

Vzorce dostupné u zkoušky

Histogram

- Scottovo pravidlo pro doporučenou šířku sloupce histogramu

$$h_{opt} = \left(\frac{24\sqrt{\pi}}{n} \right)^{1/3} s_x$$

s_x je výběrová směrodatná odchylka, n je počet pozorování

- Sturgesovo pravidlo pro doporučený počet sloupců

$$k = 1 + 3.3 \log_{10} n$$

n je počet pozorování

Normovaná proměnná

-

$$U_i = \frac{X_i - \bar{X}}{\sigma_x}$$

σ_x je směrodatná odchylka, $\bar{X}$ je průměr

Normální rozdělení

- Pro absolutní četnost n_i hodnoty x_i platí

$$n_i \approx c_0 \exp - \frac{(x_i - \mu)^2}{2\sigma^2}$$

σ je směrodatná odchylka, μ je střední hodnota, c_0 je konstanta

- hustota normálního rozdělení

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

σ je směrodatná odchylka, μ je střední hodnota

Hledání modu

- Jádrový odhad hustoty

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K \left(\frac{x - x_i}{h} \right)$$

K je jádro, n je počet pozorování, h je šířka okna
příklady jader

- Epanechnikovo $K(u) = \frac{3}{4}(1 - u^2)I_{\{|u| \leq 1\}}$
- Trojúhelníkové $K(u) = (1 - |u|)I_{\{|u| \leq 1\}}$
- Rovnoměrné $K(u) = \frac{1}{2}I_{\{|u| \leq 1\}}$
- Gaussovo $K(u) = \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}u^2$
- I značí identifikátor množiny
- Modus pro číselnou diskrétní proměnnou

$$\hat{x} = A + h \frac{d_0}{d_0 + d_1}$$

A je počáteční bod, h je šířka sloupce, d_0 je rozdíl výšky nejvyššího a od něj levého sloupce, d_1 je rozdíl výšky nejvyššího a od něj pravého sloupce

Odlehle hodnoty

- Klasické pravidlo založené na IQR
od bližšího kvartilu dál než 1.5 násobek mezikvartilového rozpětí
- Striktnější pravidlo založené na IQR
od bližšího kvartilu dál než 3 násobek mezikvartilového rozpětí
- Základní pravidlo založené na SD (3 sigma)
od průměru dále než 3 násobek směrodatné odchylky
- Striktnější pravidlo založené na SD (4 sigma)
od průměru dále než 4 násobek směrodatné odchylky
- transformace dat
 - logaritmická transformace

$$y = \ln x$$

- mocninná transformace

$$y \equiv \left(\frac{x - \min x}{\max x - \min x} \right)^h$$

$h > 0$ se určuje experimentálně

Popisné statistiky polohy

- Useknutý průměr – usekne se $\alpha\%$ nejmenších a největších hodnot a zbytek se zprůměruje
- Huberův odhad – hodnoty menší než $\bar{x} - K$ se nahradí hodnotou $\bar{x} - K$ a hodnoty větší než $\bar{x} + K$ se nahradí hodnotou $\bar{x} + K$, pak se vše zprůměruje

Popisné statistiky variability

- Variační koeficient

$$CV = \frac{\sigma_x}{\bar{X}}$$

σ_x je směrodatná odchylka, $\bar{X}$ je průměr

- Mediánová absolutní odchylka od mediánu

$$MAD = \text{median}(|X_1 - \tilde{X}|, |X_2 - \tilde{X}|, \dots, |X_n - \tilde{X}|) / 0.6745$$

$\tilde{X}$ je medián, n je počet pozorování

Popisné statistiky tvaru rozdělení

- Klasická šikmost

$$Skew = \frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma_x} \right)^3$$

σ_x je směrodatná odchylka, $\bar{X}$ je průměr, n je počet pozorování

- Klasická špičatost

$$Kurt = \frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma_x} \right)^4 - 3$$

σ_x je směrodatná odchylka, $\bar{X}$ je průměr, n je počet pozorování

- Bowleyho kvartilový koeficient šikmosti

$$S_{Q_4} = \frac{(Q_3 - Q_2) + (Q_1 - Q_2)}{(Q_3 - Q_2) - (Q_1 - Q_2)}$$

- Moorsův oktilový koeficient špičatosti

$$K_{Q^8} = \frac{(Q_7^8 - Q_5^8) + (Q_3^8 - Q_1^8)}{(Q_6^8 - Q_2^8)} - 1.233$$

Informační kritéria

- Akaikeho informační kritérium (AIC):

$$AIC = 2k - 2 \ln(L)$$

L je věrohodnost, k je počet parametrů v modelu

- Bayesovské informační kritérium (BIC):

$$BIC = \ln(n)k - 2 \ln(L)$$

Intervaly spolehlivosti

- interval spolehlivosti pro podíl

$$p \pm z(1 - \alpha/2) \sqrt{\frac{p(1-p)}{n}}$$

p je odhad podílu, z je kvantil normálního rozdělení, α je hladina významnosti, n je počet pozorování

- interval spolehlivosti pro rozdíl podílů

$$(p_1 - p_2) \pm z(1 - \alpha/2) \sqrt{\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right)}$$

- interval spolehlivosti pro průměr se známou variabilitou σ

$$\bar{X} \pm z(1 - \alpha/2) \frac{\sigma}{\sqrt{n}}$$

$\bar{X}$ je průměr, z je kvantil normálního rozdělení, α je hladina významnosti, n je počet pozorování

- interval spolehlivosti pro průměr s neznámou variabilitou

$$\bar{X} \pm t_{n-1}(1 - \alpha/2) \frac{\text{sd}(\mathbf{X})}{\sqrt{n}}$$

$\bar{X}$ je průměr, t_{n-1} je kvantil t-rozdělení, α je hladina významnosti, $\text{sd}(\mathbf{X})$ je výběrová směrodatná odchylka, n je počet pozorování

- interval spolehlivosti pro rozdíl průměrů

$$(\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2}(1 - \alpha/2) \sqrt{\frac{(n_1-1)\text{sd}(\mathbf{X})_1^2 + (n_2-1)\text{sd}(\mathbf{X})_2^2}{n_1+n_2-2}} \sqrt{\frac{n_1+n_2}{n_1n_2}}$$

- interval spolehlivosti pro rozptyl

$$\left(\frac{(n-1)\text{sd}(\mathbf{X})^2}{\chi_n^2(1-\alpha/2)}, \frac{(n-1)\text{sd}(\mathbf{X})^2}{\chi_n^2(\alpha/2)} \right)$$

χ_n^2 je kvantil χ^2 -rozdělení, α je hladina významnosti, $\text{sd}(\mathbf{X})$ je výběrová směrodatná odchylka, n je počet pozorování

Korelační koeficient

- Pearsonův korelační koeficient

$$\text{Cor}(\mathbf{X}, \mathbf{Y}) = \frac{\text{Cov}(\mathbf{X}, \mathbf{Y})}{\sqrt{\text{Var}(\mathbf{X})\text{Var}(\mathbf{Y})}} = \frac{\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{Y}_i - \bar{\mathbf{Y}})}{\sqrt{\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})^2 \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})^2}}$$

Lineární regrese

- Metoda nejmenších čtverců

$$\min \sum_{i=1}^n R_i^2 = \min \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \min_{b_0, b_1} \sum_{i=1}^n (Y_i - (b_0 + b_1 X_i))^2$$

- Absolutní člen

$$b_0 = \bar{Y} - b_1 \bar{X}$$

- Lineární člen

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

- Koeficient determinace

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \text{cor}(\mathbf{X}, \mathbf{Y})^2$$

Pravděpodobnostní rozdělení

- Alternativní rozdělení

$$P(X = k) = p^k (1 - p)^{1-k}, \quad k \in \{0, 1\}$$

- Binomické rozdělení

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k \in \{0, 1, \dots, n\}$$

- Hypergeometrické rozdělení

$$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}, \quad k \in \{\max(0, n + M - N), \dots, \min(M, n)\}$$

- Poissonovo rozdělení

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \{0, 1, 2, \dots\}$$

- Geometrické rozdělení

$$P(X = k) = p(1 - p)^k, \quad k \in \{0, 1, 2, \dots\}$$

- Rovnoměrné rozdělení

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}$$

- Normální rozdělení

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

- Exponenciální rozdělení

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$