Лекция 3
Доверительные интервалы

Алина Жукова
Head of Analytics
Novakid

План лекции

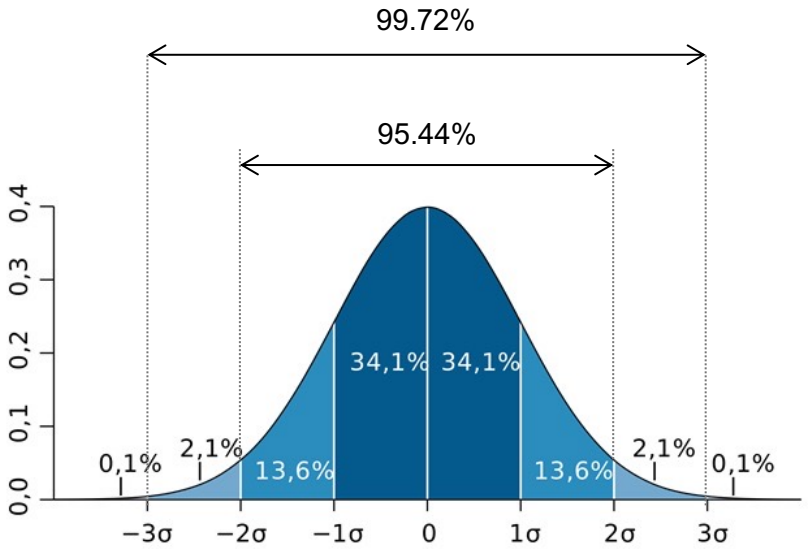
- Предсказательный интервал
- Доверительный интервал
- Точный доверительный интервал для нормальных выборок
- Распределения Хи-квадрат и Стьюдента
- Построение доверительных интервалов для среднего
- Построение доверительных интервалов для доли
- Построение доверительных интервалов на основе бутстрепа

Правило трёх сигм

Случайная величина $X \sim N(\mu, \sigma^2)$

$$\mathbb{P}(\mu - 3 \cdot \sigma \leq X \leq \mu + 3 \cdot \sigma) \approx 0.99$$

Правило трёх сигм (3σ)
С крайне высокой вероятностью случайная величина не отклонится от своего среднего значения более, чем на 3σ. Практически все значения нормально распределённой случайной величины лежат в интервале **(μ − 3 σ ; μ + 3 σ)**, где μ = E (ξ) — математическое ожидание случайной величины.



Предсказательный интервал

Случайная величина $X \sim F(x)$

Для квантиля порядка α $-$
Предсказательный интервал порядка $1 - \alpha$:

$$\mathbb{P}\left(X_{\frac{\alpha}{2}} \leq X \leq X_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

α – уровень значимости

Предсказательный интервал

Случайная величина $X \sim F(x)$

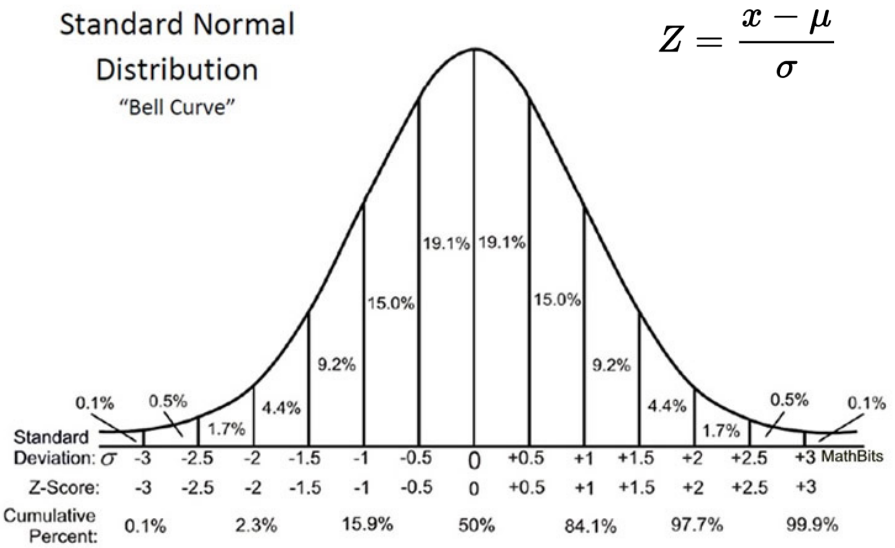
Для квантиля порядка α –
Предсказательный интервал порядка $1 - \alpha$:

$$\mathbb{P}\left(X_{\frac{\alpha}{2}} \leq X \leq X_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

α – уровень значимости

Для $X \sim N(\mu, \sigma^2)$ предсказательный интервал:

$$\mathbb{P}\left(\mu - z_{1-\frac{\alpha}{2}} \cdot \sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}} \cdot \sigma\right) = 1 - \alpha$$



Предсказательный интервал

Предсказательный интервал для среднего $\bar{X}$:

$$\mathbb{P}\left(\mu - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \bar{x} \leq \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

Доверительный интервал для μ:

$$\mathbb{P}\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

Доверительный интервал оценивается по выборке

Доверительный интервал

$X \sim F(x, \theta)$

$\mathbb{P}(\theta_L \leq \theta \leq \theta_U) \geq 1 - \alpha$

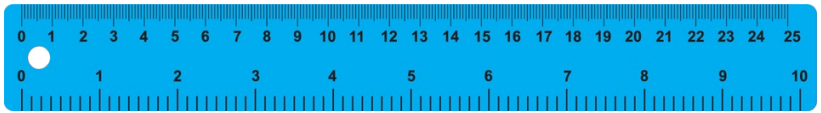
Интервал $[\theta_L; \theta_U]$ – **доверительный интервал**
с уровнем доверия $1 - \alpha$, который с вероятностью $1 - \alpha$
покрывает истинное значение параметра θ
при многократном повторении эксперимента

Доверительный интервал

$$X \sim F(x, \theta)$$

$$\mathbb{P}(\theta_L \leq \theta \leq \theta_U) \geq 1 - \alpha$$

Интервал $[\theta_L; \theta_U]$ – **доверительный интервал**
с уровнем доверия $1 - \alpha$, который с вероятностью $1 - \alpha$
покрывает истинное значение параметра θ
при многократном повторении эксперимента



Когда мы много раз измеряем
лягушку, то с вероятностью $1 - \alpha$
наш доверительный интервал
покрывает её истинную длину

Резюме

Точечная оценка делается по случайной выборке

Неопределённость

Доверительный интервал даёт нам диапазон уверенности в точечной оценке

Чем более узкий доверительный интервал, тем точнее

Примеры

1. Медицинские исследования: Для оценки эффективности нового лекарства проводят испытания на выборке пациентов. Доверительный интервал помогает определить диапазон, в котором может находиться истинная эффективность препарата в общей популяции.

2. Опросы и социологические исследования: При проведении опросов общественного мнения на выборке населения рассчитывается доверительный интервал, чтобы понять, с какой точностью результаты опроса отражают мнение всей популяции.

3. Качество продукции: В производстве доверительные интервалы используются для оценки качества продукции. Например, если измеряется средний вес партии продукции, доверительный интервал позволяет оценить, насколько близко эта оценка к истинному среднему весу всей продукции.

4. Финансовый анализ: Аналитики используют доверительные интервалы, чтобы оценить будущую доходность инвестиций или финансовых инструментов, принимая во внимание рыночную волатильность и другие факторы.

Точный доверительный интервал для нормальных выборок

$X_1, \dots, X_n \sim iid\ N(\mu, \sigma^2)$

Доверительный интервал для μ

σ^2 известна

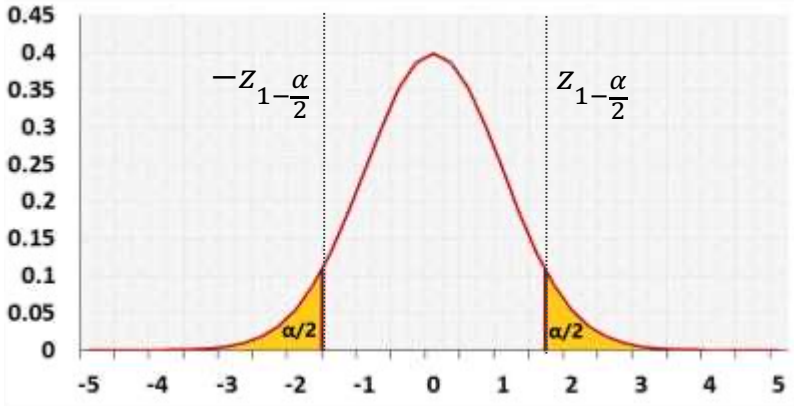
σ^2 неизвестна

Дисперсия известна

$X_1, \dots, X_n \sim iid\ N(\mu, \sigma^2)$ с известной дисперсией σ^2

$$\hat{\mu} = \bar{x} = \frac{X_1 + \dots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$



Доверительный интервал строится по аналогии с асимптотическим, но является точным:

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

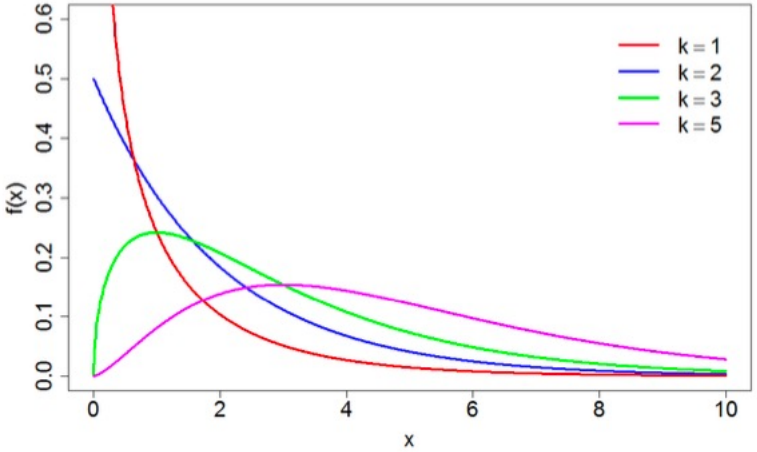
Дисперсия неизвестна

$X_1, \dots, X_n \sim iid\ N(\mu, \sigma^2)$ с неизвестной дисперсией σ^2

$$\hat{\mu} = \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim ?$$

Распределение χ^2



$$X_1, \dots, X_k \sim N(0,1)$$

$$X = \sum_{i=1}^{k} X_i^2 = \chi_k^2$$

Плотность

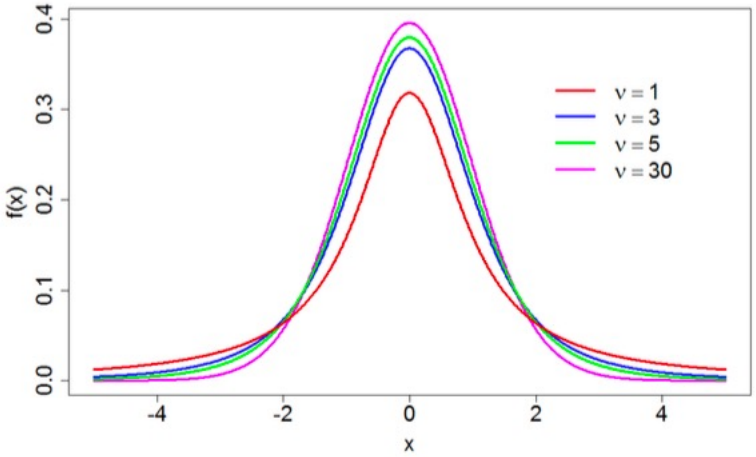
$$f(x) = \frac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot x^{\frac{k}{2}-1} \cdot e^{-\frac{x}{2}}, x \geq 0$$

Характеристики

$$\mathbb{E}(X) = k$$

$$\mathrm{Var}(X) = 2k$$

Распределение Стьюдента



$$X_1 \sim N(0,1), \quad X_2 \sim \chi^2_\nu.$$

$$X = \frac{X_1}{\sqrt{X_2/\nu}} \sim St(\nu),$$

Плотность

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\cdot\Gamma\left(\frac{k}{2}\right)}\cdot\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

Характеристики

$$\mathbb{E}(Z) = 0$$

$$\mathrm{V}ar(Z) = \frac{k}{k-2}, k > 2$$

Теорема Фишера

Если $X_1, \dots, X_n \sim iid\ N(0,1)$, тогда

1. Выборочное среднее $\bar{x}$ и дисперсия s^2 независимы

2. $\frac{(n-1)\cdot s^2}{\sigma^2}$ имеет χ^2 – распределение с $n-1$ степенью свободы

Дисперсия неизвестна

$X_1, \dots, X_n \sim iid\ N(\mu, \sigma^2)$ с неизвестной дисперсией σ^2

$$\hat{\mu} = \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

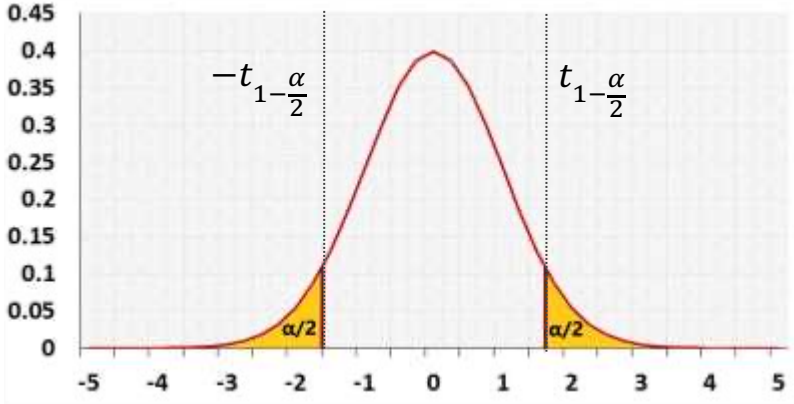
$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim ?$$

$\sigma^2 \longleftarrow s^2$

Дисперсия неизвестна

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$ с неизвестной дисперсией σ^2

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t(n-1)$$



Точный доверительный интервал

$$P\left(\bar{x} - t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Логика


Размер выборки
Дисперсия неизвестна
большое число наблюдений n
любое число наблюдений n
выборка независимая, без аномалий
выборка независимая, из нормального распределения
ЦПТ
Теорема Фишера
Асимптотический доверительный интервал для среднего
Точный доверительный интервал для среднего

Пример

Измерили длину лягушек на озере рядом с дачей:

$\bar{x} = 16$ см, $s = 1.7$ см, $n = 12$

Из ~~wikipedia~~ мета-анализа исследований озёрных лягушек мы знаем:

$\sigma = 2$ см

Пример

Измерили длину лягушек на озере рядом с дачей:

$\bar{x} = 16$ см, $s = 2.7$ см, $n = 12$

Из ~~wikipedia~~ мета-анализа исследований озёрных лягушек мы знаем:

$\sigma = 3$ см

Асимптотический доверительный интервал

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \qquad 16 \pm 1.96 \cdot \frac{2.7}{\sqrt{12}}$$

Точный доверительный интервал с известной дисперсией

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \qquad 16 \pm 1.96 \cdot \frac{3}{\sqrt{12}}$$

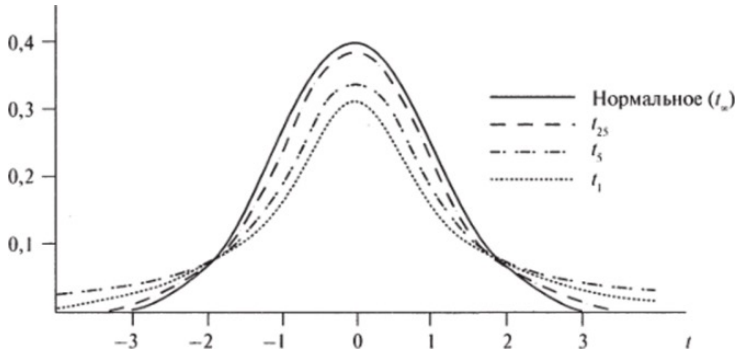
Точный доверительный интервал с известной дисперсией

$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \qquad 16 \pm 2.26 \cdot \frac{2.7}{\sqrt{12}}$$

Что такое большая выборка

Распределение Стьюдента сходится к нормальному распределению при росте числа степеней свободы:

$$t(n) \xrightarrow{d} N(0,1) \text{ при } n \to \infty$$



Доверительные интервалы для доли

Генеральная совокупность состоит из бинарных событий

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

Доверительный интервал Уилсона

Когда доля очень близка к 0 или к 1

$$\frac{1}{1+\frac{z^2}{n}} \left(\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n} + \frac{z^2}{4n^2}} \right), \quad z \equiv z_{1-\frac{\alpha}{2}}.$$

Пример

Опрос об удовлетворённости клиентов

Предположим, что магазин провёл опрос среди своих клиентов, чтобы узнать, сколько из них довольны обслуживанием. Из 400 опрошенных клиентов 320 ответили, что они довольны. Нужно найти 95%-ный доверительный интервал для истинной доли удовлетворённых клиентов.

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

Пример

Опрос об удовлетворённости клиентов

Предположим, что магазин провёл опрос среди своих клиентов, чтобы узнать, сколько из них довольны обслуживанием. Из 400 опрошенных клиентов 320 ответили, что они довольны. Нужно найти 95%-ный доверительный интервал для истинной доли удовлетворённых клиентов.

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

n = 400

$\hat{p} = \frac{320}{400} = 0.8$

z = 1.96

Пример

Опрос об удовлетворённости клиентов

Предположим, что магазин провёл опрос среди своих клиентов, чтобы узнать, сколько из них довольны обслуживанием. Из 400 опрошенных клиентов 320 ответили, что они довольны. Нужно найти 95%-ный доверительный интервал для истинной доли удовлетворённых клиентов.

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

n = 400

$\hat{p} = 320/400$

z = 1.96

[76.1; 83.9%]