Мат. стат. ликзбез

Интеллектуальный анализ данных, 20!7

Малютин Евгений Алексеевич

Планчик

Сегондя в программе:

- Ликбез
- Ликбез
- Ликбез
- Чуть-чуть о гипотезах

- $X \sim N(\mu, \sigma^2) \Rightarrow P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \simeq 0.95$
- $X \sim N(\mu, \sigma^2) \Rightarrow P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \simeq 0.9974$

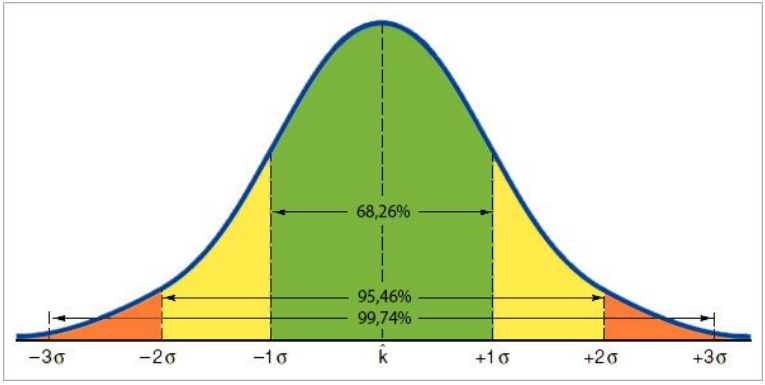


Рис.: Правило 2-3х сигм

Квантиль

- определение 1: $\exists\ \alpha - X_\alpha$

$$P(X \leq X_\alpha) \geq \alpha \quad P(X \geq X_\alpha) \geq 1 - \alpha$$

- определение 2:

$$F(x) = P(X \leq x) \Rightarrow \quad X_\alpha = F^{-1} = inf\{x : F(x) \geq \alpha\}$$

- $P(X_{0.025} < X < X_{0.975}) = 0.975$

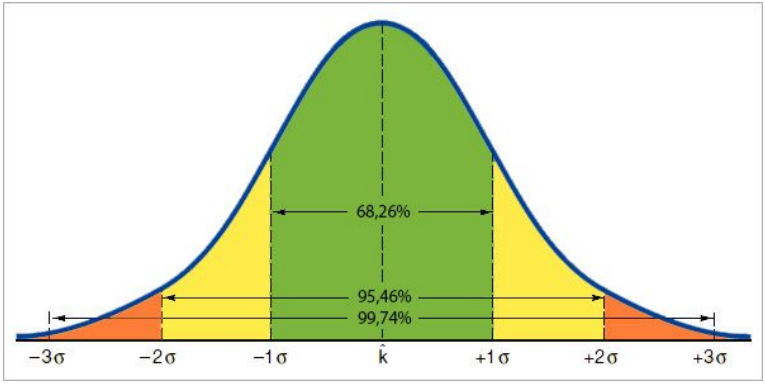


Рис.: Правило 2-3х сигм

T – трюк

- $X \sim F(x) \Rightarrow \quad P(X_{\frac{\alpha}{2}} \leq X \leq X_{(1-\frac{\alpha}{2})}) = (1-\alpha)$
  $[X_{\frac{\alpha}{2}}, X_{1-\frac{\alpha}{2}}]$ – доверительный интервал порядка $1-\alpha$

T – трюк

- $X \sim F(x) \Rightarrow \quad P(X_{\frac{\alpha}{2}} \leq X \leq X_{(1-\frac{\alpha}{2})}) = (1-\alpha)$
  $[X_{\frac{\alpha}{2}}, X_{1-\frac{\alpha}{2}}]$ – доверительный интервал порядка $1-\alpha$
- $X \sim N(\mu, \sigma^2) \Rightarrow$
  $P(\mu - z_{1-\frac{\alpha}{2}}\sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}}\sigma) = 1-\alpha$
  z_α – квантиль нормального распределения $N(0,1)$

T – трюк

- $X \sim F(x) \Rightarrow \quad P(X_{\frac{\alpha}{2}} \leq X \leq X_{(1-\frac{\alpha}{2})}) = (1-\alpha)$
  $[X_{\frac{\alpha}{2}}, X_{1-\frac{\alpha}{2}}]$ – доверительный интервал порядка $1-\alpha$
- $X \sim N(\mu, \sigma^2) \Rightarrow$
  $P(\mu - z_{1-\frac{\alpha}{2}}\sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}}\sigma) = 1-\alpha$
  z_α – квантиль нормального распределения $N(0,1)$
- $z_{0.975} \simeq 1.95966 \simeq 2$

- $X \sim F(x, \theta)$; θ – неизвестный параметр
- $\theta = ?$

- $X \sim F(x, \theta)$; θ – неизвестный параметр
- $\theta = ?$
- $X^n = (X_1, ...X_n)$
- $\widehat{\theta}$ – оценка θ по выборке

- $X \sim F(x, \theta)$; θ – неизвестный параметр
- $\theta = ?$
- $X^n = (X_1, ...X_n)$
- $\widehat{\theta}$ – оценка θ по выборке
- Например, для $\theta = E[X]$:

$$\widehat{\theta} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ – хорошая оценка}$$

- Доверительный интервал для параметра θ – пара таких статистик C_L и C_U, что:

$$P(C_L \leq \theta \leq C_U) \geq 1 - \alpha$$

- Доверительный интервал для параметра θ – пара таких статистик C_L и C_U, что:
$$P(C_L \leq \theta \leq C_U) \geq 1 - \alpha$$
- Как оценить C_L и C_U по выборке?

- Доверительный интервал для параметра θ – пара таких статистик C_L и C_U, что:
$$P(C_L \leq \theta \leq C_U) \geq 1 - \alpha$$
- Как оценить C_L и C_U по выборке?
- Если $\widehat{\theta}$ – оценка θ и мы знаем её распределение $F_{\widehat{\theta}}$, то:
$$P(F^{-1}_{\widehat{\theta}}(\frac{\alpha}{2}) \leq \widehat{\theta} \leq F^{-1}_{\widehat{\theta}}(1 - \frac{\alpha}{2})$$

- $X \sim N(\mu, \sigma^2), X^n = (X_1, ...X_n)$

- $X \sim N(\mu, \sigma^2), X^n = (X_1, ...X_n)$
- $\widehat{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow$

- $X \sim N(\mu, \sigma^2), X^n = (X_1, ...X_n)$
- $\widehat{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow$
- Пресдказательный интервал: $P(\mu - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \widehat{X}_n \leq \mu + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$

Ликбез

Для нормального распределения

- $X \sim N(\mu, \sigma^2), X^n = (X_1, ...X_n)$
- $\widehat{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow$
- Пресдказательный интервал: $P(\mu - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \widehat{X}_n \leq \mu + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- Доварительный интервал для μ: $P(\widehat{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \widehat{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$

Для нормального распределения

- Предсказательный интервал для X:

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

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

- Доверительный интервал для μ:

$$P(\widehat{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \widehat{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

- Логично - оценка величины и мат.ожидания

Ликбез

Для любого другого распределения:

- $X \sim F(X), X^n = (X_1, ..., X_n)$
- $\widehat{X}_n$ – оценка $E[X]$

Ликбез

Для любого другого распределения:

- $X \sim F(X), X^n = (X_1, ..., X_n)$
- $\widehat{X}_n$ – оценка $E[X]$
- $\widehat{X}_n \simeq N(E[X], \frac{D[X]}{n})$(ЦПТ $\Rightarrow$)

Ликбез

Для любого другого распределения:

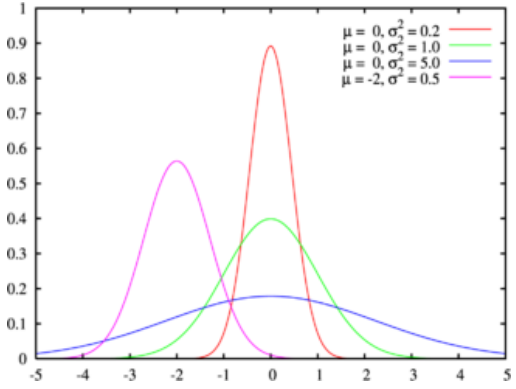
- $X \sim F(X), X^n = (X_1, ..., X_n)$
- $\widehat{X}_n$ – оценка $E[X]$
- $\widehat{X}_n \simeq N(E[X], \frac{D[X]}{n})$(ЦПТ $\Rightarrow$)
- Доверительный интервал для μ:

$$P(\widehat{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{D[X]}{n}} \leq \mu \leq \widehat{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{D[X]}{n}}) = 1 - \alpha$$

Ликбез. Другие распределения

Нормальное распределение

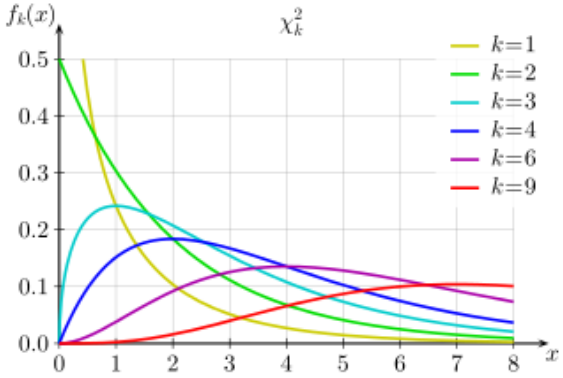
- $X \sim N(\mu, \sigma^2)$
- $f(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- $F(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} exp(-\frac{(t-\mu)^2}{2\sigma^2}) dt$



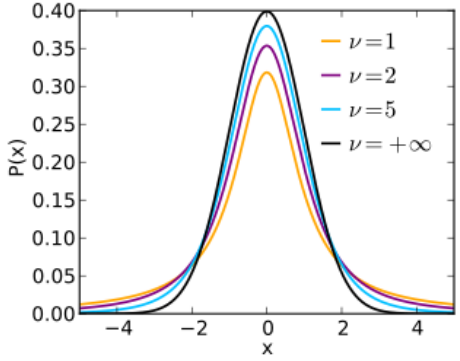
Ликбез. Другие распределения

χ^2 распределение

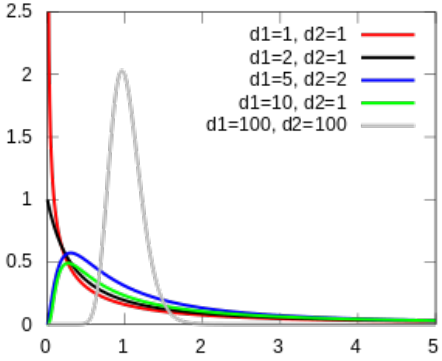
- $X_1, X_2, ...X_n \sim N(0,1)$ – независимы
- $X = \sum_{i=1}^{k} X_i^2 \sim \chi_k^2$ – распределение "хи-квадрат"с k степенями свободы



- $X_1 \sim N(0,1), \quad X_2 \sim \chi^2_v$ – независимы
- $X = \frac{X_1}{\sqrt{X_2/v}} \sim St(v)$ – распределение Стьюдента с v степенями свободы



- $X_1 \sim \chi^2_{d_1}, \quad X_2 \sim \chi^2_{d_2}$ – независимы
- $X = \frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$ – распределение Стьюдента с v степенями свободы



- $X \sim N(\mu, \sigma^2), \quad X^n = (X_1, ..., X_n)$

- $X \sim N(\mu, \sigma^2), \quad X^n = (X_1, ..., X_n)$
- $X_n = \frac{1}{n} \sum\limits_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$

- $X \sim N(\mu, \sigma^2), \quad X^n = (X_1, ..., X_n)$
- $X_n = \frac{1}{n} \sum\limits_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$
- $S_n^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - \widehat{X}_n)^2 \Rightarrow (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

- $X \sim N(\mu, \sigma^2), \quad X^n = (X_1, ..., X_n)$
- $X_n = \frac{1}{n} \sum\limits_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$
- $S_n^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - \widehat{X}_n)^2 \Rightarrow (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$
- $T = \frac{\widehat{X_n} - \mu}{S_n/\sqrt{n}} \sim St(n-1)$

- $X \sim N(\mu, \sigma^2), \quad X^n = (X_1, ..., X_n)$
- $X_n = \frac{1}{n} \sum\limits_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$
- $S_n^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - \widehat{X}_n)^2 \Rightarrow (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$
- $T = \frac{\widehat{X_n} - \mu}{S_n/\sqrt{n}} \sim St(n-1)$
- Если есть две X_1, X_2 из двух нормальных распределений: $\frac{S_1^2/\sigma_1^2}{S_1^1/\sigma_1^1} \sim F(n_1 - 1, n_2 - 1)$

Z и T

- Мы знаем дисперсию выборки – z-интервал: $X_n \pm z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$

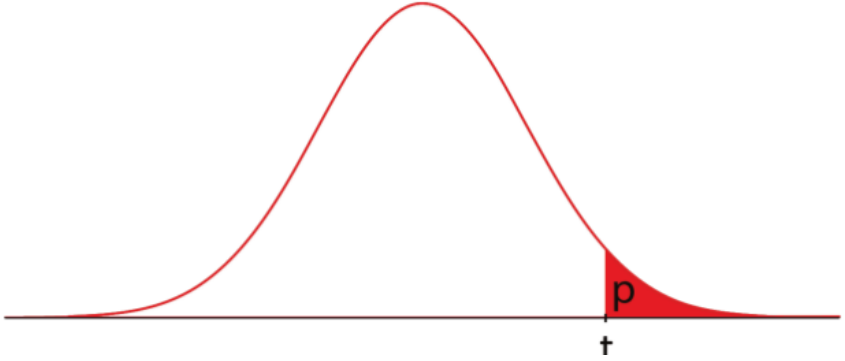
Z и T

- Мы знаем дисперсию выборки – z-интервал: $X_n \pm z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
- Мы не знаем дисперсию выборки – t-интервал: $X_t \pm t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$

Гипотезы

Формальные определения

- выборка: $X^n = (X_1, ...X_n), X \sim P$
- нулевая гипотеза: $H_0 : P \in \omega$
- альтернатива: $H_1 : P \notin \omega$
- статистика: $T(X^n)$,
  $T(X^n) \sim F(x)$ при H_0
  $T(X^n) \nsim F(x)$ при H_1



- $T(X) = t$
- $P(T(X) \geq t) =$ p-value
- p-value $\leq \alpha$ – H_0 отвергается

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка II рода
H_0 отвергается	Ошибка I рода	H_0 верно отвергнута

Рис.: Ошибки

- $P(H_0 \text{ отвергнута} | H_0 \text{ верна}) = P(p \leq \alpha | H_0) \leq \alpha$
- $pow = P(\text{ отвергаем } H_0 | H_1) = 1 - P(\text{ принимаем } H_0 | H_1)$

- $p = P(T \geq t|H_0)$
- $p = P(T \geq t|H0) \neq P(H_0) = P(H_0|T \geq t)$