ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ

- Все единицы изучаемого явления называются *генеральной совокупностью,* а отдельная часть этих единиц, отобранных из генеральной совокупности для непосредственного наблюдения, именуется *выборочной совокупностью.*

- Таким образом, выборочная совокупность *репрезентует* (представляет) всю генеральную совокупность.

научно обоснованные *способы отбора* единиц выборочной совокупности

а) выборка из генеральной совокупности должна быть проведена *случайно,* то есть каждая ее единица должна иметь такую же вероятность попасть в выборку, как и остальные (так, например, отобранные наилучшие или наихудшие единицы не отображают действительное распределение признака в генеральной совокупности);

б) выборка должна быть осуществлена из *однородной* совокупности, так как при других обстоятельствах результаты выборки будут не точными и не могут в полной мере представлять генеральную совокупность.

- Различают два принципиально разных способа формирования выборочной совокупности:

а) *повторная выборка,* когда отобранная из генеральной совокупности занумерованная единица фиксируется и снова возвращается на свое место, после чего пачка номеров единиц генеральной совокупности тщательным образом перемешивается; этот способ отбора на практике является ограниченным из-за нецелесообразности, а иногда и невозможности повторного

б) *бесповторная выборка,* когда отобранный из пачки номер единицы генеральной совокупности откладывается в сторону и не возвращается обратно в пачку; этот способ отбора характеризуется повышенной степенью точности, надежности выборки и чаще всего используется на практике.

В статистической практике различают такие *разновидности* выборки:

- по способу организации выборочного обследования:

- простая случайная выборка;
- механическая выборка;
- районированная (типическая) выборка;
- серийная выборка;
- ступенчатая выборка.

по *степени охватывания* единиц обследуемой совокупности выборки:

- большие (при $n = 30$);
- малые (при $n < 30$).

Характеристики генеральной и выборочной совокупностей

Рассматриваем изучение признака *X* в гене ральной совокупности объема *N* единиц.

Генеральная совокупность представляется вариационным рядом, но это распределение неизвестно и стоит задача его определения.

Обобщающими характеристиками этого ряда будут:

- *генеральная средняя:*

$$\bar{x} = \frac{\sum_{i-1}^{M} x_i F_i}{N};$$

- *генеральная дисперсия :*

$$\bar{\sigma}^2 = \frac{\sum_{i-1}^{M} (x_i - \bar{x})^2 F_i}{N}$$

- *генеральное среднее квадратическое отклонение*

$$\overline{\sigma} = \sqrt{\overline{\sigma}^2} = \sqrt{\frac{\sum_{i-1}^{M} (x_i - \overline{x})^2 F_i}{N}}$$

- *доля единиц признака* генеральной совокупности *p*, то есть часть единиц M, которая обладает данным значением признака в общем объеме N единиц генеральной совокупности:

$$p = \frac{M}{N}$$

- Ц*ель выборочного исследования* заключается в том, чтобы, отобрав из генеральной совокупности *n* единиц, обследовать их и на этой основе оценить неизвестные нам генеральные характеристики. Вариация признака *x* в выборочной совокупности объемом *n* может быть представлена в виде вариационного ряда, который1 в общем случае отличаеся от вариационного ряда, представляющего генеральную совокупность, но характеристики которого могут быть определены.

Обобщающими характеристиками выборочной совокупности будут:

1) *выборочная средняя*

$$\tilde{x} = \frac{\sum_{i=1}^{m} x_i f_i}{n}$$

2) *выборочная дисперсия*

$$\sigma_в^2 = \frac{\sum_{i=1}^{m} (x_i - \tilde{x})^2 f_i}{n}$$

3) выборочное среднее квадратическое отклонение δ;

4) доля единиц признака выборочной совокупности w, то есть отношение количества единиц выборочной совокупности m, которая обладает данным признаком, к объему выборочной совокупности n:

$$w = \frac{m}{n}$$

5) часть выборки $w_в$ как отношение объема выборки к объему генеральной совокупности

$$w_в = \frac{n}{N}$$

Ошибки выборочного наблюдения

- *Ошибками выборки* называются некоторые расхождения характеристик генеральной и выборочной совокупности. Они включают ошибки регистрации и репрезентативности.

- *Ошибками регистрации* называют такие, которые возникают в результате получения неточных или неверных сведений от отдельных единиц совокупности из-за несовершенства измерительных приборов, недостаточной квалификации наблюдателя, недостаточной точности расчета и т. п. Эти ошибки должны быть исключены или сведены к минимуму.

Ошибки репрезентативности разделяют на

- систематические
- случайные.

Систематические ошибки репрезентативности возникают в результате особенностей принятой системы накопления и обработки данных наблюдения или из условий несоблюдения правил отбора в выборочную совокупность.

Такие ошибки также должны быть исключены

- *Случайные ошибки* репрезентативности возникают прежде всего из-за того, что выборочная совокупность при ее малом объеме не всегда точно воспроизводит характеристики генеральной совокупности. Поэтому этот вид ошибок выборки является основным, и задание выборочного метода заключается в получении таких выборочных характеристик, которые бы как можно точнее воспроизводили характеристики генеральной совокупности, то есть давали наименьшие ошибки репрезентативности.

Закон больших чисел

Выборочный метод наблюдения основан на вероятном подходе, теоретической базой для которого является *закон больших чисел.*

Сущность закона больших чисел заключается в том, что при увеличении численности единиц совокупности постепенно уменьшается элемент случайности в обобщенных характеристиках совокупности.

На основе закона можно утверждать, что при достаточно большом объеме выборки (n=30) выборочные характеристики мало отличаются от генеральных, в результате чего используются приближенные зависимости для средней, доли, дисперсии, среднем квадратическом отклонении:

$$\bar{x} \approx \tilde{x};\ \ p \approx w;\ \sigma^2 \approx \sigma_в^2;\ \sigma \approx \sigma_в$$

Теорема Чебышева

- при неограниченном увеличении количества независимых наблюдений в генеральной совокупности при ограниченной дисперсии с вероятностью, сколь угодно приближенной к единице, можно утверждать, что выборочные характеристики (средняя, доля) будут достаточно мало отличаться от соответствующих генеральных характеристик, то есть

$$P(|\tilde{x} - \bar{x}| < \varepsilon) \to 1 \ \ при \ \ n \to \infty$$

Теорема Ляпунова

- при достаточно большом количестве независимых наблюдений в генеральной совокупности с ограниченной дисперсией вероятность того, что величина отличия между выборочной и генеральной средней не превышает по абсолютной величине некоторого значения Δ и равняется интегралу Лапласа, то есть

$$P(|\tilde{x} - \bar{x}| \leq \Delta) = \Phi(t)$$

где Δ — *предельная ошибка выборки,* или максимально возможная для принятой вероятности *P:*

$$\Delta = t\mu$$

μ — *средняя квадратическая (стандартная) ошибка выборки;*

t — коэффициент доверия, который показывает соотношение предельной и стандартной ошибок и зависит от значения вероятности P;

Ф(t) — *интеграл Лапласа*

$$\Phi(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-\frac{t^2}{2}} dt$$

- Из теоремы Ляпунова следует, что при достаточно большом количестве независимых наблюдений распределение выборочных средних и их отклонение от генеральной средней приближено к *нормальному закону распределения.*

Простая случайная выборка

При *простой случайной выборке* отбор единиц осуществляется из всей массы единиц генеральной совокупности без предварительного распределения ее на любые группы и единицы отбора совпадают с единицами наблюдения.

С практической точки зрения преимущество отдается простой бесповторной выборке

Важным условием репрезентативности случайного отбора является то, что каждой единице генеральной совокупности предоставляется одинаковая возможность попасть в выборочную совокупность. Именно принцип случайности попадания любой единицы генеральной совокупности в выборку предотвращает возникновение систематических ошибок отбора.

При простой случайной выборке (как и в других видах выборочного наблюдения) возможно решение таких *задач:*

- определение *ошибки* выборочного наблюдения;

- определение *границ генеральных характеристик* на основе выборочных с заданной доверительной вероятностью (степенью надежности);

- определение *доверительной вероятности* того, что генеральные характеристики могут отличаться от выборочных не более определенной заданной величины;

- нахождение необходимой *численности выборки,* которая с практической достоверностью обеспечивала бы заданную точность выборочных характеристик.

Решение первой задачи

Средняя квадратическая ошибка бесповоротной выборки m определяется по формулам:

а) для средней $\mu_x = \sqrt{\frac{\sigma_s^2}{n}(1-\frac{n}{N})}$

б)для доли $\mu_w = \sqrt{\frac{w(1-w)}{n}(1-\frac{n}{N})}$

На основе теоремы Ляпунова предельная ошибка выборки равна

$$\Delta = t\mu$$

Коэффициент доверия t при определении предельной ошибки зависит от принятого уровня вероятности Р:

так, при t=1,0 значение вероятности Р=0,683;

t=1,96— для вероятности P = 0,950;

t=2,0 — для вероятности P = 0,954;

t = 3,0 — для вероятности P=0,997 .

Решение второй задачи

Оценка по данным выборки характеристик генеральной совокупности

а) для средней

$$\tilde{x} - \Delta_x \le \bar{x} \le \tilde{x} + \Delta_x;$$

б) для доли

$$w - \Delta_w \le \bar{p} \le w + \Delta_w$$

Эти формулы устанавливают границы, в которых при заданной доверительной вероятности находится неизвестная величина оцениваемого параметра: средней $\bar{x}$ или доли *p* в генеральной совокупности. Вероятность того, что величина генеральной средней или доли выйдет за доверительные границы, равняется

$$\alpha = 1 - P$$

и называется *уровнем значимости.*

Решение третьей задачи

Доверительная вероятность *P,* которую необходимо вычислить по теореме Ляпунова, является функцией от коэффициента t:

$$P = \Phi(t),$$

где Ф(t) — интеграл Лапласа.

Значение *t,* в свою очередь, может быть определено через предельную и стандартную ошибки

$$t = \frac{\Delta}{\mu}$$

вычисленными относительно средней или доли.

Наконец, по найденным значениям *t* из справочных таблиц находится интеграл Лапласа, отвечающий разыскиваемой веро ятности *P,* которая сравнивается с заданной величиной.

Решение четвертой задачи

а) для средней

$$n = \frac{t^2 \sigma_s^2 n}{\Delta_x^2 N + t^2 \sigma_s^{\,2}}$$

б) для доли

$$n = \frac{t^2 N(1-w)w}{\Delta_w^2 N + t^2 w(1-w)}$$

Механическая выборка

Механической называется такая выборка, при которой генеральная совокупность объемов N единиц, расположенных в определенном порядке (по увеличению или уменьшению, по алфавиту, географическому положению и т. п.), разделяется на *n* равных частей, и из каждой части обследуется одна единица.

Отношение

$$\frac{N}{n}$$

называется *интервалом* выборки.

Например, если отбор составляет 5% от генеральной совокупности работающих на предприятии, размещенных в списке в алфавитном порядке, то обследуют каждого 20-го работающего (5% — это 1/20 списочного состава работающих).

Интервал выборки будет равняться

$$\frac{100}{5} = 20\%$$

За начало отсчета при обследовании генеральной совокупности принимают или *начальную единицу,* определенную случайным отбором (при неблагоприятном размещении единиц генеральной совокупности)

или *середину первого интервала* (если единицы в списке размещены по определенному признаку — увеличению или уменьшению).

Механическая выборка очень удобна в случаях, когда уже есть списки единиц, составленные в том или другом порядке, или тогда, когда мы не можем предварительно составить список единиц генеральной совокупности, которые появляются постепенно в течение какого-то периода (например: при изучении покупок в магазине обследовать каждого 10-го покупателя; при контроле качества продукции — проверить каждую 5-ую деталь, которая сошла со станка).

Ошибки выборки при механическом отборе единиц вычисляют по формулам простой случайной бесповторной выборки.

С целью экономии времени и средств иногда бывает удобно обследовать не всю выборочную совокупность, а часть ее, то есть осуществить *подвыборку* из единиц первичной выборки.

Этот способ называют *двухфазным,* а при наличии нескольких подвыборок — *многофазным.*

Многофазный способ чаще всего используют в тех случаях, когда количество необходимых для определения показателей имеет разную точность (например, в случаях разной степени вариации показателей).

Ошибки при многофазной выборке рассчитываются на каждой фазе отдельно.

Иногда бывает целесообразным взять из совокупности две или больше независимых между собой выборок, используя для каждой из них одинаковый способ отбора.

Такие выборки называют *взаимопроникаемыми выборками.* Преимущество таких выборок заключается в том, что они позволяют получить отдельные и независимые оценки тех или других признаков совокупности.

Районированная (типическая) выборка

Районированной выборкой называют такой способ отбора, который осуществляется на основе распределения количества отобранных единиц и между районами (группами), которые присутствуют в генеральной совокупности.

В качестве районов, в зависимости от характера генеральной совокупности, могут быть приняты территориальные области, отрасли производства, отдельные предприятия, социальные группы населения и т. п.

Если генеральная совокупность разделяется на *m* частей, групп, районов, то есть $N=N_1+N_2+...+N_i+...+N_m$, то и выборочная совокупность должна формироваться из *m* частей так, чтобы $n=n_1+n_2+...+n_i+...+n_m$.

Способы распределения между районами

а) *пропорциональный,* когда количество отобранных в выборку единиц является пропорциональным к удельному весу района в генеральной совокупности, то есть количество наблюдений в каждом районе рассчитывается по формуле:

$$n_i = \frac{N_i}{N}$$

б) непропорциональным, если из каждого района отбирают одинаковое количество единиц:

$$n_i = \frac{n}{k}$$

где k— количество выделенных районов;

в) *оптимальным,* которое учитывает и численность района N_i, и среднее квадратическое отклонение признака в районе y_i; тогда численность каждого района выборки n_i рассчитывается по формуле:

$$n_i = \frac{\sigma_i N_i}{\sum_{i=1}^{m} \sigma_i N_i} n$$

На практике в большинстве случаев применяют первый и третий способы распределения между районами. Но использование оптимального размещения осложняется тем, что мы не всегда имеем данные о величинах y_i в генеральной совокупности. Поэтому в таких случаях используется наиболее часто применяемое пропорциональное распределение между районами.

Формулы расчета *средней квадратической ошибки выборки* при бесповторном отборе внутри районов для пропорционального способа распределения между районами

а) *для средней*

$$\mu_{\bar{x}} = \sqrt{\frac{\overline{\sigma}_i^2}{n}\left(1 - \frac{n}{N}\right)}$$

где $\overline{\delta}_s^2$ — средняя из дисперсий районов выборки

$$\overline{\sigma}_s^2 = \frac{\sum_{i=1}^{m} \sigma_i^2 n_i}{n}$$

б) для доли

$$\mu_w = \sqrt{\frac{\overline{w(1-w)}}{n}(1-\frac{n}{N})}$$

где $\overline{w(1-w)}$ - средняя из частей районов

$$w(1-w) = \frac{\sum_{i=1}^{m} w_i(1-w_i)n_i}{\sum_{i=1}^{m} n_i}$$

Необходимая численность выборки при бесповторном отборе внутри районов

а)*для средней*

$$n = \frac{t^2 \overline{\sigma}_6^2 n}{\Delta^2 N + t^2 \overline{\sigma}_6^{\,2}}$$

б) *для доли*

$$n = \frac{t^2 \overline{w}(\overline{1} - \overline{w})}{\Delta^2 N + t^2 \overline{w}(\overline{1} - \overline{w})}$$

Разновидностью районированной выборки является *типическая выборка.* При таком отборе районы генеральной совокупности выделяются по признаку, который изучается. Так, например, для определения среднего возраста студентов можно разделить их на группы, которые имеют или не имеют производственного стажа. Таким образом получаем «тип» с точки зрения принятого признака группы и увеличиваем точность выборки.

Серийная выборка

При серийной выборке отбору подлежат отдельные серии (группы, гнезда) единиц генеральной совокупности.

На практике часто встречается отбор с равными сериями. В отобранных сериях методом случайного бесповторного или механического отбора проводят сплошное наблюдение всех единиц, которые в них вошли.

Поскольку при серийной выборке каждая серия выступает как самостоятельная единица наблюдения, то дисперсия внутри серий в случае определения средней ошибки и численности выборки должна быть исключена и учитывается только межсерийная дисперсия .

При равных сериях *средняя квадратическая ошибка* бесповторной выборки и ее *численность* определяются по формулам:

$$\mu = \sqrt{\frac{\delta^2}{r}\left(1 - \frac{r}{R}\right)}$$

$$r = \frac{t^2 \delta^2 R}{R\Delta^2 + t^2 \delta^2}$$

где r - количество отобранных серий; R — общее количество серий в генеральной совокупности.

Межсерийная дисперсия рассчитывается:

а) для средней

$$\delta^2 = \frac{\sum_{i-1}^{r} (\tilde{x}_i - \tilde{x}_0)}{r}$$

б) для доли

$$\delta^2 = \frac{\sum_{i-1}^{r} (w_i - \overline{w})^2}{r}$$

где $\tilde{x}_i$ - среднее в сериях; $\tilde{x}_0$ - общая средняя для серий; w_i – доли в сериях (группах); $\overline{w}$ - средняя доля признака для всей выборочной совокупности.

Чем меньше групповые средние и доли отличаются одна от другой, то есть чем ближе одна от другой серии за уровнем принятого признака, тем точнее серийная выборка.

Ступенчатая выборка

Серийную выборку можно рассматривать как *одноступенчатую выборку,* где в случайно отобранных сериях генеральной совокупности проводят сплошное обследование всех единиц, которые в них включены.

Но возможно сформировать выборочную совокупность в два этапа:

на первом этапе методом случайного бесповторного отбора формируют серии, которые подлежат обследованию;

на втором этапе в каждой серии случайным бесповторным отбором формируется определенное количество единиц для последующего обследования.

Средняя квадратическая ошибка выборки будет зависеть от ошибки серийного отбора и ошибки индивидуального отбора:

$$\mu_{\bar{x}} = \sqrt{\frac{\delta_r^2}{r}\left(1 - \frac{r}{R}\right) + \frac{\bar{\sigma}_{вх}^2}{mr}\left(1 - \frac{mr}{N}\right)}$$

где m - количество отобранных единиц в каждой серии;

$\bar{\sigma}_{вх}^2$ - средняя из внутрисерийных дисперсий.

Такая выборка называется

Многоступенчатый отбор характеризуется тем, что на всех ступенях, за исключением последней, осуществляется наблюдение только за последней ступенью. Этот отбор отличается от многофазного отбора тем, что используется в механической выборке: при многоступенчатом отборе на разных ступенях используют единицы отбора разных порядков, а при многофазном отборе пользуются на каждой фазе одними и теми же единицами отбора.

Малые выборки

Теорема Ляпунова доказывает, что ошибки выборки являются случайными величинами и распределены по нормальному закону распределения.

В том случае, когда выборка малая данное утверждение будет уже не справедливо, то есть закон распределения отклонений выборочных характеристик от генеральных будет отличаться от нормального

Английский ученый В. Госсет (Стьюдент) (1908). Определил характеристики этого закона, который и был назван его именем *t-распределение Стьюдента,* которое подобно нормальному закону.

Отклонение выборочной средней $\bar{x}$ от генеральной средней $\bar{X}$ Стьюдент выразил в виде *отношения Стьюдента.*

Фактически это коэффициент доверия между предельной и средней квадратической ошибкой малой выборки:

$$\Delta_{мв}=t\mu_{мв}$$

Значение *t* может быть найдено по математическим таблицам распределения Стьюдента в зависимости от уровня значимости

$a = 1 - P$

где P — уровень вероятности и числа степеней свободы

$k = n - 1$

п — объем малой выборки.

Средняя квадратическая ошибка для количеств признака малой выборки определяется по формуле:

$$\mu_{мв} = \sqrt{\frac{\sigma^2_{мв}}{n}}$$

где $\sigma^2_{мп}$ — дисперсия малой выборки

$$\sigma^2_{мп} = \frac{\sum_{i-1}^{n} (x_i - \tilde{x})^2}{n-1}$$

Вероятность того, что ошибка выборки будет не больше заданного значения

$$\left|\tilde{\bar{x}} - \bar{\bar{x}}\right| \le t\mu_{MS}$$

представляет собой функцию *S(t,n),* приведенную в *таблицах Стьюдента* в литературе по математической статистике:

$$S(t,n) = P(\left|\tilde{\bar{x}} - \bar{\bar{x}} \le t\mu_{MS}\right|)$$

Из таблиц Стьюдента следует, что при увеличении объема выборки распределение Стьюдента приближается к нормальному закону и при $n = 20$ он мало отличается от нормального распределения.

Следует учесть, что распределение Стьюдента используется только в оценке ошибок выборки, взятой из генеральной сово купности с нормальным законом распределения признака.

Ряды динамики