А. Чуриков

# Случайные и неслучайные выборки в социологических исследованиях

В специальной литературе о выборке обычно встречается много формул и математических терминов. Попробуем поговорить о выборке более простым языком. Сократим до минимума количество формул и по возможности постараемся избегать специальных терминов, не всегда понятных читателю. Для тех, кто захочет подробнее познакомиться с выборочной теорией, в списке литературы указаны три классических труда, которые до сих пор служат учебниками на данную тему. К сожалению, работа Л. Киша, в 1995 году переизданная в США, так и не переведена на русский язык.

## 1. Случайные и неслучайные выборки

Большинство людей, даже далеких от социологии, знают, что при проведении опросов применяется выборка и что отбор респондентов должен быть случайным. В обыденном понимании случайным часто считают произвольный неуправляемый отбор, типичным примером которого служит отлавливание для опроса на улице случайных прохожих. С точки зрения специалиста полученная таким способом выборка не является случайной.

Выборка называется **случайной**, если каждый человек (каждый представитель совокупности) имеет известную ненулевую вероятность быть отобранным. Корни этого определения лежат в теории вероятностей, которая обосновала выборочные методы исследования. Именно на ее основе было строго доказано, что по ответам относительно небольшого числа людей можно с высокой точностью судить о мнении всех. Обязательным условием этого является случайный характер выборки. Чтобы отличать такие выборки от случайных в обыденном понимании, их еще называют **вероятностными** выборками.

Выборки, в которых невозможно вычислить вероятность отбора людей, не являются случайными. Математическая теория к ним неприменима. Существует большое многообразие неслучайных выборок, одним из представителей которых является квотная выборка. В реальных исследованиях применяется даже больше неслучайных выборок, чем случайных. Вопрос, насколько можно доверять результатам таких исследований, будет рассмотрен ниже.

Сейчас сконцентрируем внимание на случайных выборках. Начнем с наиболее простого, хрестоматийного варианта.

## 2. Простая случайная выборка

Можно без особого труда создать случайную выборку, если в распоряжении исследователя имеется полный список всех людей, мнением которых он интересуется. Множество таких людей называется **изучаемой совокупностью**, или **генеральной совокупностью**. Создание списка представителей совокупности возможно, когда, например, планируется проведение опроса на предприятии, руководство которого готово предоставить полный список своих сотрудников. В этом случае надо сначала решить, сколько человек будет опрошено, то есть определить

**размер выборки** (иногда говорят "объем выборки"), а затем последовательно отобрать из списка нужное число людей. Для отбора необходимо использовать случайный механизм, обеспечивающий любому человеку из списка равную вероятность попасть в выборку. Полученная таким способом выборка называется **простой случайной**.

В качестве случайного механизма для отбора с равной вероятностью используют таблицы случайных чисел или датчики случайных чисел. Каждый человек в списке имеет свой порядковый номер: 1, 2, ..., $N$, где $N$ равно общему числу людей в списке. Таблица или датчик случайных чисел выдает с равной вероятностью номера в интервале от 1 до $N$. Люди с соответствующими порядковыми номерами включаются в выборку. Если какой-либо порядковый номер выпал повторно, то его просто игнорируют, поскольку человек с этим номером уже есть в выборке и второй раз в отборе участвовать не должен. Такой способ отбора называют отбором **без возвращения** (по аналогии с вытягиванием жребия из шляпы, когда вытянутые листки с номерами обратно в шляпу не возвращают). Случайные числа получают до тех пор, пока не будет выбрано нужное количество людей.

Во многих программах, предназначенных для обработки результатов опросов, есть специальная команда для получения простой случайной выборки. Такая команда есть, например, в программе SPSS. Но и без специальной команды легко получить простую случайную выборку, если в программе имеется датчик случайных чисел. Достаточно рядом с каждым человеком из общего списка записать случайное число, полученное этим датчиком (например, случайное число, равномерно распределенное на отрезке от 0 до 1), а затем пересортировать людей из списка в порядке возрастания (или убывания) значений этих случайных чисел. Теперь, чтобы получить простую случайную выборку нужного размера $n$, достаточно взять первые $n$ человек из списка, отсортированного в случайном порядке.

В простую случайную выборку всегда попадает ровно столько людей, сколько запланировал исследователь, потому что именно так устроен механизм отбора. Несложно показать, что при размере выборки $n$ каждый из $N$ людей, помещенных в список, имеет равную вероятность попасть в выборку. Эта вероятность равна $f = n/N$. Более того, в выборку могут попасть любые $n$ человек из $N$, т. е. любая комбинация $n$ людей из $N$ возможна, и даже одинаково вероятна. В математике число различных комбинаций $n$ элементов из $N$ (без повторений) обозначается символом $C_N^n$ и вычисляется по формуле $\frac{N!}{(N-n)!\,n!}$. Именно столько существует различных простых случайных выборок размера $n$. Все они одинаково вероятны.

Чем хороша простая случайная выборка? Тем, что при достаточном размере $n$ в ней будут представлены все категории людей, присутствующие в списке, из которого она отбиралась, и примерно в тех же самых пропорциях. А это значит, что исследователю не надо думать о том, сколько надо опросить мужчин и сколько женщин, сколько молодых и сколько пожилых, сколько богатых и сколько бедных. Все эти пропорции будут с большой вероятностью выдержаны в простой случайной выборке.

Например, если вся совокупность, которая насчитывает 10 тысяч человек, на 45% состоит из мужчин и на 55% – из женщин, то в выборке из 1000 человек пропорции мужчин и женщин бу-

дут примерно такие же. Конечно, нельзя рассчитывать на то, что мужчин будет ровно 450 человек, а женщин – 550. Возможны случайные отклонения от точных пропорций, но они будут невелики. Величину отклонений можно вычислить по следующей формуле:

$$\Delta = \pm 1{,}96 \sqrt{1-\frac{n}{N}} \sqrt{\frac{p(1-p)}{n-1}}\ ^{1}.$$

Кроме числа людей в выборке $n$ и в совокупности $N$ в формулу входит еще доля $p$ мужчин (или женщин). Число 1,96 соответствует принятому в социологии уровню доверия 95%. Подставим в формулу числовые значения параметров, чтобы вычислить возможные случайные отклонения для доли мужчин в выборке.

$$\Delta = \pm 1{,}96 \sqrt{1-\frac{1000}{10000}} \sqrt{\frac{0{,}45\,(1-0{,}45)}{1000-1}} =$$

$$= \pm 1{,}96 \sqrt{0{,}9} \sqrt{\frac{0{,}45 \cdot 0{,}55}{999}} \approx \pm 0{,}029\ .$$

Получилось, что доля мужчин в выборке может случайно отклоняться от правильного значения 0,45 в пределах ± 0,029, т. е. на ± 2,9%. Иными словами, доля мужчин в выборке может колебаться в интервале от 45% – 2,9% = 42,1% до 45% + 2,9% = 47,9%. Этот интервал называется **доверительным интервалом** (с уровнем доверия 95%), а величина Δ, равная ± 0,029 или ± 2,9%, называется **статистической погрешностью**.

Уровень доверия 95% означает, что в 95% из всех возможных простых случайных выборок размера $n$ (а всего их существует $C_N^n$) доля мужчин окажется внутри посчитанного доверительного интервала, а в 5% выборок – за границами этого интервала. Аналогичное утверждение справедливо для любого другого параметра изучаемой совокупности. Всегда существует небольшая вероятность (которая зависит от принятого уровня доверия), что значение, посчитанное по выборке, окажется вне границ вычисленного для данного параметра доверительного интервала. При этом вероятность больших отклонений очень близка к нулю. Например, существует ненулевая вероятность того, что в простую случайную выборку попадет 1000 мужчин и ни одной женщины, но вероятность этого выражается очень маленьким числом, в котором после запятой стоит более 300 нулей.

Рассмотренная в примере ситуация с долей мужчин справедлива и для любых других параметров. Если значение какого-либо параметра известно для всей совокупности, то несложно посчитать, какое отклонение по этому параметру возможно в выборке. Если же значение для всей совокупности неизвестно, то можно посчитать значение по выборке, и тогда истинное значение будет отличаться от значения в выборке не более чем на величину доверительного интервала ± Δ. Это можно гарантировать с доверительной вероятностью 95% (т. е. в 95 выборках из 100).

*Простая случайная выборка обладает несомненными достоинствами – такими, как простота реализации, хорошее воспроизведение структуры совокупности, возможность вычисления доверительных интервалов*

При увеличении размера выборки n уменьшается величина Δ доверительного интервала, т. е. статистическая погрешность. Можно даже подобрать размер выборки $n$ так, чтобы погрешность Δ стала такой, какой нужно исследователю. Формулу для вычисления $n$ несложно получить из приведенной выше формулы для вычисления Δ.

Простая случайная выборка всегда создает приближенную копию всей совокупности, точность которой возрастает с увеличением размера выборки. Это справедливо для всех параметров, в том числе и для места жительства. Дома и квартиры людей, включенных в выборку, будут равномерно распределены по всей территории, на которой проживают представители изучаемой совокупности. Если это город, то по всем районам города, если это Россия, то по всей

[1] Большей известностью пользуется упрощенный вариант этой формулы: $\Delta = \pm 1{,}96 \sqrt{\frac{p(1-p)}{n}}$.

территории России. Это обеспечивает хорошую географическую представительность выборки, но одновременно создает дополнительные сложности при проведении опросов. Чтобы опросить нескольких человек, интервьюеру придется совершать путешествия из одной части города

*Недостатки простой случайной выборки: необходим список всех представителей совокупности; стоимость исследования велика из-за удаленности респондентов друг от друга; статистическая погрешность возникает по всем параметрам выборки, даже по тем, для которых известны истинные пропорции*

в другую или из одного населенного пункта в другой. А это сильно повышает стоимость исследования.

Подведем итог. Простая случайная выборка, обладая несомненными достоинствами, такими как простота реализации, хорошее воспроизведение структуры совокупности, возможность вычисления доверительных интервалов, имеет также ряд недостатков:

- для реализации выборки необходимо иметь список всех представителей совокупности;
- стоимость исследования сильно возрастает из-за удаленности респондентов друг от друга;
- статистическая погрешность возникает по всем параметрам выборки, даже по тем, для которых известны истинные пропорции (например, по полу или возрасту).

Для устранения перечисленных недостатков используются два специальных приема формирования выборки – стратификация и кластеризация, к рассмотрению которых мы сейчас перейдем.

### 3. Стратификация

Совокупность, из которой формируется выборка, обычно имеет свою структуру. В соответствии с этой структурой можно разделить совокупность на части по определенному признаку – территориальному, административному, производственному, социальному и т. п. Например, крупный город делится на административные районы, среди которых есть промышленные и спальные. Россия делится на федеральные округа, на субъекты Федерации (области, края, республики) или на населенные пункты – городские (с разной численностью населения) и сельские. При проектировании выборки бывает важно, чтобы основные части, из которых состоит совокупность, были представлены в выборке в нужных пропорциях.

Простая случайная выборка не может гарантировать отбор заданного числа людей из каждой части совокупности. Она хотя и дает в среднем пропорциональное представительство в выборке людей разных групп, однако эти пропорции подвержены случайным колебаниям. Иногда они могут заметно нарушаться.

Для обеспечения в выборке нужного соотношения между разными частями совокупности применяется стратификация. Она заключается в разбиении всей совокупности на непересекающиеся части, называемые **стратами**. Для каждой страты вычисляется приходящийся на нее размер выборки, а затем производится случайный отбор нужного числа респондентов. В результате в каждой страте отбирается ровно столько респондентов, сколько запланировал исследователь. Полученная таким способом выборка называется **стратифицированной**. Иногда вместо термина “страты” применяют названия “типические районы” или “слои”, а стратифицированную выборку называют районированной или расслоенной.

Чаще всего выборка распределяется по стратам пропорционально числу людей в них. Такое распределение называется **пропорциональным**. Оно позволяет выдержать в выборке те же пропорции между стратами, что и во всей совокупности. Наряду с пропорциональным применяется также равное размещение, размещение Неймана и оптимальное размещение.

При **равном** размещении из каждой страты опрашивают одинаковое число людей, хотя число людей в стратах может заметно различаться. Равное размещение применяют, когда требуется

сравнить между собой разные части совокупности. Например, сравнивается уровень доходов или состав потребительской корзины городского и сельского населения России. Равный размер выборки для города и для села обеспечивает одинаковый уровень погрешности в обеих группах, что позволяет сравнивать их между собой. При этом суммарная погрешность для всего населения будет больше, чем в случае пропорционального размещения выборки между городом и селом.

Размещение **Неймана** основано на том, что размер выборки делают больше в тех стратах, где труднее оценить интересующий исследователя параметр. Например, если целью исследования является оценка среднедушевого дохода по России в целом, то понятно, что основные погрешности будут возникать при оценке дохода в крупных городах, и особенно в Москве, так как здесь разница в доходах людей очень велика. Оценка среднедушевого дохода будет сильно зависеть от того, какие именно москвичи попадут в выборку при случайном отборе. Разброс в уровне дохода сельских жителей существенно меньше, а значит и погрешность при оценке дохода этой части населения будет меньше. Поэтому при размещении Неймана размер выборки в тех стратах, где разброс в доходах людей велик, будет больше, чем при пропорциональном размещении, а размер выборки в стратах с небольшим разбросом по доходам будет меньше.

При **оптимальном** размещении учитывается не только разброс в стратах по оцениваемому параметру, но и разница в стоимости опроса. В тех стратах, где стоимость опроса выше (например, в труднодоступных районах), размер выборки уменьшается по сравнению с размещением Неймана. Там, где стоимость опроса ниже, размер выборки увеличивается. За счет экономии на "дорогих" стратах оптимальное размещение позволяет увеличить общий размер выборки при той же самой стоимости исследования.

Размещение Неймана применяют в тех случаях, когда нужно уменьшить статистическую погрешность по какому-то одному, наиболее важному для исследователя параметру. При этом погрешности по другим параметрам могут увеличиться по сравнению с пропорциональным размещением выборки. Оптимальное размещение позволяет еще больше уменьшить погрешность по этому параметру за счет увеличения общего размера выборки при сохранении ее стоимости. Но чтобы применить размещение Неймана, необходима информация о величине разброса параметра в стратах (т. е. о дисперсии), а для оптимального размещения требуется также информация о стоимости опроса в стратах.

Четыре рассмотренных способа распределения выборки между стратами являются типовыми. Каждый из них решает определенную задачу. Можно применять и другие способы размещения выборки в зависимости от преследуемой цели. Но только один из способов обеспечивает пропорциональное представительство в выборке людей из каждой страты, а именно – пропорциональное размещение. Означает ли это, что все другие способы размещения приводят к ошибкам?

Никаких ошибок не возникнет, если при вычислениях учитывается число людей в каждой страте. Предположим, что вся совокупность разделена на $M$ страт и что число представителей совокупности в стратах равно соответственно $N_1$, $N_2$, ..., $N_M$. Пусть требуется оценить по выборке некоторый параметр, например средний доход за последний месяц. Сначала посчитаем средний

*Для обеспечения в выборке нужного соотношения между разными частями совокупности применяется стратификация – разбиение всей совокупности на непересекающиеся части, называемые стратами*

доход в каждой страте обычным способом, как среднее арифметическое доходов респондентов из этой страты. Общий средний доход для всех страт считается по следующей формуле:

$$\overline{y}_{str} = \sum_{i=1}^{M} \frac{N_i}{N} \overline{y}_i ,$$

где $\overline{y}_{str}$ – средний доход по всей стратифицированной выборке, $\overline{y}_i$ – средний доход в $i$-ой стра-

те, $N_i$ – число людей в $i$-ой страте, $N$ – число людей во всей совокупности. Несмотря на то, что правильные пропорции между стратами в выборке могут не соблюдаться, использование множителей $N_i/N$ восстанавливает эти пропорции.

Например, если население России разделено на две страты – на городское и сельское население – и из каждой страты опрошено по 500 человек (т. е. применено равное размещение выборки), то для вычисления среднего всероссийского дохода надо знать истинную долю населения каждой страты. По данным переписи 2002 г., население России в возрасте от 18 лет и старше составляло 113,8 млн человек, в том числе городское население – 84,7 млн человек, сельское население – 29,1 млн человек. Таким образом, доли городского и сельского населения соответственно равны

$$W_1 = \frac{N_1}{N} = \frac{84{,}7}{113{,}8} = 0{,}74;$$

$$W_2 = \frac{N_2}{N} = \frac{29{,}1}{113{,}8} = 0{,}26.$$

Средний доход для России считается по формуле

$$\overline{y}_{str} = 0{,}74\,\overline{y}_1 + 0{,}26\,\overline{y}_2\,.$$

Получается, что доход городского населения будет учтен с коэффициентом 0,74, а доход сельского населения – с коэффициентом 0,26, т. е. несмотря на равные размеры выборки в городе и на селе, правильные пропорции между стратами будут восстановлены.

Величина $W_i = N_i/N$, равная доле населения страты во всем населении, называется **весом** страты. Веса страт надо учитывать при вычислении среднего значения в стратифицированной выборке. При пропорциональном размещении веса страт можно не учитывать, т. к. нужные пропорции и так выдержаны.

Стратифицированная выборка позволяет не только опросить нужное число респондентов в каждой страте, но и уменьшить статистическую погрешность. Дело в том, что статистическая погрешность стратифицированной выборки $\Delta_{str}$ зависит только от погрешностей внутри страт. Для нее справедлива следующая формула:

$$\Delta_{str}^2 = \sum_{i=1}^{M} W_i^2\,\Delta_i^2\,,$$

где $W_i$ – вес $i$-ой страты, $\Delta_i$ – погрешность в $i$-ой страте (все величины возводятся в квадрат). Если погрешности во всех стратах будут невелики, то и общая погрешность будет мала.

Это свойство стратифицированной выборки можно эффективно использовать при делении совокупности на страты. Страты надо создавать таким образом, чтобы в них попадали схожие между собой люди. Чем более похожи друг на друга будут люди внутри одной страты, тем меньше будет погрешность стратифицированной выборки.

Поясним это на примере выборочного опроса сотрудников крупного предприятия. Целью опроса является оценка среднемесячного дохода сотрудников. Предположим, что руководство предприятия предоставило список всех сотрудников, в котором указаны их пол, год рождения, должность и название подразделения. Используем в качестве страт группы сотрудников, занимающих одинаковую должность. Выборку распределим по стратам пропорционально числу людей в них и из каждой страты отберем простую случайную выборку нужного размера. Поскольку зарплата сотрудников обычно определяется принятой на предприятии тарифной сеткой, то средний доход людей, занимающих одинаковые должности, будет мало различаться. Поэтому независимо от того, кто именно попадет в выборку в каждой страте при случайном от-

боре, средние доходы в стратах будут посчитаны с очень маленькой погрешностью. А значит, и погрешность при вычислении среднего дохода всех сотрудников будет мала.

В том случае, если тарифные ставки на предприятии жестко фиксированы и все сотрудники, занимающие одинаковые должности, получают одинаковую зарплату, доходы в стратах будут вычислены абсолютно точно без всяких погрешностей. В результате и общий средний доход будет вычислен точно. В такой ситуации достаточно в каждой страте опросить всего по одному сотруднику, а затем умножить его доход на вес страты и все результаты сложить. Получим точный средний доход всех сотрудников.

Если взять для сравнения простую случайную выборку, полученную из полного списка сотрудников без его деления на страты, то результат такого опроса будет иметь гораздо большую погрешность. Причина увеличения погрешности состоит в том, что люди, занимающие высокие и низкие должности, будут отбираться с равной вероятностью из всего списка и их пропорции в выборке будут выдержаны лишь приблизительно.

Конечно, на практике редко удается провести такую идеальную стратификацию, какая была в рассмотренном примере. Для создания страт надо иметь точные данные о числе людей в каждой страте. Эти данные обычно имеются по небольшому числу социально-демографических параметров, таких как пол, возраст, тип места жительства и некоторым другим. Но даже из этих параметров не все можно использовать для стратификации, так как при случайном отборе люди из одной страты должны быть отделены от людей из другой. Поэтому страты чаще всего формируются по территориальным признакам или по признакам, с ними связанным, например, по типам населенных пунктов. В результате в одну страту попадают люди, не слишком похожие друг на друга, из-за чего статистическая погрешность выборки уменьшается незначительно.

И тем не менее при пропорциональном размещении выборки по стратам общая статистическая погрешность всегда уменьшается, либо, в крайнем случае, остается той же самой. Она в принципе не может увеличиться. По этой причине стратифицированные выборки применяют в большинстве исследований.

Стратификация устраняет только один из недостатков простой случайной выборки – она позволяет выдержать в выборке точные пропорции всей совокупности, если они известны. Но два других недостатка остаются. Для проведения случайного отбора надо иметь список людей каждой страты. Отобранные в стратах респонденты будут по-прежнему удалены друг от друга, что увеличивает стоимость опроса.

Избавиться от этих недостатков позволяет другой метод формирования выборки – кластеризация.

## 4. Кластеризация

Кластеризация позволяет включать в выборку респондентов, проживающих на небольшом расстоянии друг от друга, сохраняя при этом случайный механизм их отбора. Это достигается путем объединения людей в группы, которые участвуют в отборе как самостоятельные единицы. Такие группы называются **кластерами**. Чаще всего в качестве кластеров используют различные территориальные образования. Это могут быть административные районы, населенные пункты, городские микрорайоны, городские кварталы, территории избирательных округов или избирательных участков и т. п. В роли кластеров могут также выступать предприятия при опросе рабочих и служащих, учебные заведения при опросе учащихся, магазины при опросе продавцов.

Для получения выборки надо сначала отобрать нужное число кластеров, а затем в каждом из отобранных кластеров отобрать нужное число респондентов, т. е. отбор надо проводить в два этапа. На первом этапе в отборе участвуют кластеры, на втором – люди.

*Одна из задач кластеризации состоит в том, чтобы сократить время и затраты на перемещение интервьюера от респондента к респонденту в пределах кластера*

Прежде чем приступить к отбору кластеров, надо составить их полный список. Каждый человек, входящий в изучаемую совокупность, должен быть отнесен к какому-либо кластеру, причем только к одному. Составление полного списка кластеров представляет гораздо меньше проблем, чем составление полного списка людей. Особенно тогда, когда кластерами служат единицы административно-территориального деления. Например, списки всех административных районов, а также всех городов и поселков городского типа России (с указанием числа жителей) ежегодно публикуются Федеральной службой госстатистики. Их вполне можно использовать в качестве кластеров при опросе населения.

Одна из задач кластеризации состоит в том, чтобы сократить время и затраты на перемещение интервьюера от респондента к респонденту в пределах кластера. Желательно, чтобы это время не превышало 10–15 минут. Если отобранные кластеры имеют слишком большую территорию и не обеспечивают выполнения данного требования, приходится проводить еще один этап или ступень отбора. При этом кластеры, которые отбираются сначала, на первой ступени, называются первичными единицами отбора (ПЕО). Внутри них формируются более мелкие кластеры, которые называются единицами отбора второй ступени или вторичными единицами отбора (ВЕО)[2]. Вторая ступень отбора проводится только в тех кластерах, которые были отобраны на первой ступени. Например, если на первой ступени проводился отбор административных районов России, то на второй ступени могут отбираться населенные пункты районов, попавших в выборку.

В кластерах, отобранных на второй ступени, можно провести отбор еще более мелких кластеров. Например, в городах можно провести отбор микрорайонов, кварталов или избирательных участков. Это будет третья ступень отбора. На последней ступени отбираются люди (или другие элементы, из которых состоит изучаемая совокупность и которые являются объектом исследования). В зависимости от числа ступеней отбора выборка будет называться двухступенчатой, трехступенчатой и т.д.

Выборка, в которой на начальных этапах отбираются кластеры, а на последнем этапе – люди (представители совокупности), называется **многоступенчатой** или **кластерной**[3]. В некоторых изданиях на русском языке кластеры называются гнездами, а кластерная выборка – гнездовой.

Использование кластерной выборки избавляет исследователя от необходимости составлять полный список всех представителей совокупности. Вместо этого составляются списки кластеров: первичных единиц отбора – для всей совокупности, вторичных единиц отбора – для тех ПЕО, которые попали в выборку на первой ступени, и т.д. Списки людей нужны только для проведения последней ступени отбора. Они составляются для тех небольших по размеру кластеров, которые были отобраны на предпоследней ступени. При опросах по месту жительства списки людей заменяются списками домохозяйств. Эти списки могут быть получены на основе домовых

[2] В англоязычной литературе единицы отбора первой ступени называются primary sample units (PSU), а единицы отбора второй ступени – secondary sample units (SSU).

[3] Можно применять сразу оба названия – многоступенчатая кластерная выборка. Иногда термину "кластерная" придают более узкий смысл, называя так выборки, в которых в отобранных кластерах опрашивают всех людей поголовно, а все остальные выборки называют многоступенчатыми.

книг жилищно-эксплуатационных организаций или сельских администраций, а могут быть составлены интервьюером непосредственно на местности, что достаточно просто для небольших кластеров.

Таким образом, у кластерной выборки отсутствуют два главных недостатка простой случайной выборки – не требуется список всех представителей совокупности и интервьюер имеет возможность опросить нескольких человек, проживающих на небольшом расстоянии друг от друга.

Кластерная выборка является случайной, т. е. для каждого человека (элемента совокупности) должна быть обеспечена определенная (желательно – равная) вероятность попасть в выборку. Для этого кластеры должны отбираться с вероятностью, пропорциональной числу элементов в кластере. Такой способ отбора часто называют **ВПР-отбором** (по первым буквам слов "вероятность, пропорциональная размеру") или **PPS-отбором** (от аналогичного английского выражения "probability proportional to the size"). Число элементов совокупности в кластере называют **размером** кластера.

Если отбирать кластеры с вероятностью, пропорциональной размеру, а людей внутри кластера – с равной вероятностью, то для любого человека из изучаемой совокупности будет обеспечена одинаковая вероятность попадания в выборку. Это следует из того, что итоговая вероятность отбора получается путем умножения вероятности отбора кластера $p_1$ на вероятность отбора человека внутри кластера $p_2$. Если в $i$-ом кластере содержится $A_i$ элементов, а во всей совокупности – $N$ элементов, то вероятность PPS-отбора $i$-го кластера равна $p_1 = A_i/N$. Вероятность отбора одного человека в кластере, состоящем из $A_i$ людей, равна $p_2 = 1/A_i$. Итоговая вероятность попадания человека в выборку получается после умножения $p_1$ на $p_2$, она равна

$$p = p_1 p_2 = \frac{A_i}{N}\,\frac{1}{A_i} = \frac{1}{N}.$$

Эта вероятность в итоге не зависит от размера кластера $A_i$ и будет одинаковой для любого человека из совокупности.

Когда последовательно отбирается $n_1$ кластеров и в каждом кластере отбирается по $n_2$ элементов, то общая вероятность отбора будет равна

$$p = p_1\, p_2 = n_1 \frac{A_i}{N}\, n_2 \frac{1}{A_i} = \frac{n_1 n_2}{N} = \frac{n}{N},$$

где $n$ – общий размер выборки.

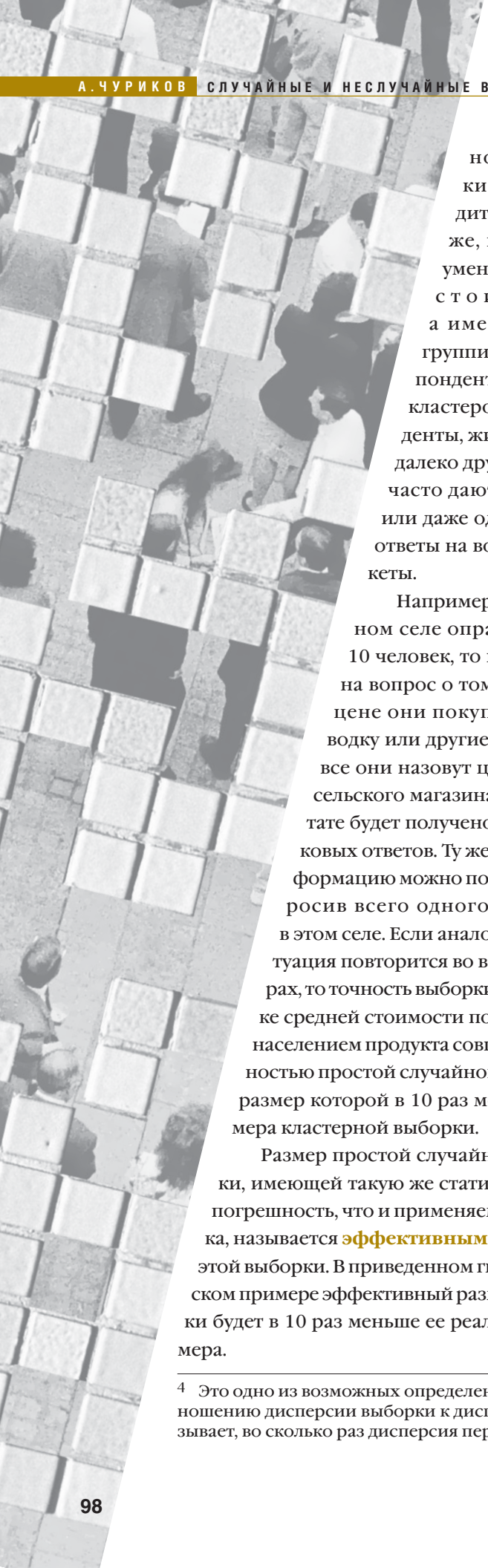
Аналогично вычисляется вероятность и при многоступенчатом отборе. Например, при трехступенчатом отборе вероятность равна

$$p = p_1\, p_2\, p_3 = \left(n_1 \frac{A_i}{N}\right)\left(n_2 \frac{B_j}{A_i}\right)\left(n_3 \frac{1}{B_j}\right) = \frac{n_1 n_2 n_3}{N} = \frac{n}{N}.$$

Отметим, что кластеры надо отбирать **"с возвращением"**, т. е. все кластеры, в том числе и уже попавшие в выборку, участвуют в каждом из $n_1$ отборов. Поэтому один и тот же кластер может попасть в выборку два и более раз. Повторное попадание кластера в выборку означает, что внутри кластера тоже должен проводиться повторный отбор. Если в кластере, попавшем в выборку один раз, отбирается $n_2$ человек, то в попавшем в выборку два раза – дважды по $n_2$ человек, в попавшем в выборку три раза – трижды по $n_2$ человек, и т. д. То же самое относится и к многоступенчатому отбору, когда на второй ступени отбираются не люди, а более мелкие кластеры. Их количество тоже увеличивается в соответствующее число раз.

*У кластерной выборки отсутствуют два главных недостатка простой случайной: не нужен список всех представителей совокупности и можно опросить нескольких человек, проживающих на небольшом расстоянии друг от друга. При этом уменьшается стоимость исследования, но растет статистическая погрешность*

Кластерная выборка получается дешевле простой случайной. За уменьшение стоимости приходится платить увеличением статистической погрешности. Потеря точности кластер-



ной выборки происходит из-за того же, из-за чего уменьшается ее стоимость, а именно из-за группировки респондентов внутри кластеров. Респонденты, живущие недалеко друг от друга, часто дают похожие или даже одинаковые ответы на вопросы анкеты.

Например, если в одном селе опрашивается 10 человек, то при ответе на вопрос о том, по какой цене они покупают хлеб, водку или другие продукты, все они назовут цену своего сельского магазина. В результате будет получено 10 одинаковых ответов. Ту же самую информацию можно получить, опросив всего одного человека в этом селе. Если аналогичная ситуация повторится во всех кластерах, то точность выборки при оценке средней стоимости покупаемого населением продукта совпадет с точностью простой случайной выборки, размер которой в 10 раз меньше размера кластерной выборки.

Размер простой случайной выборки, имеющей такую же статистическую погрешность, что и применяемая выборка, называется **эффективным размером** этой выборки. В приведенном гипотетическом примере эффективный размер выборки будет в 10 раз меньше ее реального размера.

Понятие "эффективный размер" используется при сравнении выборок между собой, поскольку реальный размер выборки не отражает величину ее статистической погрешности. Чем меньше величина статистической погрешности, тем больше эффективный размер выборки, и наоборот. Про реальный размер выборки этого сказать нельзя.

Для измерения качества выборки используют параметр, который называется дизайн-эффектом и обозначается *deff*. Он получается в результате сравнения выборки произвольного типа с простой случайной выборкой такого же размера. Простая случайная выборка играет здесь роль эталона. **Дизайн-эффект** показывает, во сколько раз реальный размер выборки $n$ больше или меньше ее эффективного размера $n_{эф}$ [4].

$$n = deff\, n_{эф}\,.$$

Связь между статистической погрешностью выборки $\Delta$ и статистической погрешностью $\Delta_0$ простой случайной выборки такого же размера выражает следующее соотношение:

$$\Delta = \sqrt{deff}\,\Delta_0\,.$$

Если $deff > 1$, то погрешность будет больше погрешности простой случайной выборки, т. е. применяемая выборка "хуже". Если $deff < 1$, то ее погрешность меньше, т. е. она "лучше". Если $deff = 1$, то выборки одинаковы по точности.

Кластерная выборка всегда менее точна, чем простая случайная, т. е. ее дизайн-эффект всегда больше единицы. Потеря в точности происходит из-за наличия зависимости между ответами респондентов одного кластера. Для измерения степени этой зависимости используют показатель, который называется коэффициентом внутрикластерной корреляции и обозначается ***rob*** (rate of homogeneity). Он принимает значения от нуля до единицы; 0 означает полное отсутствие зависимости внутри кластеров, 1 – максимальную зависимость (внутри каждого кластера все рес-

[4] Это одно из возможных определений. Классическое определение следующее: дизайн-эффект равен отношению дисперсии выборки к дисперсии простой случайной выборки такого же размера (т. е. он показывает, во сколько раз дисперсия первой выборки больше или меньше дисперсии второй).

понденты отвечают одинаково). На практике *roh* принимает всегда промежуточное значение между 0 и 1. Для разных параметров совокупности *roh* может принимать разные значения.

По результатам опроса можно оценить величину roh по любому измеряемому признаку. Для этого существуют специальные программные средства.

Дизайн-эффект кластерной выборки зависит от двух факторов: от коэффициента внутрикластерной корреляции *roh* и от размера подвыборки в кластере $n_c$. Эта зависимость выражается формулой

$$deff = 1 + roh(n_c - 1).$$

При малых значениях *roh* в каждом кластере можно опрашивать больше респондентов. При больших *roh* число опрашиваемых в кластерах надо сокращать, а необходимый размер выборки достигается за счет увеличения числа отбираемых кластеров.

Исследователь не может повлиять на величину *roh*, это свойство кластеров, которое можно измерить, но нельзя изменить. Единственное, что он может сделать, это использовать в качестве единиц отбора другой тип территориальных единиц с другим значением *roh*. А вот количество человек $n_c$, которое будет опрашиваться в кластерах, полностью зависит от исследователя. При большом $n_c$ стоимость выборки уменьшается, но погрешность растет. При малом $n_c$ стоимость растет, а погрешность уменьшается.

Существует некоторое оптимальное значение $n_c$. Чтобы его определить, надо из общих затрат на исследование попытаться выделить затраты, связанные с кластерами, и непосредственные затраты на проведение интервью. К первым относятся время и транспортные расходы интервьюера на то, чтобы добраться до места нахождения кластера. Если кластером является городской квартал, то это время и стоимость проезда интервьюера до квартала и обратно. Если кластером является село, то стоимость кластера определяется временем, которое интервьюер затрачивает на дорогу туда и обратно, а также стоимостью проезда на электричке, автобусе, попутной машине и т. п. В затраты, относящиеся к кластеру, входит также стоимость проживания интервьюера в гостинице, когда опрос проходит в удаленной местности и его не удается завершить в течение одного дня.

Стоимость проведения интервью определяется длиной вопросника и средней длительностью интервью, а также временем, затрачиваемым интервьюером на поиск респондента после прибытия на место проведения опроса (в нужный квартал, село и т. п.).

Для определения величины $n_c$ важны не столько сами стоимости, сколько их отношение. Если обозначить через $C$ средние затраты на один кластер, а через $I$ – средние затраты на одно интервью, то оптимальное значение размера выборки в кластере можно вычислить по следующей формуле:

$$n_c = \sqrt{\frac{C}{I}\,\frac{1-roh}{roh}}.$$

Полученное значение $n_c$ будет оптимальным в том смысле, что обеспечит минимальную статистическую погрешность при фиксированной стоимости исследования, или, что эквивалентно, обеспечит заданную погрешность при минимальных затратах.

Для вычисления $n_c$ надо знать помимо отношения стоимостей $C/I$ еще и коэффициент внутрикластерной корреляции *roh*. Обычно его определяют по результатам предыдущих исследований, в которых были использованы

кластеры того же типа (районы, населенные пункты или городские кварталы).

В тех случаях, когда оптимальное значение $n_c$ вычислить не удается из-за отсутствия нужной информации или по другим причинам, размер подвыборки в кластерах определяется исходя из числа имеющихся интервьюеров, максимально допустимой нагрузки на одного интервьюера и других подобных соображений.

Подведем итог. Кластеризация уменьшает стоимость выборочного исследования, позволяя отбирать респондентов, проживающих на небольшом расстоянии друг от друга. При этом увеличивается статистическая погрешность. При изучении общественного мнения больших масс населения, проживающих на обширной территории, это единственный способ создать случайную выборку.

## 5. Различия между стратифицированной и кластерной выборками

Нами были рассмотрены два приема, используемые для создания случайной выборки, – стратификация и кластеризация. Между ними есть существенные различия.

Статистическая погрешность кластерной выборки тем больше, чем сильнее зависимость в ответах респондентов внутри кластеров, т. е. чем более похожи по своим взглядам люди, входящие в кластер. И наоборот, чем более непохожи друг на друга люди внутри кластера, тем погрешность будет меньше.

*Страты должны содержать как можно более однородные элементы, кластеры – как можно более разнородные*

Поэтому при формировании кластерной выборки лучше использовать такие единицы отбора, которые содержат более разнородные элементы совокупности. В этом состоит одно из отличий кластеров от страт. Страты должны содержать как можно более однородные элементы, кластеры – как можно более разнородные.

Это различие между стратами и кластерами вполне объяснимо. В стратифицированную выборку обязательно входят элементы из каждой страты. Каждый отобранный элемент представляет в выборке элементы только своей страты, чем лучше он их репрезентирует, тем точнее выборка. Если каждая страта будет состоять из очень похожих элементов, то страты будут представлены с минимальной погрешностью, а значит, будет минимальна и статистическая погрешность всей выборки.

В кластерной выборке, в отличие от стратифицированной, каждый отобранный кластер должен репрезентировать все элементы совокупности. Чем больше отдельные кластеры похожи на всю совокупность, тем точнее кластерная выборка. Если каждый кластер будет являться маленькой копией всей совокупности, в которой, как в капле воды, отражается все многообразие имеющихся мнений, то результат исследования будет мало зависеть от того, какие именно кластеры попадут в выборку, в этом случае погрешность кластерной выборки будет минимальной.

На погрешность стратифицированной выборки влияет только погрешность измерения внутри страт, степень различия страт между собой на погрешности не отражается. На погрешность кластерной выборки больше всего влияют различия между кластерами, результат сильно зависит от того, какие именно кластеры попали в выборку. Погрешность измерения внутри кластеров тоже влияет на общую погрешность, но значительно меньше.

Еще одно различие между стратифицированной и кластерной выборками состоит в том, что стратификация уменьшает статистическую погрешность, а кластеризация – увеличивает. Поэтому дизайн-эффект стратифицированной выборки всегда меньше или равен единице (если страты в выборке представлены пропорционально), а дизайн-эффект кластерной выборки всегда больше единицы.

## 6. Стратифицированные кластерные выборки

Несмотря на различие между стратификацией и кластеризацией, оба эти метода формирования выборки могут применяться одновременно.

В результате их совместного применения получается стратифицированная кластерная выборка. Для создания выборки такого типа надо формировать страты не из отдельных элементов совокупности, а из кластеров. При этом требование однородности страт сохраняется. Это означает, что в одну страту следует помещать похожие между собой кластеры.

Число создаваемых страт обычно равно либо числу кластеров $n_1$, которые надо отобрать, либо в два раза меньше, т. е. равно $n_1/2$. В первом случае из каждой страты отбирается по одному кластеру, во втором случае – по два. Отбор кластеров проводится независимо в каждой страте с вероятностью, пропорциональной размеру кластеров.

При формировании многоступенчатой выборки стратификация может применяться на любой ступени отбора. На каждой ступени страты создаются из соответствующих данной ступени единиц отбора, на первой ступени – из ПЕО, на второй – из ВЕО, и т. д. На последней ступени отбора страты формируются непосредственно из элементов совокупности, принадлежащих данному кластеру.

Большинство социологических исследований проводятся по стратифицированным кластерным выборкам. Стратификация и кластеризация – два основных методических приема, используемых при создании всего многообразия случайных выборок. Конечно, в арсенале разработчиков имеются также разнообразные технические приемы для проведения случайного отбора элементов и кластеров. К ним относится **систематический отбор**, при котором первый элемент отбирается случайно, а каждый последующий получается путем прибавления к порядковому номеру предыдущего некоторого фиксированного числа, называемого **шагом отбора**. При шаге отбора 10 отбирается каждый 10-й элемент, при шаге отбора 100 – каждый 100-й и т. д. Систематический отбор применяется также для отбора кластеров – в выборку попадает весь кластер, содержащий систематически отобранный элемент (при этом обеспечивается отбор кластеров с вероятностью, пропорциональной размеру). В число технических приемов входит также **контролируемый отбор**, который позволяет увязывать между собой результаты отбора в разных стратах или кластерах и способствует повышению уровня контроля над выборкой. Есть и другие технические приемы. Но все они служат скорее вспомогательным инструментом для создания случайных выборок, основанных на стратификации и кластеризации.

*Несмотря на различие между стратификацией и кластеризацией, оба эти метода могут применяться одновременно. В результате их совместного применения получается стратифицированная кластерная выборка*

## 7. Неслучайные выборки

В категорию неслучайных попадают все выборки, для которых невозможно вычислить вероятность отбора людей. Классификацию неслучайных выборок можно встретить во многих работах, но разные авторы по-разному их группируют и порой используют для одного и того же типа выборки разные названия. Поэтому, не претендуя на полноту и однозначность списка, рассмотрим несколько наиболее распространенных видов неслучайных выборок.

**Выборка добровольцев**, или **стихийная выборка**, характеризуется тем, что исследователь обращается с предложением принять участие в опросе ко всем желающим, а люди сами решают, стоит им откликнуться или нет. Призыв высказать свое мнение может прозвучать в эфире теле- или радиопередачи, а также быть опубликованным в газете или журнале в виде анкеты. Инициатор такого опроса обычно не знает, сколько людей услышали его призыв. А среди услышавших далеко не каждый надумает отозваться. Отреагирует, скорее всего, специфическая часть аудитории, не очень-то похожая на большинство зрителей, слушателей или читателей.

**Квотная выборка** состоит в том, что исследователь задает определенные пропорции между разными категориями респондентов, которые

необходимо выдержать. Обычно требуется воспроизвести в выборке известные из статистики пропорции всей совокупности по некоторым параметрам, например, по полу и возрасту респондентов, по уровню образования, по типу места жительства и т. п. Эти пропорции называются квотами. Интервьюеру предоставляется определенная свобода при отборе людей, лишь бы они удовлетворяли заданным квотам. Эта свобода может в большей или меньшей степени ограничиваться правилами, которые он должен соблюдать. Например, интервьюер может проводить опрос во всем населенном пункте или только на его части (на указанной улице или в указанном квартале), в любом месте (на улице, в магазинах, в транспорте) или только по месту жительства респондентов и т. п. Но в пределах заданных ограничений интервьюер сам решает, кого ему опросить. (В этом отличие квотной выборки от случайной стратифицированной, при которой интервьюер должен опросить определенных, заранее отобранных людей.)

**Целевая (экспертная) выборка** строится по принципу принадлежности респондентов к группе людей, интересующих исследователя. Эти люди часто называются **целевой группой**. Примерами целевых групп могут служить владельцы автомобилей определенных марок, покупатели корма для собак, слушатели радио "Эхо Москвы", читатели "Коммерсанта", служащие банков и т. п. У исследователя обычно нет надежных статистических данных о составе и структуре целевой группы, поэтому он не может задать точные квоты. Интервьюер может опросить любого человека, удовлетворяющего заданному критерию принадлежности к целевой группе. Где и как искать таких людей, интервьюер, как правило, решает самостоятельно, он проводит **целенаправленный отбор**. Поиск представителей малочисленных целевых групп проще всего проводить там, где они чаще бывают, в **местах скопления**. Автовладельцев можно опрашивать на бензозаправках, покупателей корма для собак – у специальных магазинов, служащих банков – на месте их работы.

**Доступная выборка** получается тогда, когда опрашиваются только те представители совокупности, которые легко доступны для исследования. Например, изучение заболеваемости проводится на тех людях, которые обратились к врачу, изучение преступности – только по жертвам зарегистрированных преступлений, в выборку руководителей предприятий попадают только те, чьи предприятия включены в справочник, изданный два года назад, мнение родителей учеников выясняют у тех, кто пришел на родительское собрание, и т. д. Различие между доступной и целевой выборками весьма условно. О доступной выборке можно говорить в тех случаях, когда отсутствует четкое описание изучаемой совокупности и исследователь не озабочен вопросом, кого именно представляют опрошенные им люди.

Особой разновидностью целевой выборки является **выборка типичных единиц**. В нее входят "типичные" представители совокупности. Например, типичный город и типичное село каждого региона, типичные представители разных социальных групп и т. д. Выбор типичных представителей проводится на основе экспертных оценок или с применением специальных математических методов. Но и эксперты, и матметоды опираются на имеющуюся информацию о совокупности, которая либо получена в предыдущих исследованиях, либо основана на данных статистики. Будут ли типичные по этим данным элементы оставаться типичными для вновь изучаемых параметров, еще большой вопрос.

**Выборка методом "снежного кома"** служит еще одной разновидностью целевой выборки. Она применяется тогда, когда представите-

лей изучаемой совокупности трудно отобрать другими методами. Трудности возникают либо из-за малочисленности самой совокупности, либо из-за сложности выявления тех, кто в нее входит. Вот несколько примеров таких совокупностей: эксперты в некоторой области (по демографическим проблемам, по утилизации ядерных отходов), представители сексуальных меньшинств, люди определенной национальности (грузины, евреи, китайцы), люди с очень высоким уровнем доходов. Метод получения выборки основан на том, что почти каждый представитель целевой группы может назвать еще одного или нескольких человек, которые в эту группу входят. Поэтому сначала интервьюеры любыми методами ищут первых респондентов, часто среди своих знакомых, а те, в свою очередь, подсказывают, кого еще можно опросить. В результате число опрошенных растет, как снежный ком.

Мы уже отмечали, что неслучайные выборки применяются на практике гораздо чаще, чем случайные. Это связано с тем, что они, как правило, проще и дешевле. Преимущество неслучайных типов выборки особенно сильно проявляется при исследовании малочисленных и трудновыделяемых целевых групп. Хотя и для таких групп можно применять случайные методы отбора.

Общей чертой всех неслучайных выборок является то, что состав отбираемых респондентов существенно зависит от пристрастий и предпочтений отдельных людей. Решение о том, кто именно будет опрошен, принимают либо исследователи (выборка типичных единиц), либо интервьюеры в рамках заданных исследователем ограничений (квотная, целевая и доступная выборки), либо сами респонденты (выборка добровольцев и “снежный ком”). Пристрастия и предпочтения людей обычно вносят в выборку неслучайные искажения. Интервьюеры вольно или невольно отбирают более симпатичных и приятных для себя респондентов, избегая людей угрюмых, озлобленных, неопрятно одетых. Исследователи при отборе типичных представителей руководствуются своими научными гипотезами, для проверки которых как раз и проводится опрос. О существенных различиях между добровольными участниками опросов и остальными людьми уже говорилось раньше.

Искажения, возникающие из-за влияния неслучайных факторов отбора, могут иметь хаотический характер и в значительной степени компенсировать друг друга. Но иногда они направлены в одну и ту же сторону, их влияние складывается. В этом случае возникают систематические смещения.

## 8. Смещенные и несмещенные выборки

Что такое смещение выборки и в чем его отличие от статистической погрешности? Чтобы разобраться в этом, вернемся опять к простой случайной выборке размера $n$. В такой выборке возможны любые комбинации $n$ элементов из всех $N$ элементов совокупности. Общее число разных выборок равно $C_N^n$, это все возможные выборки такого размера. Обозначим их число через $L$.

При оценивании по выборке какого-либо параметра, например среднемесячного дохода, в разных выборках будут получаться разные значения. По одной выборке средний доход равен, скажем, 4563 рублям, по другой – 4687 рублям и т. д. Теоретически можно посчитать средний доход для каждой из всех $L$ возможных выборок (практически это, конечно же, неосуществимо). В одних выборках средние доходы совпадут, в других – будут различаться. Каждому значению дохода будет соответствовать своя точка на горизонтальной оси “средний доход”. Над этой точкой на оси нарисуем небольшой кружок, обозначающий соответствующую выборку. Если

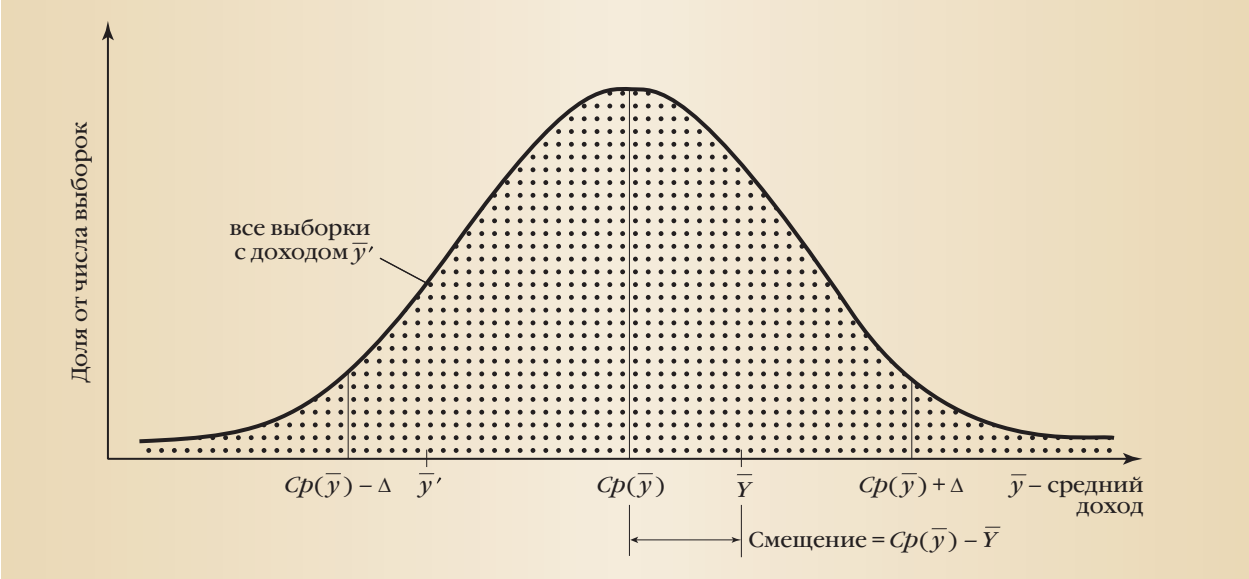



*Рис. 1.* ***Распределение оценки среднего дохода $\bar{y}$ для всего множества выборок фиксированного размера***

один и тот же доход получается в нескольких выборках, то нарисуем несколько кружочков друг над другом. Такими кружочками обозначим все возможные выборки. Картинка, которая получится в результате, изображена на *рис. 1*[5].

*В категорию неслучайных попадают все выборки, для которых невозможно вычислить вероятность отбора людей*

Верхняя граница кружочков, обозначающих выборки, напоминает очертания холма. Если принять число всех выборок за единицу, то высота столбика над некоторой точкой $\bar{y}'$ на горизонтальной оси показывает долю выборок, в которых средний доход равен $\bar{y}'$ рублей. Вершина холма лежит над некоторой средней точкой, которая на рисунке обозначена $Cp(\bar{y})$, это среднее значение дохода по всем $L$ выборкам[6]. Выборки группируются симметрично вокруг этой средней точки, чем дальше от нее, тем число выборок меньше. 95% всех выборок расположены в интервале $Cp(\bar{y}) \pm \Delta$. Это доверительный интервал, а $\Delta$ – статистическая погрешность. И только 5% выборок лежат справа или слева от границ интервала.

Исследователю надо стремиться к тому, чтобы средняя точка $Cp(\bar{y})$ совпадала со значением $\bar{Y}$, которое получится, если бы опросили все население. Будем называть величину $\bar{Y}$ истинным средним доходом, хотя это и не совсем так[7].

Когда среднее по всем выборкам $Cp(\bar{y})$ совпадает с истинным значением параметра $\bar{Y}$, то такой способ получения выборки называется **несмещенным**. Если же эти значения не совпадают, $Cp(\bar{y}) \neq \bar{Y}$, то способ получения выборки называют **смещенным**, а разность $Cp(\bar{y}) - \bar{Y}$ называется **смещением** выборки.

[5] Изображенная кривая близка к плотности нормального распределения.

[6] Значение $Cp(\bar{y})$ можно было бы вычислить по формуле $Cp(\bar{y}) = \sum_{i=1}^{L} \bar{y}_i \Big/ L$, как среднее арифметическое средних доходов $\bar{y}'$, полученных в каждой из $L$ выборок.

[7] Даже если спросить всех людей и все они ответят, величина $\bar{Y}$ будет отличаться от истинного среднего дохода, поскольку далеко не все ответы будут правдивы. Возникнет так называемое невыборочное смещение, которое никак не связано с выборкой. Смещения могут также возникать из-за невозможности опросить кого-то из респондентов, включенных в выборку (из-за того, что интервьюеру не удается с ними встретиться, или из-за их отказа от участия в опросе). Эти смещения связаны не с особенностями случайной выборки, а с особенностями респондентов и с уровнем профессионализма интервьюеров. Хотя их обязательно надо учитывать при планировании исследования.

В теории вероятности есть теорема, доказывающая несмещенность простой случайной выборки. Это означает, что картинка, изображенная на *рис. 1*, справедлива для любого параметра, оцениваемого по простой случайной выборке, и для любой совокупности, из которой эта выборка получена. Все выборки будут несмещенными (т. е. точки $Ср(\overline{y})$ и $\overline{Y}$ совпадут), а различия возможны только в высоте холма и крутизне его склонов. Очертания холма зависят от размера выборки и от степени различия между всеми людьми совокупности по оцениваемому параметру. Чем больше выборка – тем выше и круче холм, чем выборка меньше – тем холм ниже и положе[8]. Когда размер выборки одинаков, то для параметров, по которым различия между всеми людьми невелики, холм будет высокий и крутой; чем больше между ними различия – тем ниже и положе холм.

Простая случайная выборка позволяет получить максимальное число различных выборок данного размера, а именно все существующие выборки. Применение стратификации или кластеризации сокращает число потенциально возможных выборок. В стратифицированной выборке всегда выдержаны заданные пропорции между стратами (пропорции зависят от способа размещения выборки). Те выборки, где эти пропорции нарушены, недопустимы. В кластерной выборке в каждом кластере отбирается заданное число людей. Выборки, в которых на кластер приходится другое число людей, также не попадут в число возможных.

Таким образом, когда исследователь собирается применить определенный тип стратифицированной или кластерной выборки, он тем самым исключает из числа возможных огромное число выборок, которые могли бы возникнуть при простом случайном отборе. Однако число оставшихся выборок все равно будет очень большим.

Как изменится вид рисунка, если на нем оставить только те кружочки, которые соответствуют типу применяемой выборки, а остальные стереть?

Начнем с выборки, стратифицированной по уровню дохода людей. Предположим, что у исследователя есть достоверные статистические данные о доле людей с низким, средним и высоким уровнем доходов и что он выдерживает эти пропорции в выборке. Раз пропорции выдержаны, значит не могут получиться такие выборки, в которых больше чем нужно людей имеют низкий доход или, наоборот, слишком много людей имеют высокий доход. На *рис. 1* этим выборкам

*Преимущество неслучайных типов выборки особенно сильно проявляется при исследовании малочисленных и трудновыделяемых целевых групп*

соответствуют множество кружочков, расположенных на левом и на правом краях. Все их надо стереть как недопустимые. Недопустимые выборки могут оказаться и в центре. Например, если выборка на 99% состоит из очень бедных людей и на 1% – из очень богатых, то посчитанный по такой выборке средний доход может совпасть с истинным средним доходом, то есть эта выборка будет расположена точно по центру рисунка. Однако в ней не выдержаны правильные пропорции между тремя доходными группами (в частности, полностью отсутствуют представители средней группы), значит, выборка недопустима. То есть недопустимым выборкам соответствуют все крайние кружочки, а также часть кружочков в центральной части. После их удаления с *рис. 1* холм станет значительно уже и выше. Напомним, что общая площадь, занятая кружочками, не меняется и остается равной 1, поскольку за единицу принято число всех допустимых выборок.

Если бы исследователь создавал страты не по уровню доходов, а, например, по возрасту людей, то изменения в рисунке были бы аналогичные, хотя и не такие существенные. Недопустимым выборкам соответствовали бы кружочки как по краям, так и по центру рисунка, но недопустимых выборок по краям было бы больше. В результате холм стал бы поуже и повыше, но не так значительно, как при стратификации по

[8] При небольших размерах выборки – обычно менее нескольких десятков человек – гладкость и симметричность склонов холма может нарушаться, к малым выборкам математическая теория уже неприменима.

уровню доходов.

Увеличение крутизны холма означает уменьшение величины доверительного интервала Δ, а значит, и уменьшение статистической погрешности. Поскольку в пропорциональной стратифицированной выборке погрешность никогда не увеличивается, следовательно, и ширина холма тоже не увеличивается, чаще всего он становится выше и у́же, в крайнем случае, остается таким же. При этом выборка всегда остается несмещенной, т. е. середина холма совпадает с истинным значением оцениваемого параметра. (При непропорциональном размещении выборки между стратами ширина холма в принципе может увеличиться, но выборка по-прежнему останется несмещенной, поскольку при вычислении среднего учитываются веса страт.)

Теперь перейдем к кластерным выборкам. Для них множество недопустимых кружочков будет расположено по всей площади холма, однако в центральной части их все же будет больше, чем по краям. Поэтому после удаления всех лишних кружочков холм станет ниже и шире, чем для простой случайной выборки. Это связано с тем, что дизайн-эффект кластерной выборки на практике всегда больше единицы, и она всегда приводит к увеличению статистической погрешности. Однако, как и стратифицированная выборка, кластерная выборка будет несмещенной, если все процедуры отбора проведены правильно.

Получается, что при стратификации исключается больше выборок, расположенных по краям холма, при кластеризации – больше выборок, расположенных в центре, но при этом и стратифицированная, и кластерная выборки всегда остаются несмещенными[9]. Соотношение между простой случайной, пропорциональной стратифицированной и кластерной выборками показано на *рис. 2*.

Несмещенность случайных выборок обеспечивается применяемым способом отбора, при котором каждому представителю совокупности стараются обеспечить равную вероятность попадания в выборку. В тех типах выборки, где при отборе людей применяются неравные вероятности, различия в вероятностях обязательно учитываются при вычислении средних значений (как это делается, например, в непропорциональных стратифицированных выборках).

Несмещенность каждого типа случайной выборки обосновывается в специальной литературе по выборочным методам (иногда в форме теорем с доказательствами). Для оценивания параметров применяются специальные формулы, вид которых зависит не только от типа выборки, но и от параметра. Есть формулы для оценки среднего (например, среднего дохода), для оценки доли людей (в процентах), для оценки отношения двух величин, для оценки параметров линейных регрессионных моделей и т. п. У каждого типа случайной выборки есть также свои особые формулы для вычисления величины статистической погрешности. Широко известные формулы для простой случайной выборки неприменимы к выборкам других типов.

А что можно сказать про неслучайные выборки? В разных типах неслучайных выборок имеются свои ограничения на способ отбора респондентов. В квотных выборках должны быть выдержаны заданные квоты. В целевых выбор-

[9] Напомним, что несмещенной является вся выборка целиком. Невозможность опросить кого-то из респондентов, включенных в выборку (из-за того, что интервьюеру не удается с ними встретиться, или из-за их отказа от участия в опросе), может приводить к смещениям, которые напрямую от выборки не зависят.

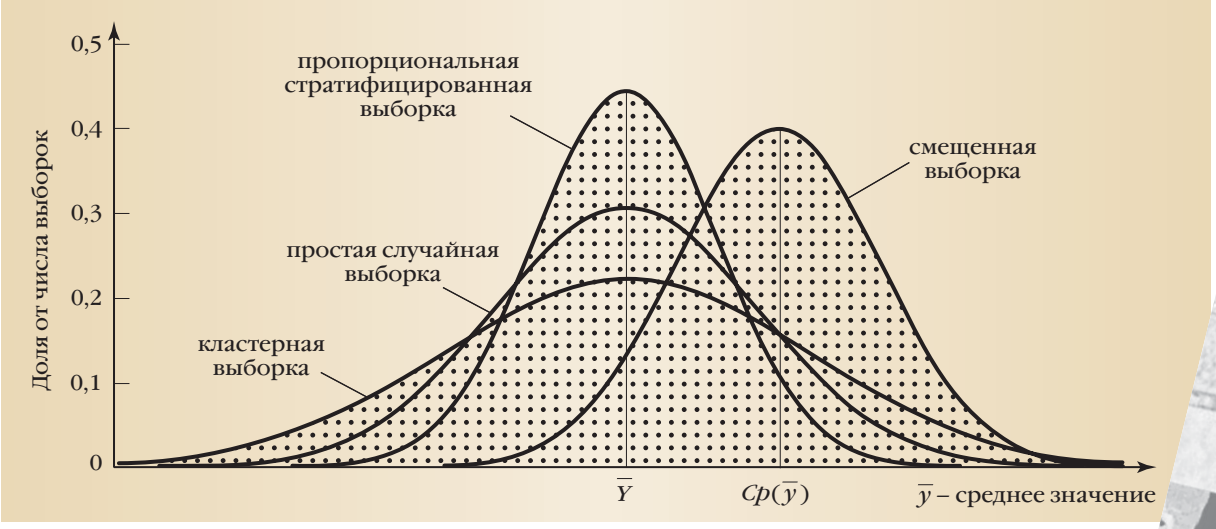



*Рис. 2.* ***Распределение оценки среднего значения параметра ȳ для разных типов выборок***

ках могут регламентироваться точки, в которых надо проводить опрос, например, может быть задан перечень "мест скопления". В выборке типичных единиц есть свои правила, по которым определяют, кого можно отнести к типичным представителям совокупности. Но в рамках формализованных правил отбора у интервьюера всегда остается свобода самому решить, кого именно ему опрашивать. Предпочтения интервьюеров, участвующих в опросе, тоже накладывают свои ограничения на множество возможных выборок. В выборке добровольцев и в опросах методом снежного кома вместо предпочтений интервьюеров в качестве ограничений действуют желания и настроения респондентов.

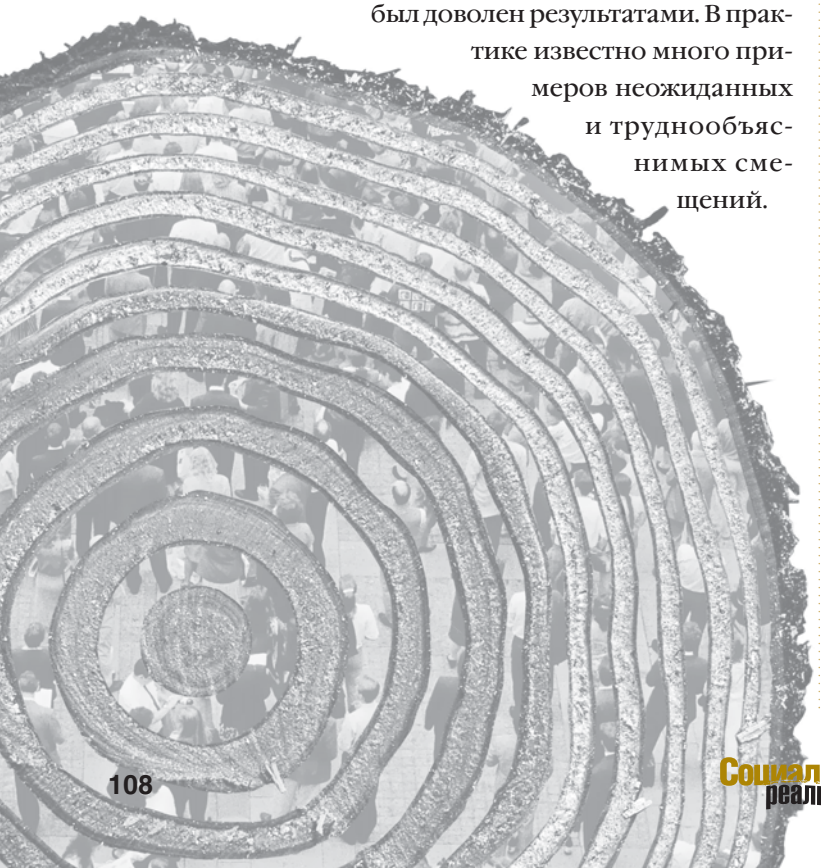
Для каждого типа неслучайной выборки существуют свои картинки, показывающие, как распределяются возможные выборки по осям разных параметров. Но поскольку здесь есть влияние предпочтений интервьюеров или респондентов, вид картинки точно предсказать невозможно. Какие именно кружочки на *рис. 1* окажутся недопустимыми и должны быть удалены, а какие останутся, зависит от многих субъективных факторов. В одних случаях выборки получатся несмещенными, а в других – могут очень сильно сместиться влево или вправо. Для неслучайных выборок не существует доказанных теорем и готовых формул для расчета погрешности. Поэтому исследователь может полагаться только на свой опыт и интуицию.

Строгое выполнение квот не гарантирует несмещенность выборки по другим, не квотируемым параметрам. Например, исследователь может добиться, чтобы квоты по полу и возрасту респондентов были выдержаны с точностью до одного человека. Это обеспечит несмещенные картинки по осям "доля мужчин" ("доля женщин") и "возраст". Все крайние точки на соответствующих картинках исчезнут, останутся только узкие и высокие столбики выборок, совпадающие с истинным процентом мужчин и истинным средним возрастом в совокупности. Но по оси "средний доход" картинка может оказаться иной, весь холм с выборками может сместиться влево относительно истинного среднего дохода. Это произойдет, например, если у большинст-

ва интервьюеров осознанно или неосознанно будет присутствовать установка: не опрашивать тех, чей социальный статус выше, чем у них самих (им может не нравиться немного пренебрежительное отношение к ним респондентов). Тогда в каждой половозрастной группе будут опрошены люди, чей доход ниже или примерно такой же, как у интервьюера, что приведет к смещению выборки по доходу.

Многие исследователи, применяющие неслучайные выборки, используют комбинацию случайных и неслучайных методов отбора. Например, при опросе населения сначала создается случайная выборка населенных пунктов, она формируется как кластерная выборка. Внутри этих населенных пунктов отбираются домохозяйства случайным маршрутным методом. Этот метод является реализацией случайного систематического отбора с определенным шагом. И только при отборе респондента в домохозяйстве применяются квоты, т. е. неслучайный отбор. Такие комбинированные методы отбора позволяют снизить вероятность смещения выборки, но все равно не дают полной гарантии.

Неслучайная выборка всегда может оказаться смещенной, даже в том случае, когда до этого исследователь уже неоднократно применял данный способ отбора респондентов и каждый раз был доволен результатами. В практике известно много примеров неожиданных и труднообъяснимых смещений.



## 9. Репрезентативные выборки

Термин “репрезентативная выборка” применяется довольно часто, однако разные люди вкладывают в это понятие различный смысл. Но практически все придают ему положительный оттенок, подразумевая, что репрезентативная выборка – это прежде всего хорошая, правильная выборка.

Если переводить этот термин буквально, то он означает выборку, репрезентирующую, то есть представляющую, изучаемую совокупность. Причем представляющую правильно, в отличие от нерепрезентативной выборки, которая не представляет изучаемую совокупность или представляет ее неправильно. Правильное представление совокупности означает в первую очередь отсутствие смещений, поэтому термин “репрезентативность” ближе всего примыкает к понятию несмещенности выборки. Ведь результаты опросов, проведенных по несмещенной выборке, можно смело распространять на всю совокупность. Именно это имеют в виду люди, когда говорят, что выборка репрезентирует все население или какую-то его часть – сельских жителей, молодежь, автовладельцев и т. п.

Еще один смысл, который вкладывают в понятие “репрезентативность”, заключается в том, что в выборке должны быть представлены все категории респондентов, что ни одна из категорий не должна быть пропущена при отборе. Такое понимание репрезентативности очень близко к определению случайной (вероятностной) выборки как выборки, в которой каждый представитель совокупности имеет известную ненулевую вероятность быть отобранным. Использование термина в таком контексте, на первый взгляд, нисколько не противоречит пониманию репрезентативности в смысле несмещенности выборки, поскольку случайные выборки являются несмещенными, – но все же некоторые различия есть. Например, в непропорциональной стратифицированной выборке соотношение между отдельными категориями людей может не совпадать с их пропорциями в совокупности. Размер выборки по некоторым категориям может быть сознательно увеличен или уменьшен, а несмещенность сред-

них оценок обеспечивается за счет использования при расчетах весов страт. Некоторые авторы называют такие выборки нерепрезентативными, хотя и признают необходимость и обоснованность нарушения пропорций между разными категориями респондентов. Отсутствие в выборке представителей отдельных малочисленных категорий людей, доля которых в совокупности не превышает величину статистической погрешности, тоже порой интерпретируется как нерепрезентативность выборки, хотя с точки зрения теории вероятности это вполне допустимо. Таким образом, в число нерепрезентативных во втором смысле попадают некоторые несмещенные случайные выборки, что свидетельствует о различии между таким пониманием репрезентативности и ее пониманием в смысле несмещенности.

Бывает, что после слов "репрезентативная выборка" идет уточнение, по каким именно параметрам она репрезентативна. Например, выборка репрезентативна по полу и возрасту, по типу места жительства, по национальному составу респондентов и т. д. Независимо от того, что имеет в виду автор, эти слова часто понимаются буквально. А именно, что пропорции в выборке и во всей совокупности по указанным параметрам совпадают. Нетрудно заметить, что это не совсем то же самое, что несмещенность выборки по этим параметрам. В простой случайной выборке возможны небольшие отклонения по любому параметру в пределах статистической погрешности. В начале статьи была посчитана величина погрешности для доли мужчин в выборке из 1000 человек – она равна ± 2,9%. Отклонения в этих пределах не означают смещенности выборки, однако могут восприниматься отдельными людьми как отсутствие ее репрезентативности по полу, то есть как признак некачественной выборки. И наоборот, репрезентативность по ряду параметров, в которой можно легко убедиться самостоятельно, часто воспринимается как надежное доказательство несмещенности выборки по всем параметрам, что, как известно, справедливо не всегда. Поэтому "репрезентативность по параметру" отличается от рассмотренных ранее понятий репрезентативности и не позволяет провести четкую границу между смещенными и несмещенными выборками, случайными и неслучайными.

Иногда говорят о статистической репрезентативности выборки, имея в виду, что ее точность должна соответствовать целям исследования. Для решения одних задач статистическая погрешность должна быть небольшой, например, не более 3%, для других задач приемлемая погрешность может составлять 10% и даже больше. Когда величина погрешности отвечает целям исследования, выборку называют статистически репрезентативной. Такое определение применимо только к случайным выборкам, для которых можно вычислить статистическую погрешность. Случайные выборки, имеющие недостаточную точность, попадают в категорию статистически нерепрезентативных.

Таким образом, понятие "репрезентативная выборка" имеет много значений, не совпадающих одно с другим. Бывает трудно понять, в каком смысле этот термин употребляется. Порой плохие выборки могут быть названы репрезентативными, и наоборот, хорошие, правильные выборки могут попасть в категорию нерепрезентативных.

Наверное, если подсчитать, какие слова чаще всего встречаются вместе со словом выборка, то сочетание "репрезентативная выборка" займет одно из первых мест. По распространенности его сможет опередить разве что сочетание "случайная выборка". Возможно, такое широкое употребление этих словосочетаний объясняется многообразием значений, которые в них вкладываются. Об этом важно помнить, встречая или используя эти выражения. ■

## Литература

*Kish L.* Survey Sampling. John Wiley and Sons, Inc., New York, 1965.

*Йейтс Ф.* Выборочный метод в переписях и обследованиях. М.: Статистика, 1965.

*Кокрен У.* Методы выборочного исследования. М.: Статистика, 1976.