Лекция 11

Определение размера выборки

Еще одним вопросом, непосредственно связанным с построением доверительных интервалов, является определение размера выборки. Достаточно часто исследователь оказывается перед проблемой - насколько большой должна быть выборка, чтобы ошибка с заданной доверительной вероятностью p_0 не превысила заданного максимального значения E?

Например, если оценивается средняя величина покупки, то мы должны задать максимальную величину ошибки, с которой мы готовы примириться (10 копеек, 1 рубль, 5 рублей) и доверительную вероятность - степень уверенности в получаемом результате (95%; 99%; 90%). Если среднее квадратическое отклонение для генеральной совокупности σ известно, то искомый размер выборки легко получить из выражения для максимальной ошибки выборки E:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Перенося $\sqrt{n}$ на одну сторону, все остальное - на другую и возводя обе части в квадрат, получаем:

$$\sqrt{n} = z_{\alpha/2}\frac{\sigma}{E}, \quad n = \left(z_{\alpha/2}\frac{\sigma}{E}\right)^2$$

Последнее выражение и есть искомая оценка размера выборки. В случае необходимости оно округляется в большую сторону до ближайшего целого.

Пример. Начальник учебного отдела института хочет оценить, сколько часов занятий в семестр в среднем проводят преподаватели-совместители. При этом он хочет получить ответ с точностью 1 час и быть уверенным в результате на 99%. Сколько человек ему нужно опросить, если ему известно по предшествующим годам, что среднее квадратическое отклонение этой величины составляет 5 часов.

Решение. По условиям задачи $\sigma = 5$ часов, $E = 1\,час$, доверительная вероятность $p_0 = 99\%$, $\alpha = 0.01$, $z_{\alpha/2} = 2.58$. По формулу для определения размера выборки получаем:

$$n = \left(z_{\alpha/2} \frac{\sigma}{E} \right)^2 = \left(2.58 \times \frac{5}{1} \right)^2 \approx 167 \text{ человек.}$$

Учебный отдел должен опросить не менее 167 человек.

При рассмотрении точечных оценок разбирался вопрос о поправке, связанной с конечностью выборки, с учетом которой средняя и максимальная ошибки выборки записывается следующим образом:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}}, \quad E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}}$$

Конечность выборки легко учесть и в формуле расчета длины:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}} = z_{\alpha/2} \sigma \times \sqrt{\frac{1}{n} - \frac{1}{N}},$$

тогда $n = \dfrac{z_{\alpha/2}^2 \sigma^2 N}{E^2 N + z_{\alpha/2}^2 \sigma^2}$.

Интервальная оценка для доли (биномиального параметра)

Предположим, что признак, исследуемый у единиц генеральной совокупности, является альтернативным, то есть каждая единица либо обладает этим признаком или нет. Генеральная совокупность в целом характеризуется величиной p, равной доле единиц, обладающих признаком. При рассмотрении

вероятностных задач, связанных с этой совокупностью, величина p будет являться параметром биномиального распределения. Если генеральная совокупность неизвестна, то величину доли можно оценить по выборке с помощью точечной несмещенной оценки $p \approx \omega = \frac{k}{n}$, где n - длина выборки, k - доля единиц, обладающих признаком, ω - выборочная доля.

Интервальная оценка параметра p основана на том, что выборочная доля представляет собой выборочное среднее случайных величин с одинаковым распределением $B(1, p)$, математическим ожиданием p и средним квадратическим отклонением $\sqrt{p(1-p)}$. Следовательно, для ω выполнены условия центральной предельной теоремы и при достаточно большой длине выборки она будет приближенно подчиняться нормальному распределению с параметрами $m_\omega = p$, $\sigma_\omega = \sqrt{\frac{p(1-p)}{n}}$. Зная вид распределения и заменяя точное значение дисперсии ее оценкой $\sigma_\omega = \sqrt{\frac{\omega(1-\omega)}{n}}$, получаем выражение для доверительного интервала:

$$\omega - z_{\alpha/2}\sqrt{\frac{\omega(1-\omega)}{n}} < p < \omega + z_{\alpha/2}\sqrt{\frac{\omega(1-\omega)}{n}}$$

Пример. Выборочное обследование 100 девущек - подростков показало, что 30% из них красят волосы. Построить 95% интервал для Доли всех девушек, которые красят волосы.

Решение. Согласно условиям задачи длина выборки $n = 100$, выборочная доля $\omega = 0.3$, доверительная вероятность $p_0 = 95\%$, $\alpha = 0.05$, $z_{\alpha/2} = 1.96$. Подставляя значения в формулу доверительного интервала, имеем:

$$0.3 - 1.96\sqrt{\frac{0.3(1-0.3)}{100}} < p < 0.3 + 1.96\sqrt{\frac{0.3(1-0.3)}{100}};$$

$0.3 - 0.0898 < p < 0.3 + 0.0898;$

$0.2102 < p < 0.3898$

Истинное значение доли лежит в интервале [21.02; 38.98]% .

Интервальная оценка генерального среднего при неизвестной σ.

Рассмотренная в начале лекции интервальная оценка для генерального среднего $m \in \left[\overline{x} \mp z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right]$ зависит от среднего квадратического отклонения генеральной совокупности σ, которая должна быть известна. В случае, когда параметр σ неизвестен, его заменяют на несмещенную оценку:

$$\sigma \approx s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}{n-1}} .$$

Как уже отмечалось ранее, построение любого доверительного интервала основано на законе распределения значений оценки, полученных по различном выборкам, относительно истинного значения. При известной σ, $\frac{\overline{x} - m}{\sigma / \sqrt{n}}$ подчинялось стандартному нормальному распределению $N(0,1)$. При неизвестной с определить распределение удается только в частном случае - когда значения признака в генеральной совокупности ***приближенно*** подчиняются нормальному распределению.

Свойство. Если значения x в генеральной совокупности приближенно описываются нормальным распределением, то величина $\frac{\overline{x} - m}{s / \sqrt{n}}$ подчиняется распределению Стьюдента с числом степеней свободы $k = n - 1$.

Распределение Стьюдента было открыто английским исследователем У. Госсетом в 1908 году. Название Стьюдент является псевдонимом автора, под которым он публиковал свои работы, так как владельцы пивоваренного завода, где он работал, не разрешали своим работника выступать в средствах массовой информации. На рисунке 1 представлены кривые плотности распределения при $k = 3$ и 20.

Свойства распределения Стьюдента.

1) Распределение является семейством кривых, зависящим от числа степеней свободы.
2) Распределение имеет колоколообразную форму и обладает осевой симметрией относительно оси ординат Oy.
3) Кривая распределения приближается к оси абсцисс Ox, но никогда не касается ее.
4) Величины среднего арифметического, медианы и моды совпадают и равны 0, дисперсия и среднее квадратическое отклонение больше 1.
5) При увеличении k распределение приближается к нормальному и при $k > 30$ практически совпадает с ним.

Понятие "число степеней свободы", определяющее распределение Стьюдента, используется и во многих других статистических распределениях. Степень свободы представляет собой количество изменяемых, варьируемых значений, после того как для данной выборки рассчитаны и зафиксированы некоторые обобщающие значения. Пусть, например, в выборке с 5 элементами мы рассчитали и зафиксировали среднее значение, равное 10: $\frac{1}{5}\left(x_1 + x_2 + x_3 + x_4 + x_5\right) = 10$. Тогда из пяти величин можно произвольно менять любые четыре, но пятое будет автоматически вычисляться из условия постоянства среднего. Следовательно, число степеней свободы равно 4. Например, если варьируются x_1, x_2, x_3, x_4 то x_5 определяется как $x_5 = 10 \cdot 5 - \left(x_1 + x_2 + x_3 + x_4\right)$. Если же для этой выборки дополнительно

зафиксировать и величину выборочного отклонения, то число степеней свободы равнялось бы 3.

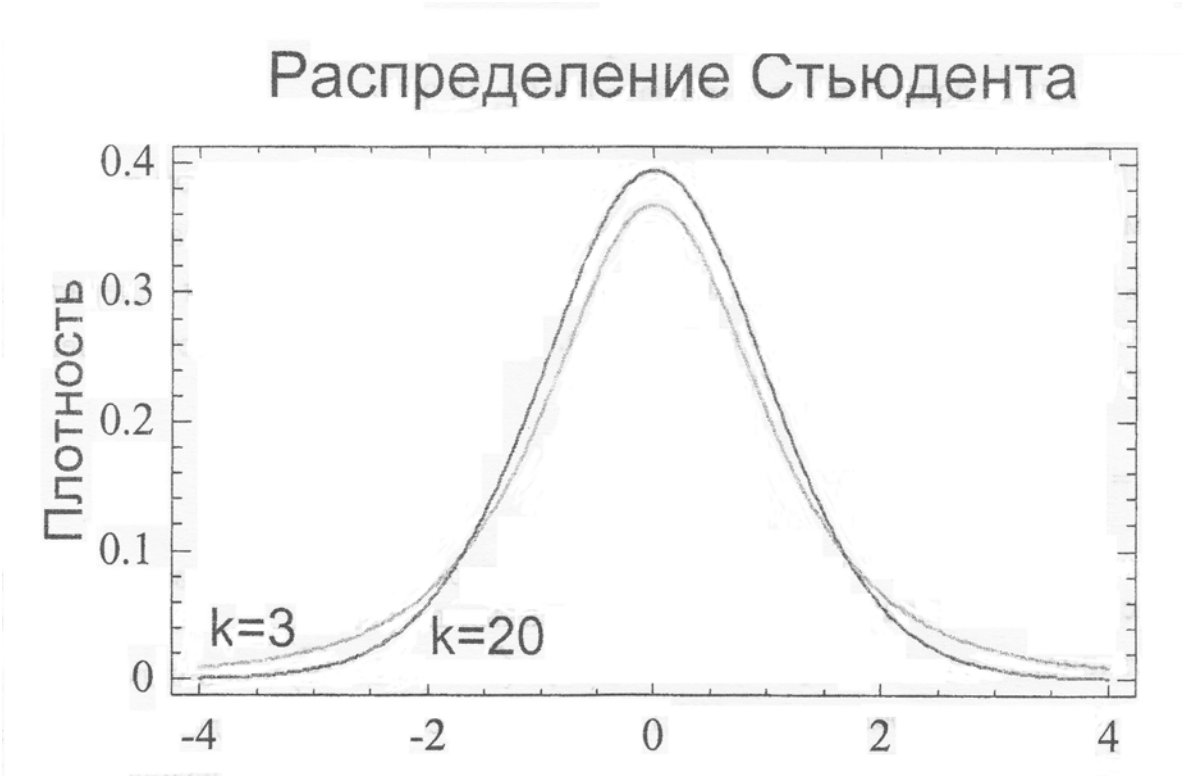

Рис. 1. Распределение Стьюдента

Основываясь на распределении Стьюдента для заданной выборки й доверительной вероятности p_0, можно выписать вид интервальной оценки:

$$m \in \left[\overline{x} - t_{\alpha/2} \times \frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2} \times \frac{s}{\sqrt{n}} \right].$$

Критические значения $t_{\alpha/2}$, определяемые из условия $P\left(t > t_{\alpha/2}\right) = P\left(t < -t_{\alpha/2}\right) = \frac{\alpha}{2}$, где $\alpha = 1 - p_0$ находятся по таблице распределения Стьюдента. Нужная строка таблицы определяется числом степеней свободы, а нужная колонка - по уровню значимости $\frac{\alpha}{2}$. Для длины выборки $n > 30$ критическое значение $t_{\alpha/2}$ находится по таблице нормального распределения $t_{\alpha/2} \approx z_{\alpha/2}$. К примеру, выпишем доверительные интервалы для выборки из $n = 11$ элементов и доверительной вероятности 95%. Вычисляем число

степеней свободы $k = n - 1 = 11 - 1 = 10$, $\frac{\alpha}{2} = \frac{1}{2}(1 - 0.95) = 0.025$ и по таблице распределения находим $t_{\alpha/2} = 2.23$, тогда 95% доверительный интервал равен $[-2.23, 2.23]$. Для сравнения, для стандартного нормального распределения 95% доверительный интервал более узкий $[-1.96, 1.96]$.

Пример. На дороге для контроля остаточной высоты протектора шин проверено 10 автомобилей. Установлено, что средняя остаточная величина протектора составляет 8 мм при среднем квадратическом отклонении 2 мм. Построить 99% доверительный интервал для генерального среднего, предполагая, что остаточная высота протектора приближенно описывается нормальным распределением.

Решение. В силу приближенной нормальной распределенности генеральной совокупности для построения доверительного интервала можно использовать распределение Стьюдента. По размеру выборки $n = 10$ определяем число степеней свободы, $k = 9$, по доверительной вероятности 99% - уровень значимости $\alpha = 0.01$ и находим $t_{\alpha/2} = 3.25$. Следовательно:

$$m \in \left[\overline{x} - t_{\alpha/2} \times \frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2} \times \frac{s}{\sqrt{n}} \right] = 8 \mp 3.25 \times \frac{2}{\sqrt{10}} \approx 8 \mp 2.05 \approx [6, 10].$$

Остаточная высота протектора с 99% вероятностью лежит в интервале $[6, 10]$мм.

Интервальная оценка дли дисперсии генеральной совокупности

Точечной оценкой для дисперсии D генеральной совокупности является величина

$$D \approx s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

Интервальная оценка для дисперсии основывается на следующем утверждении:

Свойство. Если значения x в генеральной совокупности приближенно описываются нормальным распределением, то величина $\frac{(n-1)\times s^2}{D}$ подчинена распределению χ^2 (хи-квадрат) с числом степеней свободы $k = n-1$.

Свойства распределения χ^2 (рис. 2)

1) Распределение является семейством кривых, зависящим от числа степеней свободы.

2) Распределение принимает только положительные значения и положительно-асимметрично.

3) Для числа степеней свободы более 100 распределение становится приближенно-симметричным.

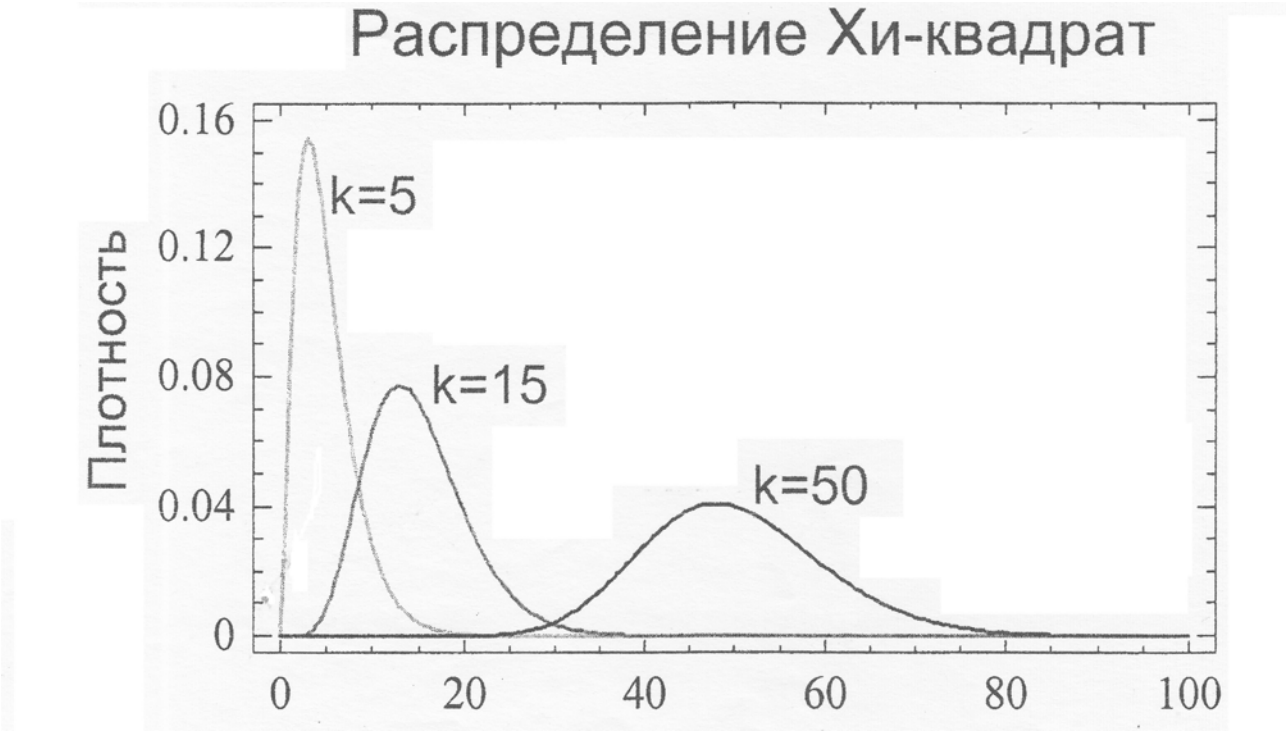

Рис. 2. Распределение χ^2

При построении доверительного интервала для χ^2 распределения в силу его асимметричности нужно при заданном уровне значимости находить два критических (граничных) значения - нижнее и верхнее, для двух интервалов:

$$P\left(\chi^2 < \chi^2_{\alpha/2,\,н}\right) = \frac{\alpha}{2} \text{ и } P\left(\chi^2 > \chi^2_{\alpha/2,\,в}\right) = \frac{\alpha}{2}.$$

Для их определения используется таблица распределения χ^2. Строка таблицы

определяется числом степеней свободы, нижнее критическое соответствует колонке $1-\frac{\alpha}{2}1$, верхнее $\frac{\alpha}{2}$.

С учетом введенных обозначений интервальная оценка для дисперсии имеет вид:

$$\frac{(n-1)\times s^2}{\chi^2_{\alpha/2,\,в}} < D < \frac{(n-1)\times s^2}{\chi^2_{\alpha/2,\,н}}.$$

Пример. Производитель сигарет провел исследование с целью оценить дисперсию содержания никотина в сигаретах. Проведен анализ 20 сигарет, и выборочное отклонение s составило 1мг. Выписать интервальную оценку дисперсий при доверительной вероятности 95%.

Решение. Доверительный интервал конструируем с помощью распределения χ^2 с числом степеней свободы $k = n-1 = 20-1 = 19$. Для уровня значимости $\alpha = 0.05$ граничные значения составляют $P\left(\chi^2 < \chi^2_{\alpha/2,\,н}\right) = 8.9$ и $P\left(\chi^2 > \chi^2_{\alpha/2,\,в}\right) = 32.9$, тогда

$$\frac{(n-1)\times s^2}{\chi^2_{\alpha/2,\,в}} = \frac{(20-1)\times 1^2}{32.9} < D < \frac{(n-1)\times s^2}{\chi^2_{\alpha/2,\,н}} = \frac{(20-1)\times 1^2}{8.9},$$

$0.58 < D < 2.13$ и $0.76 < \sigma < 1.46$

Дисперсия содержания никотина находится в интервале $[0.76, 1.46]$ мг^2.