социологический клуб **город**

Методическое пособие

Выборочный метод в социологии

Редактор-составитель:
Андрей Ермолаев

Москва 2000

Содержание

1. Введение.

Настоящее методическое пособие является результатом серии семинаров по теме «Выборочное исследование в социологии», проведенной Социологическим клубом «Город» Государственного университета – Высшей Школы Экономики.

Цель работы – составить общее представление о выборочном методе и о возможностях его применения в социологии. *Работа содержит* классификацию типов случайной и неслучайной выборки, описание каждого метода, их преимущества и недостатки. Для каждого типа случайной выборки приведены формулы расчета ошибки репрезентативности (выборочного среднего) и объема выборки.

Предполагается, что читатель знаком с основами теории вероятности и математической статистики.

Авторы выражают огромную благодарность *Ю. Н. Толстовой* и *А.О. Крыштановскому*, за помощь в проведении и организации семинаров, подборе материала и постоянные консультации.

2. Суть выборочного метода и его роль в социологии.

Одной из задач, которые стоят перед социологом при проведении исследования, является сбор необходимых эмпирических данных об объекте исследования. Множество элементов, составляющих объект исследования называют *генеральной совокупностью* (ГС). Наиболее простым, на первый взгляд, способом сбора данных является *сплошное обследование* ГС. Однако применение сплошного обследования не всегда представляется возможным. В этом случае применяется *выборочное обследование*. Суть выборочного метода заключена в том, что обследованию подвергается только часть элементов ГС, которая называется *выборочной совокупностью* (ВС). Как писал профессор *А. Кауфман* , «изобретателем выборочного была сама жизнь» [4]. Действительно, еще до теоретического обоснования возможностей применения выборочного метода, статистики были вынуждены проводить выборочные обследования. Основными причинами для этого были *отсутствие времени* и *средств*.

Выборочный метод позволяет не только *сократить временные и материальные затраты* на проведения исследования, но и *повысить достоверность результатов исследования* [6, 16]. Это утверждение может вызвать недоумение: как можно получить более достоверные данные, обследовав менее половины ГС? Достоверность полученной информации может быть не только не ниже, чем при сплошном обследовании, но и выше вследствие возможности привлечения персонала более высокого класса и применения различных процедур контроля качества получаемой информации.

Кроме того выборочный метод имеет *более широкую область применения* [там же]. Широта области применения выборочного метода объясняется тем, что небольшой (по сравнению с ГС) объем выборки позволяет использовать более сложные методы обследования, включая использование различных технических средств (например, видео- и аудиоаппаратуры).

Следует различать единицы отбора и единицы наблюдения. *Единицами отбора* являются единицы или группы единиц ГС отбираемые на каждом этапе формирования ВС. *Единицы наблюдения* – это отобранные единицы ГС, характеристики которых непосредственно измеряются. Если выборка проходит в несколько этапов (многоступенчатая выборка), то единицы отбора и единицы наблюдения могут не совпадать. Мы будем рассматривать только одноступенчатую выборку, т.е. выборку, проходящую в один этап.

Развитие *теории вероятностей* позволило теоретически обосновать возможность применения выборочного метода. В основе теоретического обоснования выборочного метода лежит так называемый *закон больших чисел*. Физический смысл этого закона можно выразить следующим образом:

«при очень большом числе случайных явлений средний их результат практически перестает быть случайным и может быть предсказан с большой степенью определенности» [, 399].

Также это дало возможность определять ошибку репрезентативности. *Репрезентативностью ВС* называется ее способность адекватно представлять (репрезентировать) характеристики ГС. *Ошибкой репрезентативности,* как правило, называют отклонение выборочного среднего значения признака от генерального. Важно учитывать, что при помощи выборочного метода никогда нельзя получить абсолютно точную оценку наблюдаемого признака, *всегда существует вероятность ошибки*, но, если вероятность ошибки мала, то она скорее всего не произойдет.

Разделяют два типа ошибок. *Случайная (статистическая) ошибка* – это ошибки, которые возникают вследствие случайной вариации значений, вызванной тем, что наблюдается только часть единиц, а не вся ГС [6, 379]. Случайные ошибки уменьшаются с увеличением объема ВС. Случайную ошибку можно измерить методами математической статистики, если при формировании ВС соблюдался принцип случайности. *Принцип случайности* заключается в следующем: каждый элемент ГС имеет равную и отличную от нуля вероятность попасть в ВС [9, 100]. Иными словами, термин «случайный» употребляется здесь и далее как синоним слова «равновероятный». Для соблюдения принципа случайности формирование выборочной совокупности должно проходить по строго определенным правилам, которые составляют метод формирования выборочной совокупности.

На практике принцип случайности соблюсти очень сложно, а иногда просто невозможно, что приводит к появлению систематической ошибки. *Систематическая ошибка* – это неконтролируемые перекосы в распределении выборочных наблюдений[1] [1, 132]. Число опрошенных не влияет на величину систематической ошибки.

Общая типология методов отбора представлена на рис. 1. Рассмотрим их.

[1] О причинах систематических ошибок см., например, [1, 132-168].

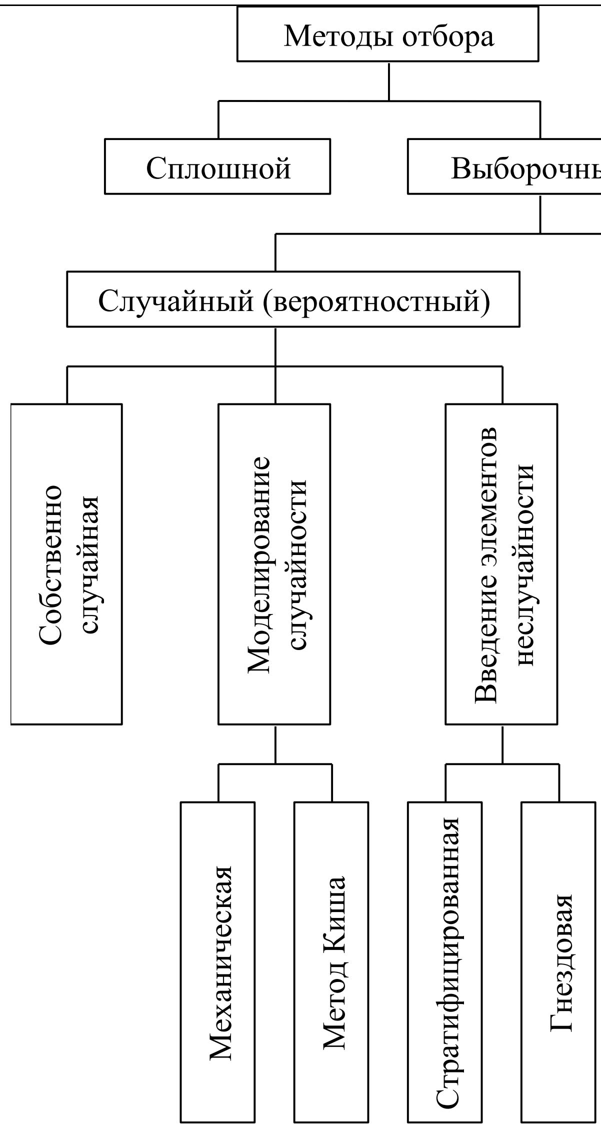

Рисунок 1. Типология методов отбора.

3. Случайные (вероятностные) методы отбора.

3.1 Собственно случайная выборка.

Собственно случайная выборка лежит в основе всех остальных типов выборки, которые будут рассмотрены далее.

3.1.1 Определение собственно случайной выборки.

Выборка называется *собственно случайной*, если при извлечении выборки объема n все возможные комбинации из n элементов, которые могут быть получены из генеральной совокупности объема N, имеют равную вероятность быть извлеченными [16]

По определению, при собственно случайной выборке выполняется принцип случайности.

3.1.2 Способы практической реализации собственно случайной выборки.

Отбор производится с помощью жеребьевки, таблицы (либо генератора) случайных чисел. *Главный принцип* – случайность, т.е. все единицы генеральной совокупности имеют равную вероятность попасть в выборочную совокупность.

1. *Принцип жеребьевки*. Каждый элемент генеральной совокупности заносится на бумажку (это могут быть фамилии, адреса, просто номера (в этом случае выпавшие номера ставят в

соответствие с людьми в списках) и т.д.), затем бумажки помещаются в барабан, перемешиваются и не глядя вытаскиваются.

2. *Принцип таблицы случайных чисел*. Начиная с любого места таблицы, берем четыре следующих друг за другом числа. Эти числа и будут номерами людей в списке, которых следует отобрать в выборку (числа, превышающие численность генеральной совокупности, опускаются) [1, 101].
3. *Принцип генератора случайных чисел*. Это то же самое, что и таблицы случайных чисел, только числа вырабатываются компьютером (для этого существует специальная программа).

Различают *повторную* и *бесповторную* выборку. При повторном отборе каждый выбранный элемент возвращается в ГС. При бесповторном отборе выбранный элемент не возвращается в ГС[2].

Также используются различные методы *моделирования случайности*.

1. *Механическая выборка* требует список характеристик респондентов (фамилии, адреса, телефоны и т.д.). Из этого списка через равные промежутки люди отбираются в выборку. Этот промежуток называется *шагом выборки*.

   $шаг = \frac{N}{n}$ [3, 19].

   Начало отбора выбирается случайным образом в пределах шага выборки. Например, если шаг выборки равен 20, то начинать отбор надо с любого числа от 1 до 20.
2. *Территориальный отбор* используется, когда нет основы выборки или ее составление сопряжено с большими трудностями [9, 104-111].

3.1.3 Вычисление ошибки репрезентативности для собственно случайной выборки.

Пусть нам необходимо оценить средний возраст некоторой группы людей по ограниченному числу наблюдений *n*. Оценкой среднего значения непрерывной случайной величины является *математическое ожидание*:

$$\overline{X} = M(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx.$$

Естественной оценкой математического ожидания является среднее арифметическое:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

От оценки необходимо потребовать следующие *свойства*:

1. *состоятельность* – оценка называется состоятельное, если при увеличении числа опытов оценка сходится по вероятности с искомым параметром,
2. *несмещенность* – оценка называется несмещенной, если выполнялось условие
$$M(\bar{x}) = \overline{X} \ ,$$
3. *эффективность* – оценка называется эффективной, если ее дисперсия минимальна по сравнению с другими.

Среднее арифметическое обладает этими свойствами[3].

Оценка параметра является функцией от случайных величин x_1, x_2, … , x_n, поэтому сама является случайной величиной. Другими словами, мы можем сделать множество выборок, для каждой из которых значение оценки будет различно. По закону больший чисел распределение оценки является нормальным с математическим ожиданием

$$M(\bar{x}) = \overline{X}$$

и дисперсией

[2] Следует заметить, что бесповторный отбор не отвечает принципу случайности. Это нарушение тем существеннее, чем меньше ГС. Однако на практике как правило применяется бесповторный отбор.

[3] Мы не будем приводить доказательство.

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}\ ^{4},$$

где σ^2 - генеральная дисперсия.

Тогда можно рассчитать вероятность того, что $\bar{x}$ попадет в интервал $[\bar{X} - \Delta, \bar{X} + \Delta]$. Поскольку нам неизвестна величина $\bar{X}$, то мы будем говорить о вероятности, с которой интервал $[\bar{x} - \Delta, \bar{x} + \Delta]$ накроет $\bar{X}$. Эта которая равна площади под графиком функции распределения случайной величины $\bar{x}$ (см. рис. 2):

$$P(|\bar{x} - \bar{X}| < \Delta) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \int_{\bar{X}-\Delta}^{\bar{X}+\Delta} e^{-\frac{(\bar{x}-\bar{X})^2}{2 \cdot \sigma^2}} d\bar{x} .$$

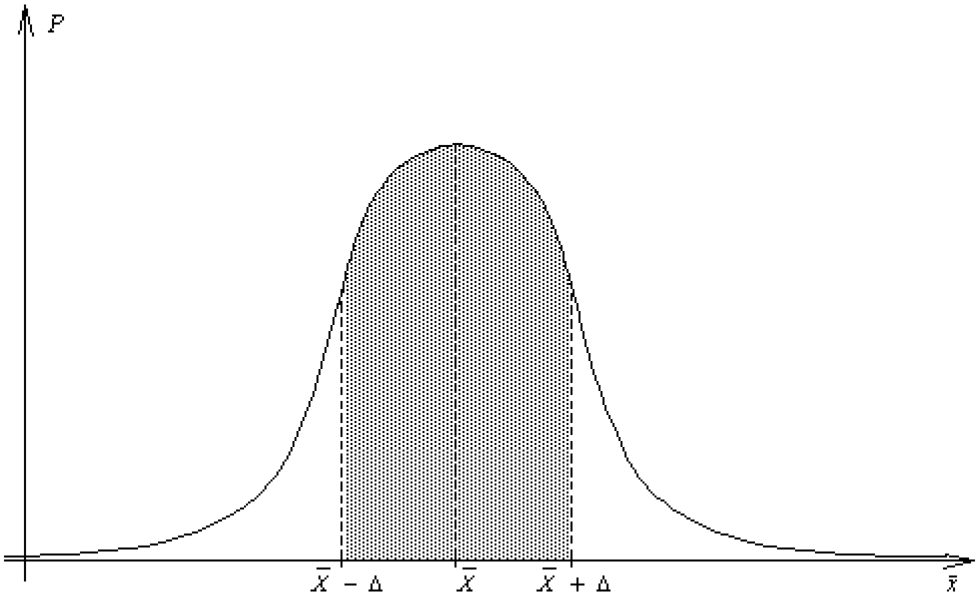

Рисунок 2. Распределение выборочной оценки среднего.

Приведем это распределение к стандартному виду.

$$z = \frac{\bar{x} - \bar{X}}{\sigma_{\bar{x}}}$$

$$\bar{x} = z \cdot \sigma_{\bar{x}} + \bar{X}$$

$$d\bar{x} = \sigma_{\bar{x}} \cdot dz$$

Произведем замену переменной:

$$P(|\bar{x} - \bar{X}| < \Delta) = \frac{1}{\sqrt{2 \cdot \pi}} \int_{-z}^{z} e^{-\frac{z^2}{2 \cdot}} dz = \frac{1}{\sqrt{2 \cdot \pi}} \int_{-z}^{0} e^{-\frac{z^2}{2}} dz + \frac{1}{\sqrt{2 \cdot \pi}} \int_{0}^{z} e^{-\frac{z^2}{2}} dz = 2 \cdot \Phi(z) .$$

Справа получили функцию Лапласа, которая табулирована (см. Приложение):

$$\Phi(z) = \frac{1}{\sqrt{2 \cdot \pi}} \int_{0}^{z} e^{-\frac{z^2}{2}} dz .$$

$$\Delta = z \cdot \sigma_{\bar{x}} = z \cdot \sqrt{\frac{\sigma^2}{n}}$$

---

[4] Мы не будем приводить доказательство этих соотношений.

Нам не известно значение σ, поэтому заменим его на s. Но в этом случае нужно использовать не нормальное распределение, а распределение Стьюдента.

$$\Delta = t \cdot \sqrt{\frac{s^2}{n}},$$

где $s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n-1}$

При больших объемах выборки вид распределения Стьюдента приближается к виду нормального распределения, поэтому для больших выборок также можно использовать функцию Лапласа.

Для повторной выборки

$$\Delta = t \cdot \sqrt{\frac{s^2}{n}} \text{ (1).}$$

Для бесповторной выборки необходимо внести поправку на конечность ГС

$$\Delta = t \cdot \sqrt{\frac{s^2 \cdot (1 - \frac{n}{N})}{n}} \text{ (2).}$$

Для большой ГС (объем ВС составляет менее 5% от ГС) поправкой на конечность совокупности можно пренебречь.

Про коэффициент доверия z следует сказать отдельно. Этот коэффициент исследователь выбирает сам. Чем меньше z, тем меньше доверительный интервал, но тем меньше и вероятность того, что оценка не выйдет за пределы доверительного интервала.

Пример 1. Пусть была произведена выборка 1600 человек. Средний возраст по выборке – 30 лет, среднеквадратическое отклонение – 10 лет. Необходимо найти доверительный интервал.

Прежде всего, необходимо задать надежность оценки. Возьмем 95% надежность. Поскольку выборка большая, воспользуемся таблицей значений функции Лапласа и найдем коэффициент доверия z - 1,96.

Тогда

$$\Delta = 1{,}96 \cdot \sqrt{\frac{10^2}{1600}} = 0{,}49 .$$

С вероятностью 95% истинное средний возраст по ГС находится в интервале от 29,51 лет до 30,49 лет.

Для биномиального распределения

$$s^2 = p \cdot q,$$

где p – доля признака, $q = (1 - p)$.

Тогда для повторной выборки из (1)

$$\Delta = t \cdot \sqrt{\frac{p \cdot q}{n}} \text{ (3),}$$

для бесповторной выборки из (2)

$$\Delta = t \cdot \sqrt{\frac{p \cdot q \cdot (1 - \frac{n}{N})}{n}} \text{ (4).}$$

Пример 2. Из 200 опрошенных 55% - женщины. Действуем аналогично примеру 1. Выборку также можно считать большой. Тогда z =1,96 для 95% надежности.

$$\Delta = 1{,}96 \cdot \sqrt{\frac{0{,}55 \cdot 0{,}45}{200}} = 0{,}07 .$$

С вероятностью 95% доля женцин в ГС находится в интервале от 48% до 62%.

Таблица 1.

Формулы ошибки репрезентативности для собственно случайного отбора.[3, 16]

Предмет изучения.	Повторный отбор.	Бесповторный отбор.
Среднее значение признака.	$\Delta = z\sqrt{\frac{s^2}{n}}$	$\Delta = z\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$
Доля признака.	$\Delta = z\sqrt{\frac{pq}{n}}$	$\Delta = z\sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}$

Где:
z – коэффициент доверия,
n – объем выборки,
s^2 - выборочная дисперсия,
N – объем генеральной совокупности,
p - доля признака в выборочной совокупности.

3.1.4 Определение объема выборки.

Определение объема выборки – это задача, обратная решенной выше задачи вычисления ошибки выборки.

Формулы для вычисления объема выборки при случайном отборе – просто преобразованные формулы ошибки репрезентативности. Они представлены в таблице 4.

Таблица 2.

Формулы для определения объема выборки при собственно случайном отборе.

Предмет изучения.	Повторный отбор.	Бесповторный отбор.
Среднее значение признака.	$n = \frac{z^2 s^2}{\Delta^2}$	$n = \frac{z^2 s^2 N}{\Delta^2 N + z^2 s^2}$
Доля признака.	$n = \frac{z^2 pq}{\Delta^2}$	$n = \frac{z^2 pqN}{\Delta^2 N + z^2 pq}$

Из (1) легко получить искомое n

$$n = \frac{1}{\frac{\Delta^2}{t^2 \cdot s^2}}.$$

Для нахождения объема выборки необходимо знать выборочное значение дисперсии признака. Его можно оценить несколькими способами [6, 95].

1. Отобрать некоторое количество n_1 единиц из ГС. Рассчитать по полученной ВС s^2. Рассчитать необходимый объем n ВС и добрать недостающее число элементов $n_2 = n - n_1$.
2. Воспользоваться результатами предыдущих исследований (если таковые проводились).
3. Для биномиального распределения $s^2 = p \cdot q$, где p - доля признака, $q = (1 - p)$. Тогда из (3)

$$n = \frac{1}{\dfrac{\Delta^2}{t^2 \cdot p \cdot q}}.$$

Произведение $p \cdot q$ максимально, когда $p = 0{,}5$. Таким образом, мы получаем выборку с некоторым запасом [10]:

$$n = \frac{t^2}{4 \cdot \Delta^2}$$

Точность и надежность выборки мы задаем, исходя из целей исследования. Например, насколько важное управленческое решение будет принято на основе результатов исследования.

3.1.5 Плюсы и минусы собственно случайной выборки.

Плюсом данного метода является полное соблюдения принципа случайности и, как следствие – избежание систематических ошибок.

Случайная выборка обладает рядом недостатков, которые затрудняют ее применение на практике. Эти недостатки можно представить в трех пунктах:

1. *Необходимость наличия списка элементов генеральной совокупности.* Обычно элементами генеральной совокупности являются люди; в этом случае в качестве списка могут выступать адреса, телефоны и т.д. Трудность здесь заключается в том, что получить такой список далеко не всегда представляется возможным. Следовательно, в тех случаях, когда невозможно получить список элементов генеральной совокупности, невозможно проводить и случайный отбор.

2. *Сложность проведения опроса.* Процедура опроса при случайном отборе является очень громоздкой и требующей много времени. Ведь в результате случайного отбора исследователь получает на выходе список фамилий респондентов (телефонов, адресов и т.д.), которых необходимо опросить. Иными словами, интервьюерам приходится «бегать» за каждым респондентом и добиваться от него согласия ответить на «парочку вопросов».

Осложняет дело и то, что респондентов порой бывает не так просто достать; в случае отсутствия респондента его приходится посещать по нескольку раз (по крайней мере не менее трех раз).

Все вышеперечисленное ведет к повышенным временным затратам на проведение опроса. Временные затраты можно уменьшить только благодаря привлечению дополнительных интервьюеров, т.е. только за счет дополнительных денежных расходов. Помимо этого возникает еще так называемая *проблема неответивших.*

3. *Сравнительно большой объем выборки.* Для получения результатов со сравнительно высокой степенью точности собственно случайный отбор требует достаточно большого объема выборки по сравнению с другими видами отбора. Другими словами, случайный отбор обладает меньшей степенью точности, что в конечном счете является причиной его меньшей *эффективности*[5].

Существует два способа повышения эффективности выборки, которые :

1. корректировка выборочных показателей,
2. использование методов построения выборки с внедрением *элемента неслучайности* [6, 34].

Рассмотрим их.

3.2 Корректировка выборочных показателей.

Как было сказано выше, корректировка выборочных показателей является одним из способов повышения эффективности выборки.

---

[5] Выборка считается более эффективной, если:

1. при одинаковых расходах она более точна.
2. при одинаковой точности она более дешевая. [6, 34]

В данном параграфе кратко проиллюстрируем принцип корректировки выборочных показателей. Корректировка применяется для повышения точности выборки при существующей методике отбора в выборочную совокупность. Это возможно только за счет привлечения дополнительной информации о генеральной совокупности за предыдущие периоды времени[6].

Необходимо отметить, что исследователю важно не столько, чтобы средняя всех выборочных показателей была равна генеральному показателю (т.е. была несмещенной), сколько, чтобы стандартная ошибка (дисперсия) всех возможных выборок была наименьшей, т.е., чтобы риск ошибиться в отдельной выборке был как можно меньше [6, 35].

Следующий пример продемонстрирует возможность применения корректировки результатов выборки[7].

Пример [6, 36-39]**:** Допустим, что нам необходимо выяснить средний доход на этот год. В качестве генеральной совокупности выступают 12 человек, представленные в таблице 3.

Таблица 3.

Распределение дохода гипотетических респондентов [6, стр.13].

Респондент.	Доход.
A.	1300
B.	6300
C.	3100
D.	2000
E.	3600
F.	2200
G.	1800
H.	2700
I	1500
J.	900
K.	4800
L.	1900
Общий доход.	**32100**
Средний доход.	**2675**

Пусть выборочная совокупность составляет 3 человека. Задача, таким образом, сводится к тому, чтобы по этим 3 людям оценить средний доход всей генеральной совокупности.

Допустим, что в нашу выборку, проведенную случайным методом, попали респонденты B,C,J. Средний доход в этом случае (равный 3433) оказывается намного выше аналогичного показателя в генеральной совокупности (2675). Это говорит о том, что в нашу выборку попали в основном богатые люди.

Применяя упомянутый принцип корректировки выборочных показателей, можно скорректировать результат расчета средней по случайной выборке. Для этого необходимо иметь

[6] Вообще любое повышение точности выборки возможно только благодаря наличию дополнительной информации о генеральной совокупности.

Эту информацию можно получить из постоянно проводящихся обследований Госкомстата РФ, а также из отдельных статистических, демографических и социологических исследований.

[7] В данном параграфе рассматривается лишь один из многих возможных способов. Задача состояла в том, чтобы показать саму возможность подобной методики.

информацию об уровне доходов за прошлый период (например, за прошлый год). Если предположить, что уровни доходов данного и предыдущего периодов коррелированы, то можно скорректировать выборочный показатель на показатель прошлого периода[8].

Для этого нам необходимо знать средний доход всей генеральной совокупности за прошлый год и доходы респондентов B,C и J за прошлый год. Допустим, что эти доходы оказались соответственно 5500, 3500, 1200 (средняя соответственно = 3400), а генеральная средняя = 2883.

Корректировка, следовательно, будет выглядеть следующим образом:

2883*3433/3400 = 2911.

2911 и будет скорректированным средним доходом всей генеральной совокупности в этом году согласно рассматриваемому принципу корректировки выборочных показателей. Как видно, он гораздо более «похож» на истинное значение средней.

Данную операцию можно провести для всех возможных выборок из 3 человек и получить среднюю и ошибку выборки. Эти данные представлены в следующей таблице.

Таблица 4.

Показатели случайной выборки и скорректированные показатели.

Показатель.	Случайная выборка.	Скорректированные показатели.
Средняя.	2675	2658
Стандартная ошибка.	786	240

Как видно из таблицы, средняя скорректированных средних является смещенной (не совпадает с генеральной средней), но зато ошибка выборки намного меньше. Это значит, что шанс получить «хорошую» выборочную оценку повышается.

Однако необходимо отметить, что в данном случае корректировку можно применять только в случае сохранения структуры доходов за данный и предыдущий годы. В противном случае данный метод может дать искаженные результаты.

И здесь опять происходит знакомый парадокс. Дело в том, что для того, чтобы узнать, пропорционально или непропорционально изменилась эта структура, нам нужно иметь данные о генеральной совокупности за этот год. А это как раз то (и даже больше), что мы хотим выяснить нашим исследованием. Иными словами, мы не можем достоверно узнать, насколько связаны структуры доходов за данный и прошлый год. Мы можем только предполагать (на основании статистических данных за много лет и тому подобным показателям), что структура доходов не претерпела значительных изменений за год.

4. Выборочные методы с внедрением элемента неслучайности.

Итак, рассмотрев вкратце один из методов корректировки, можно перейти непосредственно к рассмотрению типов (модификаций) собственно случайного отбора.

Использование различных типов случайного отбора позволяет несколько сгладить некоторые из вышеупомянутых трудностей, возникающих при проведении собственно случайного отбора. Например, некоторые типы случайного отбора позволяют упростить организацию опроса, но главное – это то, что они увеличивают эффективность выборки.

Так при случайном отборе ошибка выборки контролируется только за счет изменения объема выборки. В рассматриваемых же нами типах случайного отбора эффективность выборки можно повысить за счет моделирования выборки без увеличения ее объема.

---

[8] Под корреляцией здесь понимается то, что удельный вес каждой единицы генеральной совокупности остался примерно тот же, что и в прошлом году. Иными словами, имеется в виду пропорциональный рост или снижение доходов.

Под моделированием выборки понимается проведение случайного опроса с учетом информации о генеральной совокупности. Это означает, что по некоторым параметрам составляется модель генеральной совокупности для того, чтобы уже на стадии, предшествующей стадии случайного отбора, повысить соответствие этих параметров в выборке и генеральной совокупности[9].

Однако модификации случайного отбора не могут преодолеть всех трудностей, связанных со случайной выборкой. Это связано с тем, что все они являются *разновидностями именно случайного отбора* и в них используется принцип случайности.

Из этого следует, что проводить любой случайный отбор невозможно без списка элементов генеральной совокупности. Более того, большинство типов случайного отбора приводят к тем же трудностям при организации опроса, что и при собственно случайной выборке. Главное, чего достигают эти модификации случайного отбора, так это увеличения точности выборки.

Однако при формальном сходстве с собственно случайной выборкой, любая ее вариация есть все же некоторое отклонение от принципа случайности. Эти отклонения могут приводить к *систематическим ошибкам*, которые невозможны при собственно случайной выборке. Теперь непосредственно перейдем к рассмотрению типов случайного отбора.

4.1 Механическая выборка.

Наиболее близкой к собственно случайной выборке является механическая выборка. Однако даже она может приводить к систематическим ошибкам.

4.1.1 Практическая реализация.

Проведение механической выборки требует список характеристик респондентов (фамилии, адреса, телефоны и т.д.). Из этого списка через равные промежутки люди отбираются в выборку. Этот промежуток называется *шагом выборки*.

$шаг = \frac{N}{n}$ [3, 19], где

N – объем генеральной совокупности

n – объем выборочной совокупности.

Начало отбора выбирается случайным образом в пределах шага выборки. Например, если шаг выборки равен 20, то начинать отбор надо с любого числа от 1 до 20.

4.1.2 Вычисление ошибки выборки.

При определении ошибки репрезентативности используются те же формулы, что и при случайной выборке.

4.1.3 Определение объема выборки.

Как следствие, при определении объема выборки так же используются те же формулы, что и при случайной выборке.

4.1.4 Плюсы и минусы механического отбора.

Процедура проведения механической выборки менее громоздка, чем проведение случайной выборки. Хотя применение компьютеров практически нивелирует это преимущество.

Механическая выборка может быть как более точной, так и менее точной по сравнению со случайной выборкой. Это продемонстрирует следующий пример.

Пример: [6, 51-52]. Воспользуемся данными таблицы 1. Из всех респондентов проведем механическую выборку путем отбора каждого четвертого респондента, начиная с первого. В таблице 5 представлены четыре возможные выборки.

---

[9] Принцип моделирования выборки станет более понятен при непосредственном рассмотрениии модификаций случайной выборки.

Таблица 5.

Возможные выборки при механическом отборе.

№ группы.	**Респонденты, попавшие в группу.**
1.	A,E,I.
2.	B,F,J.
3.	C,G,K.
4.	D,H,L.

Если посчитать стандартное отклонение для этих четырех выборок и для всех возможных выборок при случайном отборе, то механическая выборка окажется точнее (510 против 786)[10].

Если же отбирать каждого третьего человека, то число возможных выборок окажется равным трем. Они представлены в следующей таблице.

Таблица 6.

Возможные выборки при механическом отборе.

№ группы.	**Респонденты, попавшие в группу.**
1.	A,D,G,J.
2.	B,E,H,K.
3.	C,F,I,L.

Здесь механическая выборка оказывается менее точной, чем случайная (1216 против 642).

Механическая выборка может обнаружить систему, что может привести к систематическим ошибкам. Возможности допущения систематической ошибки проиллюстрированы следующими примерами.

Пример: Если неправильно выбрать шаг выборки, то можно получить серьезные искажения полученных результатов. Например, если мы имеем список жителей г. Москвы в алфавитном порядке, то маленький шаг выборки приведет к перебору людей с фамилиями, начинающимися на букву «А» если мы начинаем отбор с начала списка. А если принять во внимание, что среди армян часто встречаются фамилии, начинающиеся на букву «А», то налицо смещение выборки (т.е. число армян в выборке будет завышенным).

Отсюда следует, что шаг выборки нельзя брать произвольно, а надо рассчитывать по указанной выше формуле. В нашем случае это обеспечит пропорциональное попадание в выборку людей с фамилиями, начинающимися на любую букву. Однако даже при правильно рассчитанном шаге выборки нельзя гарантировать невозможность систематической ошибки, т.к. уже в одной процедуре механического отбора заложена система. Это проиллюстрирует следующий пример.

Пример: Например, у нас есть списки всех жителей какого-то города по избирательным участкам. Тогда, делая механическую выборку из каждого списка, мы опять набираем слишком много людей с фамилией на букву «А», т.к. по обыкновению начинаем отбор с начала списка. [2, 169].

Чтобы исправить это обстоятельство, необходимо четко определить начало отбора на каждом избирательном участке. Начало отбора, например, может быть рассчитано по формуле: (к+6)/7, где к-номер избирательного участка (в данном примере от 1 до 700).

[10] При случайном отборе существует 495 возможных выборок при генеральной совокупности в 12 человек и при объеме выборки 4 человека. В нашем случае (при объеме выборки в 3 человека) число возможных выборок еще больше.

Таким образом, по мере роста номера избирательного участка, начало отбора будет сдвигаться «вглубь» списка.

Пример: Допустим, мы имеем город, состоящий из микрорайонов, и у нас есть адреса жителей микрорайонов, причем в списках адреса упорядочены по микрорайонам. Вроде бы ничто не мешает нам сделать механическую выборку.

Однако если предположить, что микрорайоны неоднородны (состоят из центра с элитными квартирами и окраин), объем выборки не очень большой и микрорайоны невелики, то механический отбор может привести к систематической ошибке.

При таких допущениях шаг выборки может «перескакивать» из центрального адреса одного микрорайона в центральный адрес другого, что приведет к тому, что в выборку попадут лишь состоятельные люди (возможен и противоположный вариант).

Из этого следует основной вывод о том, что при отклонении от принципа случайности необходимо четко отслеживать любую возможность возникновения систематической ошибки.

4.2 Стратифицированная (районированная) выборка.

4.2.1 Практическая реализация.

При проведении стратифицированного отбора, генеральная совокупность сначала разбивается на группы (страты) по какому-либо признаку. Далее уже в этих выделенных группах проводится случайная или механическая выборка.

Стратифицированная выборка может быть пропорциональной объему группы (в этом случае каждая страта имеет одинаковую долю в выборке) или непропорциональной (в этом случае доля страты в выборке зависит от доли этой страты в генеральной совокупности); также она может проводиться пропорционально колебанию признака в группах[11].

Например, всех представителей генеральной совокупности можно разделить по полу, и затем провести случайный отбор среди мужчин и женщин.

Если мы отберем 50% мужчин и 50% женщин, то это будет пропорциональный отбор. В данном случае мы исходим из того, что мужчин и женщин в генеральной совокупности примерно поровну, а большей точности для нашего исследования не требуется.

Если же мы отберем такой же процент мужчин и женщин, как в генеральной совокупности (например, 49% мужчин и 51% женщин), то это будет непропорциональный отбор.

А если мы знаем, что рассматриваемый нами признак (например, количество выкуриваемых за день сигарет), среди мужчин колеблется несильно, т.е. среди мужчин достаточно мало совсем не курящих и злостных курильщиков, в то время как у женщин наблюдается обратная ситуация, то, чтобы добиться необходимой точности оценки количества выкуриваемых за день сигарет при тех же затратах на проведение опроса, можно опросить меньше мужчин, и за счет этого увеличить число опрашиваемых женщин. Это делается потому, что в данном случае получить оценку количества выкуриваемых за день сигарет у женщин с необходимой точностью является более трудной задачей (из-за сильного колебания признака), чем для мужчин. Этот пример - иллюстрация отбора пропорционально колебанию признака в группах.

4.2.2 Вычисление ошибки выборки.

Формулы для расчета ошибки репрезентативности при пропорциональном стратифицированном отборе даны в таблице 7.

Таблица 7.

Формулы ошибки репрезентативности для стратифицированной выборки (пропорциональный отбор). [3, 22]

Предмет изучения.	Повторный отбор.	Бесповторный отбор.

[11] То есть пропорционально дисперсии признака в группах.

Среднее значение признака.	$\Delta_x = t\sqrt{\dfrac{\overline{\sigma}^2}{n}}$	$\Delta_x = t\sqrt{\dfrac{\overline{\sigma}^2}{n}\left(1-\frac{n}{N}\right)}$
Доля признака.	$\Delta_x = t\sqrt{\dfrac{\overline{pq}}{n}}$	$\Delta_x = t\sqrt{\dfrac{\overline{pq}}{n}\left(1-\frac{n}{N}\right)}$

Где:

$\overline{\sigma}^2 = \dfrac{\sum \sigma_i^2 n_i}{\sum n_i}$ - средняя из внутригрупповых дисперсий, где σ_i^2 - дисперсия в группе i, а n_i - численность группы i.

$\overline{pq} = \dfrac{\sum p_i q_i n_i}{\sum n_i}$ - средняя величина доли признака,

p_i - доля признака в группе i,

$q_i = 1 - p_i$

Ясно, что доверительный интервал при стратифицированной выборке будет меньше (выборка точней), чем при случайной выборке, т.к. <u>средняя из внутригрупповых дисперсий меньше общей дисперсии[12]</u>.

Строгое математическое доказательство того, почему при стратифицированной выборке мы имеем право вместо общей дисперсии ставить среднюю внутригрупповых дисперсий и тем самым уменьшать величину доверительного интервала при сохранении той же надежности, можно найти в [5, 104-107].

На «качественном» же уровне можно сказать следующее. Если представить доверительный интервал как дисперсию средней или как ошибку оценки этой средней ($\frac{\sigma}{\sqrt{n}}$), то при стратифицированном отборе эта ошибка оценки может быть выражена как «взвешенное среднее ошибок, сделанных при оценивании по отдельным слоям» [5, 106], что и будет средней из внутригрупповых дисперсий.

То есть нам достаточно обеспечить несмещенную оценку всех групповых средних, чтобы обеспечить несмещенную оценку общей средней. А точность оценки групповых средних зависит только от дисперсии внутри наших групп и количества опрошенных.

Другая составляющая общей дисперсии (межгрупповая дисперсия) не играет здесь никакой роли, т.к. если мы обеспечим попадание групповых средних в свои доверительные интервалы (которые зависят от внутригрупповых дисперсий), то мы автоматически добиваемся попадания общей средней в свой доверительный интервал.

Иными словами, за счет моделирования выборки мы «покрываем» межгрупповую дисперсию (исключаем возможность случайной ошибки в оценке межгрупповой дисперсии). Если же наше конструирование не будет соответствовать реальности, либо группы в самой генеральной совокупности окажутся размытыми[13], то величина межгрупповой дисперсии будет минимальной, что сводит на нет преимущества стратифицированной выборки.

Таким образом, получаем, что дисперсия средней и, значит, величина доверительного интервала зависит лишь от внутригрупповых дисперсий.

---

[12] То же самое справедливо и для доли, т.к. $\overline{pq}$ есть не что иное, как дисперсия доли.

[13] Под «размытостью» понимается большая внуригрупповая и маленькая межгрупповая дисперсии. Иными словами, рассматриваемый нами признак примерно равномерно распределен в выделенных группах.

При пропорциональном отборе вместо общей дисперсии берется средняя внутригрупповых дисперсий, а при непропорциональном отборе – сумма взвешенных по объему всей генеральной совокупности внутригрупповых дисперсий.

Теперь перейдем к непропорциональной выборке, т.е. выборке с неодинаковой удельной долей страт. В следующей таблице даны формулы ошибки репрезентативности для такой выборки.

Таблица 8.

Формулы ошибки репрезентативности для стратифицированной выборки (непропорциональный отбор). [3, 24]

Предмет изучения.	**Повторный отбор.**	**Бесповторный отбор.**
Среднее значение признака.	$\Delta_x = t\frac{1}{N}\sqrt{\sum\frac{\sigma_i^2}{n_i}N_i^2}$	$\Delta_x = t\frac{1}{N}\sqrt{\sum\frac{\sigma_i^2}{n_i}N_i^2\left(1-\frac{n_i}{N_i}\right)}$
Доля признака.	$\Delta_x = t\frac{1}{N}\sqrt{\sum\frac{p_i q_i}{n_i}N_i^2}$	$\Delta_x = t\frac{1}{N}\sqrt{\sum\frac{p_i q_i}{n_i}N_i^2\left(1-\frac{n_i}{N_i}\right)}$

Где:

N_i - объем страты в генеральной совокупности.

n_i - объем страты в выборке.

Как видно из формул, при непропорциональном отборе вместо средней внутригрупповых дисперсий берется сумма взвешенных по объему генеральной совокупности внутригрупповых дисперсий.

Стратифицированная выборка может проводиться пропорционально дисперсии признака в группах. Формулы ошибки репрезентативности для этого случая представлены в таблице 9.

Таблица 9.

Формулы ошибки репрезентативности для стратифицированной выборки (пропорционально колеблемости признака в группах). [3, 26]

Предмет изучения.	**Повторный отбор.**	**Бесповторный отбор.**
Среднее значение признака.	$\Delta_x = t\frac{1}{N}\sum\frac{\sigma_i N_i}{\sqrt{n_i}}$	$\Delta_x = t\frac{1}{N}\sum\frac{\sigma_i N_i}{\sqrt{n_i}}\sqrt{1-\frac{n}{N}}$
Доля признака.	$\Delta_x = t\frac{1}{N}\sum\frac{pqN_i}{\sqrt{n_i}}$	$\Delta_x = t\frac{1}{N}\sum\frac{pqN_i}{\sqrt{n_i}}\sqrt{1-\frac{n}{N}}$

Эти формулы являются просто преобразованными формулами ошибки репрезентативности для непропорционального отбора. Преобразование производится путем подстановки вместо n_i выражения, которое будет представлено немного ниже.

4.2.3 Определение объема выборки.

Формулы для вычисления объема выборки за исключением отбора пропорционально дисперсии — легко получаются путем элементарных преобразований из формул ошибки репрезентативности.

Таблица 10.

Формулы для определения объема выборки при пропорциональном стратифицированном отборе.

Предмет изучения.	Повторный отбор.	Бесповторный отбор.
Среднее значение признака.	$n = \dfrac{t^2 \overline{\sigma}^2}{\Delta_x^2}$	$n = \dfrac{t^2 \overline{\sigma}^2 N}{\Delta_x^2 N + t^2 \overline{\sigma}^2}$
Доля признака.	$n = \dfrac{t^2 \overline{pq}}{\Delta_x^2}$	$n = \dfrac{t^2 \overline{pq} N}{\Delta_x^2 N + t^2 \overline{pq}}$

Для непропорционального отбора число опрашиваемых в каждой страте определяется отдельно, исходя из их численности в генеральной совокупности.

Отбор пропорционально колеблемости признака в группе вносит и другой критерий для определения величены страт в выборке – внутригрупповую дисперсию.

Таблица 11.

Формулы для определения объема выборки при стратифицированном отборе пропорционально колеблемости признака в группе.

Среднее значение признака.	$n_i = n \dfrac{N_i \sigma_i}{\sum N_i \sigma_i}$
Доля признака.	$n_i = n \dfrac{N_i \sqrt{p_i q_i}}{\sum N_i \sqrt{p_i q_i}}$

4.2.4 Плюсы и минусы стратифицированного отбора.

Стратифицированная выборка в любом случае оказывается точнее собственно-случайной. Этот метод особенно хорош, когда генеральная совокупность неоднородна. В этом случае собственно-случайный отбор крайне неэффективен (требует большого объема выборки).

Однако стратифицированная выборка может быть применена лишь при наличии дополнительной информации о генеральной совокупности (например, нам необходимо процентное соотношение мужчин и женщин, в случае, если мы хотим стратифицировать выборку по полу). Отсутствие такой информации делает применение стратифицированной выборки невозможным. Еще один недостаток стратифицированного отбора – это возможность систематической ошибки. Далее на примерах попытаемся проиллюстрировать различные способы применения стратифицированной выборки.

Пример: [6, 41] Возьмем опять данные о доходах из таблицы 1. Только теперь из этого же массива произведем не случайную, а стратифицированную выборку из четырех человек.

Для возможности проведения стратифицированной выборки сделаем допущение, что из предыдущей переписи мы знаем доходы респондентов за предыдущий период и имеем основания предполагать, что к этому периоду они не изменились или изменились пропорционально.

На основании этих данных можно стратифицировать генеральную совокупность по доходу. Результаты этого деления представлены в таблице 12.

Таблица 12.

Распределение респондентов по стратам.

№ группы.	Респонденты, попавшие в группу.
1.	A,J,D.
2.	L,G,F.
3.	I,H,C.
4.	E,K,B.

Здесь в каждой группе (или страте) находятся люди с максимально близкими доходами. Необходимые нам четыре человека мы отбираем путем случайного отбора одного человека из каждой группы (т.е. проводим пропорциональную выборку).

В итоге мы получаем вполне репрезентативную по доходу выборку, т.к. в нашей выборке будут в нужной пропорции представлены люди с различным материальным достатком.

Более того, подобный отбор является надежнее случайного, т.к. при таком отборе не могут быть выбраны «плохие» выборки, т.е. выборки, содержащие только бедных или богатых (например, ADJL или BCEK). Однако стратифицированная выборка не всегда приносит выгоду, что показано в следующем примере.

Пример:[6, 42] Допустим, что перед нами опять встала необходимость провести стратифицированную выборку из респондентов, представленных в таблице 1. Далее предположим, что за период между переписью и опросом доходы респондентов претерпели значительные изменения, то выделенные нами группы на основе данных переписи могут получиться не гомогенными (в одну страту попадут люди с разными доходами). Например, такими, которые представлены в таблице 13.

Таблица 13.

Распределение респондентов по стратам.

№ группы.	Респонденты, попавшие в группу.
1.	A,F,C.
2.	L,J,D.
3.	K,E,G.
4.	H,B,I.

В этом случае практически никакой выгоды по сравнению с собственно-случайной выборкой стратифицированный отбор не принесет, т.к. вероятность отбора нами плохих выборок сохранится и здесь.

Пример:[6, 42-46] Теперь рассмотрим случай с неоднородной генеральной совокупностью, т.е. совокупностью, в которой существуют отдельные резкие отклонения от средней тенденции. Например, генеральная совокупность, представленная в таблице 1, будет неоднородной, если, к примеру, респондент B вместо дохода в 6300 будет получать доход в 20000.

Стратифицировать в этом случае можно двумя путями: так, как было стратифицировано в таблице 12 и так, как это сделано в таблице 14.

Таблица 14.

Распределение респондентов по стратам.

№ группы.	Респонденты, попавшие в группу.
1.	A,J,D,L.
2.	G,F,I,H.
3.	C,E,K.
4.	B.

В первом случае, как уже было сказано, был проведен пропорциональный отбор, а в данном случае - соответственно непропорциональный. Пропорциональный отбор в данном случае (в случае с неоднородной генеральной совокупностью) не годится, т.к. он не обеспечит однородность страт (в какую бы страту ни попал В, эта страта сразу же станет неоднородной).

Однако при непропорциональном отборе нарушается принцип равной вероятности попадания в выборку, т.к. респондент В попадает в нашу выборку при любом исходе.

Для исправления этого обстоятельства при непропорциональном отборе применяется процедура взвешивания. Взвешивание призвано «восстановить» Взвешиванием мы увеличиваем удельный вес респондентов из больших страт. Для нашего примера средняя будет рассчитываться следующим образом:

$$\bar{x} = \frac{4x_1 + 4x_2 + 3x_3 + x_4}{12}$$

Если отобрать все возможные выборки при данной стратификации (таб. 14), то без взвешивания мы получим смещенную выборочную оценку (генеральная средняя = 3817, а выборочная средняя = 6852). Однако если произвести взвешивание, то в данном случае непропорциональный отбор (стандартное отклонение = 277) будет эффективнее пропорционального (стандартное отклонение = 1878).

Пример:

Рассмотрим далее рассмотренный выше пример с микрорайонами. Решение проблемы может состоять в увеличении объема выборки или в проведении случайного опроса. Но последнее может привести к некоторым искажениям и неточностям, т.к. микрорайоны отличаются также и между собой. Первое же связано с дополнительными расходами.

Здесь нам на помощь может прийти стратифицированная выборка. Ее можно осуществить если нам удастся объединить несколько микрорайонов как сходные. Разделив все микрорайоны на несколько относительно однородных групп, можно построить репрезентативную выборку, не увеличивая ее объем.

Стратифицированная выборка может помочь нам и тогда, когда нет полного списка жителей города. В этом случае можно из каждой страты выбрать один микрорайон, в котором будет проводиться опрос, и собрать информацию о его жителях.

Однако если мы неправильно выберем критерии для объединения микрорайонов, то это приведет к серьезной систематической ошибке.

4.3 Гнездовая (серийная) выборка.

4.3.1 Практическая реализация.

Здесь отбираются не люди, а группы. Группы отбираются случайным образом, а внутри них проводится сплошной опрос. Например, в ВУЗе с большим количеством студенческих групп отбор можно проводить путем случайного отбора этих групп и дальнейшего сплошного опроса в этих группах.

4.3.2 Вычисление ошибки выборки.

Формулы для расчета ошибки репрезентативности при гнездовом отборе даны в таблице 15.

Таблица 15.

Формулы ошибки репрезентативности для стратифицированной выборки. [3, 29]

Предмет изучения.	Повторный отбор.	Бесповторный отбор.
Среднее значение признака.	$\Delta_x = t\sqrt{\frac{\delta^2}{r}}$	$\Delta_x = t\sqrt{\frac{\delta^2}{r}\frac{R-r}{R-1}}$

Доля признака.	$\Delta_x = t\sqrt{\frac{p_r q_r}{r}}$	$\Delta_x = t\sqrt{\frac{p_r q_r}{r}\frac{R-r}{R-1}}$

Где:

$\delta^2 = \frac{\sum (\tilde{x}_i - \tilde{x})^2}{r}$ - межгрупповая дисперсия.

r – число групп в выборке.

$\tilde{x}_i$ - групповая средняя.

$\tilde{x}$ - общая средняя.

R – число групп в генеральной совокупности.

$p_r = \frac{\sum p_i}{r}$ - межгрупповая доля.

Ясно, что доверительный интервал при гнездовой выборке будет меньше (выборка точней) при той же надежности чем при случайной, т.к. межгрупповая дисперсия меньше общей дисперсии.

Внутригрупповая дисперсия нам не нужна, т.к. мы опрашиваем все гнездо целиком и поэтому отклонения выборочного показателя от генерального внутри этой группы не имеем. Следовательно, нас должно волновать то, правильно ли мы выбрали сами группы. Поэтому мы и учитываем лишь межгрупповую дисперсию.

4.3.3 Определение объема выборки.

Формулы для вычисления объема выборки – преобразованные формулы ошибки репрезентативности. Они даны в следующей таблице.

Таблица 16.

Формулы для определения объема выборки при гнездовом отборе.

Предмет изучения.	Повторный отбор.	Бесповторный отбор.
Среднее значение признака.	$r = \frac{t^2\sigma^2}{\Delta_x^2}$	$r = \frac{t^2\sigma^2 R}{\Delta_x^2 R + t^2\sigma^2}$
Доля признака.	$r = \frac{t^2 p_r q_r}{\Delta_x^2}$	$r = \frac{t^2 p_r q_r R}{\Delta_x^2 R + t^2 p_r q_r}$

4.3.4 Плюсы и минусы этого метода.

Разные источники по-разному оценивают точность гнездовой выборки по сравнению со случайной. [3, 29; 6, 49-50].

Главный «козырь» этого типа отбора в том, что он гораздо проще в организационном плане. Действительно, гораздо проще выбрать несколько групп и опросить их целиком, чем бегать за каждым респондентом. Это дает нам выигрыш в средствах и во времени.

Но при этом необходимо следить, чтобы количество групп в генеральной совокупности было достаточно большим, иначе ни о каком принципе случайности не может быть и речи. Более того, возможны перекосы из-за того, что на момент опроса не удается застать всех членов группы. К тому же объем выборки при гнездовом отборе обычно больше, чем при случайном отборе.

Пример:[6, 48-50]

Возьмем опять все ту же генеральную совокупность из таблицы 1 и сделаем из нее гнездовую выборку. Вопрос заключается в том, какие гнезда наиболее подходящие (здесь остается за скобками тот очевидный факт, что респонденты одного гнезда должны быть доступны в единый промежуток времени).

Для ответа на этот вопрос сначала разделим генеральную совокупность по принципу наибольшего сходства (в реальности эти группы нам, конечно же, уже заданы), т.е. в страты попадут люди с максимально близкими доходами. Результаты представлены в таблице 17.

Таблица 17.

Распределение респондентов по группам.

№ группы.	Респонденты, попавшие в группу.
1.	A,D,J,L.
2.	F,G,H,I.
3.	B,C,E,K.

В выборку попадет какая-либо из этих трех групп целиком. Естественно, что при таком разделении на группы, мы, скорее всего получим плохие результаты, т.к. исключаются “хорошие” выборки. Мы отбираем лишь людей с близкой величиной дохода.

Отсюда следует вывод, что при гнездовом отборе мы должны выбирать не максимально гомогенные, а максимально гетерогенные гнезда, т.к. эти гнезда должны представлять собой генеральную совокупность в миниатюре. Подобное разделение можно видеть в таблице 18.

Таблица 18.

Распределение респондентов по группам.

№ группы.	Респонденты, попавшие в группу.
1.	E,C,G,J.
2.	A,H,K,L.
3.	B,F,D,I.

Здесь люди сгруппированы так, что их доходы максимально различаются.

5. Неслучайные (невероятностные) методы отбора.

5.1 Почему применяют неслучайный отбор?

1. Невозможность проведения случайного отбора вследствие:
   - ограниченности ресурсов (в широком смысле: ограниченность денежных средств, ограниченность времени, отведенного на проведение исследования, отсутствие списков единиц генеральной совокупности и т.д.);
   - этических проблем (мы не можем заставить респондента отвечать, если он отказывается).
2. Отсутствие необходимости проведения случайного отбора.

5.2 Классификация методов неслучайного отбора.

Классификация, которой мы будем придерживаться в данном докладе, строится на основе критериев, предложенных В.Э. Шляпентохом. Хотя эта классификация не является абсолютно строгой, и часто трудно провести границу между двумя методами, она отражает основные отличия методов и позволяет их некоторым образом структурировать. Основными факторами, определяющими природу неслучайного отбора, являются:

1. Фактор доступности (насколько включение в выборку зависит от респондента). Этот фактор отражает искажение случайности, идущее со стороны респондентов.
2. Фактор целенаправленности (насколько состав выборки контролируется исследователем). Этот фактор определяет искажение случайности, идущее от исследователя.

Классификация методов неслучайного отбора на основе названных факторов представлена в таблице 1.

Таблица 1.

Классификация методов неслучайного отбора.

<table>
<tr><th colspan="2" rowspan="2">Факторы, определяющие природу неслучайного отбора</th><th colspan="3">Целенаправленность: контроль выборки исследователем:</th></tr>
<tr><th>минимальный</th><th>на среднем уровне</th><th>максимальный</th></tr>
<tr><td rowspan="2">Доступность: включение в выборку от респондента:</td><td>не зависит</td><td>A</td><td>C</td><td>E</td></tr>
<tr><td>зависит</td><td>B</td><td>D</td><td>X</td></tr>
</table>

Рассмотрим подробнее каждый из типов неслучайного отбора.

5.2.1 Доступная выборка

Как следует из названия, в этом случае проводится отбор доступных единиц. Одним из плюсов этого метода являются сравнительно низкие издержки на поиск респондентов.

A: доступные респонденты выделены заранее;

B: респонденты выявляются в процессе опроса, поэтому действительное число доступных объектов определяется апостериори.

Сферы применения доступной выборки:

1) тестирование анкет
2) отработка процедур опроса
3) изучение интимных сторон жизни людей
4) изучение здоровья населения на основе данных об обращениях в больничные учреждения
5) монографические обследования.

C – соответствует принципам случайного отбора.

Этот тип выборки был подробно рассмотрен в предыдущих докладах.

5.2.2 Стихийная выборка.

Исследователь при применении данного метода в некоторой степени контролирует выборку (например, публикуя анкету в журнале, он обращается только к читателям этого журнала), но решение о включении в выборку принимает сам респондент.

Сферы применения стихийной выборки:

1) анкеты, публикуемые в периодическом издании
2) почтовые опросы.

5.2.3 Направленный отбор.

Можно выделить три основных метода направленного отбора:

1) метод типичных единиц (E)
2) целевая выборка (X)
3) квотный отбор (X).

Остановимся на каждом из этих методов.

1. Метод типичных единиц.

При использовании данного метода отбираются единицы генеральной совокупности, обладающие средним (или типичным) значением признака. Однако в таком случае встает проблема выбора признака и определения его типичного значения. Субъективный характер оценки вполне может привести к систематической ошибке. Данный метод целесообразно применять для изучения таких объектов, о которых мы уже обладаем некоторой информацией, например, территориальных общностей, предприятий, учреждений и т.п.

2. Целевая выборка.

Сферы применения целевой выборки:

1) формирование состава участников эксперимента (например, формирование контрольных групп точечным методом, когда для каждого участника основной группы подбирается

участник контрольной группы, обладающий сходными признаками). Это один из тех редких случаев, когда нет необходимости в проведении случайного отбора.
2) отбор экспертов, который может проводиться на основе следующих критериев:
   - объективные характеристики экспертов, содержащиеся в документах
   - тестирование кандидатов в эксперты
   - взаимный отбор
   - самооценка кандидатов в эксперты.

3. Квотный отбор.

Остановимся на описании этого метода более подробно, т.к. это один из самых распространенных методов неслучайного отбора.

При использовании данного метода отбирают один или несколько признаков, по которым будет контролироваться выборка. Количество единиц в выборке, обладающих определенными характеристиками, должно быть пропорционально количеству таких единиц в генеральной совокупности.

Виды квотного отбора.

Можно выделить две разновидности метода квот:
1) априорный отбор
2) апостериорный отбор.

Априорный отбор осуществляется интервьюером на стадии сбора первичной информации.

Апостериорный отбор проводится для корректировки выборки. Например, когда в газету приходят письма с заполненными читателями анкетами, часто среди ответивших имеется перекос по некоторым важным параметрам (возраст, пол и т.п.). В таком случае можно взвесить полученные результаты, а можно провести выборку из выборки квотным методом.

Почему используют квотный отбор?
1) отсутствие необходимости в повторных посещениях
2) достижение заданной точности результатов при меньшем объеме выборки.

Объем выборки при квотном отборе.

Считается, что при использовании метода квот можно делать выборку меньшего объема, чем при случайном отборе, так как квотный отбор дает почти полное совпадение выборочной и генеральной совокупностей по заданным параметрам. Однако это утверждение невозможно подтвердить при помощи математических методов. Единственный способ проверить его справедливость – провести эксперимент.

Такой эксперимент, например, поставил Ф.Э. Шереги. Сначала он случайным методом отобрал 300 рабочих текстильной фабрики в Узбекистане и доказал высокую репрезентативность выборки. Затем он произвел квотным методом выборку из этих 300 человек. Контролировалось два параметра: возраст и национальность. Сначала были отобраны 200 человек, затем – 100. Оказалось, что в выборке 200 человек около 80% всех параметров имели относительную ошибку не более 3%. В выборке 100 человек такую ошибку имели 55% параметров. В выборке 200 человек 6% параметров имели ошибку более 5%, а в выборке 100 человек такую ошибку имели 25% показателей [5, с.126].

Выбор признаков.

Во-первых, выбранные признаки должны быть тесно связаны с изучаемыми характеристиками, иначе полученные результаты могут оказаться сильно искаженными.

Во-вторых, признаки должны быть независимыми, иначе расход средств на их контроль будет нерациональным.

Требования к выборке могут быть жесткими и пониженными***. Жесткие требования*** означают совпадение пропорций генеральной и выборочной совокупностей по ***сочетаниям признаков***. В этом случае структура выборочной и генеральной совокупностей по заданным параметрам точно совпадают. При использовании ***пониженных требований*** контролируют лишь совпадение пропорций ***по каждому параметру отдельно***.

Например, если исследователи решили контролировать выборку по четырем параметрам: пол (2 градации), возраст (7 градаций), образование (6 градаций) и род занятий (12 градаций), то при предъявлении пониженных требований они получат 2+6+7+12=27 групп, а при предъявлении жестких требований они получат 2*6*7*12=1008 групп.

Обычно к выборке предъявляют пониженные требования, так как в обратном случае теряется основное преимущество квотного отбора – малый объем выборки, и увеличиваются затраты на поиск респондентов, обладающих определенными характеристиками.

Чаще всего используются социально – демографические признаки, так как:

- они часто носят ключевой характер
- легко получить информацию о распределении по этим признакам единиц в генеральной совокупности.

Обычно используют не более трех – четырех признаков, так как при увеличении их числа растет число ограничений и, соответственно, растут затраты на поиск респондентов.

Трудности, возникающие при применении метода квот.

1. Необходимо предварительное изучение объекта для выявления в нем пропорций единиц с различными характеристиками и связей между характеристиками.
2. Необходима свежая информация о генеральной совокупности. Например. Если активно происходят какие-то демографические процессы, например, миграция, то применение данных переписи населения, проведенной несколько лет назад, может дать большую систематическую ошибку.
3. Некоторые проблемы могут возникнуть на полевом этапе проведения исследования:

3.1 Интервьюер, скорее всего, будет проводить отбор среди наиболее доступных ему лиц, поэтому выборка имеет тенденцию превращаться в доступную. При этом проблема «крепких орешков» не решается, а обходится, так как даже в группе труднодоступных, «дефицитных» респондентов будет происходить смещение в сторону тех, кто наиболее охотно идет на контакт с интервьюером.

3.2 Ближе к концу полевого этапа часто возникает группа «дефицитных» признаков, поэтому повышается соблазн для интервьюера сфальсифицировать результаты.

Пути усовершенствования квотного метода.

1. Часто метод квот применяется не в чистом виде, а в смеси со случайным. Например, интервьюер получает список лиц, с которыми он должен вступить в контакт, и проводит интервью только с теми, кто оказался носителем необходимых параметров. Для внесения элементов случайности интервьюеру может быть задан определенный маршрут, который он обязан соблюдать при поиске респондентов.
2. Квотный метод можно применять в многоступенчатой случайной выборке (на последней ступени отбора). Используемая на предшествующих ступенях случайная стратифицированная выборка обеспечит самовзвешивание по важнейшим признакам.
3. Квотный метод может применяться для замены труднодоступных единиц при использовании случайного отбора.

6. Литература.

1. Венецкий И.Г. Виды выборки.//Вестник статистики, 1974, №2.
2. Кокрен У. Методы выборочного исследования. М., «Статистика», 1976.
3. Ноэль Э. Массовые опросы. Введение в методику демоскопии. М., Ава-Эстра, 1993.
4. Шереги Ф.Э. Применение метода квот в выборочном социологическом исследовании.//СОЦИС, 1975, №3.
5. Шляпентох В.Э. Проблемы репрезентативности социологической информации (случайная и неслучайная выборки в социологии). М., «Статистика», 1976.
6. Sudman S. Reducing The Cost of Surveys. Chicago, 1967.
7. Ноэль. Э. «Массовые опросы». М. Ава-Эстра, 1993г. (с.85-111)
8. Батыгин. Г. С. «Методология социологического исследования» (с. 169)

9. Королев. Ю. Г. «Выборочный метод в социологии». М. 1975г. (с.8-33)
10. «Методика выборочного обследования миграции сельского населения» (под ред. Заславской, Миркиной, Ершовой). Новосибирск. 1969г. (с.58-68)
11. Кокрен. У. «Методы выборочного исследования» М. Статистика. 1976г. (с.104-107)
12. Hansen, Hurwitz, Madow. “Sample Survey [Methods and Theory]” v.1. Wiley Classics Library. 1953. (c. 13, 34-52).
13. Батыгин Г. С. Лекции по методологии социологических исследований. М.: «Аспект-Пресс», 1995.
14. Венецкий И. Г. Теоретические и практические основы выборочного метода. М., 1971.
15. Вестник статистики. 1921, №№ 1-4.
16. Гмурман В. Г. Теория вероятностей и математическая статистика. М.: «Высшая школа», 1998.
17. Кокрен У. Методы выборочного исследования. М.: Статистика, 1976.
18. Ноэль Э. Массовые опросы. Введение в методику демоскопии. М.: «АВА-ЭСТРА», 1993.
19. Паниотто В. И. Качество социологической информации. Киев: «Наукова Думка», 1986.
20. Рабочая книга социолога. М.: «Наука», 1976.
21. Чурилов Н. Н. Проектирование выборочного социологического исследования. Киев: «Наукова Думка», 1986.
22. Шаповалов В. И., Гаскаров Д. В. Малая выборка. М.: Статистика, 1978.
23. Hansen H., Hurwitz N., Madow G. Sample survey. Methods and theory. New York, 1993.
24. Kalton G. Introduction to survey sampling.