

Лекция 5. Доверительные интервалы

Курбацкий А. Н.

МШЭ МГУ

16 марта 2020

Содержание

1 Интервальное оценивание

2 Доверительный интервал для среднего

- Случай известной дисперсии
- Случай неизвестной дисперсии и объём выборки $n > 30$
- Случай малой выборки и неизвестной дисперсии
- Минимальный объём выборки

3 Доверительный интервал для доли и дисперсии

- Доверительный интервал для доли (асимптотический)
- Доверительный интервал для дисперсии

4 Более подробно

Содержание

1 Интервальное оценивание

2 Доверительный интервал для среднего

- Случай известной дисперсии
- Случай неизвестной дисперсии и объём выборки $n > 30$
- Случай малой выборки и неизвестной дисперсии
- Минимальный объем выборки

3 Доверительный интервал для доли и дисперсии

- Доверительный интервал для доли (асимптотический)
- Доверительный интервал для дисперсии

4 Более подробно

Важно!

В этой теме мы научимся оценивать параметры не только одним числом, а сможем находить интервал, который накрывает его с заданной вероятностью.

Пример

Пусть мы провели опрос, с целью оценить долю избирателей некоторого кандидата в президенты страны. Опросили 1000 человек и 275 его поддерживают, значит выборочная доля сторонников кандидата равна 27.5%. Это точечная оценка для доли его сторонников во всей стране. Но может так оказаться, что она малоинформативна.

Намного лучше знать не просто точечную оценку 27.5%, а интервал, в котором с большой вероятностью находится доля сторонников кандидата. Скажем, с 95% вероятностью доля сторонников лежит в пределах от 25% до 30%.

Доверительный интервал

Определение

Точечной оценкой называется число, которое используют для оценки параметра ГС.

Для того, чтобы делать содержательные выводы, стараются находить не точечные, а интервальные оценки.

Определение

Доверительный интервал - это интервал, который с заданной вероятностью накрывает оцениваемый параметр ГС.

Важно!

Имейте ввиду, что для разных выборок одной и той же ГС могут получаться разные доверительные интервалы!

Уровни значимости и доверия

При работе с доверительными интервалами часто используют два термина.

- Уровень значимости α - это вероятность, с которой значение параметра не попадает в доверительный интервал.
- Уровень доверия $\beta = 1 - \alpha$ - это вероятность того, что доверительный интервал покрывает значение параметра.

Обычно уровень значимости равен 0.01, 0.05, 0.1, что соответствует уровню доверия 0.99, 0.95, 0.9. Очень часто уровни значимости и доверия измеряются в процентах, то есть уровень доверия 0.99 и 99% - это одно и то же.

Определение

Пусть задано малое число $0 < \alpha < 1$. Интервал со случайными концами

$$(\theta_1; \theta_2)$$

называется *доверительным интервалом* для параметра θ с уровнем доверия $1 - \alpha$, если для любого $\theta \in \Theta$

$$P(\theta_1 < \theta < \theta_2) \geq 1 - \alpha$$

Замечание

Если последнее неравенство выполняется при $n \rightarrow \infty$, то интервал называется *асимптотическим*.

Имеется в виду не один какой-то интервал, а последовательность интервалов, зависящих от n .

Содержание

1 Интервальное оценивание

2 Доверительный интервал для среднего

- Случай известной дисперсии
- Случай неизвестной дисперсии и объём выборки $n > 30$
- Случай малой выборки и неизвестной дисперсии
- Минимальный объем выборки

3 Доверительный интервал для доли и дисперсии

- Доверительный интервал для доли (асимптотический)
- Доверительный интервал для дисперсии

4 Более подробно

Оттуда ноги растут

Важнейшей характеристикой генеральной совокупности является среднее значение. Что же необходимо сделать, чтобы построить для него доверительный интервал?

По ЦПТ среднее значение одинаково распределенных случайных величин стремится к нормальному распределению. Более того, верна

Теорема

Если распределение генеральной совокупности имеет конечные математическое ожидание и дисперсию, то при $n \rightarrow \infty$ основные выборочные характеристики (среднее, дисперсия, эмпирическая функция распределения) являются нормальными.

Важно!

Далее мы часто будем предполагать, что генеральная совокупность имеет нормальный закон распределения.

Доверительный интервал для среднего

Рассмотрим случайную выборку объема n , вычислим среднее значение \bar{x} по выборке и зададим уровень доверия β .

- Доверительный интервал для среднего имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$, где Δ - это точность интервальной оценки.

Важно!

Правило для вычисления точности зависит от того, что мы знаем о генеральной совокупности и с какой выборкой мы имеем дело.

- Пусть нам известно стандартное отклонение σ генеральной совокупности.
- Тогда $\Delta = \frac{\sigma}{\sqrt{n}} z_{\alpha}$, где z_{α} - это квантиль нормального распределения уровня $1 - \frac{\alpha}{2}$.

Теорема

Доверительный интервал для среднего с известной дисперсией имеет вид $(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha}; \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha})$

Пример

Пример

Дана выборка 9, 5, 7, 7, 4, 10, дисперсия $\sigma^2 = 1$. Постройте 99% доверительный интервал.

Пример

Пример

Дана выборка 9, 5, 7, 7, 4, 10, дисперсия $\sigma^2 = 1$. Постройте 99% доверительный интервал.

Решение

- Среднее значение равно $\bar{x} = \frac{9+5+7+7+4+10}{6} = 7$.
- Доверительный интервал имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$. По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.995$ и определяем квантиль $z_\alpha = 2.58$. Теперь можем найти точность $\Delta = \frac{\sigma}{\sqrt{n}} z_\alpha = \frac{1}{\sqrt{6}} 2.58 \approx 1.05$ (здесь мы воспользовались тем, что известна дисперсия генеральной совокупности).
- Искомый 99%-доверительный интервал имеет вид $(7 - 1.05; 7 + 1.05) = (5.95; 8.05)$.

Пример

Пример

Пусть для выборки объема $n = 25$ вычислено среднее $\bar{x} = 130$. Из предыдущих исследований известно стандартное отклонение $\sigma = 12$. Постройте 98% доверительный интервал для среднего значения.

Решение

- Доверительный интервал имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$. Уровень доверия равен $\beta = 0.98$, поэтому $\alpha = 0.02$. По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.99$ и определяем квантиль $z_{\alpha} = 2.33$. Теперь можем найти точность $\Delta = \frac{\sigma}{\sqrt{n}} z_{\alpha} = \frac{12}{\sqrt{25}} 2.33 \approx 5.59$.
- Искомый 98%-доверительный интервал имеет вид $(130 - 5.59; 130 + 5.59) = (124.41; 135.59)$.

Содержание

1 Интервальное оценивание

2 Доверительный интервал для среднего

- Случай известной дисперсии
- Случай неизвестной дисперсии и объём выборки $n > 30$
- Случай малой выборки и неизвестной дисперсии
- Минимальный объём выборки

3 Доверительный интервал для доли и дисперсии

- Доверительный интервал для доли (асимптотический)
- Доверительный интервал для дисперсии

4 Более подробно

Доверительный интервал

Если выборка больше 30, но стандартное отклонение нам неизвестно, то вместо σ мы будем использовать выборочное стандартное

отклонение $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$.

Теорема

Доверительный интервал для среднего при неизвестной дисперсии, но большой выборке ($n > 30$), имеет вид $\left(\bar{x} - \frac{s}{\sqrt{n}} z_{\alpha}; \bar{x} + \frac{s}{\sqrt{n}} z_{\alpha} \right)$.

Содержание

1 Интервальное оценивание

2 Доверительный интервал для среднего

- Случай известной дисперсии
- Случай неизвестной дисперсии и объём выборки $n > 30$
- Случай малой выборки и неизвестной дисперсии
- Минимальный объем выборки

3 Доверительный интервал для доли и дисперсии

- Доверительный интервал для доли (асимптотический)
- Доверительный интервал для дисперсии

4 Более подробно

Сложный случай

Самый проблемный случай для любого исследователя, когда выборка маленькая и про её параметры ничего неизвестно. Если дисперсия неизвестна и объем выборки небольшой ($n \leq 30$), тогда вместо нормального распределения теперь используется t -распределение.

Теорема

Доверительный интервал в этом случае имеет вид

$$(\bar{x} - \frac{s}{\sqrt{n}} t_{\alpha}(n-1); \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha}(n-1))$$

Здесь $t_{\alpha}(n-1)$ - это квантиль распределения Стьюдента уровня $1 - \frac{\alpha}{2}$ с $n-1$ степенью свободы (мы ищем это число в таблице t -распределения).

Замечание

Распределение Стьюдента стремится к нормальному распределению при $n \rightarrow \infty$, поэтому при больших выборках доверительные интервалы для среднего, посчитанные по любой из наших формул, будут почти совпадать.

Что за степени свободы и кто такой Стьюдент?

Замечание

Число степеней свободы зависит от того, сколько имеется связей между наблюдениями. Так как мы знаем среднее, то наблюдения связаны одним равенством и степеней свободы становится на одну меньше. То, что других связей нет, надо доказывать, но их действительно нет. Честное слово.

Замечание

Распределение Стьюдента было введено в 1908 году В.С.Госсетом, ирландским служащим пивоваренного завода, который участвовал в разработке новых технологий производства пива и никаким студентом не был. Придавать известности результаты исследований означало открыть корпоративную тайну, поэтому Госсет напечатал свои материалы под псевдонимом Стьюдент. Фишер ввёл для него обозначение t -распределение.

Доверительный интервал

Пример

Пусть объем выборки $n = 16$, выборочное среднее $\bar{x} = 5$, выборочная дисперсия $s^2 = 4$. Постройте 99% доверительный интервал.

Решение

- Среднее значение равно $\bar{x} = 5$, а выборочная дисперсия $s^2 = 4$.
- Так как неизвестна дисперсия генеральной совокупности и $n < 30$, поэтому точность интервальной оценки $\Delta = \frac{s}{\sqrt{n}} t_{\alpha}$.
- По таблице распределения Стьюдента находим $1 - \frac{\alpha}{2} = 0.995$ и, так как у нас $n - 1 = 16 - 1 = 15$ степеней свободы, определяем квантиль $t_{\alpha} = 3.29$. Теперь можем найти точность $\Delta = \frac{s}{\sqrt{n}} t_{\alpha} = \frac{2}{\sqrt{16}} 3.29 \approx 1.645$.
- Искомый 99%-доверительный интервал имеет вид $(5 - 1.645; 5 + 1.645) = (3.355; 6.645)$.

Минимальный объем выборки

Благодаря тому, что мы знаем формулу для доверительного интервала, можно решить интересную задачу: найти минимальный необходимый объем выборки для того, чтобы с заданной точностью и уровнем доверия найти среднее значение.

Важно!

Для того чтобы найти минимальный необходимый объем выборки для построения доверительного интервала для среднего значения с заданной точностью Δ и уровнем значимости α , достаточно применить формулу

$$n = \left(\frac{z_{\alpha} \sigma}{\Delta} \right)^2$$

Теперь понятно, как определить объем выборки при проведении собственных исследований!

Пример

Найдите минимально необходимый объем выборки для построения интервальной оценки среднего с точностью $\Delta = 3$, дисперсией $\sigma^2 = 225$ и уровнем доверия $\beta = 0.95$.

Содержание

1 Интервальное оценивание

2 Доверительный интервал для среднего

- Случай известной дисперсии
- Случай неизвестной дисперсии и объём выборки $n > 30$
- Случай малой выборки и неизвестной дисперсии
- Минимальный объем выборки

3 Доверительный интервал для доли и дисперсии

- Доверительный интервал для доли (асимптотический)
- Доверительный интервал для дисперсии

4 Более подробно

Доверительный интервал для доли

Следующим популярным параметром, который часто требует оценивания, является доля признака p в ГС.

По выборке мы можем определить долю \hat{p} того или иного признака, просто посчитав число объектов m с этим признаком и поделив на объем выборки n , то есть $\hat{p} = \frac{m}{n}$. Долю объектов, не обладающих этим признаком, обозначают $\hat{q} = 1 - \hat{p}$.

Теорема

Асимптотический доверительный интервал для доли имеет вид

$$\left(\hat{p} - \sqrt{\frac{\hat{p}\hat{q}}{n}}z_{\alpha}; \hat{p} + \sqrt{\frac{\hat{p}\hat{q}}{n}}z_{\alpha} \right).$$

Обратите внимание, что для использования этой формулы требуют выполнения условий $n\hat{p} \geq 5$ и $n\hat{q} \geq 5$.

Доверительный интервал

Пример

Объем выборки $n = 100$, выборочная доля $\hat{p} = 0.2$, $\alpha = 0.05$.

Решение

- Выборочная доля $\hat{p} = 0.2$, поэтому $\hat{q} = 1 - \hat{p} = 0.8$. Убеждаемся, что выполнены условия надёжности использования этих формул $n\hat{p} = 20 \geq 5$ и $n\hat{q} = 80 \geq 5$.
- По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.975$ и определяем квантиль $z_{\alpha} = 1.96$. Теперь можем найти точность $\Delta = \sqrt{\frac{\hat{p}\hat{q}}{n}}z_{\alpha} = \sqrt{\frac{0.2 \cdot 0.8}{100}}1.96 \approx 0.078$.
- Искомый 95%-доверительный интервал имеет вид $(0.2 - 0.078; 0.2 + 0.078) = (0.122; 0.278)$.

Минимальный объем выборки

Если мы хотим узнать минимально необходимый объем выборки для того, чтобы с заданными точностью и уровнем доверия оценить долю признака в ГС, то сделать это можно по формуле $n = \hat{p} \cdot (1 - \hat{p}) \left(\frac{z_{\alpha}}{\Delta}\right)^2$.

Важно!

Имейте ввиду, что выборочная доля \hat{p} может быть неизвестна. В таких случаях её кладут равной 0.5, потому что при этом выражение $\hat{p} \cdot (1 - \hat{p}) \left(\frac{z_{\alpha}}{\Delta}\right)^2$ принимает наибольшее значение.

При остальных значениях \hat{p} объём выборки был бы меньше, но если мы её не знаем, то берём крайний вариант.

Доверительный интервал для дисперсии

Перейдём к следующему важнейшему параметру генеральной совокупности - к дисперсии.

Важно!

Доверительный интервал для дисперсии имеет вид

$$\left(\frac{(n-1)s^2}{\chi_r^2(\alpha)}; \frac{(n-1)s^2}{\chi_l^2(\alpha)} \right).$$

Здесь значения $\chi_r^2(\alpha)$ и $\chi_l^2(\alpha)$ находятся по таблицам χ^2 -распределения с $n - 1$ степенью свободы, причем в таблице мы ищем $\alpha/2$ и $1 - \frac{\alpha}{2}$.

Пример

Построить 90% доверительный интервал для дисперсии по выборке объема $n = 20$, если выборочная дисперсия $s^2 = 196$.

Решение

- Выборочная дисперсия $s^2 = 196$.
- По таблице χ^2 -распределения находим $\alpha/2 = 0.05$, $1 - \alpha/2 = 0.95$, число степеней свободы $n - 1 = 20 - 1 = 19$ и определяем критические точки $\chi_l^2 = 10.12$, $\chi_r^2 = 30.14$.
- Искомый 90%-доверительный интервал имеет вид

$$\left(\frac{(20 - 1)196}{30.14}; \frac{(20 - 1)196}{10.16} \right) = (123.56; 366.54).$$

Полезный факт

Часто для поиска дисперсии вручную удобнее использовать формулу

$$s^2 = \frac{1}{n-1} \left(\sum x_i^2 - n\bar{x}^2 \right).$$

Замечание

Для доказательства формулы достаточно раскрыть скобки в исходном выражении для дисперсии. Для смещённой оценки дисперсии формула ещё более укорачивается $s_{\text{смещ}}^2 = \bar{x}^2 - \bar{x}^2$.

Пример

По данным выборки объема $n = 12$ было найдено, что $\sum x_i = 216$, $\sum x_i^2 = 4046$. Постройте 90% доверительный интервал для теоретической дисперсии. В ответ укажите длину доверительного интервала, округлив до целого числа.

Решение

- Выборочное среднее равно $\bar{x} = \frac{216}{12} = 18$, а выборочная дисперсия $s^2 = \frac{1}{n-1}(\sum x_i^2 - n\bar{x}^2) = \frac{1}{11}(4046 - 12 \cdot 18^2) \approx 14.36$.
- По таблице χ^2 -распределения находим $\alpha/2 = 0.05$, $1 - \alpha/2 = 0.95$, число степеней свободы $n - 1 = 12 - 1 = 11$ и определяем критические точки $\chi_l^2 = 4.57$, $\chi_r^2 = 19.675$.
- Искомый 90%-доверительный интервал имеет вид $\left(\frac{(12-1)14.36}{19.675}; \frac{(12-1)14.36}{4.57} \right) = (8; 34.56)$. Длина интервала приблизительно равна 27.

Внимание!

Причины неверных выводов:

- зависимость наблюдений;
- нарушение однородности данных (обычно при их большом количестве);
- систематическая ошибка измерений.

"... студенты должны понимать, что не следует вообще обрабатывать статистически получаемые в окончательном виде макроэкономические данные: применительно к любым макроэкономическим рядам, каждый из которых имеется в единственном экземпляре, вероятностно-статистический подход смысла не имеет¹."

¹Тутубалин В.Н. "Эконометрика: Образование, которого нам не нужно"

- Что означает фраза "истинное значение того или иного параметра лежит в определенных пределах [...] с вероятностью 0.95"? А она означает следующее: если представить себе, что делается не одна серия наблюдений с целью уточнения параметров, а много таких серий (и все в одинаковых условиях), и результаты каждой серии обрабатываются одним методом, то для 95% серий получающийся доверительный интервал действительно содержит истинные значения параметра, а для 5% серий не содержит (т.е. доверительный интервал является ошибочным).
- Касательно расстояния от Земли до Солнца можно заметить, что каждое новое, более точное определение этой величины не укладывается в доверительный интервал, построенный по старым наблюдениям.

Содержание

1 Интервальное оценивание

2 Доверительный интервал для среднего

- Случай известной дисперсии
- Случай неизвестной дисперсии и объём выборки $n > 30$
- Случай малой выборки и неизвестной дисперсии
- Минимальный объем выборки

3 Доверительный интервал для доли и дисперсии

- Доверительный интервал для доли (асимптотический)
- Доверительный интервал для дисперсии

4 Более подробно

Где и что почитать?

Тема. Доверительные интервалы для математического ожидания нормального распределения при известной и неизвестной дисперсии. Доверительные интервалы для среднего нормального распределения при известной и неизвестной дисперсии. Доверительный интервал для доли признака. Доверительный интервал для дисперсии. ([И-М], §11; [Ф,Л], глава 14).



Фадеева Л. Н., Лебедев А. В., Теория вероятностей и математическая статистика: учебное пособие. - 2-е изд., перераб. и доп. - М.: Эксмо, 2010. - 496 с. – (Новое экономическое образование).



Ивашев-Мусатов О. С., Теория вероятностей и математическая статистика: учеб. пособие. - 2-е изд., перераб. и доп. - М.: ФИМА, 2003. - 224 с.