Мат. стат. ликзбез

Интеллектуальный анализ данных, 20!7

Малютин Евгений Алексеевич

Планчик

Сегондя в программе:

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

Правило 2-х(3-х) сигм

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

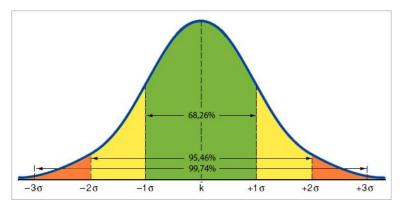


Рис.: Правило 2-3х сигм

Квантиль

• определение 1: $\exists \ \alpha - X_{\alpha}$

$$P(X \le X_{\alpha}) \ge \alpha \quad P(X \ge X_{\alpha}) \ge 1 - \alpha$$

• определение 2:

$$F(x) = P(X \le x) \Rightarrow X_{\alpha} = F^{-1} = \inf\{x : F(x) \ge \alpha\}$$

Правило двух-трёх сигм

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

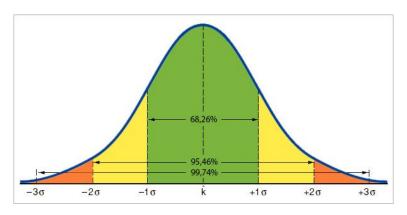


Рис.: Правило 2-3х сигм

Предсказательный интервал

Т – трюк

ullet $X\sim F(x)\Rightarrow P(X_{rac{lpha}{2}}\leq X\leq X_{(1-rac{lpha}{2})})=(1-lpha)$ $[X_{rac{lpha}{2}},X_{1-rac{lpha}{2}}]$ — доверительный интервал порядка 1-lpha

Т – трюк

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

Т – трюк

- $X \sim F(x) \Rightarrow P(X_{\frac{\alpha}{2}} \leq X \leq X_{(1-\frac{\alpha}{2})}) = (1-\alpha)$ $[X_{\frac{\alpha}{2}}, X_{1-\frac{\alpha}{2}}]$ 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)
- $z_{0.975} \simeq 1.95966 \simeq 2$

Точечные оценки

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

Точечные оценки

- ullet $X \sim F(x, heta)$; heta неизвестный параметр
- $\theta = ?$
- $X^n = (X_1, ... X_n)$
- $oldsymbol{\hat{ heta}}$ оценка heta по выборке

Точечные оценки

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

$$\widehat{ heta} = \overline{X_n} = rac{1}{n} \sum_{i=1}^n X_i$$
 — хорошая оценка

Доверительный интервал

• Доверительный интервал для параметра θ — пара таких статистик C_L и C_U , что:

$$P(C_L \le \theta \le C_U) \ge 1 - \alpha$$

Доверительный интервал

• Доверительный интервал для параметра θ — пара таких статистик C_L и C_U , что:

$$P(C_L \le \theta \le C_U) \ge 1 - \alpha$$

• 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?
- ullet Если $\widehat{ heta}$ оценка heta и мы знаем её распределение $F_{\widehat{ heta}}$, то:

$$P(F_{\widehat{\theta}}^{-1}(\frac{\alpha}{2}) \leq \widehat{\theta} \leq F_{\widehat{\theta}}^{-1}(1 - \frac{\alpha}{2})$$

Для нормального распределения

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

Для нормального распределения

$$\bullet \ \ X \sim \textit{N}(\mu,\sigma^2), X^n = (X_1,...X_n)$$

•
$$\widehat{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow$$

Для нормального распределения

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

•
$$\widehat{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow$$

ullet Пресдказательный интервал: $P(\mu-z_{1-rac{lpha}{2}rac{\sigma}{\sqrt{n}}} \leq \widehat{X}_n \leq \mu+z_{1-rac{lpha}{2}rac{\sigma}{\sqrt{n}}}) = 1-lpha$

Для нормального распределения

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

Для нормального распределения

• Предсказательный интервал для X:

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

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

ullet Доверительный интервал для μ :

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

• Логично - оценка величины и мат.ожидания

Для любого другого распределения:

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

Для любого другого распределения:

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

Для любого другого распределения:

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

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



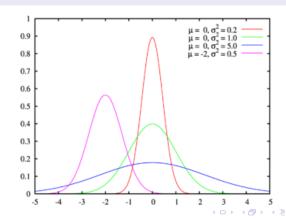
Ликбез. Другие распределения

Нормальное распределение

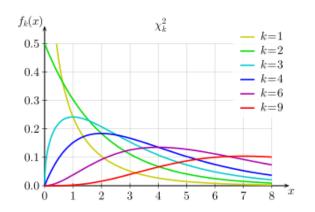
•
$$X \sim N(\mu, \sigma^2)$$

•
$$f(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

•
$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} exp(-\frac{(t-\mu)^2}{2\sigma^2})dt$$



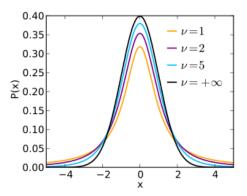
- ullet $X_1, X_2, ... X_n \sim \mathcal{N}(0,1)$ независимы
- ullet $X=\sum\limits_{i=1}^k X_i^2\sim \chi_k^2$ распределение "хи-квадрат"с k степенями свободы



Ликбез. Другие распределения

Распределение Стьюдента

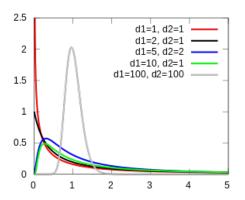
- ullet $X_1 \sim {\it N}(0,1), \quad X_2 \sim \chi^2_{\it v}$ независимы
- ullet $X=rac{X_1}{\sqrt{X_2/
 u}}\sim St(
 u)$ распределение Стьюдента с u степенями свободы



Ликбез. Другие распределения

Распределение Стьюдента

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



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



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

•
$$X_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$$

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

$$\bullet X_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$$

•
$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \widehat{X}_n)^2 \Rightarrow (n-1) \frac{S^2}{\sigma^2} \sim \chi_{n-1}^2$$



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

•
$$X_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$$

•
$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \widehat{X}_n)^2 \Rightarrow (n-1) \frac{S^2}{\sigma^2} \sim \chi_{n-1}^2$$

•
$$T = \frac{\widehat{X}_n - \mu}{S_n / \sqrt{n}} \sim St(n-1)$$

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

$$\bullet X_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$$

•
$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \widehat{X}_n)^2 \Rightarrow (n-1) \frac{S^2}{\sigma^2} \sim \chi_{n-1}^2$$

•
$$T = \frac{\widehat{X_n} - \mu}{S_n / \sqrt{n}} \sim St(n-1)$$

ullet Если есть две X_1, X_2 из двух нормальных распределений:

$$\frac{S_1^2/\sigma_1^2}{S_1^1/\sigma_1^1} \sim F(n_1-1,n_2-1)$$

Построение доверительных интервалов для среднего

ΖиΤ

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

Построение доверительных интервалов для среднего

ΖиТ

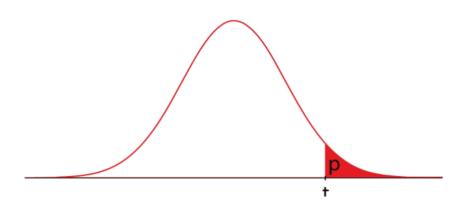
- ullet Мы знаем дисперсию выборки z-интервал: $X_n \pm z_{1-lpha} rac{\sigma}{\sqrt{n}}$
- ullet Мы не знаем дисперсию выборки t-интервал: $X_t \pm t_{1-rac{lpha}{2}}rac{S}{\sqrt{n}}$

Гипотезы

Формальные определения

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

Нулевое распределение



- T(X) = t
- $P(T(X) \ge t) = \text{p-value}$
- p-value $\leq \alpha H_0$ отвергается

Гипотезы

Ошибки I и II рода

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

Рис.: Ошибки

- $P(H_0$ отвергнута $|H_0|$ верна $)=P(p\leq lpha|H_0)\leq lpha$
- ullet роw=P(отвергаем $H_0|H_1)=1-P($ принимаем $H_0|H_1)$

Гипотезы

Достигаемый уровень значимости

•
$$p = P(T \ge t | H_0)$$

•
$$p = P(T \ge t | H0) \ne P(H_0) = P(H_0 | T \ge t)$$