Лекция №10

Интервальные оценки

Понятие доверительного интервала

В предыдущей лекции были рассмотрены примеры того, как с помощью выборки приближенно определять значение показателя генеральной совокупности. Недостатком примененных для этой цели точечных оценок является то, что, имея единственную выборку и рассчитав по ней единственное оценочное значение, ничего нельзя сказать о точности результата, оценить отклонение от истинного значения. Преодолеть этот недостаток и оценить точность результата позволяют интервальные оценки.

Определение. ***Интервальной оценкой*** называется диапазон значений, используемый для оценки параметра генеральной совокупности.

При использовании интервальной оценки наше утверждение относительно истинного значения заключается в том, что оно попадает в некоторый диапазон значений.

Например, утверждения, что средний возраст студентов составляет $19.9 < m < 20.7$ лет или $m = 20.3 \pm 0.4$ года, являются интервальными оценками.

Построение интервальной оценки основано на задании вероятности того, что истинное значение показателя лежит в данном интервале. Например, исследователь может поставить задачу, чтобы истинное значение попало в заданный интервал с вероятностью 95% (доверительной вероятностью), - он хочет быть уверенным на 95%. Возникает вопрос - почему именно 95%, а не 99% или 99.5%. Если требуется большая степень уверенности, то и интервал, который мы используем для оценки, придется сделать шире.

Например, интервал для среднего возраста, в который оцениваемое значение попадет с вероятностью 99%, будет 19.7 $< m <$ 20.9, m = 20.3 ±0.6 года.

Определение. ***Доверительный интервал*** - это интервальная оценка параметра генеральной совокупности, определяемая с помощью выборки для заданной доверительной вероятности.

Определение. ***Доверительная вероятность*** p_0 интервальной оценки - это вероятность, что истинное (генеральное) значение характеристики попадет в заданный интервал.

Определение. ***Уровень значимости*** α - величина, равная вероятности, что реальное значение не попадает в заданный интервал, $\alpha = 1 - p_0$.

Доверительная вероятность обычно задается в процентах, а уровень значимости - в долях от единицы. Наиболее часто используются 95% и 99% доверительные интервалы, которым соответствуют уровни - значимости α =0.05 и 0.01. Реже встречается 90% доверительный интервал, $\alpha = 0.1$.

При построении доверительных интервалов по известному распределению оценки мы будем пользоваться одним и тем же приемом, который проиллюстрируем на примере нормального распределения.

Задача. Случайная величина z описывается стандартным нормальным распределением $N(0;1)$. Найти значение $z_{\alpha/2}$ $\alpha = 1 - p_0$, такое, что вероятность $P\left(-z_{\alpha/2} \le z \le z_{\alpha/2}\right)$ равна p_0: $p_0 = 95\%, 99\%, 90\%$.

Решение. Отметим, что задача является обратной по отношению к тем, которые уже рассматривались для нормального распределения. Раньше по заданному интервалу значений случайной величины определялась вероятность попадания в этот интервал. Теперь задана вероятность, а интервал, ей соответствующий, должен быть определен. Рассмотрим график плотности стандартного нормального распределения, рис 1.

Рис. 1 Нахождение интервала по заданной вероятности

По основному свойству плотности распределения площадь под графиком в интервале $-z_{\alpha/2} \le z \le z_{\alpha/2}$ равняется p_0. С другой стороны, величина $\alpha = 1 - p_0$ представляет вероятность того, что величина z не попадет на этот интервал и будет находится на одном из "хвостов" распределения в области $z < -z_{\alpha/2}$ или $z_{\alpha/2} < z$. В силу симметрии плотности распределения, вероятности этих двух событий равны между собой и равны $\alpha/2$, рис. 1. Этот факт одновременно служит пояснением обозначения $z_{\alpha/2}$, использованного для границ интервала. Пользуемся определением функции Лапласа *F*, записываем:

$$P\left(z \le -z_{\alpha/2}\right) = P\left(z \ge z_{\alpha/2}\right) = 0.5 - F\left(z_{\alpha/2}\right) = \alpha/2 \quad \Rightarrow \quad F\left(z_{\alpha/2}\right) = 0.5 - \alpha/2.$$

С помощью таблицы функции Лапласа нужно найти значение аргумента $z_{\alpha/2}$ функции, такое, чтобы ее значение при этом приближенно равнялось $0.5 - \alpha/2$.

$$p_0 = 90\%; \quad \alpha = 1 - p_0 = 0.1, \quad \alpha/2 = 0.05; \quad F\left(z_{\alpha/2}\right) = 0.5 - \alpha/2 = 0.45; \quad z_{\alpha/2} = 1.64;$$

$$p_0 = 95\%; \quad \alpha = 1 - p_0 = 0.05, \quad \alpha/2 = 0.025; \quad F\left(z_{\alpha/2}\right) = 0.5 - \alpha/2 = 0.475; \quad z_{\alpha/2} = 1.96;$$

$$p_0 = 99\%; \quad \alpha = 1 - p_0 = 0.01, \quad \alpha/2 = 0.005; \quad F\left(z_{\alpha/2}\right) = 0.5 - \alpha/2 = 0.495; \quad z_{\alpha/2} = 2.58;$$

Искомые интервалы построены:

$P\left(-1.64 \le z \le 1.64\right) = 90\%$;

$P\left(-1.96 \le z \le 1.96\right) = 95\%$;

$P\left(-2.58 \le z \le 2.658\right) = 99\%$

Применим полученный результат для анализа выборочной средней $\overline{x}$. Согласно результату предыдущей лекции эта величина приближенно подчиняется нормальному распределению $N\left(m;\frac{\sigma}{\sqrt{n}}\right)$где m - генеральное среднее, σ - генеральное отклонение. Пользуясь преобразованием общего нормального распределения к стандартному, получаем, что величина $\frac{\overline{x}-m}{\sigma/\sqrt{n}}$ имеет распределение $N(0;1)$. Тогда, согласно результатам задачи,

$$P\left(-z_{\alpha/2} \le \frac{\overline{x}-m}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = p_0.$$

Умножаем все части неравенства на $\frac{\sigma}{\sqrt{n}}$:

$$P\left(-z_{\alpha/2}\,\sigma/\sqrt{n} \le \overline{x}-m \le z_{\alpha/2}\,\sigma/\sqrt{n}\right) = p_0$$

и после этого прибавления m получаем окончательно:

$$P\left(m - z_{\alpha/2}\,\sigma/\sqrt{n} \le \overline{x} \le m + z_{\alpha/2}\,\sigma/\sqrt{n}\right) = p_0$$

Таким образом, для вероятности $p_0 = 1-\alpha$ выполняется соотношение

$$m - z_{\alpha/2}\,\sigma/\sqrt{n} \le \overline{x} \le m + z_{\alpha/2}\,\sigma/\sqrt{n} \qquad (*)$$

Это соотношение еще не является интервальной оценкой, так как в нем положение *известной выборочной* величины $\overline{x}$ выражается через *неизвестную* величину m генеральной совокупности и генеральное отклонение σ.

Интервальная оценка генерального среднего при известном генеральном отклонении σ.

Приведем кратко те рассуждения, с помощью которых строится доверительный интервал для генерального среднего в случае, когда значение генерального среднего квадратического отклонения σ известно, например, из предшествующих исследований. Запишем соотношение (*) для вероятности $p_0 = 95\%$, $z_{\alpha/2} = 1.96$:

$$m - 1.96 \times \sigma/\sqrt{n} \le \bar{x} \le m + 1.96 \times \sigma/\sqrt{n}$$

Это означает, что для произвольно сформированной выборки из n единиц выборочное среднее с 95% вероятностью отклонится от истинного значения m *не более* чем на $1.96 \times \sigma/\sqrt{n}$. Взаимное положение выборочных средних относительно генерального иллюстрирует рис. 2,

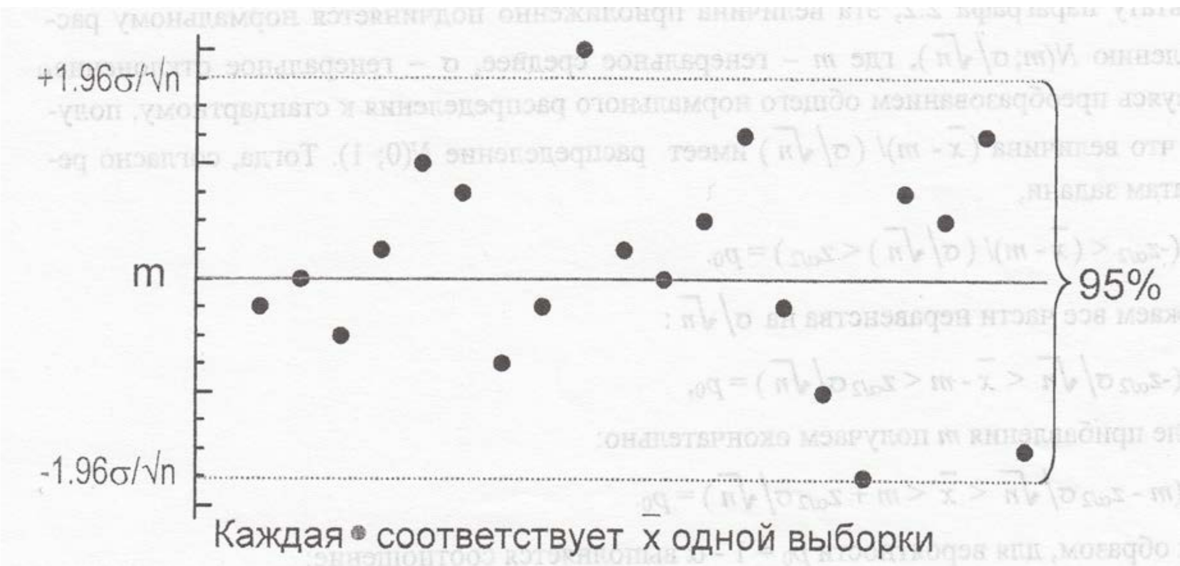

Рис. 2. 95% доверительные интервалы для выборочных средних

Но это же утверждение можно переформулировать по-другому. Имея в распоряжении конкретное значение $\bar{x}$ построим вокруг него 95-процентный доверительный интервал $\bar{x} \mp 1.96 \times \sigma / \sqrt{n}$, рис. 3.

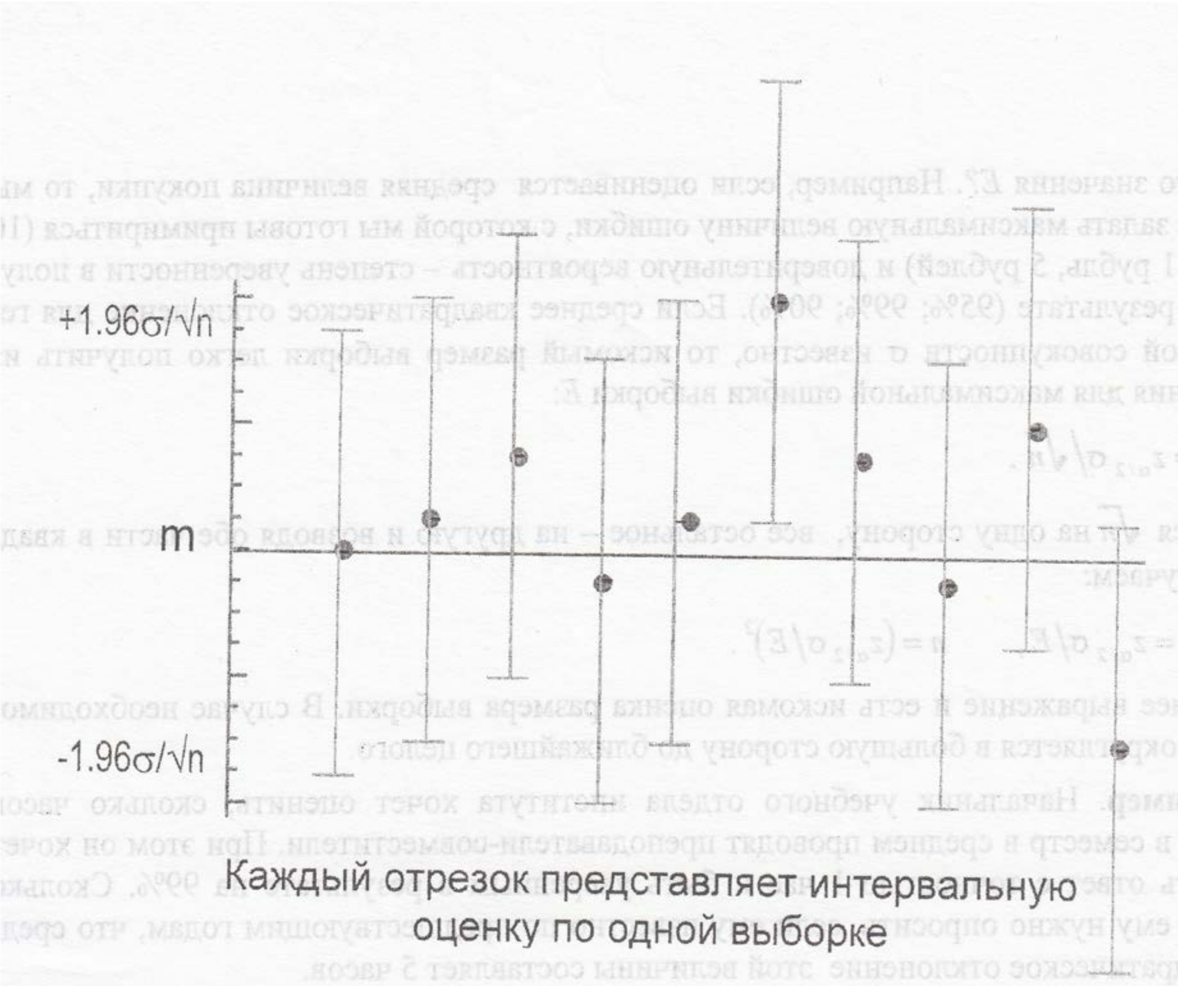

Рис. 3. 95%интервальные оценки для различных выборочных средних

В большинстве случаев (в 95%) доверительные интервалы, построенные по выборочным средним, захватывают истинное значение. Однако в оставшихся 5% случаев генеральное среднее не захватывается. Таким образом, в частном случае $p_0 = 95\%$ интервальная оценка имеет вид:

$$\bar{x} - 1.96 \times \sigma / \sqrt{n} \le m \le \bar{x} + 1.96 \times \sigma / \sqrt{n}$$

а общая формула для произвольной доверительной вероятности записывается следующим образом:

$$\overline{x} - z_{\alpha/2} \times \sigma/\sqrt{n} \le m \le \overline{x} + z_{\alpha/2} \times \sigma/\sqrt{n}$$

Величина $E = z_{\alpha/2} \times \sigma/\sqrt{n}$ называется *максимальной ошибкой* оценки.

Полученная оценка включает в себя значение среднего квадратического отклонения генеральной совокупности σ. Как она может быть известна, если проводится только выборочное наблюдение и данных по всей генеральной совокупности нет? Генеральное отклонение σ может быть известно:

- Из данных предшествующих наблюдений, которые могли быть сплошными или большими по охвату, предполагая, что за прошедший период отклонение не изменилось, $\sigma = \sigma_0$. В других случаях неизменной полагают коэффициент вариации $V = \frac{\sigma_0}{\overline{x}_0} \times 100\%$, а текущее отклонение оценивают как $\sigma = \frac{V \cdot \overline{x}_0}{100\%}$.
- Исходя из закона распределения. Если распределение близко к нормальному, то размах вариации выборки $R = x_{\max} - x_{\min}$ примерно в 6 раз больше среднего квадратического отклонения, $R \approx 6 \times \sigma$. В таком случае, зная минимальное и максимальное выборочные значения, можно оценить σ: $\sigma \approx \frac{R}{6} = \frac{x_{\max} - x_{\min}}{6}$.

Пример. При обследовании 40 пациентов терапевт определил, что среднее время приема составило 16.2 минуты. Из предшествующих исследований известно, что среднее квадратическое отклонение σ этой величины составляет 3 минуты. Построить 95% доверительный интервал для среднего времени приема всех пациентов.

Решение. Последовательно рассчитываем следующие величины:

Средняя ошибка выборки: $\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{40}} \approx 0.47$;

Уровень значимости для $p_0 = 95\%$: $\alpha = 1 - p_0 = 1 - 0.95 = 0.05$ и $z_{\alpha/2} = 1.96$;

Максимальная ошибка выборки $z_{\alpha/2} \times \sigma/\sqrt{n} \approx 1.96 \times 0.47 \approx 0.92$;

Применяя общую формулу, получаем: $m \in \left[\overline{x} - z_{\alpha/2} \times \sigma/\sqrt{n}, \overline{x} + z_{\alpha/2} \times \sigma/\sqrt{n}\right] = [16.2 - 0.92, 16.2 + 0.92] =$

$= [15.28, 17.12] = 16.2 \mp 0.92$.