NMAI059 Pravděpodobnost a statistika 1
12. přednáška

Robert Šámal

Statistika – Co už víme

- základní nastavení: uvažujeme náhodný výběr $X_1, \ldots, X_n$ z distribuce F_ϑ — popisuje proces měření, jak mohlo měření proběhnout
- naměříme data – konkrétní čísla, tzv. realizaci náhodného výběru $x_1, \ldots, x_n$ — jak naše měření skutečně proběhlo

1. bodové odhady: máme určit co nejlepší číslo, odhad pro parametr ϑ, nebo nějakou jeho funkci $g(\vartheta)$.
2. intervalové odhady: máme určit interval, ve kterém parametr ϑ pravděpodobně leží
3. testování hypotéz

Přehled

Testování hypotéz – ilustrace

- Chceme testovat, zda je mince spravedlivá.
- H_0: je spravedlivá *očekávaný stav světa*
- H_1: není spravedlivá *překvapivé zjištění* („Vědci objevili, že v kasinu byla použita falešná mince.“)
- Výsledky: zamítneme H_0/nezamítneme H_0
- Chyba 1. druhu: chybné zamítnutí. Zamítneme H_0, i když platí. Trapas.
- Chyba 2. druhu: chybné přijetí. Nezamítneme H_0, ale ona neplatí. Promarněná příležitost.
- Potřebujeme určit k takové, že budeme zamítat H_0 pokud $|S - n/2| > k$.

Testování hypotéz – obecný postup

- Vybereme vhodný statistický model.
- Volíme *hladinu významnosti (significance level)* α: pravd. chybného zamítnutí H_0. Typicky $\alpha = 0.05$.
- Určíme *testovou statistiku* $T = h(X_1, \ldots, X_n)$, kterou budeme určovat z naměřených dat.
- Určíme *kritický obor (rejection region)* – množinu W.
- Naměříme hodnoty $x_1, \ldots, x_n$ náh. veličin $X_1, \ldots, X_n$.
- Rozhodovací pravidlo: zamítneme H_0 pokud $h(x_1, \ldots, x_n) \in W$.
- $\alpha = P(h(X) \in W; H_0)$
- $\beta = P(h(X) \notin W; H_1)$ … $1 - \beta$ je tzv. *síla testu*

- často α nevolíme předem, ale spočítáme tzv. p*-hodnotu:* minimální α, pro které bychom H_0 zamítli.

Testování hypotéz – příklad

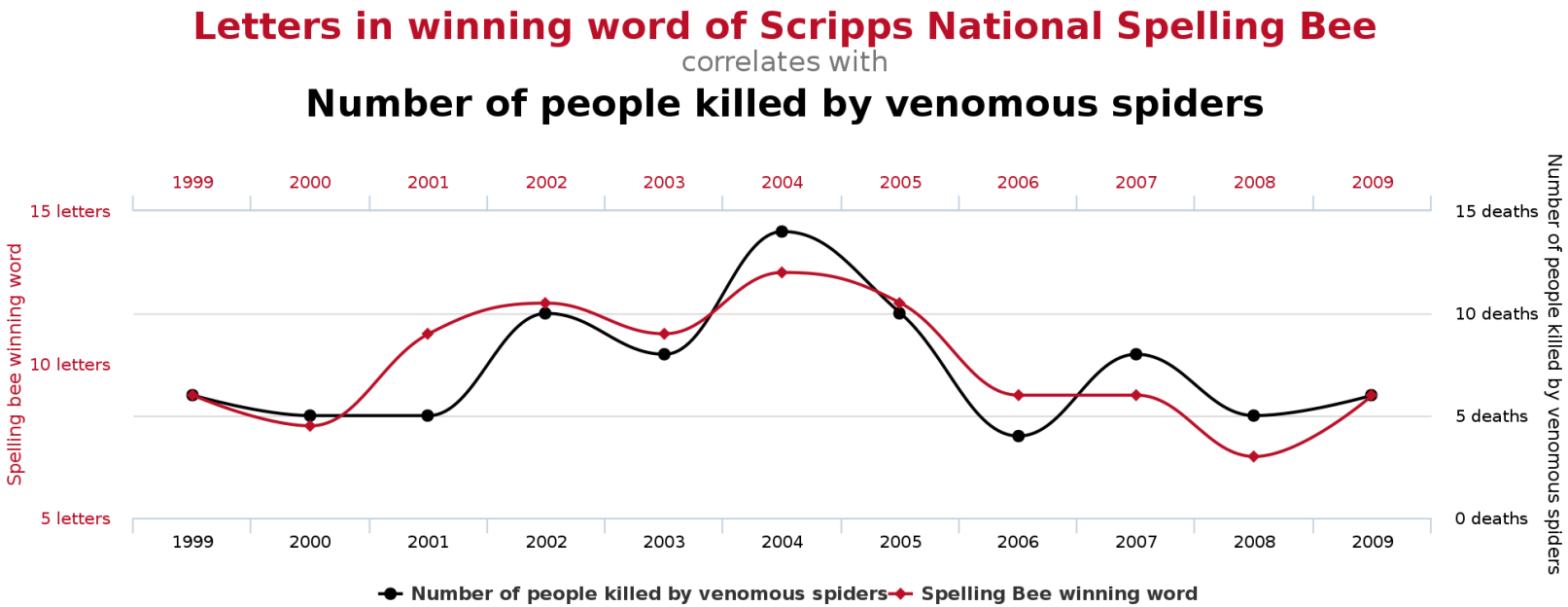
- $X_1, \ldots, X_n$ náhodný výběr z $N(\vartheta, \sigma^2)$
- σ^2 známe, μ dáno
- $H_0 : \vartheta = \mu \qquad H_1 : \vartheta \neq \mu$

Testování hypotéz – příklad dvojvýběrového testu

- $X_1, \ldots, X_{n_1}$ náhodný výběr z $Ber(\vartheta_X)$
- $Y_1, \ldots, Y_{n_2}$ náhodný výběr z $Ber(\vartheta_Y)$
- $H_0 : \vartheta_X = \vartheta_Y \qquad H_1 : \vartheta_X \neq \vartheta_Y$

p-hacking

- napřed získáme data, pak v nich hledáme zajímavosti
- když máme dost dat, tak tam nějaké budou „shodou okolností“
- *reprodukovatelnost* – po explorační analýze dat uděláme nezávislý sběr dat a ten analyzujeme konfirmačně
- nebo dopředu náhodně rozdělíme data na část pro tvorbu hypotéz a část pro jejich potvrzení . . . jednoduchý případ křížové validace (cross validation)



Přehled

χ^2_k – rozdělení χ-kvadrát

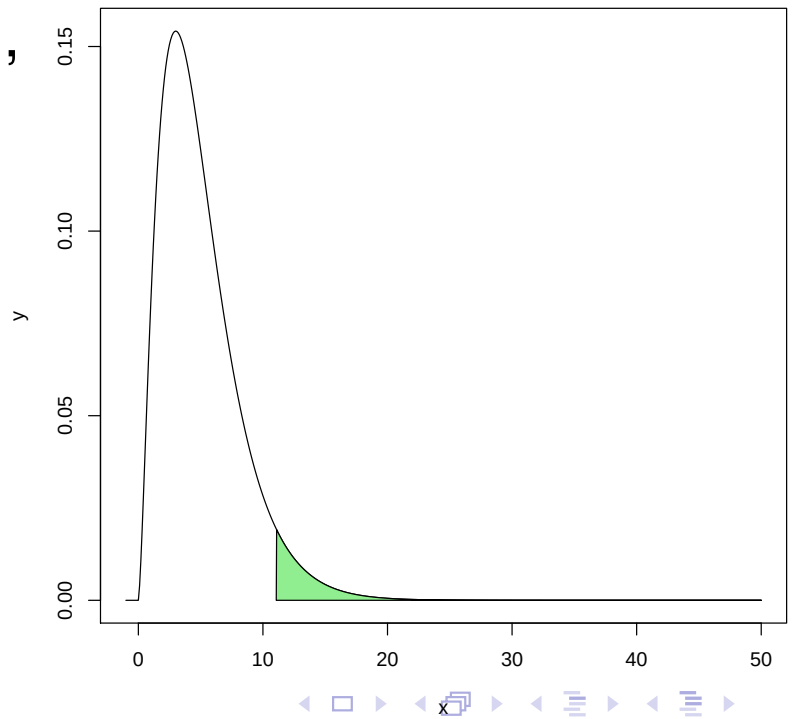
Definice

$Z_1, \ldots, Z_k \sim N(0,1)$ *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
- $Q \doteq N(k, 2k)$ pro velká k (CLV)



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

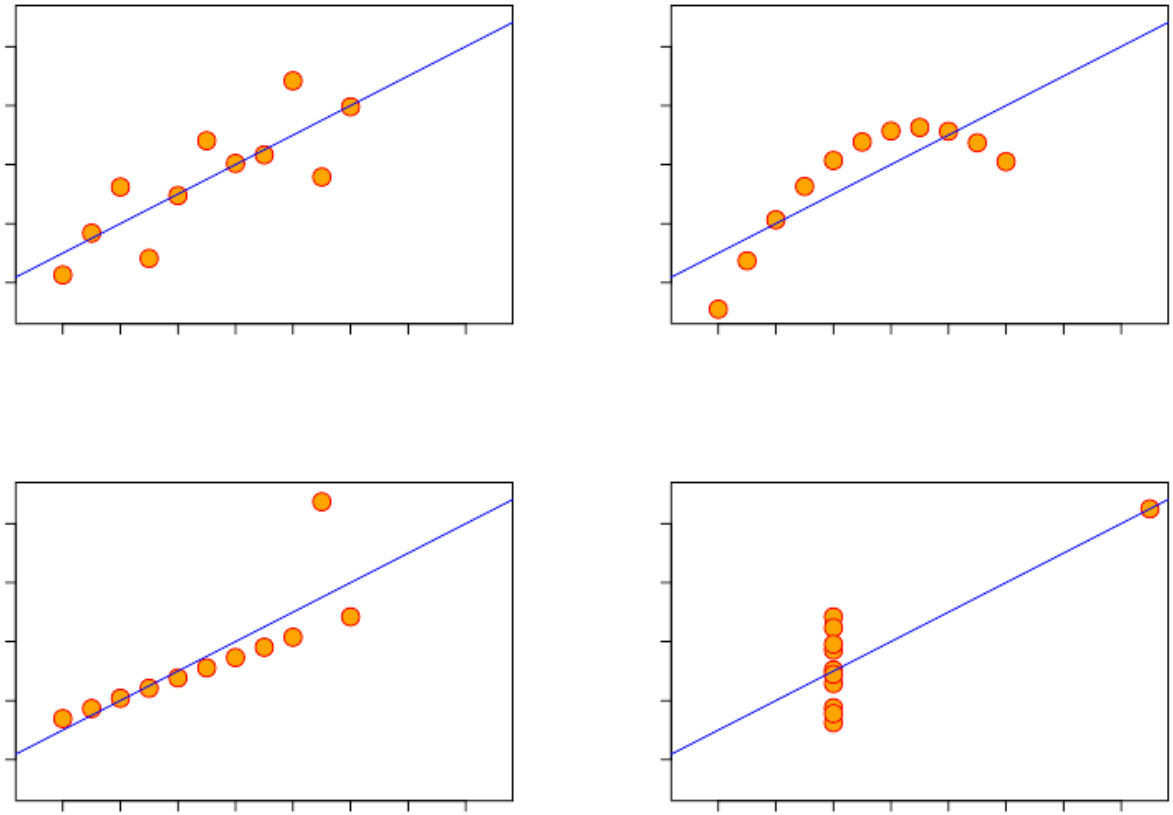
$$Y_i = \vartheta_0 + \vartheta_1 x_i + W_i$$

- $W_i \sim N(0, \sigma^2)$ 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

Simpsonův paradox

Treatment / Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**