Отбор хозяйствующих субъектов с вероятностью, пропорциональной их размеру

С.В. Степанов, *кандидат социологических наук, консалтинговая компания Планова-Системз*. postmaster@planova.ru

Аннотация

В настоящей работе рассматривается группа алгоритмов формирования выборочной совокупности фиксированного размера с неравными вероятностями отбора (probability proportional to size - PPS), а именно с вероятностями, пропорциональными размеру единиц совокупности. Алгоритмы предназначены для реализации их в выборочных обследований крупных и средних предприятий, субъектов малого предпринимательства и индивидуальных предпринимателей в практике статистической работы для повышения качества результатов выборочного наблюдения и минимизации информационной нагрузки на респондентов.

Совокупность инструментальных средств, используемых, чтобы получить выборку из совокупности обычно называют *выборочной методологией*, а специфические процедуры и алгоритмы отбора единиц наблюдения в выборку называют *дизайном выборки* или *планом выборки*. Объединение процедуры отбора, плана выборки и процедуры получения оценок показателей по выборочным данным называют *выборочной стратегией*.

Термин вероятность отбора определяется как «вероятность, распределённая всем и каждой единице совокупности, с которой она может быть отобрана в некотором определённом отборе». Как правило, единицам совокупности распределяются не нулевые вероятности отбора.

Термин вероятность включения единицы совокупности определяется как «*полная вероятность, назначенная единице совокупности, с которой она может быть включена в выборку во всех отборах*». Вероятность включения *i*-ой единицы, как правило, обозначают π_i.

Вероятностные выборки подразделяются также на отборы с возвращением и без возвращения. В случае отбора *с возвращением* выборка размера *n* формируется с помощью *n* отборов из полной совокупности, поэтому в выборке могут быть отобраны одни и те же единицы совокупности несколько раз. Выборка *без возвращения* гарантирует, что в выборке размера *n* будут присутствовать *n* уникальных единиц совокупности. Нас в этой работе будут больше интересовать алгоритмы отборов без возвращения, которые обеспечивают меньшую дисперсию оценок показателей.

Выводы и вычисление оценок по плану выборки при использовании идеи рандомизации опираются на вероятности отбора в выборку. Как правило, это вероятности отбора первого и второго порядков, которые изменяются от единицы к единице для первого порядка и от пары к паре - для второго порядка. Основные свойства плана выборки – это базовые

оценки, которые определяются в терминах и на основе повторных выборок.

Мы рассмотрим группу методов формирования выборочной совокупности фиксированного размера с неравными вероятностями отбора (probability proportional to size - PPS) - а именно: с вероятностями, пропорциональными размеру единиц совокупности, и представим их в виде понятных алгоритмов отбора, которые можно легко перенести в статистическую практику или превратить в соответствующие модули выборочного программного обеспечения.

1. Алгоритмы отбора с вероятностью, пропорциональной размеру единицы наблюдения

Методы отбора с вероятностью, пропорциональной размеру единицы совокупности (ВПР-отбор или πps-отбор) могут быть особенно эффективны в применении их для государственного статистического наблюдения за предприятиями и индивидуальными предпринимателями. Как показывает российская практика обследований предприятий и, в особенности, предприятий-субъектов малого предпринимательства, чем меньше размер предприятия, тем меньше вероятности получить от него достоверные данные экономического характера по следующим причинам:

- Возникновения случаев полного неответа,
- Получение ответа, не отражающего реальные экономические показатели,
- Получение пустого ответа, или как говорят «нулевого» отчёта,

В условной подсовокупности мелких по размеру (численности работников) предприятий сосредотачивается наибольшая активность демографических процессов предприятий: прекращение экономической активности, появление новых предприятий, исчезновение закрывшихся предприятий. Основа выборки в этой условной подсовокупности быстро

теряет актуальность и предприятия, попадающие в выборку из этой подсовокупности реально малопригодны для обследования.

С учётом этого контекста, формирование выборочных совокупностей методами отбора вероятностью, пропорциональной размеру единицы совокупности имеет важные для нас свойства: в выборку с большей вероятностью попадут наиболее крупные предприятия, которые несут на себе основную экономическую нагрузку и, соответственно, большую долю в экономических показателях, которые мы собираемся оценивать по результатам выборочного наблюдения.

Основные понятия отбора с пропорциональными вероятностями рассмотрим на следующем примере, представленном в Таблице 1.

Таблица 1 – Отбор одного предприятия с вероятностью, пропорциональной размеру из гипотетической совокупности из шести предприятий

Номер предприятия	**Идентификатор предприятия**	**Размер предприятия** y_i	**Вероятность отбора** $\pi_i = y_i / \sum_i^N y_i$
1	A	8	8/30
2	B	6	6/30
3	C	3	3/30
4	D	5	5/30
5	E	4	4/30
6	F	4	4/30
Итог	-	30	1

Представим гипотетическую совокупность из 6-ти предприятий с таким показателем их размера, как численность работников. Допустим, мы хотим отобрать только одно предприятие из совокупности с вероятностью, пропорциональной размеру для оценки суммарного размера по совокупности (общего числа работников). Если, конечно, размеры всех предприятий совокупности нам известны, то нет нужды производить отбор в выборку для оценки общего размера совокупности, но это нам нужно для

иллюстрации логики отбора с вероятностью, пропорциональной размеру (ВПР). Если предприятия отбираются с ВПР, вероятность отбора одного предприятия есть его размер, делённый на общий размер совокупности (30). Эти вероятности показаны в последней колонке Таблицы 1. Заметим, что сумма вероятностей равна 1, как это и должно быть.

В простой случайной выборке вероятность отбора одной единицы в одном отборе равна 1/*N*, где *N* есть общее количество единиц совокупности, а несмещённая оценка суммы по совокупности *Y* для исследуемой переменной получается умножением значения для отобранной единицы (y_i) на общее количество единиц совокупности *N*, или другими словами, при делении y_i на вероятность отбора (=1/*N для простой случайной выборки*) единицы *i*. Подобно этому, при отборе с неравными вероятностями, несмещённая оценка *Y* получается из *i*-го отбора делением значения *i*-ой единицы (y_i) на вероятность отбора π_i (которая меняется в совокупности от единице к единице). Эта несмещённая оценка суммы по совокупности *Y* при отборе с неравными вероятностями есть

$$\hat{y}_i = y_i / \pi_i \tag{1}$$

Если значения y_i были известны для всех единиц совокупности перед проведением отборов и отбор был произведён с вероятностью, пропорциональной y_i , а именно

$$\pi_i = y_i / \sum_i^N y_i \tag{2}$$

Тогда несмещённая оценка $\hat{y}_i$ с учётом выражения (1) получается

$$\hat{y}_i = y_i / \pi_i = \sum_i^N y_i = Y$$

для суммы по совокупности. В нашем примере пусть единственный отбор даёт предприятие *c* , вероятность отбора для него 3/30, тогда несмещённая оценка общего размера совокупности по этому отобранному

предприятию 3/(3/30)=30, что соответствует актуальному итогу по совокупности.

Часто случается, что вместо отбора с вероятностью, пропорциональной актуальному признаку, мы хотим осуществлять отбор с вероятностью, пропорциональной некоторой вспомогательной переменной размера z_i , которая связана с признаком единицы (y_i) точным отношением

$$z_i = \beta\, y_i$$

где β – положительная константа. Тогда вероятность отбора

$$p\ z_i = z_i \Big/ \sum_i^N z_i = \beta y_i \Big/ \beta \sum_i^N y_i = y_i \Big/ \sum_i^N y_i = p\ y_i$$

останется той же самой и даст те же результаты, как и при отборе с вероятностью, пропорциональной размеру изучаемого признака (y_i), то есть не приведёт к ошибке выборки.

В связи с изложенным, обосновано предпочтение выбора метода отбора в пользу отбора с ВПР. Как правило, мы не можем знать актуальную величину изучаемого признака, но мы можем найти вспомогательную переменную или признак, значения которого нам известны и также известно, что он связан с изучаемым признаком примерно пропорционально. В этом случае мы можем применить отбор с вероятностями, пропорциональными размеру вспомогательного признака и получаемые по такой выборке оценки будут значительно эффективнее, чем при простой случайной выборке. Вспомогательную переменную следует искать среди таких, значения которых известны до проведения отбора, во-первых, и во-вторых линия регрессии которых со связанной изучаемой переменной проходит через начало координат (0, 0). Если имеет место существенная положительная корреляция между исследуемой и вспомогательной переменными, но линия регрессии не проходит через точку (0, 0), отбор с ВПР по вспомогательной переменной может вовсе не

дать большей эффективности оценок по сравнению с простой случайной выборкой.

Для выборки *s* фиксированного размера рассмотрим π-оценку Горвица-Томпсона

$$\hat{Y}_{\pi} = \sum_s y_k / \pi_k \tag{3}$$

Предположим, что удалось построить план выборки фиксированного размера без возвращения и реализующий его алгоритм отбора, такой, что

$$y_k / \pi_k = c, \quad k=1, \ldots, N \tag{4}$$

где *c* есть константа. Тогда, для любой выборки *s* мы имели бы

$$\hat{Y}_{\pi} = nc$$

где *n* это фиксированный объём выборки *s*.

Так как $\hat{Y}_{\pi}$ не варьируется от выборки к выборке, то величина дисперсии этой оценки была бы равна 0.

Так как в случае плана с фиксированным объемом дисперсия оценки:

$$V(\hat{Y}_{\pi}) = -\frac{1}{2} \sum \sum_U \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \text{ , где } \Delta_{kl} = \pi_{kl} - \pi_k \pi_l \text{ ,}$$

тогда, если пропорциональность (4) строго выполняется, то это выражение равно 0.

Очевидно, что план и соответствующий ему алгоритм отбора, удовлетворяющий (4) найден быть не может, поскольку это требует полных данных по всем y_k . Однако, предположим, что нам доступна вспомогательная переменная *x*, про которую известно, что она приблизительно пропорциональна *y*. Тогда выбор π_k, пропорциональных известным значениям x_k приводит нас к выводу о том, что отношение y_k / π_k есть приблизительно константа. Как результат получаем, что дисперсия π-оценки тоже будет малой.

Для совокупностей с большой вариацией x не всегда может быть получена выборка в строгом соответствии с правилом пропорциональности ВПР $\pi_k \sim x_k$. Необходим фактор пропорциональности $n / \sum_U x_k$ (поскольку $\sum_U \pi_k = n$ - фиксированный объём выборки). Другими словами,

$$\pi_k = \frac{n\, x_k}{\sum_U x_k} \tag{5}$$

Теперь должно удовлетворяться условие $\pi_k \le 1$. Если n = 1 это справедливо для всех k. Если $n > 1$ и некоторые x_k имеют очень большие значения, для некоторых единиц может оказаться справедливым

$$\frac{n\, x_k}{\sum_U x_k} > 1$$

В противоречие с требованием $\pi_k \le 1$.

Есть метод, чтобы избежать этого конфликта. Мы можем положить π_k = 1 для всех k, таких, для которых $n\, x_k > \sum_U x_k$ и положить π_k пропорциональными x для оставшихся элементов k. Таким образом, для желательного фиксированного объёма выборки n возьмём

$$\pi_k = (n - n_A)\frac{x_k}{\sum_{U-A} x_k}$$

для $k \in U - A$, где A есть множество из n_A элементов, таких, что $n\, x_k > \sum_U x_k$. Если необходимо, эта процедура повторяется до тех пор , пока не станут все $\pi_k \le 1$. (Далее будем просто предполагать, что $\frac{nx_k}{\sum_U x_k} \le 1$, для всех единиц совокупности $k \in U$).

Ниже перечислены желаемые свойства, которыми должна обладать схема отбора с вероятностями пропорциональными размеру единиц.

- Относительно простая реализация схемы на практике.

- Вероятности включения первого порядка π_k $(k = 1,...,N)$ строго пропорциональны x_k.
- Вероятности включения второго порядка $\pi_{kl} > 0$ для всех $k \neq l$ (необходимо для получения несмещенной оценки дисперсии.)
- π_{kl} в соответствии со схемой могут быть получены без привлечения сложных вычислений.
- $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l < 0$ для всех $k \neq l$ для гарантии, что оценка дисперсии по известной формуле Сена–Йетса-Гранди, задаваемая выражением:

$$\hat{V}(\hat{Y}_\pi) = -\frac{1}{2}\sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \qquad (6)$$

всегда принимает неотрицательное значение.

1.1. Алгоритм систематического отбора

Метод систематического отбора с ВПР является самым популярным, так как его реализация в качестве схемы отбора или алгоритма достаточно проста. Метод базируется на разделении совокупности на подсовокупности, основанные на некотором интервале отбора, связанном с объёмом выборки и отборе с ВПР из каждой подсовокупности одной единицы.

Положим $T_0 = 0$ и $T_k = T_{k+1} + x_k$, $k = 1, \ldots, N$, то есть посчитаем нарастающие итоги по совокупности. Определим интервал отбора a, где a положительное целое. Пусть n есть целая часть от T_N / a, где $T_N = \sum_U x_k$. Тогда

$$T_N = \sum_U x_k = na + c$$

где c удовлетворяет $0 \leq c < a$. Если $c = 0$, будет отобрана выборка размером n. Если $c > 0$, размер выборки будет n или $n + 1$.

Примем что $nx_k \leq T_N - c = na$ для всех k, которые связаны предположением, как обсуждалось выше. Для простоты мы примем, что

каждый x_k представлен целым числом или, с практически незначительными последствиями, как округлённое до ближайшего целого.

Описание процедуры систематического ВПР-отбора может быть представлено так:

Выберем с равной вероятностью 1 / a целое число, скажем, r , в интервале между 1 и a включительно.

Сформированная выборка тогда будет

$$s=\{k:T_{k-1}<r+(j-1)a\leq T_k \text{ для некоторых } j=1,2,...,n_s\}=s_r$$

где объём выборки n_s есть $n + 1$ (если $r \leq c$) или n (если $c < r \leq a$).

Мы можем описать эту схему с помощью изображения интервалов x_k следующих один за другим по горизонтальной шкале, начинающейся с нулевой точки и заканчивающейся значением $T_N = \sum_U x_k$. Если $c = 0$, вся дистанция T_N делится на n интервалов одинаковой длины a. Начало случайным образом выбирается в первом интервале, и мы систематически продолжаем двигаться дальше, отбирая каждый следующий элемент k , определяемый в точке попадания в него через постоянный интервал a.

Систематический ВПР-отбор подразумевает план выборки с по существу фиксированным объёмом выборки, поскольку n_s есть n или $n + 1$ и вероятности включения в выборку

$$\pi_k = \frac{n\, x_k}{T_N - c}$$

Т.е. мы имеем план с вероятностью включения, пропорциональной размеру.

Однако при всей своей простоте систематический ВПР-отбор порождает те же проблемы, что и систематический отбор с равными вероятностями:

- управление объемом выборки,
- выбор подходящего способа ранжирования основы выборки,

- проблема оценки дисперсии для π-оценки.

Почти все остальные существующие схемы для *πps*-выборки фиксированного объема $n>2$ (с π_k строго пропорциональными x_k) сложны. Большинство из них основывается на последовательных извлечениях, и при возрастании объема выборки вычисление вероятностей второго порядка быстро становится громоздким, в ряде случаев совершенно не осуществимом на практике. Даже для наиболее применимых схем остаётся фактом огромное число вычислений, включая построение и сохранение π_{kl}, вероятностей включения второго порядка, необходимые для оценки дисперсии при *πps* отборе.

Исходя из этих соображений, для прикладной реализации целесообразно опираться на схему отбора единиц с неравными вероятностями без возвращения с заданными вероятностями включения первого порядка, которая бы не содержала указанных недостатков. К числу таких схем относится предложенный И. Тийе метод на основе процедуры исключения единиц из совокупности на каждом шаге отбора, который обеспечивает точный фиксированный объем выборки, независящий от порядка единиц в совокупности, при этом совместные вероятности включения (вероятности включения второго порядка) необходимые для расчета оценки дисперсии вычисляются относительно просто.

1.2. Алгоритм отбора Тийе с исключением единиц совокупности

Основная идея метода Тийе заключается в последовательном исключении из совокупности отбора единицы, отобранной с ВПР на каждом шаге отбора, до тех пор, пока не будет достигнут установленный объём выборки. Вероятности отбора, пропорциональные размеру единиц пересчитываются после каждого исключения отобранной единицы.

1. Последовательные шаги нумеруются, начиная с номера *N*-1.

2. На шаге N-1 для каждой единицы совокупности рассчитываются величины

$$\pi(i/N-1)=\frac{(N-1)x_i}{\sum_{l\in U}x_l},(i\in U) \tag{7}$$

где x_i - значения размера единицы. Из совокупности отбирается одна единица с вероятностью $1-\pi(i|N-1)(i\in U)$ из U.

3. На начало N-2-го шага оставшаяся совокупность (или выборка) состоит из N-1 элемента. На N-2 шаге i-ая единица отбирается с вероятностью, задаваемой следующим выражением:

$$r_{N-1i}=1-\frac{\pi(i|N-2)}{\pi(i|N-1)}(i\in U) \tag{8}$$

Выбранная единица исключается из совокупности.

Таким образом, на начало k-го шага выборка состоит из k+1 единиц и на этом шаге очередная единица отбирается с вероятностью, задаваемой следующим выражением:

$$r_{ki}=1-\frac{\pi(i|k)}{\pi(i|k+1)}(i\in U)$$

Выбранная единица исключается из совокупности и к концу k-го шага только k единиц остается в выборке. Алгоритм останавливается в конце n –го шага.

Если на некотором k шаге выражение $\frac{kx_i}{\sum_U x_i}>1$, что противоречит требованию $\pi_i\le 1.$, то $\pi_i(k)=1$ для всех i, у которых $kx_i>\sum_U x_i$ и для всех остальных элементов совокупности, то есть:

$\pi_i=(k-k_A)\frac{x_i}{\sum_{U-A}x_i}$ для всех $k\in U-A$, где A – множество элементов, для которых $kx_i>\sum_U x_i$. Если необходимо, то процедура повторяется до тех пор, пока все π_i не станут меньше или равными 1.

Тогда (8) можно переписать следующим образом:

$$r_{kl} = \begin{cases} 0, i \in A_k \\ 1 - \pi(i|k), i \in B_k \\ \dfrac{1 - \sum_{i \in B_k} \{1 - \pi(i|k)\}}{k + 1 - \# A_k - \# B_k}, i \in C_k \end{cases}$$

Для $i \in U$, где $A_k = \{i : \pi(i|k) = 1\}$, $B_k = \{i : \pi(i|k) < 1 \ u \ \pi(i|k+1) = 1\}$, $C_k = \{\pi(i|k+1) < 1\}$. # - размер множества.

Отбор единицы:

Для того чтобы отобрать единицу из совокупности с заданной вероятностью r_i можно использовать метод накопленных сумм, следующим образом:

Присвоим $T_0 = 0$ и вычислим $T_i = T_{i-1} + r_i, i = 1,..., N$

Используя равномерное распределение на отрезке [0,1] извлечем случайное число ε, *i*-ый элемент отбирается, если выполняется следующее неравенство: $T_{i-1} \le \varepsilon T_N \le T_i$

Единица будет извлечена с заданной вероятностью r_i, так как $\Pr(T_{i-1} < \varepsilon T_N \le T_i) = \dfrac{T_i - T_{i-1}}{T_N} = \dfrac{r_i}{T_N} = r_i,$ так как в нашем случае $T_N = 1$.

Пример.

Предположим, что мы имеем совокупность, состоящую из N=5 элементов с заданными вероятностями отбора $\pi_1 = 0.1$ $\pi_2 = 0.25$ $\pi_3 = 0.05$ $\pi_4 = 0.4$ $\pi_5 = 0.2$

Для извлечения единицы построим следующую таблицу:

I	π_i	T_i	***Интервал отбора***
1	0.1	0.1	0-0.1
2	0.25	0.35	0.11-0.35

3	0.05	0.4	0.36-0.4
4	0.4	0.8	0.41-0.8
5	0.2	1	0.8-1

Извлечём случайное число ε, равномерно распределённое на [0, 1]. Предположим, что ε=0.56. Тогда из совокупности, будет выбрана единица с номером 4, так как 0.41<ε=0.56<0.8.

Оценивание: Оценка (Горвица–Томпсона) суммарного значения исследуемого признака рассчитывается по следующей формуле

$$\hat{Y}_\pi = \sum_h \sum_{s_h} \breve{y}_i = \sum_h \sum_{s_h} y_i / \pi_i$$

Оценка дисперсии.

Для расчёта оценки дисперсии по формуле Сена–Йетса-Гранди, задаваемой выражением:

$$\hat{V}(\hat{Y}_\pi) = -\frac{1}{2} \sum \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \qquad (9)$$

необходимо предварительно рассчитать совместные вероятности включения второго порядка. Согласно И. Тийе их можно получить исходя из вероятностей исключения:

$$\pi(i_1, i_2 | n) = \prod_{k=n}^{N-1} (1 - r_{ki_1} - r_{ki_2}) \qquad (10)$$

1.3. Алгоритм отбора Пуассона

Отбор Пуассона (PO) с выборкой случайного размера является обобщением отбора Бернулли. Пусть π_k есть предопределённая положительная вероятность включения k-й единицы, где $k = 1,\ldots,N$. Отбор Пуассона может быть определён следующим образом: пусть заданы индикаторы принадлежности I_k единицы к выборке, независимые, распределённые как

$$\Pr(I_k = 1) = \pi_k , \qquad \Pr(I_k = 0) = 1 - \pi_k$$

$k = 1,\ldots,N$. План выборки *PO* это такая выборка s , которая имеет вероятность

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U-s} (1-\pi_k)$$

где $s \in S$, множество из всех 2^N подмножеств U. Вследствие независимости, $\pi_{kl} = \pi_k \pi_l$ для всех $k \neq l$. Поскольку π_k может быть определено различными путями, метод отбора Пуассона порождает целый класс выборочных планов и соответствующих им процедур.

Для заданного множества вероятностей включения $\pi_1, \ldots, \pi_N$, план Пуассона имеет простую реализацию с помощью последовательного списка. Пусть $\varepsilon_1, \ldots, \varepsilon_N$ есть независимые случайные числа, равномерно распределенные на отрезке [0,1] Если $\varepsilon_k < \pi_k$, единица отбирается в выборку, в противном случае – нет.

При отборе Пуассона выборка размера n_s случайна, со средним

$$E_{PO}(n_s) = \sum_U \pi_k$$

и вариацией

$$V_{PO}(n_s) = \sum_U \pi_k (1 - \pi_k),$$

оценка суммы по совокупности

$$\hat{Y}_\pi = \sum_s \breve{y}_k = \sum_s y_k / \pi_k,$$

дисперсия оценки

$$V_{PO}(\hat{Y}_\pi) = \sum_U \pi_k (1 - \pi_k) \breve{y}_k^2 = \sum_U (\frac{1}{\pi_k} - 1) y_k^2 , \qquad (11\text{-}14)$$

несмещённая оценка дисперсии

$$\hat{V}_{PO}(\hat{Y}_\pi) = \sum_s (1-\pi_k) \breve{y}_k^2$$

Здесь уместно отметить, что $\hat{V}_{PO}(\hat{Y}_\pi)$ может быть необычно большой, потому, что размер выборки изменяется. При отборе Бернулли план полностью определён, потому что мы устанавливаем ожидаемый размер выборки. По контрасту при отборе Пуассона больше значения придается выбору π_k , по сравнению с ожидаемым размером выборки. Какой же вариант лучше? Ответ достигается минимизацией дисперсии (11)

для заданного объема выборки, $n = \sum_U \pi_k$. Это эквивалентно минимизации выражения

$$(\sum_U y_k^2 / \pi_k)(\sum_U \pi_k)$$

Но по неравенству Коши-Шварца,

$$(\sum_U y_k^2 / \pi_k)(\sum_U \pi_k) \geq (\sum_U \pi_k)^2$$

С равенством в случае, если и только $y_k / \pi_k = \lambda$ - константа. Учитывая, что $y_k > 0$ для всех k, мы имеем $\pi_k = y_k / \lambda$. Наконец, поскольку $n = \sum_U \pi_k$, мы имеем

$$\pi_k = n\, y_k / \sum_U y_k \qquad (12)$$

$k = 1,\ldots,N$, добавляя также, что $y_k \leq \sum_U y_k / n$ для всех k.

Теперь, поскольку y_k неизвестны, решение, данное в формуле (12) имеет исключительно академический интерес. Однако в некоторых обследованиях мы имеем доступ к одной или больше вспомогательным переменным, значения которых известны для всех единиц совокупности. Пусть $x_1, \ldots, x_N$ известные положительные значения вспомогательной переменной x. Также следует добавить, что y *приблизительно* пропорциональна x. В этом случае мы можем получить π_k пропорциональные известным x_k. Так, для $k = 1,\ldots,N$

$$\pi_k = n\, x_k / \sum_U x_k \qquad (13)$$

С учётом, что $x_k \leq \sum_U x_k / n$ для всех k. (Если $x_k > \sum_U x_k / n$ мы должны взять $\pi_k = 1$.) Если y_k / x_k, близко к константе, результирующая оценка имеет очень маленькую дисперсию. Вероятности включения, определённые по формуле (13) полностью соответствуют правилу «*пропорциональности размеру*». Значение x_k представляет собой меру размера k-ой единицы. Типичными показателями размера предприятий являются количество работников, выручка, торговый оборот и т.п.

Несмотря на то, что эти рассуждения верны, есть одно обстоятельство, а именно, что отбор Пуассона имеет тот же недостаток, что и отбор Бернулли – это случайный размер выборки. Для иллюстрации этого недостатка представим, что оказалось возможным выбрать π_k оптимальными в соответствии с (12), с ожидаемым размером выборки $n = \sum_U \pi_k$. В этом экстремальном случае оценка равна

$$\hat{Y}_\pi = \sum_s y_k / \pi_k = (n_s / n) \sum_U y_k = (n_s / n) Y \qquad (14)$$

Следовательно, дисперсия $\hat{Y}_\pi$ от выборки к выборке будет просто состоять из дисперсии объёма выборки n_s.

Этот аргумент заставляет нас ожидать выполнения исключительно хорошей оценки вида $\hat{Y}_\pi = \sum_s \breve{y}_k$, если возможно создание плана с заданным объёмом выборки и вероятностями включения π_k, такими, что они почти точно пропорциональны y_k. Если π_k наиболее близки к пропорциональности с y_k, то при плане с заданным объёмом выборки, оценка будет иметь нулевую дисперсию.

Можно указать альтернативную оценку для $Y = \sum_U y_k$

$$\hat{Y}_{alt} = N \frac{\sum_s \breve{y}_k}{\hat{N}}$$

где $\hat{N} = \sum_s (1/\pi_k)$

Аппроксимация дисперсии тогда есть

$$V(\hat{Y}_{alt}) = \sum_U \frac{(y_k - \bar{y}_k)^2}{\pi_k} - N S_{yU}^2 \qquad (15)$$

которая обычно меньше, чем дисперсия (14). Таким образом, $\hat{Y}_{alt}$ обычно предпочитают вместо $\hat{Y}_\pi$.

Для преодоления главного недостатка отбора Пуассона, а именно, случайный характер размера выборки, предлагаются специальные схемы

модификации отбора Пуассона, например, *последовательный отбор Пуассона*. Рассмотрим его подробнее.

1.4. Алгоритм последовательного отбора Пуассона

Пусть есть совокупность U с известной для каждой единицы вспомогательной переменной размера единицы p_i , заданной так, что

$$\sum_1^N p_i = 1$$

Мы хотим осуществить выборку единиц с вероятностями, пропорциональными p_i

$$\Pr(i \in s) = n\ p_i, \quad i = 1,2,...,N \tag{16}$$

где $i \in s$ означает, что единица i включена в выборку s и n ожидаемый размер выборки. В дальнейшем будем называть обычный отбор Пуассона ОП, а последовательный отбор Пуассона ПОП. Для выполнения условия (16) мы будем считать, что

$$n p_i \le 1, \quad i = 1,2,...,N$$

На практике это всегда может быть достигнуто, если особо крупные единицы выделяются в отдельную страту, которая попадает в выборку вся.

Целью обследования является оценивание сумма исследуемой переменной $\mathbf{y} = (y_1, y_2, \ldots, y_N)$

$$Y = \sum_1^N y_i$$

Заметим, что вероятность получить пустую выборку

$$\Pr(m = 0) = \prod_{i=1}^N (1 - n p_i) \le e^{-n}$$

Нельзя рекомендовать применение ОП, если эта вероятность значительна, будем считать, что это не так. Несмещённая оценка Горвица-Томпсона для Y

$$\hat{Y}_{HT} = \frac{1}{n}\sum_{i \in s}\frac{y_i}{p_i}$$

Выражение для дисперсии $\hat{Y}_{HT}$

$$Var(\hat{Y}_{HT}) = \frac{1}{n}\sum_{i=1}^{N}(1 - n p_i)(\frac{y_i}{p_i})^2 p_i$$

Аппроксимирующее выражение для дисперсии

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{N}(1 - n p_i)(\frac{y_i}{p_i} - Y)^2 p_i \qquad (17)$$

Для преодоления недостатков простого ОП Олссоном (1999) предлагается следующий приём. Из случайных чисел X_i мы формируем *изменённые случайные числа*

$$\xi_i = X_i / p_i \qquad (18)$$

Правило отбора Пуассона выполняется просто: единица *i* включается в выборку, если и только $\xi_i \le n$. Эта формулировка ОП предлагает несколько иной способ отбора, при котором отбирается *n* единиц, имеющих наименьшие ξ_i. Такой отбор называют *последовательным отбором Пуассона* (ПОП) с объёмом выборки *n*, содержащей *n* единиц с наименьшими значениями изменённых случайных чисел ξ_i, определённых в (18).

ПОП был предложен Олссоном в 1990 году. В своих работах он на примерах показал, что, к сожалению, ПОП не является строго ВПР. По близкому отношению этого метода к отбору Пуассона естественно, однако, предположить, что он является всё же приблизительно ВПР. Результаты моделирования на гипотетических совокупностях дают существенные свидетельства в пользу этой догадки. Это приводит нас к необходимости в связи с ПОП рассматривать следующую оценку

$$\hat{Y}_S = \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i}$$

Считаем, что оценка $\hat{Y}_S$ для ПОП, при общих условиях, приблизительно нормально распределена со средним Y и дисперсией σ^2.

Таким образом, это наше суждение приводит к тому, что $\hat{Y}_S$ приблизительно является несмещённой.

Замечание 1. Если все p_i равны, ПОП представляет собой простую случайную выборку без возвращения (ПСВБВ). В этом случае выражение (17) сокращается до хорошо известной формулы для дисперсии для ПСВБВ-оценки для Y , за исключением коэффициента (N-1)/N . Чтобы «калибровать» σ^2 против известного «стандарта», достаточно умножит на коэффициент коррекции N/(N-1)

$$\sigma^2 = \frac{1}{n}\frac{N}{(N-1)}\sum_{i=1}^{N}(1-np_i)(\frac{y_i}{p_i}-Y)^2 p_i$$

ОП, с другой стороны, не эквивалентен ПСВБВ с равными вероятностями. Однако, его дисперсия аппроксимирует дисперсию ПСВБВ. Следовательно, коррекция σ^2 на N/(N-1) может быть применена и для ОП.

Оценка дисперсии. Для законченности мы представим оценки дисперсии для ОП и ПОП без теоретических деталей. Брюер и Ханиф (Brewer and Hanif, 1983, с. 83) предлагают следующую «обычную оценку» для дисперсии ОП

$$\hat{v}(\hat{Y}_R) = \frac{1}{n^2}\sum_{i\in S}(1-np_i)(\frac{y_i}{p_i}-\hat{Y}_R)^2 + \Pr(m=0)\hat{Y}_R^{\,2} \qquad (19)$$

Брюер и Ханиф утверждают, что «более устойчивая оценка может быть получена умножением первой части выражения на *n/m*». Это возможно только если $m > 0$. Мы продолжаем считать, что Pr(m=0)

пренебрежимо мало и опустим правую часть выражения. Рассуждая как в *Замечании 1*, мы корректируем $v(\hat{Y}_R)$ на величину $n/(m\text{-}1)$, считая $m>1$ (так как это необходимо для оценки дисперсии в любом случае). Это заставляет нас рассматривать следующую оценку дисперсии для ОП (оставляя её неопределённой для $m \le 1$).

$$v(\hat{Y}_R) = \frac{1}{n(m-1)} \sum_{i \in S} (1 - n p_i)\left(\frac{y_i}{p_i} - \hat{Y}_R\right)^2 \qquad (20)$$

«Обычная оценка» для ПОП была бы как в выражении (19) без правого члена. Снова, с помощью «калибровки к ПСВББ» мы умножаем выражение на величину $n(n\text{-}1)$. Таким образом, мы получаем следующую оценку дисперсии для ПОП,

$$v(\hat{Y}_S) = \frac{1}{n(n-1)} \sum_{i \in S} (1 - n p_i)\left(\frac{y_i}{p_i} - \hat{Y}_S\right)^2 \qquad (21)$$

которая в случае равных вероятностей сокращается до обычной несмещённой оценки дисперсии для ПСВББ. Исследования на моделируемых совокупностях подтверждают правомерность использования «калиброванных» оценок (20) и (21).

Некоторые замечания для применения на практике.

При выполнении ПОП на практике нам необходима не обязательно нормированная вспомогательная переменная размера, так как умножение p_i на константу оставляет ПОП-выборку неизменной. Следовательно, ПОП-выборка может быть сформирована следующим простым способом. Сначала при проходе по всей совокупности генерируется ξ_i. Если вспомогательная переменная размера не нормирована, следует предварительно вычислить и использовать её сумму *S*. Потом совокупность следует отсортировать по ξ_i. Первые *n* единиц сортированного списка составляют выборку. С помощью *S* мы проверяем $n\ p_i \le 1$. Выделение некоторых единиц в страту, которая входит в выборку

целиком, из-за нарушения этого условия не изменяет выборку. Единицы, обязательно попадающие в выборку, должны должным образом быть учтены в процессе оценивания.

Вследствие сортировки, такая процедура не является ни списко-последовательной ни отборо-последовательной в смысле классификации Эрика Сандала и др. (1992). После сортировки, ПОП может считаться последовательным в обоих отношениях, то есть мы отбираем следующую единицу, просто беря её последовательно из сортированного списка. При списко-последовательной процедуре ОП мы должны были бы пройти по всему списку ещё раз, что неудобно в случае большого регистра единиц. Именно это преимущество и заложено в названии *последовательный* ОП.

Дополнительной положительной чертой класса процедур, основанных на отборе Пуассона является то, что в алгоритмах заложена генерация и назначение единицам совокупности последовательных случайных чисел, которые потом могут быть использованы для обновления и расширения выборок, а также для положительной и отрицательной координации неоднократных выборок и управлением степени пересечения выборок.

2. Апробация методов отбора

2.1. Апробация систематического алгоритма

Проиллюстрируем этот алгоритм на следующем примере. Возьмём совокупность из 29 предприятий для которых из прошлого обследования известны величины выручки в тыс.руб, которые мы намерены использовать в качестве переменной размера единицы совокупности.

В соответствии с алгоритмом рассчитаем нарастающие итоги и результаты отобразим в Таблице 2.

Таблица 2 – Отбор 10-ти предприятий с вероятностью, пропорциональной размеру из гипотетической совокупности из 29 предприятий

Предприятие	Выручка	Накопленная выручка	Значение индекса отбора с шагом интервала	Вероятность включения

			отбора	
1	542	542		0,307622
2	245	787		0,139054
3	1032	1819	1321	0,585731
4	867	2686		0,492082
5	256	2942		0,145298
6	352	3294	3083	0,199784
7	835	4129		0,47392
8	645	4774		0,366082
9	427	5201	4845	0,242352
10	312	5513		0,177082
11	1342	6855	6607	0,761678
12	390	7245		0,221352
13	604	7849		0,342812
14	465	8314		0,26392
15	897	9211	8369	0,509109
16	476	9687		0,270163
17	365	10052		0,207163
18	967	11019	10131	0,548839
19	533	11552		0,302514
20	215	11767		0,122027
21	1590	13357	11893	0,902435
22	423	13780	13655	0,240082
23	645	14425		0,366082
24	867	15292		0,492082
25	423	15715	15417	0,240082
26	197	15912		0,111811
27	586	16498		0,332595
28	365	16863		0,207163
29	756	17619	17179	0,429082
Итого	17619			10

Сформируем выборку из 10-ти предприятий.

Накопленный итог T_N: 17619

Объём выборки n: 10

Интервал отбора a : 17619 / 10 = 1762

Случайное число r от 0 до 1762: 1321

Серия индексов отбора: r = 1321

$r + a$ = 3083

$r + 2a = 4845$

$r + 3a = 6607$

$r + 4a = 8369$

$r + 5a = 10131$

$r + 6a = 11893$

$r + 7a = 13655$

$r + 8a = 15417$

$r + 9a = 17179$

Таким образом, в выборку попадают предприятия №№ 3, 6, 9, 11, 15, 18, 21, 22, 25, 29.

Поскольку совокупность разбивается интервалом отбора на непересекающиеся группы, и из каждой отбирается по одной единице, этот алгоритм гарантирует план выборки без возвращения.

Существует несколько модификаций этого алгоритма, направленных на специфические цели обследования или адаптированные для специфических условий. Если единицы предварительно отсортировать по размеру, то это приведёт к тому, что в выборке гарантированно окажутся предприятия из всех групп по размеру, включая самые мелкие. Для неотсортированного ряда по размеру такого условия добиться нельзя.

2.2. Результаты апробации алгоритма отбора Тийе

Результаты реализации этого алгоритма подробно представлены на примере одного слоя по данным наблюдения предприятий Республики Коми, предприятий Смоленской области.

В Таблицах 3 и 6 представлены данные из Генеральной совокупности по предприятиям: показатели выручка, оборот и численность, а также данные наблюдения показателей численность занятых, выручки, фонд заработной платы, оборот. В этих же таблицах представлены оценки этих показателей, рассчитанные на основе проведенного методом Тийе выборочного отбора с вероятностью, пропорциональной показателю оборот (выручка), а также относительные ошибки выборки.

В Таблицах 4 и 7 - рассчитанные на каждом шаге алгоритма вероятности включения каждой из единиц совокупности:

В Таблице 5 и 8 представлены вероятности исключения единиц из совокупности. В соответствии с этими вероятностями на каждом шаге была исключена 1 единица (её вероятность исключения отмечена жирным шрифтом). В результате выполнения этого алгоритма остались соответственно 5 и 4 единицы, с вероятностями включения определенными в Таблицах 4 и 7. (соответствующие строки выделены темным цветом).

Для получения оценки дисперсии (9) были рассчитаны совместные вероятности включения в соответствии с формулой (10). И они представлены в Таблице 9. В Таблице 10 – оценки дисперсии и коэффициента вариации.

Таблица 3 – Расчёт оценок показателей и их точности (крупные м средние предприятия Смоленской области)

номер	ОКВЭД	Средняя численность работников	Выручка	Оборот	Оборот Обследования	π_i_4	вес	Взвешенный оборот	Взвешенный оборот обследования	Взвешенная численность	Взвешенное значение выручки
3 476	14,21	17	0	2 715	2 687,00	,	0,00	0,00	0,00	0,00	0,00
1 888	14,21	21	3 638	3 638	7 082,00	,	0,00	0,00	0,00	0,00	0,00
3 477	14,21	17	0	15 704	8 589,00	,	0,00	0,00	0,00	0,00	0,00
3 475	14,21	41	0	31 173	23 997,60	0,19	5,14	160295,57	123398,74	210,83	0,00
1 467	14,21	67	20 318	34 778	27 950,00		0,00	0,00	0,00	0,00	0,00
498	14,21	230	97 748	40 850	44 993,50		0,00	0,00	0,00	0,00	0,00
3 652	14,21	83	70 134	62 745	66 465,00		0,00	0,00	0,00	0,00	0,00
669	14,21	241	97 841	128 987	110 086,00	0,80	1,24	160296,02	136807,18	299,50	121589,95
112	14,21	337	177 590	230 989	199 988,00	1,00	1,00	230989,00	199988,00	337,00	177590,00
1 365	14,21	482	260 359	287 177	261 945,20	1,00	1,00	287177,00	261945,20	482,00	260359,00
сумма		1 536	727 628	838 756	753 783	,		838757,59	722139,12	1329,33	559538,95
относительная ошибка выборки		13,46%	23,10%	0,00%	4,20%						

Таблица 4 – Вероятности включения π(*i*/*k*)

номер	π_i_9	π_i_8	π_i_7	π_i_6	π_i_5	π_i_4
3 476	0,43	0,12	0,06	0,04	0,03	0,02
1 888	0,57	0,16	0,08	0,06	0,04	0,02
3 477	1	0,71	0,37	0,25	0,16	0,10
3 475	1	1	0,73	0,49	0,33	0,19
1 467	1	1	0,81	0,54	0,36	0,22
498	1	1	0,95	0,64	0,43	0,25
3 652	1	1	1	0,98	0,65	0,39
669	1	1	1	1	1	0,80
112	1	1	1	1	1	1,00
1 365	1	1	1	1	1	1,00

Таблица 5 – Вероятности исключения r(*i*/*k*)

номер	r_9	r_8	r_7	r_6	r_5	r_4
3 476	0,57	**0,71**	-	-	-	-
1 888	**0,43**	-	-			
3 477	0,00	0,29	**0,49**			
3 475	0,00	0,00	0,27	0,33	0,33	0,40
1 467	0,00	0,00	0,19	**0,33**		
498	0,00	0,00	0,05	0,33	0,33	**0,40**
3 652	0,00	0,00	0,00	0,02	**0,33**	
669	0,00	0,00	0,00	0,00	0,00	0,20
112	0,00	0,00	0,00	0,00	0,00	0,00
1 365	0,00	0,00	0,00	0,00	0,00	0,00

Таблица 6 – Расчёт оценок показателей и их точности (крупные и средние предприятия Республики Коми)

ОКПО	ОКВЭД	Инвестиции	Средняя численность работников	Средняя численность списочного состава	ФЗП	Выручка	Вероятность включения	Выборочный вес	Взвешенные значения				
									Инвестиции	Средняя численность работников	Средняя численность списочного состава	ФЗП	Выручка
77 893 391	18,30,2	224,0	7	7	11,7	12,7	0,04	0,00	0	0	0	0	0
16 935 098	18,22,3	320,00	7,00	6,00	28,20	59,00	0,06	0,00	0,00	0,00	0,00	0,00	0,00
51 537 368	18,22,3	345,20	13,00	4,00	5,00	0,00	0,07	0,00	0,00	0,00	0,00	0,00	0,00
12 898 118	18,22,1	362,70	3,00	2,00	32,40	89,10	0,07	0,00	0,00	0,00	0,00	0,00	0,00
16 936 258	18,22,3	418,80	18,00	14,00	101,50	124,00	0,08	0,00	0,00	0,00	0,00	0,00	0,00
12 881 632	18,24,4	472,00	5,00	5,00	73,00	120,00	0,09	0,00	0,00	0,00	0,00	0,00	0,00
57 445 153	18,23,2	492,00	14,00	15,00	77,00	108,00	0,09	0,00	0,00	0,00	0,00	0,00	0,00
16 935 081	18,22,3	505,00	10,00	10,00	59,00	112,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00
28 880 818	18,22,3	505,10	9,00	9,00	28,40	91,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00
15 094 225	18,22,3	562,40	6,00	6,00	23,00	95,10	0,11	0,00	0,00	0,00	0,00	0,00	0,00
24 953 383	18,22,3	766,00	11,00	11,00	167,60	190,30	0,15	0,00	0,00	0,00	0,00	0,00	0,00
12 881 655	18,22,2	774,30	12,00	11,00	150,00	253,00	0,15	0,00	0,00	0,00	0,00	0,00	0,00
29 668 138	18,22,3	801,00	27,00	25,00	106,20	178,00	0,15	0,00	0,00	0,00	0,00	0,00	0,00
71 093 577	18,22,2	1032,00	9,00	8,00	130,00	73,00	0,20	0,00	0,00	0,00	0,00	0,00	0,00
24 954 342	18,22,3	1069,90	9,00	9,00	219,70	245,40	0,20	4,92	5261,73	44,26	44,26	1080,48	1206,87
29 667 239	18,22,3	1474,00	15,00	15,00	247,00	349,80	0,28	0,00	0,00	0,00	0,00	0,00	0,00
24 954 419	18,22,3	1595,90	20,00	19,00	257,90	367,80	0,30	0,00	0,00	0,00	0,00	0,00	0,00
28 885 460	18,22,1	1677,00	17,00	16,00	227,80	214,00	0,32	3,14	5261,73	53,34	50,20	714,74	671,44

24 954 402	18,22,3	2387,90	29,00	29,00	529,20	570,00	0,45	2,20	5261,73	63,90	63,90	1166,09	1255,99
1 825 760	18,22,3	4672,00	14,00	18,00	121,70	3001,00	1,00	1,00	4672,00	14,00	18,00	121,70	3001,00
53 708 207	18,21	17915,00	70,00	63,00	918,00	3135,00	1,00	1,00	17915,00	70,00	63,00	918,00	3135,00
сумма		38372,20	325,00	302,00	3514,30	9388,20		оценка	38372,20	245,50	239,36	4001,01	9270,31
относительная ошибка выборки		0,0%	24,5%	20,7%	13,8%	1,3%							

Таблица 7 – Вероятности включения $\pi(i/k)$

ОКПО	π _i_20	π _i_19	π _i_18	π _i_17	π _i_16	π _i_15	π _i_14	π _i_13	π _i_12	π _i_11	π _i_10	π _i_9	π _i_8	π _i_7	π _i_6	π _i_5
77 893 391	0,54	0,43	0,37	0,32	0,27	0,24	0,21	0,18	0,16	0,13	0,12	0,10	0,09	0,07	0,06	0,04
16 935 098	0,77	0,61	0,53	0,46	0,39	0,34	0,30	0,26	0,22	0,19	0,17	0,14	0,12	0,10	0,08	0,06
51 537 368	0,83	0,66	0,57	0,49	0,42	0,37	0,32	0,28	0,24	0,21	0,18	0,15	0,13	0,11	0,09	0,07
12 898 118	0,87	0,70	0,60	0,52	0,44	0,39	0,34	0,29	0,25	0,22	0,19	0,16	0,14	0,11	0,09	0,07
16 936 258	1,00	0,80	0,70	0,60	0,51	0,45	0,39	0,34	0,29	0,25	0,22	0,19	0,16	0,13	0,11	0,08
12 881 632	1,00	0,91	0,79	0,67	0,58	0,50	0,44	0,38	0,33	0,28	0,25	0,21	0,18	0,15	0,12	0,09
57 445 153	1,00	0,94	0,82	0,70	0,60	0,53	0,46	0,40	0,34	0,29	0,26	0,22	0,19	0,16	0,12	0,09
16 935 081	1,00	0,97	0,84	0,72	0,62	0,54	0,47	0,41	0,35	0,30	0,26	0,23	0,19	0,16	0,13	0,10
28 880 818	1,00	0,97	0,84	0,72	0,62	0,54	0,47	0,41	0,35	0,30	0,26	0,23	0,19	0,16	0,13	0,10
15 094 225	1,00	1,00	0,94	0,80	0,69	0,60	0,52	0,46	0,39	0,34	0,29	0,25	0,21	0,18	0,14	0,11
24 953 383	1,00	1,00	1,00	1,00	0,94	0,82	0,71	0,62	0,53	0,46	0,40	0,34	0,29	0,24	0,19	0,15
12 881 655	1,00	1,00	1,00	1,00	0,95	0,83	0,72	0,63	0,54	0,46	0,40	0,35	0,29	0,25	0,20	0,15
29 668 138	1,00	1,00	1,00	1,00	0,98	0,86	0,74	0,65	0,56	0,48	0,42	0,36	0,30	0,25	0,20	0,15
71 093 577	1,00	1,00	1,00	1,00	1,00	1,00	0,95	0,84	0,72	0,62	0,54	0,46	0,39	0,33	0,26	0,20
24 954 342	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,87	0,74	0,64	0,56	0,48	0,41	0,34	0,27	0,20
29 667 239	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,88	0,77	0,66	0,56	0,47	0,37	0,28
24 954 419	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,95	0,83	0,71	0,61	0,51	0,40	0,30
28 885 460	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,88	0,75	0,64	0,53	0,42	0,32
24 954 402	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,91	0,76	0,61	0,45
1 825 760	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
53 708 207	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Таблица 8 – Вероятности исключения r(*i*/*k*)

ОКПО	r_20	r_19	r_18	r_17	r_16	r_15	r_14	r_13	r_12	r_11	r_10	r_09	r_08	r_07	r_06	r_05
77 893 391	**0,46**															
16 935 098	0,23	0,2	0,13	0,14	0,14	0,13	0,13	0,13	0,14	0,14	0,13	0,14	0,15	0,17	0,2	**0,25**
51 537 368	0,17	0,2	0,13	0,14	0,14	**0,13**										
12 898 118	0,13	**0,2**														
16 936 258	0	0,2	0,13	0,14	0,14	0,13	0,13	0,13	0,14	0,14	0,13	0,14	**0,15**			
12 881 632	0	0,09	0,13	0,14	0,14	0,13	0,13	0,13	0,14	0,14	0,13	0,14	0,15	**0,17**		
57 445 153	0	0,06	0,13	0,14	0,14	0,13	**0,13**									
16 935 081	0	0,03	**0,13**													
28 880 818	0	0,03	0,13	0,14	0,14	0,13	0,13	0,13	**0,14**							
15 094 225	0	0	0,06	**0,14**												
24 953 383	0	0	0	0	0,06	0,13	0,13	0,13	0,14	0,14	**0,13**					
12 881 655	0	0	0	0	**0,05**											
29 668 138	0	0	0	0	0,02	0,13	0,13	**0,13**								
71 093 577	0	0	0	0	0	0	0,05	0,13	0,14	**0,14**						
24 954 342	0	0	0	0	0	0	0,01	0,13	0,14	0,14	0,13	0,14	0,15	0,17	0,2	0,25
29 667 239	0	0	0	0	0	0	0	0	0	0,12	0,13	0,14	0,15	0,17	**0,2**	
24 954 419	0	0	0	0	0	0	0	0	0	0,05	0,13	**0,14**				
28 885 460	0	0	0	0	0	0	0	0	0	0	0,12	0,14	0,15	0,17	0,2	0,25
24 954 402	0	0	0	0	0	0	0	0	0	0	0	0	0,09	0,17	0,2	0,25
1 825 760	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
53 708 207	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Таблица 9 – Совместные вероятности включения π(k/l)

π(1,2)	0,047676936
π(1,3)	0,071721248
π(1,4)	0,201751755
π(1,5)	0,201751755
π(2,3)	0,113883264
π(2,4)	0,32035344
π(2,5)	0,32035344
π(3,4)	0,45318
π(3,5)	0,45318
π(4,5)	1

Таблица 104 – Характеристики точности, оценки

Оценка дисперсии показателя выручка	оценка коэффициента вариации
201040,441	4,84%

3. Заключение

Задача выбора метода отбора и превращение его в алгоритм должна решаться в соответствии со спецификой задач конкретного обследования и возможностями исполнительского ресурса.

Самым популярным является систематический алгоритм, так как он нагляден, убедителен и достаточно легко реализуем на практике. Однако при всей своей простоте систематический отбор с неравными вероятностями порождает те же проблемы, что и систематический отбор с равными вероятностями: управление объёмом выборки, вопрос о подходящим упорядочиванием основы выборки, проблема оценки дисперсии для π-оценки.

Достаточно прост в освоении и реализации алгоритм последовательного отбора Пуассона. Кроме того, в его алгоритме

заложена генерация и назначение единицам совокупности последовательных случайных чисел, которые потом могут быть использованы для обновления и расширения выборок, а также для положительной и отрицательной координации неоднократных выборок и управлением степени пересечения выборок. Однако, как показал Олссен в своих теоретических разработках, он не соответствует точному определению отбора с вероятностями, пропорциональными размеру единиц.

Метод Тийе более сложен в реализации по сравнению с вышеуказанными методами. Как всякий шаговый алгоритм он требует существенных затрат вычислительного ресурса, однако в нём алгоритмически заложен и поэтому проще обеспечивается принцип отбора без возвращения. Важным преимуществом метода Тийе является возможность получать выборку заданного объёма и, кроме, того, он не требует дополнительных процедур для специального отслеживания нетипичных единиц, отделений их в специальную страту с вероятностью отбора, равной единице, так как такой результат уже заложен непосредственной в алгоритм последовательного исключения единиц.

Почти все остальные существующие схемы для *πps*- выборки фиксированного объема $n>2$ (с π_k строго пропорциональными x_k) сложны. Большинство из них основывается на последовательных извлечениях, и при возрастании объема выборки вычисление вероятностей второго порядка быстро становится громоздким, в ряде случаев совершенно не осуществимом на практике. Даже для наиболее применимых для них остаётся фактом огромное число вычислений, включая построение и сохранение π_{kl}, вероятностей включения второго порядка, необходимые для оценки дисперсии при *πps* отборе.

Исходя из этих соображений, а также из анализа результатов апробации вышеперечисленных методов для реализации статистического

наблюдения предприятий комбинированным методом, по нашему мнению, целесообразно опираться на предложенный И. Тийе метод на основе процедуры исключения единиц из совокупности на каждом шаге отбора, который обеспечивает ***точный фиксированный объем выборки, не зависящий от порядка единиц в совокупности***, и для показателей, коррелированных с размером предприятия, - ***высокую точность оценивания***. При этом данный алгоритм предлагает не слишком сложный расчет характеристик точности наблюдения: совместные вероятности включения (вероятности включения второго порядка) необходимые для расчета оценки дисперсии вычисляются относительно просто.

4. СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. **Brewer, K.R.W., and Hanif M.** (1983) Sampling with Unequal Probabilities, New York, Springer.
2. **Carl-Erik Sarndal Bengt Swensson Jan Wretman**, Model Assisted Survey Sampling, 2003 Springer-Verlag New York, Inc
3. **Deville, J.-C., Särndal, C.-E. and Sautory, O.** (1993) Generalized Raking Procedures in Survey Sampling, Journal of the American Statistical Association, Vol. 88, No. 423, 1013-1020. 2006.
4. **Ives Tille,** Sampling Algorithms 2006, Springer Science + Business Media, Inc.
5. **Hansen, M.H., and Hurwitz, W.N.** On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 1943.
6. **Horvitz, D.G., and Thompson, D.J**. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 1952.
7. **Ohlsson, E**. Sequential Poisson Sampling, Journal of Official Statistics, 1998.
8. **Renssen, R.H., and Nieuwenbroek, N.J.** Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 1997.
9. **Renssen, R.H., Kroese, A.H. and Willeboordse, A.J.** Aligning estimates by repeated weighting. *Report, Central Bureau of Statistics*, 2001, The Netherlands.
10. **Zheng, H., and Little, R.J.A.** Penalized spline model-based estimation of the finite population total from probability-proportional-to-size-samples. *Journal of Official Statistics*, 2003, 19.
11. **Zieschang, K.D.** Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 1990.
12. **Методологические положения по статистике**. Вып. 3 / М 54 Госкомстат России. – М., 2000.