NMAI059 Pravděpodobnost a statistika 1
11. přednáška

Robert Šámal

Přehled

χ^2_k – rozdělení χ-kvadrát

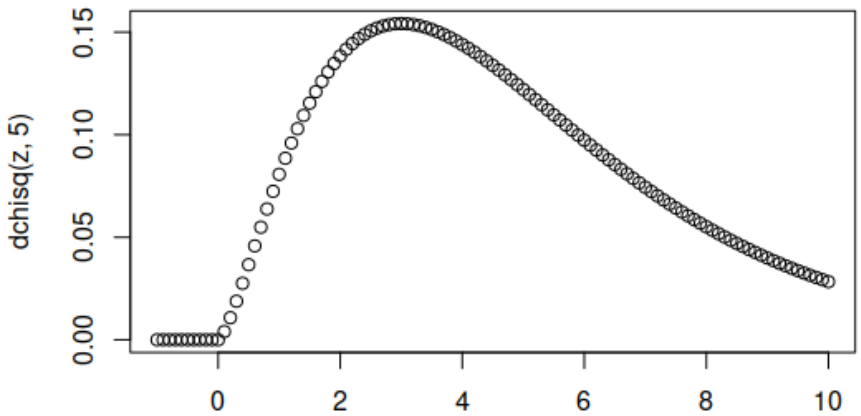
Definice

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

$$Q = Z_1^2 + \cdots + Z_k^2$$

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)



χ^2_5

Multinomické a kategoriální rozdělení

Definice

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

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

Pearsonova χ^2 statistika

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

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

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

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 $T := \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}$
- Hypotézu H_0 zamítneme, pokud $T > \gamma$
- $\gamma := F_Q^{-1}(1-\alpha)$, kde $Q \sim \chi^2_{k-1}$
- $P(\text{chyba prvního druhu}) = P(T > \gamma; H_0) \to P(Q > \gamma) = \alpha$

Test dobré shody – příklad

- Házíme opakovaně kostkou. Jednotlivá čísla padla s četností 92, 120, 88, 98, 95, 107.
- Je kostka spravedlivá?

Další rozšíření

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

Přehled

Lineární regrese – zadání

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

- měříme pomocí kvadratické odchylky

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

Lineární regrese – řešení

- Minimalizujeme výraz

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

- řešení: Optimální parametry jsou

$$\hat{\vartheta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\vartheta}_0 = \bar{y} - \vartheta_1 \bar{x},$$

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

Lineární regrese – proč součet čtverců?

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

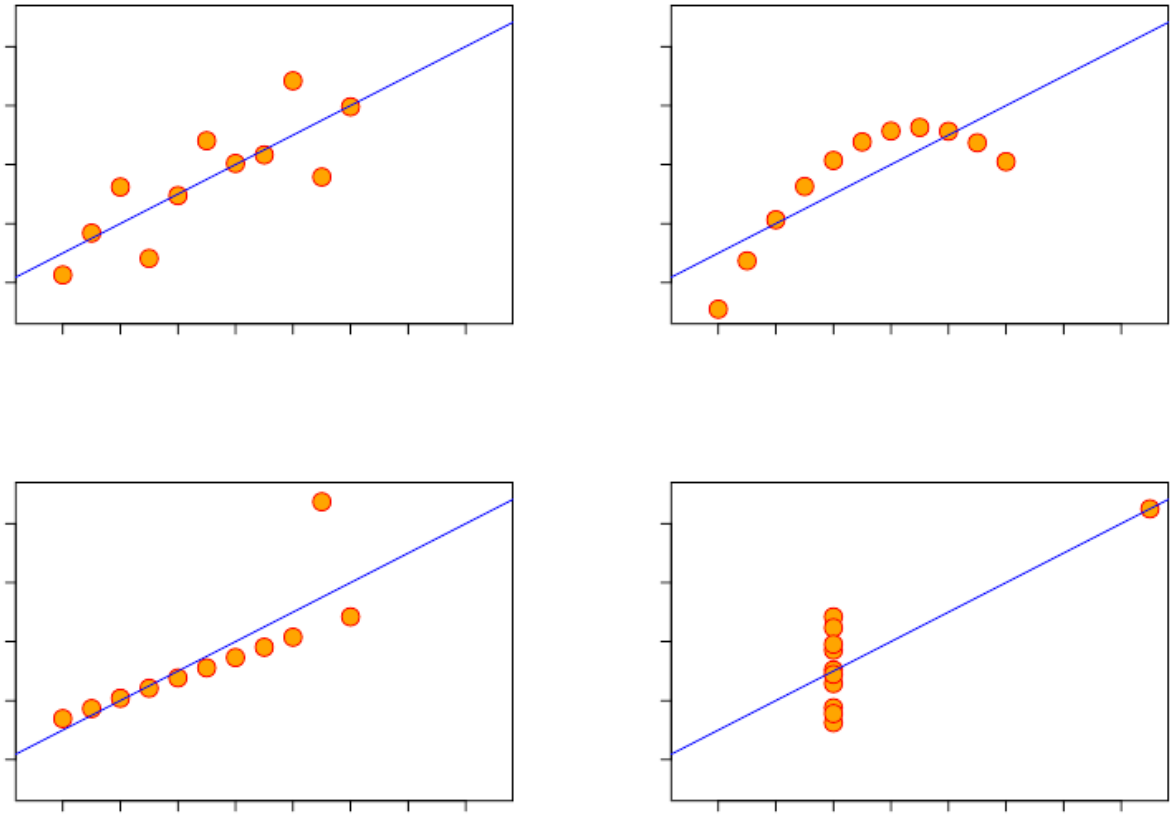
$$Y_i = \vartheta_0 + \vartheta_1 x_i + W_i$$

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

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

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

Limity regrese



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

- nelineární regrese

Přehled

Bayesova věta

Věta (Bayesova pro jevy)

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

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

(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.*

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

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

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).

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

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