NMAI059 Pravděpodobnost a statistika 1
11. přednáška

Robert Šámal

Přehled

Testy dobré shody

Lineární regrese

Bayesovská statistika

χ^2_k – rozdělení χ-kvadrát

Z_i^2 vel. se stř. h. 1, rozptyl. 2

$Z_i : \Omega \to \mathbb{R}$

Definice

$Z_1, \dots, Z_k \sim N(0,1)$ *n.n.v. Rozdělení náhodné veličiny*

$\operatorname{var} Z_i = 1 = \mathbb{E}Z_i^2 - (\mathbb{E}Z_i)^2 = \mathbb{E}Z_i^2$

$$Q = Z_1^2 + \cdots + Z_k^2 \quad : \Omega \to \mathbb{R}$$

se nazývá χ-kvadrát s k stupni volnosti. (Opravdu k!)

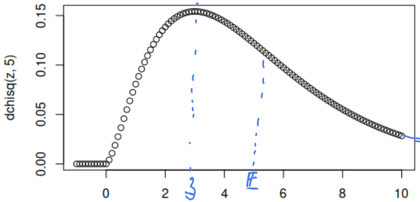
- $\mathbb{E}(Q) = k$ (lehké)
- $var(Q) = 2k$ (pro info, netřeba pamatovat)
- hustota jde napsat vzorcem, jde najít např. na Wikipedii
- Q je pro velké n blízké $N(k, 2k)$ (CLV)

$\mathbb{E}Q = \mathbb{E}\sum_{i=1}^k Z_i^2 = \sum^k \mathbb{E}Z_i^2 = \sum 1 = k$

$\mathbb{E}Q^2 = \mathbb{E}\left(\sum_i Z_i^2\right)\left(\sum_j Z_j^2\right) = \mathbb{E}\sum_{i,j} Z_i^2 \cdot Z_j^2 = \sum \mathbb{E}Z_i^2 Z_j^2$

$i \neq j$: $\mathbb{E}Z_i^2 \cdot \mathbb{E}Z_j^2 = 1 \cdot 1$

$i = j$: $\mathbb{E}Z_i^4 = ?$



χ^2_5

Multinomické a kategoriální rozdělení

$p_1 \to 1$, $p_2 \to 2$, …, $p_k \to k$ — kategoriální n.v.

Definice

Dána $p_1, \dots, p_k \geq 0$ tak, že $p_1 + p_2 + \dots + p_k = 1$.
n-krát zopakuji pokus, kde může nastat právě *jedna z k možností, i-tá má pravděpodobnost p_i.*
$X_i :=$ kolikát nastala i-tá možnost. $(X_1, \dots, X_k)$ má multinomické rozdělení s parametry $n, (p_1, \dots, p_k)$.

$n = \sum_{i=1}^{k} X_i$

$0 \leq X_i \leq n$

- triviální případ: $X_i =$ počet hodů kostkou, kdy padlo i
- důležitý případ: $X_i =$ počet výskytů i-tého písmene, i-tého slovního druhu, …
- $P(X_1 = x_1, \dots, X_k = x_k) = \underbrace{\binom{n}{x_1, \dots, x_k}}_{\text{multinom. koef.}} p_1^{x_1} \dots p_k^{x_k}$

dostanu multinom. rozd. $n, (p_1+p_2, p_3, \dots, p_k)$

multinom. koef. $\dfrac{n!}{x_1! \dots x_k!} = \#$ uspoř. $x_1 \times 1$, $x_2 \times 2$, …, $x_k \times k$ do řady

Pro kontrolu

$$\sum_{x_1 \dots x_k} P(X_1 = x_1 \dots X_k = x_k) = \sum \binom{n}{x_1, \dots, x_k} p_1^{x_1} \dots = \underbrace{(p_1 + \dots + p_k)}_{\text{kategorie}}{}^{n} = 1$$

… $0 \leq x_1, \dots, x_k \leq n$, $\sum x_i = n$

$\Rightarrow X_1 \sim Bin(n, p_1)$

$\Rightarrow$ sloučím 1 a 2 do jedné

Pearsonova χ^2 statistika

- $(X_1, \dots, X_k)$ – multinomické rozdělení s parametry n, $(p_1, \dots, p_k)$ jako minule
- $E_i := \mathbb{E}(X_i) = np_i$
- *Pearsonova χ^2 statistika* je funkce

$$\chi^2 = T := \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}$$

- **Věta** $T \xrightarrow{d} \chi^2_{k-1}$

$\mu = E X_1$

$\sigma^2 = \mathrm{Var}\, X_1 = np(1-p)$

$n \to \infty$

$T \xrightarrow{d} Z_1^2$

$Z_1 \sim N(0,1)$

$\boxed{k=2}$

$X_1 + X_2 = n$

$p_1 + p_2 = 1$

$(*) = \frac{(X_1 - np_1)^2 (p_1+p_2)}{np_1p_2}$

$T = \frac{(X_1-E_1)^2}{E_1} + \frac{(X_2-E_2)^2}{E_2}$

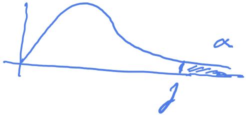
$= \frac{(X_1-np_1)^2}{np_1} + \frac{(n - X_1 - n(1-p_1))^2}{np_2} = \frac{(X_1-np_1)^2 p_2}{np_1p_2} + \frac{(X_1-np_1)^2 p_1}{np_2p_1} = (*)$

$T = \left(\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2$

$\frac{X_1 - \mu}{\sigma} \xrightarrow[n\to\infty]{d} N(0,1)$ $\boxed{CLV}$

Test dobré shody (goodness of fit)

- $(X_1, \ldots, X_k)$ – multinomické rozdělení s parametry $n, \vartheta = (\vartheta_1, \ldots, \vartheta_k)$ jako minule
- n známe, ϑ neznáme.
- Hypotéza H_0: $\vartheta = \vartheta^*$
- $E_i := n\vartheta_i^*$ pro všechna i
- Použijeme statistiku $\boxed{T := \sum_{i=1}^k \frac{(X_i - E_i)^2}{E_i}}$
- Hypotézu H_0 zamítneme, pokud $T > \gamma$
- $\gamma := F_Q^{-1}(1 - \alpha)$, kde $Q \sim \chi^2_{k-1}$
- $P(\text{chyba prvního druhu}) = P(T > \gamma; H_0) \xrightarrow{n\to\infty} P(Q > \gamma) = \alpha$



Test dobré shody – příklad (600×)

- Házíme opakovaně kostkou. Jednotlivá čísla padla s četností 92, 120, 88, 98, 95, 107.
  $X_1 \quad X_2 \quad \dots \quad X_6$
- Je kostka spravedlivá? (hlas: 8:1 ano)

$n = 600 \qquad \theta^* = \left(\frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6}\right)$

$E_i = \frac{1}{6} \cdot 600 = 100$

$$T = \sum_{i=1}^{6} \frac{(X_i - 100)^2}{100} = \frac{8^2}{100} + \frac{20^2}{100} + \frac{12^2}{100} + \frac{2^2}{100} + \frac{5^2}{100} + \frac{7^2}{100}$$

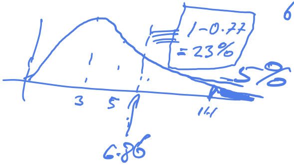
$= 6.86$

64
400
144
4
25
49
686

$Q \sim \chi^2_5 \qquad F_Q^{-1}(0.95) = 11.1 = \gamma \quad (1-\alpha)$

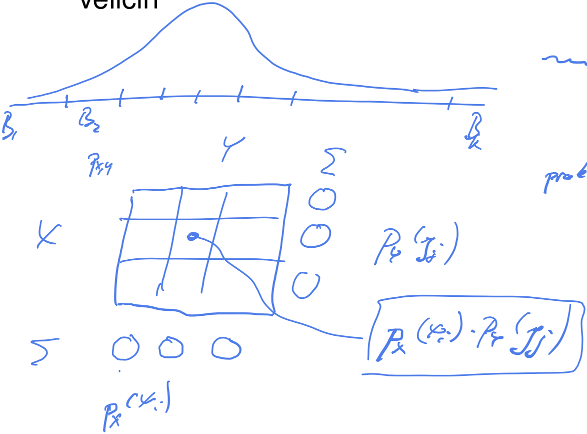
p-hodnota ... $1 - F_Q(\underbrace{6.86}_{T}) = 1 - 0.77$

$= 0.23$



Další rozšíření

- Pro zkoumání rozdělení libovolné n.v. Y můžeme vybrat „přihrádky“ $B_1, \dots, B_k$ (rozklad $\mathbb{R}$) a zkoumat, kolikrát je $Y \in B_i$
- Obdobný test pro nezávislost (diskrétních) náhodných veličin



$\rightsquigarrow$ má Hraou. rozd.

$p_i := P(Y \in B_i)$

prakt. podmínka: $E_i > 5$, jinak nefunguje

$\to T = \sum_i \frac{(X_i - E_i)^2}{E_i}$

Přehled

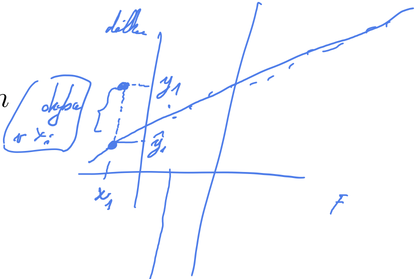
Testy dobré shody

Lineární regrese

Bayesovská statistika

Lineární regrese – zadání

- data: (x_i, y_i) pro $i = 1, \ldots, n$
- cíl: $y = \vartheta_0 + \vartheta_1 x$



- měříme pomocí kvadratické odchylky

$$\sum_{i=1}^{n} \big(y_i - (\vartheta_0 + \vartheta_1 x_i)\big)^2$$

minimalizujeme

Lineární regrese – řešení

- Minimalizujeme výraz

$$\sum_{i=1}^{n}\big(y_i - (\vartheta_0 + \vartheta_1 x_i)\big)^2$$

- řešení: Optimální parametry jsou

~ výb. kovariance

$$\boxed{\hat\vartheta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar x)(y_i-\bar y)}{\sum_{i=1}^{n}(x_i-\bar x)^2}, \qquad \hat\vartheta_0 = \bar y - \vartheta_1 \bar x,}$$

kde $\bar x := (x_1 + \cdots + x_n)/n$, $\bar y := (y_1 + \cdots + y_n)/n$.

~ rozptyl

výběrové průměry

(bez dk – viz MA/LA)

Lineární regrese – proč součet čtverců?

- Předpokládejme, že $x_1, \dots, x_n$ jsou pevná, y_i je zvoleno jako hodnota náhodné veličiny

ideální část → chyba

bodové odhady pro ϑ_0, ϑ_1

$$Y_i = \vartheta_0 + \vartheta_1 x_i + W_i$$

$W_i := Y_i - (\vartheta_0 + \vartheta_1 x_i)$

- $W_i \sim N(0, \sigma^2)$ pro všechna i; $W_1, \dots, W_k$ nezávislé.
- metoda maximální věrohodnosti:

φ_σ ... hustota W_i: $N(0, \sigma^2)$

$$L(y; \vartheta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \vartheta_0 - \vartheta_1 x_i)^2}{2\sigma^2}}$$

$\varphi_\sigma(y_i - (\vartheta_0 + \vartheta_1 x_i))$

- $\ell(y; \vartheta) = \log L(y; \vartheta) = a + b \sum_{i=1}^{n} (y_i - \vartheta_0 - \vartheta_1 x_i)^2$

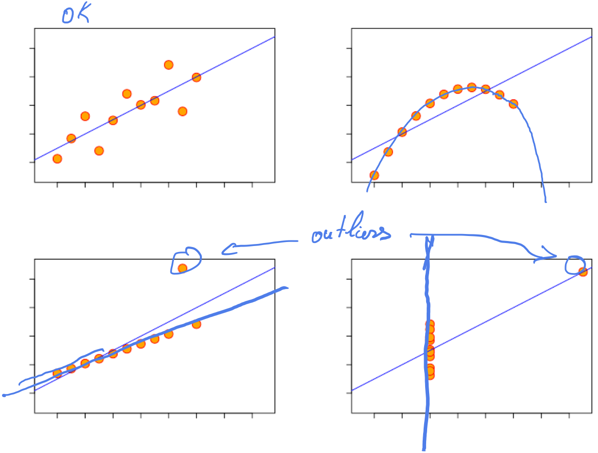
max

$= n \log \frac{1}{\sigma\sqrt{2\pi}}$ $\quad - \frac{1}{2\sigma^2} \sum (\ )^2$

min

Limity regrese

stejné příčiny, korelace



$y = g(x)$

↑

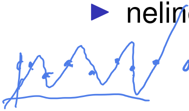
kvadratická!

(data: Francis Anscombe 1973, obrázek: wikieditor Schutz)

- nelineární regrese

$$\sum (y_i - g(x_i))^2$$

hledáme g, které minim. součet čtv.



overfitting

overtraining

Přehled

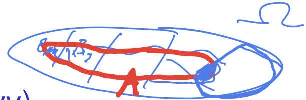
Testy dobré shody

Lineární regrese

Bayesovská statistika

klasická / frekventistická statistika

Bayesova věta



Věta (Bayesova pro jevy)

Pokud $B_1, B_2, \dots$ *je rozklad* Ω, $A \in \mathcal{F}$ *a* $P(A), P(B_j) > 0$, *tak*

měření

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_i P(A \mid B_i)P(B_i)}.$$

$$\frac{P(A \cap B_j)}{P(A)}$$

(sčítance s $P(B_i) = 0$ *považujeme za 0).*

- apriorní, posteriorní pravděpodobnost (prior, posterior)

Věta (Bayesova pro diskrétní náhodné veličiny)

X, Y *jsou diskrétní n.v.*

$P(A|B_j)$ $A = \{X = x\}$ $B_j = \{Y = y\}$

$P(B_j)$

$$P(Y=y|X=x) = P(B_j|A) = p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y'} p_{X|Y}(x|y')p_Y(y')}.$$

(sčítance s $p_Y(y') = 0$ *považujeme za 0).* $P(A|B_j) \cdot P(B_j)$

Bayesova věta

Věta (Bayesova pro spojité náhodné veličiny)

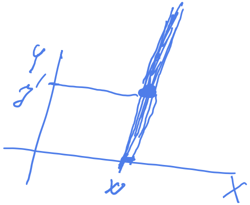
X, Y jsou spojité n.v., které mají hustotu f_X, f_Y i sdruženou hustotu $f_{X,Y}$

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_Y(y)}{\int_{y' \in \mathbb{R}} f_{X|Y}(x|y') f_Y(y') dy'}.$$

(sčítance s $f_Y(y') = 0$ považujeme za 0).

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{\boxed{f_X(x)}} = \frac{f_{X|Y}(x|y) \cdot f_Y(y)}{}$$

$$f_X(x) = \int_{y'} f_{X,Y}(x,y')\,dy' = \int_{y'} f_{X|Y}(x|y') \cdot f_Y(y')\,dy'$$



Bayesova věta – zbylé varianty

- X je spojitá Y diskrétní

$$p_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)p_Y(y)}{\sum_{y' \in ImY} f_{X|Y}(x|y')p_Y(y')}.$$

- X je diskrétní Y spojitá

$$f_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)f_Y(y)}{\int_{y' \in \mathbb{R}} p_{X|Y}(x|y')f_Y(y')dy'}.$$