S^2_{yd} \quad \textbf{4.22.}$$

где

$$f_1 = \frac{n}{N} \quad \wedge \quad f_{2,d} = \frac{m_d}{M_d}$$

Первая дробь называется долей отбора на первом этапе; вторая - долей отбора на втором этапе. Дисперсия кластерной оценки среднего значения вычисляется;

$$\mathrm{var}(\hat{\bar{Y}}_{CL2}) = \frac{1}{M^2}\mathrm{var}(\hat{Y}_{CL2}) = (\frac{1}{M})(\frac{1}{\bar{M}})(\frac{1-f_1}{f_1})S^2_{yCT} + \frac{1}{M}\sum_{d=1}^{N}(\frac{M_d}{M})(\frac{1-f_{2,d}}{f_1 f_{2,d}})S^2_{yd}$$

Влияние выборки на втором этапе на дисперсию, указанную выше, следует из сравнения формул 4.22 с 4.9.

Дисперсии оценок состоят из двух членов. Первый член (справа от знака"=") представляет собой различия значений первичных единиц выборки. Второй член (слева от знака"=") - вариацию внутри первичных единиц выборки.

Этот член не играет никакой роли в приведенных выше разделах, поскольку кластеры наблюдались полностью ($f_{2,d} \equiv 1$). Во многих ситуациях второй член будет пренебрежимо мал по сравнению с первым членом.

При сильно меняющемся M_d ПСВ без ВО менее эффективно на первом этапе, так как значения кластеров обычно очень существенно различаются. Более эффективным методом отбора является отбор первичных единиц пропорционально их размеру или с помощью оценки по отношению.

Дисперсия (4.22) без смещения может быть оценена по формуле:

$$\hat{\text{var}}(\hat{Y}_{CL2}) = N(\frac{1-f_1}{f_1})s^2_{\hat{y}CT} + \sum_{d=1}^{n} M_d(\frac{1-f_{2,d}}{f_1 f_{2,d}})s^2_{yd}$$

$$s^2_{\hat{y}CT} = \frac{1}{n-1}\sum_{d=1}^{n}(\hat{y}_d - \bar{\hat{y}}_{CT})^2 \quad \textbf{4.23}$$

Оценка плана выборки

В двухступенчатом отборе в среднем обследуются $(n \times \bar{m})$ единиц отобранных на втором этапе. Ожидаемое среднее значение объема выборки в n - отобранных первичных единиц равно m = $\sum_{d=1}^{N} m_d / N$.

Для оценки кластерной выборки сравнивается дисперсия кластерных оценок с дисперсией соответствующих оценок ПСВ без ВО объема $(n \times \bar{m})$ из M элементов. С этой целью рассчитывается эффект плана.

Дисперсия оценки суммарного значения на основе ПСВ без ВО объема $(n \times \bar{m})$ вычисляется следующим образом:

$$\text{var}(\hat{Y}_{SWSWOR}) = M(\frac{1-f_1 f_2}{f_1 f_2})S^2_y \quad \textbf{4.24}$$

$$f_1 = \frac{n}{N} \wedge f_2 = \frac{\bar{m}}{\bar{N}}$$

На основе выражений (4.22) и (4.24) эффект плана кластерной оценки суммарного значения может быть рассчитан следующим образом:

$$DEFF(\hat{Y}_{CL2}) = (\frac{1}{\bar{M}})(\frac{f_2 - f_1 f_2}{1-f_1 f_2})(\frac{S^2_{yCT}}{S^2_y}) + (\frac{f_2}{1-f_1 f_2})\sum_{d=1}^{N}(\frac{M_d}{M})(\frac{1-f_{2,d}}{f_{2,d}})(\frac{S^2_{yd}}{S^2_y}) \quad \textbf{4.25}$$

Этот результат можно интерпретировать как общий случай выражения 4.12. Этот вывод подтверждается тем,что при условии полного обследования выбранных первичных единиц (т. е. $f_{2,d} = f_2 = 1$) обе формулы идентичны.

В целях дальнейшего преобразования формулы (4.25) предполагается, что первичные единицы выборки имеют одинаковый размер: т. е. $M_d = \bar{M}$ для всех d. Кроме того, предполагается,что из каждой первичной единицы выборки отбирается одинаковое количество единиц, то есть $m_d = \bar{m}$ для d=1,…,N. Легко определить, что:

$$M_d = \overline{M} = M / N \wedge f_{2,d} = \frac{m_d}{M_d} = \frac{\overline{m}}{\overline{M}} = f_2 \quad \textbf{4.26}$$

Подставляя (4.26) в (4.25), получаем:

$$DEFF(\hat{Y}_{CL2}) = (\frac{1}{\overline{M}})(\frac{f_2(1-f_1)}{1-f_1 f_2})(\frac{S^2_{yCT}}{S^2_y}) + (\frac{1-f_2}{1-f_1 f_2})(\frac{S^2_{\text{int}\,ra}}{S^2_y}) \quad \textbf{4.27}$$

Формула (4.13) для скорректированной дисперсии остается в силе, поскольку свойства совокупности не изменяются в зависимости от плана выборки. Поэтому верно выражение (см. также (4.13)):

$$(\frac{S^2_{yCT}}{S^2_y}) = -M(\frac{\overline{M}-1}{N-1})(\frac{S^2_{\text{int}\,ra}}{S^2_y}) + \frac{\overline{M}(M-1)}{N-1} \quad \textbf{4.28}$$

Этот результат дает желаемую связь между скорректированной кластерной дисперсией и скорректированной дисперсией совокупности. Для двухэтапной выборки эффект плана кластерной оценки показателя также может быть определен, подставляя (4.28) в (4.27):

$$DEFF(\hat{Y}_{CL2}) = (\frac{1}{1-f_1 f_2})(1 - \left[\frac{M-1}{N-1}\right] f_2 + \left\{\frac{M-N}{N-1}\right\} f_1 f_2)(\frac{S^2_{\text{int}\,ra}}{S^2_y}) + \left[\frac{f_2 - f_1 f_2}{1-f_1 f_2}\right]\left\{\frac{M-1}{N-1}\right\}$$

4.29

Объем выборки

Как и в случае с простой случайной выборкой, общий размер выборки может быть определен с учетом требуемой точности. После определения общего объема выборки возникает вопрос о том, как распределить его между первичными и вторичными единицами выборки.

Другими словами, необходимо определить сколько вторичных единиц выборки должно быть выбрано из отобранных первичных единиц и сколько первичных единиц должно быть в выборке.

Эти вопросы обычно связаны с балансом затрат и требуемой точностью. Отбор множества первичных единиц с небольшим количеством вторичных единиц в одной первичной единице обычно приводит к маленькой дисперсии.

Однако потенциальные затраты, сэкономленные при наблюдении большого числа элементов в пределах одной первичной единицы выборки, таким образом почти полностью теряются. Хотя наблюдение большого количества вторичных единиц выборки в рамках ограниченного количества первичных единиц выборки обходится дешевле, оценки будут менее точными.Ответ на вопрос о том, как распределить выборку, требует

определенных знаний о сборе данных и других затратах на обоих этапах, а также об однородности первичных единиц выборки.

С совершенно однородными первичными единицами выборки, то есть $S^2_{\text{int}\,ra} = 0$, можно было бы также выбрать только один элемент на первичную единицу выборки; наблюдение большего количества элементов приведет к более высоким затратам без выигрыша в точности. Во всех остальных случаях распределение выборки будет зависеть от затрат на отбор единиц на двух этапах. Простая функция затрат вычисляется:

$$C = nc_1 + n\overline{m}c_2$$

где параметры c_1 и c_2 соответствуют, например, расходам на интервьюера и/или командировочные расходы на первом и втором этапах отбора. Принимая во внимание эту функцию затрат и фиксированную общую стоимость C, дисперсия кластерной оценки показателя минимизируется путем отбора:

$$n = \frac{C}{c_1 + c_2 m}$$

при этом первичные выборочные единицы в выборке и:

$$m = \sqrt{\frac{\overline{M}S^2_{\text{int}\,ra}}{(\overline{M}S^2_{\text{int}\,er} - S^2_{\text{int}\,ra})}\frac{c_1}{c_2}} = \sqrt{\frac{c_1 / c_2}{S^2_{\text{int}\,er} / S^2_{\text{int}\,ra} - (1/\overline{M})}}$$

вторичные единицы на первичную выборочную единицу. Для определения этих величин требуется некоторое представление об отношении $S^2_{\text{int}\,er} / S^2_{\text{int}\,ra}$, или, другими словами, об однородности первичных единиц. На практике часто происходит наоборот. Обычно выполняется перебор различных значений n и m, с учетом соответствующего значения дисперсии оценки.

V. Систематическая выборка

Систематическая выборка часто используется в качестве альтернативы ПСВ без ВО. В систематической выборке при отборе n элементов из совокупности N элементов сначала определяется длина шага L = N / n.

Для удобства предположим, что элементы пронумерованы от 1 до N. Затем стартовое число R выбирается случайным образом из интервала (0, L).

Первым выбранным элементом в выборке является элемент с номером k_1 для которого $k_1\text{-}1 < R \leqslant k_1$.

Второй элемент, отобранный в выборку, обозначается k_2, для которого $k_2 - 1 < R + L \leqslant k_2$, и так далее.

Последний n^{th}-й элемент в выборке k_n, для которого $k_n - 1 < R + (n-1)L \leqslant k_n$.

Другими словами, элементы с номерами k_1, …, k_n образуют выборку и номера в выборке равны (округленные значения) R, R + L,R + 2L,…, R + (n - 1)L. Ясно, что случайно выбранное стартовое число R определяет всю выборку, и что каждый элемент ссовокупности может включаться только в одну выборку.

Теоретически, предполагая, что L-целое число, различные выборки возможны для фиксированных совокупностей L, так что вероятность отбора систематической выборки равна:

$$P(s) = \frac{1}{L} = \frac{n}{N}$$

Поскольку каждый элемент может встречаться только в одной выборке, вероятность включения элемента k равна вероятности отбора выборки, т. е.:

$$\pi_k = \frac{1}{L} = \frac{n}{N}$$

При такой форме систематической выборки число возможных выборок меньше, чем при ПСВ без ВО того же самого объема выборки, так как не каждая комбинация n элементов возможна. Однако вероятности включения первого порядка такие же, как и в случае ПСВ без ВО.

Хотя систематическая выборка, описанная здесь, фактически представляет собой кластерную выборку, дисперсия оценки часто аппроксимируется на практике дисперсией оценки ПСВ без ВО. Предполагается, что (фиксированная) последовательность элементов не приводит к систематическому отличию от оценки ПСВ без ВО.

Оценки, несмещенность и дисперсия

Систематическая выборка, таким образом, сводится к отбору одного кластера размера n из L кластеров. Параметры совокупности оцениваются на основе выборочной статистики этого единственного выбранного кластера. Соответствующие кластерные оценки для совокупности и среднего значения совокупности:

$$\hat{Y}_{sys} = N\bar{y} \quad \wedge \quad \hat{\bar{Y}}_{sys} = \bar{y}$$

Смещение оценок следует из свойств кластерной выборки. Дисперсия кластерной оценки среднего значения по выборке из одного кластера определяется по формуле (4.10):

$$\mathrm{var}(\hat{\bar{Y}}_{sys}) = (1 - \frac{1}{L})\frac{S^2_{yCT}}{n^2} \quad \textbf{4.30}$$

Дисперсия выражается в определениях, используемых для систематической выборки. При этом N и n относятся здесь к элементам, а не к кластерам. Размер кластера (систематическая выборка) теперь равно n;

общее число элементов равно N; число кластеров равно L. Дисперсия кластерной оценки суммарного значения рассчитывается следующим образом: умножением дисперсии кластерной оценки среднего значения на N^2.

План выборки

Наконец, эффект плана кластерной оценки совокупности может быть рассчитан по формуле (4.19) следующим образом:

$$DEFF(\hat{\bar{Y}}_{sys}) = \frac{N-1}{N-n}(1+(n-1)\rho_c)$$

Поэтому систематическая выборка лучше простой случайной выборки тогда и только тогда, когда коэффициент внутриклассовой корреляции удовлетворяет:

$$-\frac{1}{n-1} \le \rho_c \le -\frac{1}{N-1}$$

Когда дисперсия внутри возможных систематических выборок больше, чем дисперсия совокупности, кластерные средние значения будут мало отличаться друг от друга, то систематическая выборка будет более точной, чем простая случайная выборка.

В противном случае, если дисперсия внутри систематических выборок относительно мала по сравнению с дисперсией совокупности ($0 < p_c$), то все элементы в выборке будут давать одинаковую информацию, и дисперсия оценки систематической выборки будет больше, чем дисперсия простой случайной выборки.

Однако существенным недостатком систематической выборки является отсутствие несмещенных оценок различных дисперсий. Поскольку выбран только один кластер, выборочная дисперсия кластера s^2_{yCT} не может быть определена. Тем не менее, дисперсия может быть аппроксимирована, если мы обладаем информацией о структуре совокупности.

Можно выделить следующие три ситуации:

1. случайные последовательность в основе выборки;
2. положительная автокорреляция;
3. периодические колебания.

Во многих ситуациях верно, что значения целевых переменных не соответствуют последовательности единиц в основе выборки. Примером может служить основа выборки, в которой люди расположены в алфавитном порядке. В этом случае систематическая выборка будет практически совпадать с простой случайной выборкой [$p_c \approx -1/(N-1)$], и обе выборки будут давать одинаковые результаты. Поэтому мы можем использовать формулу дисперсии для простой случайной выборки для оценки $\text{var}(\hat{\bar{Y}}_{sys})$.

Иногда в основе выборки может отмечаться эффект увеличения или уменьшения значений. Например, компании могут появляться в списке в порядке уменьшения их размера. В этих случаях говорят, что существует положительная автокорреляция: последовательные элементы похожи друг на друга, если они расположены ближе друг к другу, и, наоборот, если элементы находятся дальше друг от друга.

Систематическая выборка обеспечит больший разброс элементов выборки и, следовательно, будет более точной, чем простая случайная выборка [-1/(n -1)≤ p_c< -1/(N-1)]. Формула дисперсии для простой случайной выборки дала бы в этом случае завышенную оценку.

Проблемы могут возникнуть в том случае, если основа выборки демонстрирует периодические вариации с точки зрения целевой переменной. Систематическая выборка тогда будет значительно менее точной, чем простая случайная выборка, в частности, когда длина шага равна или кратна периоду колебаний.

Дисперсия оценки тогда будет чрезвычайно большой, что не очевидно при использовании формулы простой случайной дисперсии для оценки дисперсии.

Предположим, что элементы совокупности расположены так, что целевая переменная распределена следующий образом: 1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,...

При длине шага 5 все элементы в выборке, будут одинаковыми, и формула дисперсии для простой выборки будет равна $\hat{\text{var}}(\hat{\bar{Y}}) = 0$.

Однако фактическое значение $\text{var}(\hat{\bar{Y}}_{sys})$ для этой совокупности равно $\sigma_y^2 = 2$; заметим, что p_c =1. Другими словами,систематическая выборка в этом случае так же эффективна как и выборка ПСВ без ВО размером n = 1.

Показатели качества

Важным показателем качества (многоступенчатой) кластерной выборки является снижение затрат в результате кластеризации наблюдаемых единиц. Цена, которую придется заплатить за эту кластеризацию в виде увеличения дисперсии по сравнению с простой случайной выборкой без возвращения, является еще одним важным критерием.

VI. Выборки с равными и неравными вероятностями включения

Краткое описание

Преимущества

При формировании выборки не всем элементам совокупности нужно присваивать одинаковую вероятность включения, как это было сделано до сих пор. Особенно, когда целевая переменная Y_k совокупности примерно пропорциональна некой вспомогательной переменной X_k (k =1…N). В этом случае предпочтительна выборка с неравными вероятностями включения.

Последняя переменная должна быть известна для всех элементов совокупности. На практике вспомогательная переменная X_k часто представляет собой размер или релевантность элемента k во всей совокупности.

Существует множество способов выборки без возвращения с неравными вероятностями включения. В этом материале основное внимание уделяется систематическим выборкам "с вероятностью пропорциональной размеру" (ВПР), где элементы находятся в случайном порядке, а вероятности включения пропорциональны заданной вспомогательной переменной X_k.

Большое преимущество неравных вероятностей включения состоит в том, что дисперсии оценок могут быть существенно уменьшены, что также уменьшает соответствующую степень неопределенности. Еще одно преимущество заключается в том, что стратификация по размеру больше не требуется, что позволяет избежать риска иметь малое количество наблюдений в некоторых слоях.

VIII. Систематическая выборка с вероятностью пропорциональной размеру (ВПР)

После случайного упорядочения элементов совокупности каждому элементу k присваивается интервал длиной X_k (на оси) k = 1,…, N.

Первому элементу по определению присваивается интервал $(0,X_1]$, второму- $(X_1, X_1 + X_2]$ и т. д. В результате интервал (0, X], где X обозначает суммарный итог вспомогательной переменной, делится на N последовательных интервалов. Затем (0, X] делится на n равных частей отрезка длины L (L = X/n); L также называется длиной шага.

Схема отбора теперь выглядит следующим образом. Случайным образом отбирается число r между 0 и L . Это число попадает ровно в один из интервалов, составляющих (0,X].Элемент совокупности, соответствующий этому интервалу, является первым элементом в выборке. Элемент

совокупности, заданный интервал которого содержит число (r+L), становится 2-м элементом в выборке, и так далее. Последний элемент в выборке - это тот, где интервал содержит число [r+(n-1)L]. Данная схема выборки может быть представлена в следующим образом:

$S_k=X_1+\ldots X_k$ $(k=1,\ldots N)$

$S_o=0$

$J(x)=k$ где $S_{k-1}<x\leqslant S_k$ $(0<x\leq X)$

Систематическая выборка ВПР тогда состоит из n элементов:

$J(r), J(r+L),\ldots J(r+(n-1)L)$

Чтобы избежать многократного включения элемента в выборку, элементы, для которых $L<X_k$, удаляются из совокупности и включаются в отдельный слой, который наблюдается в полном объеме. Этот слой содержит само отбирающиеся элементы. При этом X должен быть пересчитан, определен новый L и объем выборки n. При необходимости этот процесс следует повторять до тех пор, пока не исчезнут само отбирающиеся элементы.

Элементы в выборке, описанной здесь, упорядочены случайным образом. Однако иногда существуют причины для сохранения некоторой фиксированной последовательности элементов, например по возрастанию X_k

Если Y_k/X_k и X_k тесно связаны, то такая последовательность может давать оценки с особенно малой дисперсией. Тем не менее, поиск хорошей оценки дисперсии может быть проблематичным.Стандартной оценки дисперсии не существуют, но при определенных допущениях могут быть получены достаточно надежные оценки дисперсии.

Заметим , что нет необходимости произвольно упорядочивать элементы, когда нет значимой связи между Y_k/X_k и X_k, или, точнее, когда нет связи между Y_k/X_k и k, $k = 1,\ldots,N$. В этих ситуациях случайная последовательность элементов не будет иметь систематического влияния на результаты.

Наконец, необходимо упоминуть частный случай, когда все элементы имеют одинаковую вероятность включения, то есть всем элементам присваивается интервал одинаковой длины, в то время как X_k различаются.

С элементами, которые расположены случайным образом, этот подход эквивалентен ПСВ без ВО, и в этом случае применяются формулы, приведенные ранее. Если же, наоборот, элементы расположены в фиксированной последовательности, то опять же гораздо сложнее оценить дисперсию оценки суммарного значения совокупности.

Однако можно показать на практике, что дисперсия уменьшается, если элементы расположены таким образом, чтобы все возможные выборки похожи друг на друга (а следовательно, и в совокупности) с точки зрения распределения X_k. Это достигается путем упорядочивания элементов в соответствии с размером X_k .

Условием уменьшения дисперсии оценки является наличие достаточно высокой корреляции между Y_k и X_k .

Другой вариант состоит в том, чтобы расположить элементы в соответствии с регионом так, чтобы каждая выборка имела достаточный региональный разброс (разнородность). Другими словами, (фиксированная) последовательность должна быть сформирована таким образом, чтобы отобранные выборки становились однородными. Как и в случае с неравными вероятностями включения, проблемой является оценка дисперсии.

Применимость

ВПР часто используется в социальной статистике, например в двухэтапной выборке индивидуумов. На первом этапе муниципалитеты отбираются с вероятностью отбора, пропорциональной количеству жителей. На втором этапе отдельные лица отбираются с применением ПСВ без ВО из каждого муниципалитета, выбранного на первом этапе. Доля отбора на втором этапе, в свою очередь, обратно пропорциональна числу жителей соответствующего муниципального образования, так что в целом все индивидуумы имеют одинаковую вероятность включения.

Подробное описание

Отправной точкой является совокупность из N элементов, из которой выборка размера n должна быть отобрана без возвращения, если не указано иное.

В этой главе предполагается, что размер выборки n является фиксированным. Как и в предыдущих главах, нас интересует оценка среднего значения $\bar{Y}$ и суммарного значения Y. n выборочных наблюдений целевой переменной обозначаются: $y_1,...,y_n$.

Индикатор бинарного отбора a_k – это случайная величина, которая определяет, был ли элемент k совокупности в конечном счете включен в выборку.

Вероятность того, что в выборке будет выбран элемент k, то есть вероятность включения, обозначается символом π_k. Формальным определением вероятности включения является:

$$\pi_k = P(a_k = 1) \qquad k=1,\ldots N \quad \textbf{5.1}$$

$$a_k = \begin{cases} 1 \\ 0 \end{cases}$$

если элемент k включен в выборку
если элемент k не включен в выборку

Вероятность включения второго порядка для двух элементов k и l, обозначаемая как π_k, представляет собой вероятность того, что элементы k и l находятся вместе в выборке. Это можно записать так:

$$\pi_{kl} = \begin{cases} P(a_k = 1 \wedge a_1 = 1) \\ \pi_k \end{cases}$$

k ≠ l =1, …N **5.2**
k = l =1, …N

Учитывая (неравные) вероятности включения π_k для всех N элементов совокупности, суммарное значение Y может быть оценено с помощью оценки Гурвица-Томпсона (ГТ).

Оценка ГТ, обозначаемая как $\hat{Y}_{HT}$ и определяется как:

$$\hat{Y}_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} \quad \textbf{5.3.}$$

Необходимым условием применения этой оценки является то,что вероятность включения π_k больше 0 для k=1,…,N. Определение (5.3) приводит к определению оценки Гурвица-Томпсона для среднего значения обозначаемого $\hat{\bar{Y}}_{HT}$:

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\hat{Y}_{HT} = \frac{1}{N}\sum_{k=1}^{n} \frac{y_k}{\pi_k} \quad \textbf{5.4}$$

Оценка ГТ является общей и может быть применена к любой выборке без возвращения.

Чтобы продемонстрировать несмещенность оценок ГТ (5.3) и (5.4), мы переходим к другим обозначениям для (5.3) и (5.4). Используя индикатор отбора, формулы оценок ГТ можно переписать следующим образом:

$$\hat{Y}_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \sum_{k=1}^{N} \left(\frac{a_k}{\pi_k}\right) Y_k \quad \textbf{5.5}$$

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{k=1}^{n} \frac{y_k}{\pi_k} = \frac{1}{N}\sum_{k=1}^{N} \left(\frac{a_k}{\pi_k}\right) Y_k$$

Поскольку $E(a_k) = \pi_k$, оценки в (5.5) являются несмещенными, т. е.

$$E(\hat{Y}_{HT}) = Y \ \wedge \ E(\hat{\bar{Y}}_{HT}) = \bar{Y} \quad \textbf{5.6}$$

При расчете дисперсии оценок ГТ (5.3) и (5.4) используются следующие результаты для вариации/ковариации показателя индикатора отбора:

$$\mathrm{var}(a_k) = \pi_k - (1 - \pi_k)$$

$$\text{cov}(a_k, a_l) = \pi_{kl} - \pi_k \pi_l$$

$1 \le k,l \le N$ **5.7.**

Этот результат является расширением свойства (2.9) индикатора отбора. Теперь можно вывести следующую формулу для дисперсии оценки ГТ:

$$\text{var}(\hat{Y}_{HT}) = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l}\right) Y_k Y_l \quad \textbf{5.8}$$

Другими словами, дисперсия оценки ГТ суммарного значения является линейной комбинацией всех возможных $Y_k Y_l$, где k,l=1,…,N.

Для определения дисперсии 5.8 на основе наблюдений вводится следующая оценка, называемая оценкой ГТ дисперсии:

$$\hat{\text{var}}_{HT}(\hat{Y}_{HT}) = \sum_{k=1}^{n}\sum_{l=1}^{n}\left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{lk} \pi_k \pi_l}\right) y_k y_l \quad \textbf{5.9}$$

Если n фиксировано, то для этой оценки существует следующая альтернативная оценка (Сена - Йейтса - Гранди (1953)):

$$\hat{\text{var}}_{SYG}(\hat{Y}_{HT}) = -\frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}\left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{lk}}\right)\left(\frac{y_k}{\pi_k} - \frac{y_k}{\pi_l}\right)^2 \quad \textbf{5.10}$$

Может быть доказано, что 5.9 и 5.10 являются несмещенными оценками дисперсии оценки ГТ суммарного значения.

Оценка дисперсии (5.10) обычно имеет несколько меньшую дисперсию, чем (5.9), но на практике обе оценки могут принимать отрицательное значение.

Рассмотрим, например, экстремальную ситуацию, когда π_k точно пропорционально Y_k, и в этом случае оценка ГТ равна Y с вероятностью 1. Поэтому дисперсия оценки ГТ в этой ситуации равна 0.

Однако оценка дисперсии ГТ (5.9) не равна 0 с вероятностью 1, но тем не менее является несмещенной, что означает, что оценка дисперсии ГТ принимает отрицательное значение для всех возможных выборок. И наоборот, в этой ситуации оценка дисперсии (5.10) равна 0 с вероятностью 1.

На практике точные выражения для π_{kl} обычно трудно вычислить, что затрудняет использование формул (5.9) и (5.10).Однако в некоторых ситуациях имеются приемлемые аппроксимации для дисперсий.

Хорошая аппроксимация возможна, когда n намного меньше N. В этом случае подход может быть таков, как если бы применялась выборка с возвращением (см. раздел далее).

Наконец, следует отметить, что в случае ПСВ без ВО:

$$\pi_k = \frac{n}{N} \quad \wedge \quad \pi_{kl} = \frac{n(n-1)}{N(N-1)} \quad k,l = 1,...N$$

ВПР без возвращения

В выборке ВПР без возвращения размера n вероятность включения π_k элемента совокупности k по определению пропорциональна заданной вспомогательной переменной X_k. Поскольку вероятности включения элементов совокупности обязательно суммируются до объема выборки, π_k определяется следующим образом:

$$\pi_k = n\frac{X_k}{X} \quad \text{k=1,…,N .}$$

Другими словами, выборка ВПР без возвращения размера n не что иное, как выборка без возвращения размера n с разными вероятностями включения. Таким образом, результаты, приведенные в предыдущем разделе, применимы к данному типу выборочного плана.

Выборка ВПР с возвращением

При выборке с возвращением с размером выборки n для всех n отборов элемент совокупности k имеет одинаковую вероятность отбора.Эта вероятность называется вероятностью извлечения элемента совокупности k и обозначается как p_k:

Для выборки выборке ВПР с возвращением вероятность извлечения p_k элемента совокупности k (k=1,…,N) равна:

$$p_k = \frac{X_k}{X} = \frac{\pi_k}{n} \quad \wedge \quad \sum_{k=1}^{N} p_k = 1$$

где π_k – вероятность включения элемента k ВПР без возвращения. Другими словами: $\pi_k = np_k = nX_k / X$.

Стандартная оценка суммарного значения совокупности Y в выборке ВПР с возвращением известна как оценка Хансена-Гурвица (ХГ) (1943). Оценка ХГ суммарного значения совокупности Y определяется следующим образом:

$$\hat{Y}_{HH} = \bar{z}_s \frac{1}{n}\sum_{k=1}^{n} z_k \text{ , где } z_k \equiv \frac{y_k}{p_k} = \frac{y_k}{x_k} X$$

Заметим, что zk (k=1,…, n) можно рассматривать как стохастическое явление, предполагая, что значение $Z_l = Y_l/p_l$ (l = 1,…, N) с вероятностью p_l. Несмещенность оценки ХГ тогда следует из:

$$E(z_k) = \sum_{l=1}^{N} p_l Z_l = \sum_{l=1}^{N} p_l \frac{Y_l}{p_l} = Y \quad (k=1,\ldots, n)$$

Определяем далее:

$$\sigma_z^2 = \sum_{l=1}^{N} p_l \left[\frac{Y_l}{p_l} - Y\right]^2$$

Так как дисперсия z_k:

$$\mathrm{var}(z_k) = E(z_k - Y)^2 = \sum_{g=1}^{N} p_l \left[\frac{Y_l}{p_l} - Y\right]^2 = \sigma_z^2 \quad (k=1,\ldots, n)$$

и поскольку выборка с возвращением, $\mathrm{var}(\hat{Y}_{HH})$:

$$\mathrm{var}(\hat{Y}_{HH}) = \frac{1}{n^2}\sum_{k=1}^{n} \mathrm{var}(z_k) = \frac{\sigma_z^2}{n} \quad \textbf{5.11}$$

Эта дисперсия может быть оценена без смещения с помощью:

$$\hat{\mathrm{var}}(\hat{Y}_{HH}) = \frac{s_z^2}{n} = \frac{1}{n(n-1)}\sum_{k=1}^{n}\left(\frac{y_k}{p_k} - \hat{Y}_{HH}\right)^2 \quad \textbf{5.12}$$

Несмещенность этой оценки дисперсии следует из:

$$E(s_z^2) = \frac{n}{n-1} E\left(\frac{1}{n}\sum_{k=1}^{n}(z_k - Y)^2 - (\bar{z}_s - Y)^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{k=1}^{n}\sigma_z^2 - \frac{\sigma_z^2}{n}\right) = \sigma_z^2$$

Более того, когда n намного меньше N, $\mathrm{var}(\hat{Y}_{HT})$ ВПР без возвращения может обоснованно быть оценено:

$$\hat{\mathrm{var}}_{HH}(\hat{Y}_{HT}) = \frac{s_z^2}{n} = \frac{1}{n(n-1)}\sum_{k=1}^{n}\left(\frac{y_k}{x_k / X} - \hat{Y}_{HT}\right)^2 \quad \textbf{5.13}$$

Предположение, лежащее в основе этой аппроксимации, состоит в том, что дисперсия оценки Y в случае ВПР вряд ли зависит от того, возвращаются обратно или нет отобранные элементы.

Простая оценка дисперсии систематической выборки ВПР

Хотя оценка дисперсии (5.13) предназначена для выборки ВПР с возвращением, с небольшой корректировкой она также может быть использована для выборки ВПР без возвращения, где n значительно меньше N.Эта скорректированная оценка дисперсии неявно упоминается Hájek (1964) для сопоставимой выборки, называемой «отклоняющая выборка», и определяется следующим образом:

$$\hat{\text{var}}_{Haj}(\hat{Y}_{HT}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(1-\pi_k)\left(\frac{y_k}{x_k / X} - \hat{Y}_{HT}\right)^2 \quad \textbf{5.14}$$

Формула, выведенная Hájek (1964) для данной выборки, имеет вид:

$$\hat{\text{var}}_{Haj}(\hat{Y}_{HT}) = \frac{1}{n^2}\sum_{k=1}^{n}\pi_k(1-\pi_k)\left(\frac{y_k}{x_k / X} - Y^*\right)^2$$

$$Y^* = \sum_{k=1}^{N} a_k \frac{y_k}{y_k / X} \qquad a_k = \frac{\pi_k(1-\pi_k)}{\sum_{k=1}^{N}\pi_k(1-\pi_k)}$$

Оценка дисперсии (5.14) может быть использована при ВПР без возвращения при условии, что элементы совокупности сначала размещаются в случайном порядке, прежде чем будет извлечена систематическая выборка ВПР. Она также может быть использована, когда n и N имеют один и тот же порядок размерности, при условии, что Y_k / X_k и X_k не коррелируют.

Если Y_k/X_k и X_k коррелированы, а n и N также имеют один и тот же порядок размерности, лучше использовать следующую оценку дисперсии:

$$\hat{\text{var}}_{Haj2}(\hat{Y}_{HT}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}\pi_k(1-\pi_k)\left(\frac{y_k}{x_k / X} - \hat{Y}^*\right)^2$$

$$\hat{Y}^* = \frac{\sum_{k=1}^{n}(1-\pi_k)\frac{y_k}{x_k / X}}{\sum_{k=1}^{n}(1-\pi_k)}$$

Наконец, если $\pi_k = n/N$ или $x_k = X / N$, то для (5.14) становится знакомой оценкой (2.12) для дисперсии прямой оценки $\hat{Y}_{SRSWOR} = N\bar{y}_s$.

Пример

Для демонстрации снижения дисперсии при систематической выборке ВПР в этом разделе подробно описывается оценка (комбинированного) индекса цен 70 компаний на основе выборки размером n = 9 . Дисперсия для систематической выборки ВПР сравнивается с дисперсией ПСВ без ВО.

Изменения цен получены на основе наблюдений цен обследования PRODCOM 27 за период с декабря 2004 года по декабрь 2005 года. Данные приведены в таблице далее. Две крупнейшие компании были пропущены, потому что их доля оборота была больше 1/9. Они наблюдались в полном объеме как отдельная страта.

Индекс цен для N компаний:

$$I = \sum_{k=1}^{N} W_k P_k$$

N = 70

P_k - изменение цены для компании k за отчетный период в процентах;

W_k - доля оборота в компании k в базовом году.

$$\sum_{k=1}^{N} W_k = 1$$

k	Изменение цены	Доля оборота	k	Изменение цены	Доля оборота
1	2	3	4	5	6
1	-18.4%	0.0608	36	34.8%	0.0427
2	-16.0%	0.0784	37	13.1%	0.0121
3	3.3%	0.0762	38	31.7%	0.0351
4	12.5%	0.0100	39	-24.8%	0.0074
5	0.0%	0.0029	40	55.3%	0.0009
6	8.3%	0.0006	41	40.5%	0.0066
7	-39.0%	0.0182	42	34.6%	0.0022
8	-25.1%	0.0020	43	1.7%	0.0001
9	1.1%	0.0040	44	0.0%	0.0039
10	4.4%	0.0066	45	3.9%	0.0304
11	-4.9%	0.0039	46	25.4%	0.0209
12	-8.9%	0.0070	47	25.6%	0.0062
13	-7.0%	0.0148	48	0.0%	0.0033
14	-15.0%	0.0108	49	-0.3%	0.0019
15	-10.7%	0.0087	50	66.6%	0.0346
16	-9.0%	0.1079	51	0.0%	0.0039
17	-11.3%	0.0247	52	-2.9%	0.0007
18	10.6%	0.0024	53	15.8%	0.0011
19	-23.2%	0.0001	54	0.0%	0.0026
20	-25.4%	0.0001	55	0.0%	0.0018
21	-80.7%	0.0002	56	11.6%	0.0057
22	13.4%	0.0005	57	0.0%	0.0042
23	-42.5%	0.0010	58	0.0%	0.0236
24	-34.8%	0.0014	59	-1.5%	0.0015
25	-30.0%	0.0126	60	0.0%	0.0003
26	8.0%	0.0530	61	11.7%	0.0067
27	0.0%	0.0208	62	0.0%	0.0012
28	2.1%	0.0119	63	0.8%	0.0040
29	11.3%	0.0208	64	2.0%	0.0009
30	0.7%	0.0322	65	2.3%	0.0018
31	9.5%	0.0447	66	4.7%	0.0026
32	11.5%	0.0018	67	0.9%	0.0064
33	5.8%	0.0174	68	-1.0%	0.0309
34	-6.9%	0.0197	69	-0.5%	0.0005
35	0.0%	0.0124	70	0.0%	0.0006

Для индекса цен целевая переменная $Y_k = W_k P_k$.

Этот пример будет использоваться при сравнении планов ПСВ без ВО и ВПР, а также связанные с ними оценки.

Стандартная оценка индекса цен, основанная на основе ПСВ без ВО, представляет собой оценку по отношению, определяемую как:

$$\hat{I}_{SRSWOR} = \frac{\sum_{k=1}^{n} w_k p_k}{\sum_{k=1}^{n} w_k} \quad \textbf{5.15}$$

Хотя оценка по отношению здесь не обсуждается, часто используемая аппроксимация дисперсии оценки по отношению:

$$\text{var}(\hat{I}_{SRSWOR}) = \text{var}\left[\frac{\sum_{k=1}^{n} w_k p_k}{\sum_{k=1}^{n} w_k}\right] = \text{var}\left[\frac{1/n\sum_{k=1}^{n} w_k p_k}{1/n\sum_{k=1}^{n} w_k} - I\right] = \text{var}\left[\frac{1}{n}\sum_{k=1}^{n}\frac{w_k (p_k - I)}{\overline{w}_s}\right]$$

Поскольку дисперсия знаменателя в приведенном выше выражении относительно мала по сравнению с волатильностью числителя, знаменатель обычно заменяется средним значением совокупности W_k, т.е. $1/N$. Это дает:

$$\text{var}(\hat{I}_{SRSWOR}) = N^2 \text{ var}\left[\frac{1}{n}\sum_{k=1}^{n} w_k (p_k - I)\right] = \frac{N^2}{n}(1-\frac{n}{N})(\frac{1}{N-1})\sum_{k=1}^{N} W_k^2 (P_k - I)^2 \quad \textbf{5.16}$$

Во второй части (5.16) формула (2.12) была использована с:

$$Y_k = W_k (P_k - 1) \ \wedge \ \overline{Y}_p = 0$$

Применение этой формулы к 70 компаниям в таблице выше:

$$\text{var}(\hat{I}_{SRSWOR}) = 101 \quad \textbf{5.17}$$

Если выборка с возвращением, то дисперсия без поправки на конечную совокупность равна:

$$\text{var}(\hat{I}_{SRSWR}) = 116$$

Другими словами, поправка на конечную совокупность вычисляется:

$(1 - f) = 0{,}87.$

Формулы дисперсии в (5.16) и (5.17) теперь можно сравнить с дисперсией второй оценки, полученной на основе систематической выборки ВПР, где вероятности включения пропорциональны W_k. При этом:

$$\pi_k = W_k n$$

Оценка ВПР показателя I в соответствии с (5.3) теперь равна:

$$\hat{I}_{PPS} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \sum_{k=1}^{n} \frac{w_k y_k}{w_k n} = \frac{1}{n} \sum_{k=1}^{n} p_k = \overline{p}_s$$

Другими словами, в выборке ВПР средневзвешенное значение отдельных изменений цен в совокупности оценивается с помощью невзвешенного среднего значения изменений цен в выборке.

Сначала изучим дисперсию для выборки ВПР с возвращением. По формуле (5.11) рассчитывается:

$$\text{var}_{HH}(\hat{I}_{PPS}) == \frac{\sigma_z^2}{n} = \frac{1}{n} \sum_{k=1}^{N} W_k (P_k - I)^2 = 44$$

Затем получаем результат аппроксимации дисперсии, лежащего в основе (5.14):

$$\text{var}\ (\hat{I}_{PPS}) == \frac{1}{n^2} = \sum_{k=1}^{N} \pi_k (1 - \pi_k)(P_k - I)^2 = \frac{1}{n} \sum_{k=1}^{N} W_k (1 - nW_k)(P_k - I)^2 = 29$$

Другими словами, в выборке ВПР поправка на конечную совокупность 0,66 (=29/44) опять же, намного меньше, чем 0,87 для ПСВ без ВО. В таблице далее приведены результаты расчета дисперсии.

Таблица 1

Метод	**С возвращением**	**Без возвращения**
ВПР	116	101
ПСВ	44	29

Наконец, следует отметить, что моделирование, в котором 20 000 систематических выборок ВПР объемом n = 9 были отобраны из совокупности 70 компаний в различной случайной последовательности в каждом случае, привело к дисперсии 30 для показателя $\hat{I}_{PPS}$. Другими словами, (5.14) ВПР в этой ситуации является почти несмещенной оценкой дисперсии. Этот пример также еще раз иллюстрирует, что снижение

дисперсии в случае систематической выборки ВПР по сравнению с ПСВ без ВО может быть существенным, примерно до 70% .

Показатели качества

Показателями качества отбора ВПР являются:
- пределы неопределенности соответствующих оценок;
- уровень неответов;
- случайность последовательности элементов в совокупности и степень зависимости между Y_k/X_k и X_k.

Отсутствие ответа может серьезно повлиять на качество результатов, если (1) размер неответа большой и (2) неответ является выборочным. Часть смещения в связи с выборочным неответом иногда может быть скорректировано с помощью вспомогательных переменных, которые соответствуют как вероятности ответа, так и целевой переменной.

Еще одно важное предположение при случайной выборке состоит в том, что основа, из которой извлекается выборка, тесно связана с целевой совокупностью, о которой должны быть сделаны выводы.

Все вышеизложенное было сосредоточено, в частности, на выборке ВПР, в которой элементы находятся (или могут рассматриваться как находящиеся) в случайном порядке.

Если нет заметной связи между Y_k/X_k и k (k =1,…,N), случайное упорядочение совокупности не окажет систематического влияния на результаты, рассмотренные ранее. Поэтому случайный порядок не является абсолютно необходимым в такой ситуации. Если элементы расположены в фиксированной последовательности, выборки не должны становиться однородными с точки зрения Y_k^* ($\equiv Y_k/\pi_k$).

VII. Список использованной литературы

1. Sampling theory: Sampling design and estimation methods, Statistics Netherlands, 2012
2. Sampling Techniques, 3th Edition, William G. Cochran
3. Survey sampling reference guidelines: Introduction to sample design and estimation techniques, Eurostat, 2008 edition
4. Sampling algorithms, Yves Tille, 2006, Institut de Statistique, Switzerland