ВЫБОРОЧНОЕ ИССЛЕДОВАНИЕ

Выборочное статистическое исследование – это обследование выборочной совокупности с целью получения достоверных суждений о характеристиках или параметрах генеральной совокупности.

Генеральная совокупность – это полная совокупность единиц (вся статистическая совокупность).

Выборочная совокупность (***выборка***) - это часть единиц генеральной совокупности, отобранная в случайном порядке.

Обозначения:
объем генеральной совокупности – N;
объем выборки - n

Почему выборочному наблюдению отдается предпочтение перед сплошным?

1) с целью экономии времени и средств в результате сокращения объема работы (при выборочном методе обследованию подвергается 5-10%, реже до 15-20% изучаемой совокупности);

2) чтобы свести к минимуму порчу или уничтожение исследуемых объектов (например, при определении прочности пряжи на разрыв нити, при испытании электрических лампочек на продолжительность горения, при проверке консервов на доброкачественность);

3) вследствие того, что исследуемая совокупность может быть полностью недоступна;

4) вследствие того, что исследуемая совокупность может не иметь конечного объема.

Наиболее часто исследуемые с помощью выборочного метода характеристики совокупности:

Статистическая характеристика (параметр)	В генеральной совокупности (г.с.)	В выборке (в.с.)
Среднее	$\overline{x} = \sum_{i=1}^{N} x_i / N$	$\tilde{x} = \sum_{i=1}^{n} x_i / n$
Доля альтернативного признака	$w = N_a / N$ N_a – число единиц с данным значением признака в г.с.	$\tilde{w} = n_a / n$ n_a – число единиц с данным значением признака в г.с.
Дисперсия	$\sigma^2 = \sum_{i=1}^{N} (x_i - \overline{x})^2 / N$	$\tilde{\sigma}^2 = \sum_{i=1}^{n} (x_i - \tilde{x})^2 / n$

По данным выборки мы не можем найти точное значение характеристики (параметра) генеральной совокупности, а можем только получить его приближенное значение (оценку).

Статистической оценкой (θ^*) характеристики (параметра) генеральной совокупности называют приближенное значение этой характеристики (параметра), полученное по некоторой функции от наблюдаемых в выборке значений признака X (x_1, x_2, ...x_n), т.е.:

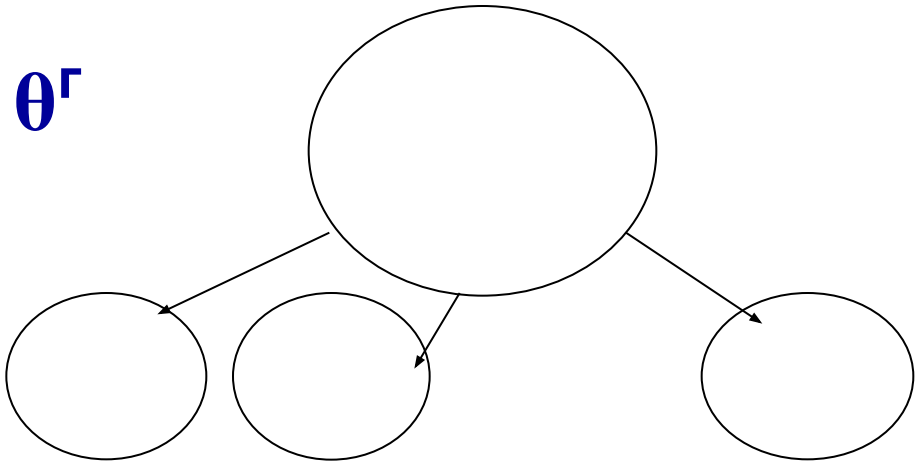
$$\theta^* = f(x_1, x_2, \ldots, x_n),$$

где n – объем выборки;

(x_1, x_2, ..., x_n) – рассматриваются как независимые случайные величины.

Функцию (f) называют **способом оценивания**.

Генеральная совокупность объемом N, $\theta^{г}$



Выборки: 1(n_1) 2 (n_2) ..... m (n_m)

θ^*_1 θ^*_2 ..... θ^*_m

m- всего выборок.

От выборки к выборке статистическая оценка (даже при одном и том же способе оценивания) меняется (θ^*_1, θ^*_2,…, θ^*_m).

Статистическая оценка (θ^*_j) представляет собой случайную переменную (т.к. сочетание значений признака X в выборке случайно, следовательно, случайным будет и значение функции от них).

Для одной и той же характеристики (параметра) генеральной совокупности может быть предложено несколько способов оценивания. Возникает проблема выбора лучшего способа оценивания.
Критерием выбора является требование состоятельности, несмещенности и эффективности оценки.

Способ оценивания дает **состоятельные** оценки, если при бесконечно большом объеме выборки значение статистической оценки стремится к искомому значению характеристики (параметра) генеральной совокупности.

Способ оценивания дает **несмещенные** оценки, если математическое ожидание оценки при данном способе оценивания тождественно искомой характеристике (параметру) генеральной совокупности (при любом объеме выборки), т.е. $M(\theta^*)=\theta^г$. Если математическое ожидание оценки не равняется характеристике генеральной совокупности, то оценка называется смещенной. И разность $M(\theta^*) - \theta^г$ называется смещением.

Способ оценивания дает **эффективные** оценки, если дисперсия оценки минимальна (при заданном объеме выборки *n*) в сравнении с другими способами отбора.

Статистическая оценка, полученная по данным выборки, отличается от генеральной характеристики (параметра) на величину ***ошибки выборки***.

Ошибка выборки состоит из двух частей: ошибки регистрации и ошибки репрезентативности.

Ошибки репрезентативности (представительности) возникают в результате того, что состав отобранной для обследования части единиц совокупности недостаточно полно отображает состав всей изучаемой совокупности (иначе говоря не все типы явления представлены в выборке).

В дальнейшем будем предполагать, что ошибка регистрации равна нулю. Следовательно, ошибка выборки равна ошибке репрезентативности.

Различают среднюю и предельную ошибки выборки.

Средняя ошибка выборки (μ) – это среднее (по выборкам) отклонение выборочной оценки от истинного значения генеральной характеристики.

В каждой конкретной выборке фактическая ошибка выборки может быть меньше средней ошибки, равна ей или больше ее. Причем каждое из этих расхождений имеет различную вероятность.

Предельная ошибка выборки (Δ) – это максимально возможная при данной вероятности ошибка выборки.

То есть мы с заданной вероятностью (Рдов) гарантируем, что оценка ,полученная по нашей конкретной выборке, будет отличаться от значения генеральной характеристики не больше, чем на величину предельной ошибки Δ.

Вероятность, с которой мы гарантируем, что ошибка нашей выборки не превысит предельную ошибку, называется ***доверительной вероятностью*** - Рдов. Предельная ошибка рассчитывается по формуле:

$$\Delta = t \cdot \mu,$$

где t- коэффициент доверия, значение которого определяется доверительной вероятностью (Рдов). Чем больше Рдов, тем больше t.

Закон больших чисел – методологическая основа выборочного метода.

Теоретической основой выборочного метода является закон больших чисел:

С увеличением объема выборки вероятность появления больших ошибок и пределы максимально возможной ошибки уменьшаются (т.е. чем больше обследуется единиц, тем меньше будет величина расхождений выборочных и генеральных характеристик).

Математически данный закон записывается через неравенство П.Л.Чебышева:

$$P\left(\left|\tilde{x}-\bar{x}\right|\le\varepsilon\right)\to 1$$

$$при\ n\to\infty\ \varepsilon\to 0$$

где ε - ошибка выборки; n – объем выборки;

$\tilde{x}$ - выборочное среднее;

$\bar{x}$ - генеральное среднее.

Следует отметить, что данное неравенство справедливо для генеральной совокупности с ограниченной дисперсией.

Центральная предельная теорема А.М.Ляпунова: При достаточно большом числе независимых наблюдений вероятность того, что расхождение между выборочной и генеральной средней не превысит по модулю некоторую величину **μ·t**, равна интегралу Лапласа Ф(t):

$$P\left(\left|\tilde{x}-\overline{x}\right|\le \mu\cdot t\right)=\Phi(t);\quad \Phi(t)=\frac{1}{\sqrt{2\pi}}\int_{-t}^{t}e^{-\frac{t^2}{2}}dt$$

(это справедливо для генеральной совокупности с конечной средней и ограниченной дисперсией).

Данная теорема позволяет указать вероятность появления ошибок определенной величины.

t	1,00	1,64	1,96	2,00
Рдов=Ф(t)	0,683	0,900	0,950	0,954

Из центральной предельной теоремы следует важный вывод:
при достаточно большом числе независимых наблюдений (объеме выборки) распределение отклонений выборочных средних от генеральной средней (а, следовательно, и самих выборочных средних) приближенно нормально.

При небольшом объеме выборки (n<30)

$$Pдов = P(|\tilde{x} - \bar{x}| \le \mu \cdot t) = F(t)$$

- интегральная функция распределения Стьюдента.

Классификация способов отбора

*1. П*овторный и бесповторный отбор

При **повторном** отборе общая численность единиц генеральной совокупности в процессе выборки остается неизменной. Единицу, попавшую в выборку, после регистрации снова возвращают в генеральную совокупность, и она сохраняет равную возможность со всеми прочими единицами на следующем шаге отбора вновь попасть в выборку. Повторная выборка в социально-экономической жизни встречается редко.

При **бесповторном** отборе единица совокупности, попавшая в выборку, в генеральную совокупность не возвращается и в дальнейшем отборе не участвует. Таким образом, при бесповторном отборе численность единиц генеральной совокупности сокращается в процессе выборки.

2. Отбор может быть организован как :
- собственно-случайный;
- механический;
- стратифицированный (типический);
- серийный

Собственно-случайный отбор – такой отбор единиц из генеральной совокупности, когда на включение (исключение) единицы в выборку (из выборки) не может повлиять какой-либо фактор кроме случая. Технически он осуществляется посредством жеребьевки или таблиц случайных чисел. При этом необходимо иметь список единиц генеральной совокупности.
Примером может служить отбор студентами на экзамене экзаменационных билетов.

Механический отбор - это бесповторный отбор элементов из генеральной совокупности, упорядоченной по нейтральному (несущественному для цели исследования) признаку через равные интервалы. Механический отбор по результатам близок к бесповторному собственно-случайному.
Примеры:
Отбор каждой 20-й детали, сходящей с конвейера для проверки ее качества. Здесь нейтральный признак – номер детали.

При исследовании успеваемости студентов вуза в качестве нейтрального признака можно взять фамилию, имя и отчество студента. Всех студентов упорядочивают по Ф.И.О. После чего отбирают заданное число студентов по фамилиям механически, через определенный интервал.

Размер интервала в генеральной совокупности равен обратному значению доли выборки. Так, при 2%-ой выборке отбирается и проверяется каждая 50-я единица (1/0,02), при 5%-ой выборке – каждая 20-ая единица (1/0,05).

Стратифицированный отбор используют для отбора единиц из неоднородной совокупности, когда все единицы генеральной совокупности можно разбить на несколько качественно однородных групп по существенным для цели исследования признакам. Из каждой такой группы собственно-случайным или механическим способом производится индивидуальный отбор единиц в выборку. Стратифицированный отбор, при котором пропорции между группами в выборке совпадают с пропорциями между группами в генеральной совокупности, называется **типическим** отбором.

Серийный отбор представляет случайный отбор из генеральной совокупности не отдельных единиц, а их равновеликих групп (серий) с тем, чтобы в таких группах подвергать наблюдению все без исключения единицы.

Серийный отбор применяют в том случае, когда исследуемый признак колеблется внутри серий незначительно.

Применение серийной выборки обусловлено тем, что многие товары для их транспортировки, хранения, продажи упаковываются в пачки, ящики и т.п. Поэтому при контроле качества упакованного товара рациональнее проверить несколько упаковок (серий), чем из всех упаковок отбирать необходимое количество товара.

Выборки также делят на ***большие*** (с объемом большим или равным 30 единицам) и ***малые*** (с объемом меньше 30 единиц).

ТОЧЕЧНОЕ И ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

Точечной называют оценку (θ*), которая определяется одним числом. При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т.е. приводить к грубым ошибкам.

Интервальной называют оценку, которая определяется двумя числами – концами интервала. Интервальные оценки позволяют установить точность оценки (величину предельной ошибки выборки) и надежность оценки (вероятность, с которой гарантирован результат оценивания). Интервальная оценка **(θ*-Δ; θ*+Δ)** представляет собой доверительный интервал.

Вероятность того, что доверительный интервал не покроет генеральную характеристику (параметр) совокупности обозначают **α** и называют **уровнем значимости**: **α =1- Рдов**.

При Рдов=0,95 α=0,05;

при Рдов=0,99 α=0,01.

Порядок расчета интервальной оценки характеристики (параметра) генеральной совокупности:

1. Определяют точечную оценку характеристики (параметра) генеральной совокупности (θ*).

Характеристика	Наилучшая точечная оценка
Среднее	$\tilde{x} = \sum_{i=1}^{n} x_i / n$ выборочное среднее
Доля альтернативного признака	выборочная доля $\tilde{w} = n_a / n$
Дисперсия	исправленная выборочная дисперсия $s^2 = \tilde{\sigma}^2 \cdot n/(n-1) = \sum_{i=1}^{n} (x_i - \tilde{x})^2 /(n-1)$

2. Рассчитывают среднюю ошибку выборки - μ. Формулы расчета средней ошибки выборки -μ зависят от способа отбора и от вида оцениваемой характеристики генеральной совокупности (среднее или доля).

Собственно –случайный отбор

Способ отбора	Среднее	Доля альтернативного признака
повторный	$\sqrt{\frac{\sigma^2}{n}} \approx \sqrt{\frac{s^2}{n}}$	$\sqrt{\frac{w(1-w)}{n}} \approx \sqrt{\frac{\tilde{w}(1-\tilde{w})}{n-1}}$
бесповторный	$\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)} \approx \sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$	$\sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$

Механический и типический способы отборы

Способ отбора	Среднее	Доля альтернативного признака
механический	$\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)} \approx \sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$	$\sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$
Типический повторный	$\sqrt{\frac{\varepsilon^2}{n}} \approx \sqrt{\frac{\tilde{\varepsilon}^2}{n-1}}; \tilde{\varepsilon}^2 = \frac{\sum_{j=1}^{k} \tilde{\sigma}_j^2 \cdot n_j}{\sum_{j=1}^{k} n_j}$ остаточная (средняя из внутригрупповых) дисперсия выборки	$\sqrt{\frac{\varepsilon_w^2}{n}} \approx \sqrt{\frac{\tilde{\varepsilon}_w^2}{n-1}}$ $\tilde{\varepsilon}_w^2 = \frac{\sum_{j=1}^{k} \tilde{w}_j(1-\tilde{w}_j)n_j}{\sum_{j=1}^{k} n_j}$
Типический бесп.	$\sqrt{\frac{\varepsilon^2}{n}\left(1-\frac{n}{N}\right)} \approx \sqrt{\frac{\tilde{\varepsilon}^2}{n-1}\left(1-\frac{n}{N}\right)}$	$\sqrt{\frac{\varepsilon_w^2}{n}\left(1-\frac{n}{N}\right)}$

Серийный отбор

Способ отбора	Среднее	Доля альтернативного признака
повторный	$\sqrt{\frac{\delta^2}{r}} \approx \sqrt{\frac{\tilde{\delta}^2}{r-1}}; \tilde{\delta}^2 = \frac{\sum_{j=1}^{r}(\tilde{x}_j - \tilde{x})^2}{r}$ Межсерийная дисперсия Число серий в выборке	$\sqrt{\frac{\delta_w^2}{r}} \approx \sqrt{\frac{\tilde{\delta}_w^2}{r-1}}$ $\tilde{\delta}_w^2 = \frac{\sum_{j=1}^{r}(\tilde{w}_j - \tilde{w})^2}{r}$
беспов торный	$\sqrt{\frac{\delta^2}{r}\left(1-\frac{r}{R}\right)} \approx$ $\approx \sqrt{\frac{\tilde{\delta}^2}{r-1}\left(1-\frac{r}{R}\right)}$ Число серий в ген. совокупности	$\sqrt{\frac{\delta_w^2}{r}\left(1-\frac{r}{R}\right)}$

3. Рассчитывают предельную ошибку выборки:
t·μ, При большом объеме выборки (≥30) значение
эффициента доверия t находим из таблиц
тегральной функции стандартного нормального
спределения по заданной доверительной вероятности
ов.
и небольшом объеме выборки (n<30) значение t
ределяют по таблицам интегральной функции
спределения Стьюдента.
начение t по таблицам Стьюдента будет чуть больше,
м по таблицам стандартного нормального
спределения.)

4. Определяют границы доверительного интервала:

 $(\theta^* - \Delta; \theta^* + \Delta)$ – интервальная оценка.

Вывод: с вероятностью Рдов данный интервал покроет генеральную характеристику (параметр).

Пример 1: Из партии готовой продукции в порядке механической выборки проверено 50 лампочек на продолжительность горения. Средняя продолжительность горения лампочки оказалась равной 840 ч. при среднем квадратическом отклонении 60 ч.

С вероятностью 0,95 определить **доверительные пределы средней** продолжительности горения лампочки в генеральной совокупности (партии продукции).

РЕШЕНИЕ:

Для построения доверительного интервала $(\theta^*-\Delta;\ \theta^*+\Delta)$ в качестве точечной оценки θ^* возьмем выборочное среднее арифметическое. По условию оно равно 840 ч.

Чтобы рассчитать предельную ошибку $\Delta=t\cdot\mu$ нужно определить среднюю ошибку μ. В случае механического отбора и оценке среднего воспользуемся формулой:

$$\mu \approx \sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)} = \sqrt{\frac{\tilde{\sigma}^2}{n-1}\left(1-\frac{n}{N}\right)} = \sqrt{\frac{60^2}{49}(1-0)} = 8{,}6(ч.)$$

Значение t найдем по таблицам стандартного нормального распределения, так как в нашем случае выборка большая (ее объем равный 50 > 30). Для Рдов=0,95 по таблице стандартного нормального распределения t=1,96.
Тогда $\Delta = 1{,}96 \cdot 8{,}6 = 16{,}86$ (ч.). То есть с вероятностью 0,95 можно утверждать, что средняя продолжительность горения лампочки в нашей выборке отличается от этой же характеристики в генеральной совокупности не более чем на 16,6 часа.
Теперь можем построить доверительный интервал:
(840 – 16,86; 840 + 16,86) или (823,14; 856,86).

Вывод: с вероятностью 0,95 можно утверждать, что средняя продолжительность горения в генеральной совокупности (т.е. во всей партии) не выйдет за пределы от 823 ч. до 857 ч.

Пример 2: За некоторый период времени рабочий изготовил 2000 деталей. Выборочно (методом собственно-случайного бесповторного отбора) проверено 120 деталей. Оказалось, что из них 4 бракованные. Требуется с вероятностью 0,90 определить доверительные пределы **доли бракованных деталей** среди всех изготовленных рабочим за этот период (т. е. в генеральной совокупности).

РЕШЕНИЕ:

В данном случае требуется построить доверительный интервал для доли альтернативного признака (w). точечной оценкой показателя доли является выборочная доля:

$$\widetilde{w} = 4/120 = 0{,}033$$

То есть среди проверенных деталей 0,033 (или 3,3%) оказалось бракованных.

Для определения границ доверительного интервала нам нужно найти предельную ошибку Δ, а чтобы найти Δ требуется определить среднюю ошибку μ.
Формула расчета в данном случае (собственно-случайный бесповторный отбор; характеристика – доля):

$$\mu \approx \sqrt{\frac{\tilde{w}(1-\tilde{w})}{n}\left(1-\frac{n}{N}\right)} = \sqrt{\frac{0{,}033\cdot(1-0{,}033)}{120}\left(1-\frac{120}{2000}\right)} =$$

$$= \sqrt{0{,}000252} = 0{,}016$$

То есть в среднем отклонение выборочной доли от генеральной составит 0,016.

Теперь найдем коэффициент доверия t по таблице стандартного нормального распределения, т.к. выборка большая (n=120>30). Для Рдов=0,90 t=1,64.

Тогда $\Delta = 1{,}64 \cdot 0{,}016 = 0{,}026$. Теперь можем построить доверительный интервал: $(0{,}03 - 0{,}026;\ 0{,}03 + 0{,}026)$ или $(0{,}004;\ 0{,}056)$.

Вывод: с вероятностью 0,9 можно утверждать, что доля бракованных деталей в общем объеме изготовленных рабочим (в генеральной совокупности) будет в пределах от 0,004 до 0,056 или от 0,4% до 5,6%

Другая задача, решаемая с помощью выборочного метода: определение необходимого объема выборки - n при заданной точности (Δ) и надежности (Рдов) оценивания.

Формулы расчета для собственно –случайного отбора:

харак-ка	повторный отбор	бесповторный отбор
Среднее	$n = \frac{s^2 \cdot t^2}{\Delta^2}$	$n = \frac{s^2 \cdot t^2 \cdot N}{s^2 \cdot t^2 + \Delta^2 \cdot N}$
Доля альтернативного признака	$n = \frac{\tilde{w}(1-\tilde{w}) \cdot t^2}{\Delta^2}$	$n = \frac{\tilde{w}(1-\tilde{w}) \cdot t^2 \cdot N}{\tilde{w}(1-\tilde{w}) \cdot t^2 + \Delta^2 \cdot N}$

Пример 3: На городской телефонной станции в порядке собственно-случайной выборки проводится обследование телефонных разговоров с целью определения сред. продолжительности разговора. Сколько телефонных разговоров требуется обследовать, чтобы с вероятностью 0,95 предельная ошибка (точность) при определении средней продолжительности разговора не превышала 1 мин. (В порядке пробного обследования исправленное среднее квадратическое отклонение длительности разговора составило 5 мин.)

РЕШЕНИЕ: Необходимый объем выборки можно определить по формуле:

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

Дисперсия (s^2) по условию равна $5^2 = 25$. При Рдов=0,95 t=1,96.

Тогда объем выборки будет равен:

$$n = \frac{s^2 \cdot t^2}{\Delta^2} = \frac{25 \cdot 1{,}96^2}{1^2} = 96$$

Вывод:

96 телефонных разговоров требуется обследовать, чтобы с вероятностью 0,95 предельная ошибка (точность) при определении средней продолжительности разговора не превышала 1 мин.

Пример 4: На основе данных примера 2, ответьте на вопрос: сколько еще деталей требуется обследовать, чтобы снизить предельную ошибку (точность) до 1% (0,01).

РЕШЕНИЕ: Необходимый объем выборки можно определить по формуле:

$$n = \frac{\tilde{w}(1-\tilde{w})\cdot t^2 \cdot N}{\tilde{w}(1-\tilde{w})\cdot t^2 + \Delta^2 \cdot N} = \frac{0{,}033(1-0{,}033)\cdot 1{,}64^2 \cdot 2000}{0{,}033(1-0{,}033)\cdot 1{,}64^2 + 0{,}01^2 \cdot 2000} \approx$$

$$\approx 605(дет.)$$

605 – 120 = 485 (дет.)

Вывод: 485 деталей требуется обследовать дополнительно, чтобы с вероятностью 0,90 предельная ошибка (точность) при определении доли брака у рабочего не превышала 1 %.