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

$\bar{X}_n - \delta \qquad \hat{\theta}_n = \bar{X}_n \qquad \bar{X}_n + \delta$

Přehled

n.n.v. $X \sim N(\mu, \sigma^2)$

$Y \sim N(\mu', \sigma'^2) \Rightarrow -Y \sim N(-\mu', \sigma'^2)$

$\Downarrow$

$X+Y \sim N(\mu+\mu', \sigma^2+\sigma'^2)$

Testování hypotéz

$$f_Z(z) = \int f_X(x)\, f_Y(z-x)$$

Testy dobré shody

Lineární regrese

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

Mince: $X_1, \dots, X_n \sim Ber(p)$

$H_0: p = \frac{1}{2}$ $H_1: p \neq \frac{1}{2}$

- Vybereme vhodný statistický model.
- Volíme *hladinu významnosti (significance level)* α: pravd. chybného zamítnutí H_0. Typicky $\alpha = 0.05$.

  F_T známe

- Určíme *testovou statistiku* $T = h(X_1, \dots, X_n)$, kterou budeme určovat z naměřených dat.

  $T = \sum_{i=1}^{n} X_i \sim Bin(n, p)$

- Určíme *kritický obor (rejection region)* – množinu W.
- Naměříme hodnoty $x_1, \dots, x_n$ náh. veličin $X_1, \dots, X_n$.

  $n = 1000$

- Rozhodovací pravidlo: zamítneme H_0 pokud $h(x_1, \dots, x_n) \in W$.

- $\alpha = P(h(X) \in W; H_0)$ $\boxed{c?}$

  $W = \{0, 500-c\} \cup \{500+c, 1000\}$

- $\beta = P(h(X) \notin W; H_1) \dots 1 - \beta$ je tzv. *síla testu*

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

volíme c: $P(T < 500 - c) \doteq \alpha/2$

$F_T(500 - c) \doteq \alpha/2$

$\lfloor F_T^{-1}(\alpha/2) \rfloor = 500 - c$

Testování hypotéz – příklad

měříme teplotu
chceme $\mu = 5\,^\circ C$

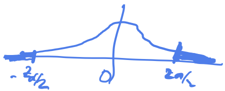
- $X_1, \dots, X_n$ náhodný výběr z $N(\vartheta, \sigma^2)$
- σ^2 známe, μ dáno
- $H_0 : \vartheta = \mu \qquad H_1 : \vartheta \neq \mu$

$$T = \frac{X_1 + \dots + X_n}{n} = \overline{X}_n \sim N(\vartheta, \sigma^2/n)$$ ... víme

1. z dřívějška známe
chování pro rozptyl var

$$S = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

2. pokud platí ... $\vartheta = \mu = 5$
a σ^2 také známe $\Rightarrow$
známe rozdělení T



ZNÁME!

$$W = \{ s \in \mathbb{R} : |s| > z_{\alpha/2} \}$$

$$z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) = 1.96$$

$\alpha = 0.05$

Testování hypotéz – příklad dvojvýběrového testu

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

dvě metody léčby

X_i, Y_j ... zda byla léčba úspěšná

$X_i = \{$ i-tý pacient, který dostal I. léčbu se uzdravil $\}$

$Y_j = \{$ j-tý pac. ... II. léčba ... se uzdravil $\}$

$\hat\Theta_X = \frac{X_1 + \dots + X_{n_1}}{n_1}$ ... odhad ϑ_X → přibl. norm. rozd. CLV

$\hat\Theta_Y = \frac{Y_1 + \dots + Y_{n_2}}{n_2}$ ... odhad ϑ_Y

$Z := \hat\Theta_X - \hat\Theta_Y$

pokud $|Z|$ je velké, tak H_0 zamítneme

zamítneme pokud $|T| > z_{\alpha/2}$

Pokud platí H_0: $E\hat\Theta_X = E\hat\Theta_Y \Rightarrow EZ = 0$

$\vartheta = \vartheta_X = \vartheta_Y$

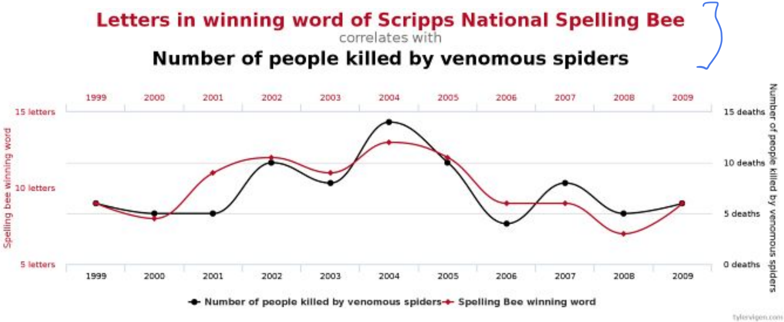
Z je přibl. $N(0, \sigma^2)$ → neznámé

$$\sigma^2 = var(Z) = var(\hat\Theta_X) + var(\hat\Theta_Y) = \frac{var X_1}{n_1} + \frac{var Y_1}{n_2} = \frac{\vartheta_X(1-\vartheta_X)}{n_1} + \frac{\vartheta_Y(1-\vartheta_Y)}{n_2} = \vartheta(1-\vartheta)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$\hat\Theta = \frac{\sum X_i + \sum Y_j}{n_1 + n_2} \text{ ... odhad } \vartheta \Rightarrow \hat\sigma^2 := \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat\Theta(1-\hat\Theta) \Rightarrow T := \frac{\hat\Theta_X - \hat\Theta_Y}{\hat\sigma} = \frac{\hat\Theta_X - \hat\Theta_Y}{\sqrt{\dots}} \sim N(0,1)$$

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

Testování hypotéz

Testy dobré shody

Lineární regrese

χ^2_k – rozdělení χ-kvadrát

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

$\mathbb{E} Q = \mathbb{E} Z_1^2 + \dots + \mathbb{E} Z_k^2 = k$

$var\, Q = var\, Z_1^2 + \dots + var\, Z_k^2$

$var\, Z_i^2 = \mathbb{E} Z_i^4 - (\mathbb{E} Z_i^2)^2$, $\mathbb{E} Z_i^4 = 3$, $\mathbb{E} Z_i^2 = 1$

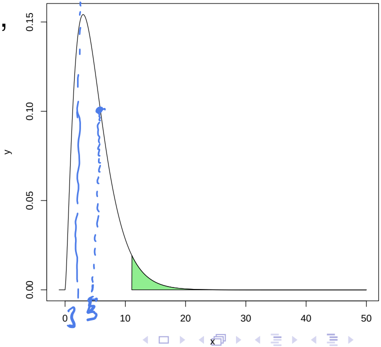
Definice

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

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

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)



χ^2_5

Multinomické a kategoriální rozdělení

p_1 výsledek 1
p_2 výsl. 2
$\vdots$
k

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

- $(X_1, \dots, X_k)$ – 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

O_i .. observed

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

k pevné, $n \to \infty$

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

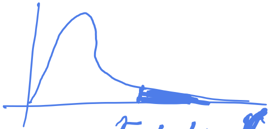
Dk pro $k=2$ $\quad X_1 + X_2 = n, \quad p_1 + p_2 = 1 \quad E_i = np_i$

$$T = \frac{(X_1 - E_1)^2}{E_1} + \dots = \frac{(X_1 - np_1)^2 p_2}{np_1 p_2} + \frac{(\overbrace{n - X_1 - n(1-p_1)}^{-X_1 + np_1})^2 p_1}{np_2 p_1} = \frac{(X_1 - np_1)^2 \cdot \overbrace{(p_1 + p_2)}^{1}}{np_1 p_2}$$

$$= \left(\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2 \qquad \frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \xrightarrow{d} N(0,1) \text{ podle CLV}$$

Test dobré shody (goodness of fit)

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



$T\ W: (\gamma, \infty)$

Test dobré shody – příklad

- Házíme opakovaně kostkou. Jednotlivá čísla padla s četností 92, 120, 88, 98, 95, 107.
  1 2 3 4 5 6
- Je kostka spravedlivá?

$n = 92 + 120 + \dots = 600 \qquad \vartheta^* = (\frac{1}{6}, \dots, \frac{1}{6}), \quad E_i = n\frac{1}{6} = 100$

$T = \sum_{i=1}^{6} \frac{(X_i - 100)^2}{100} = \frac{(92-100)^2}{100} + \frac{(120-100)^2}{100} + \frac{12^2}{100} + \frac{2^2}{100} + \frac{5^2}{100} + \frac{7^2}{100}$

$= (8^2 + 20^2 + 12^2 + 2^2 + 5^2 + 7^2)/100 = \underline{6.86}$

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

p-hodnota: $1 - F_Q(6.86) \doteq 1 - 0.77 = 0.23$

G-test, ...

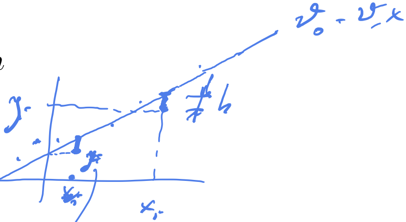
Další rozšíření

- 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

Přehled

Lineární regrese – zadání

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



skutečnost:

$$\boxed{y = \vartheta_0 + \vartheta_1 x + \text{náh. chyba}}$$

- měříme pomocí kvadratické odchylky

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

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

$$\frac{\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

--- výběrová kovariance

---> výběrový rozptyl

Lineární regrese – proč součet čtverců?

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

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

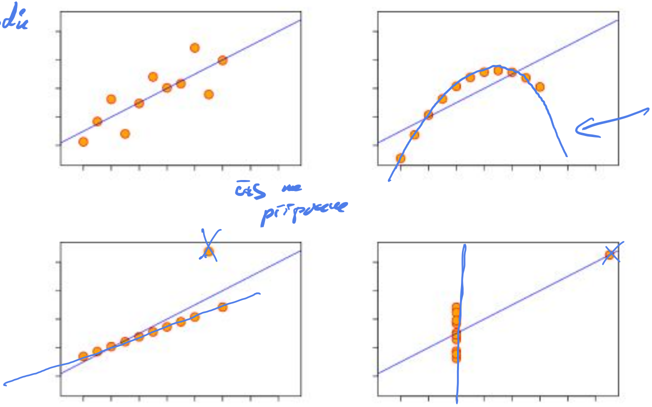
- $W_i \sim N(0, \sigma^2)$ pro všechna i; $W_1, \ldots, W_k$ nezávislé.
- metoda maximální věrohodnosti:

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

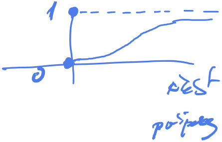
- $\ell(y; \vartheta) = \log L(y; \vartheta) = a + b \sum_{i=1}^{n} (y_i - \vartheta_0 - \vartheta_1 x_i)^2$

Limity regrese

#bodů



čas na přípravu



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

- nelineární regrese

→ logistická regrese

X_i -- reálná čísla

Y_i → 0/1

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)**