

**Katedra pravděpodobnosti a matematické statistiky**



**MATEMATICKO-FYZIKÁLNÍ  
FAKULTA**  
Univerzita Karlova

---

**prof. RNDr. Arnošt Komárek, Ph.D.**

**NMSA331 Matematická statistika 1**

---

Zimní semestr 2025–26

## Přednášky (Den HH:MM – HH:MM v Kx)

přestávka v trvání 10 minut (+ $\varepsilon$ ) přibližně uprostřed

prof. RNDr. Arnošt Komárek, Ph.D.

`komarek@karlin.mff.cuni.cz`

`https://msekce.karlin.mff.cuni.cz/~komarek`

2. patro vedle schodů

## Cvičení (Den DD:MM v Kx)

doc. Ing. Marek Omelka, Ph.D.

`omelka@karlin.mff.cuni.cz`

`https://msekce.karlin.mff.cuni.cz/~omelka`

2. patro vedle schodů

## Webpage přednášky

[https://msekce.karlin.mff.cuni.cz/~komarek/vyuka/2025\\_26/nmsa331-2025.html](https://msekce.karlin.mff.cuni.cz/~komarek/vyuka/2025_26/nmsa331-2025.html)

## Webpage cvičení

[https://msekce.karlin.mff.cuni.cz/~omelka/Vyuka\\_nmsa331\\_2526.php](https://msekce.karlin.mff.cuni.cz/~omelka/Vyuka_nmsa331_2526.php)

- **Nezákladnější materiál:** vlastní poznámky (doplnění slidů) vytvořené během duchaplné přítomnosti na přednáškách.
- **Doplňkový materiál** (pro vyjasnění pasáží, které mohl přednášející vysvětlit ne zcela dostatečně): *poznámky* k dřívějším verzím přednášky (doc. Mgr. Michal Kulich, Ph.D.).
  - Přednáška edice 2025 sleduje tyto *poznámky* v relativně vysoké míře, nicméně ne doslova. Některé pasáže z *poznámek* v přednášce (a u zkoušky) nebudou, některé pasáže přednášky v *poznámkách* nejsou, ale zkoušet se budou.
- **Učebnice:** mnohé z toho, čemu se budeme učit lze nalézt v klasických učebnicích profesora Jiřího Anděla. Jedná se zejména o *Statistické metody* a *Základy matematické statistiky*, obě dvě vydané nakladatelstvím matfyzpress. Přednáška nicméně nesleduje těsně ani jednu z těchto knih. Jedná se o podpurné texty.

- Matematická statistika je vybudována na základech teorie pravděpodobnosti.
- To nejdůležitější z této oblasti potřebné pro pochopení látky matematické statistiky je shrnuto v *dokumentu* docenta Michala Kulicha.
- Většina ze zde uvedených poznatků se probírá v předmětech **NMSA202 Pravděpodobnost a matematická statistika** a **NMSA333 Teorie pravděpodobnosti 1**.
- Některé poznatky budou probrány na cvičení k tomuto předmětu (v prvních týdnech semestru).

- **Písemná část:** 100 minut, obvykle dopoledne.
- **Ústní část:** zadání otázky/problému
  - písemná příprava (cca 20 minut)
  - rozprava nad přípravou s možností doplnění.

Obvykle odpoledne ve stejný den jako písemná část.

- Výsledná známka kombinuje výsledek písemné a ústní části.  
Kterákoliv část hodnocená známkou *nevyhověl(a)*  $\Rightarrow$  *nevyhověl(a)*.
- Obě části v rozsahu předneseného + látky ze cvičení.
- Všechny termíny zkoušky ve **zkouškovém období zimního semestru** ( $\pm \varepsilon$ , kde  $\varepsilon \leq 14$  dnů).



## Matematická statistika

≈ metody pro analýzu dat včetně matematického zdůvodnění proč „fungují“ (obvykle s využitím teorie pravděpodobnosti).

- ▶ Lze přenést do stylu Definice → Věta → Důkaz.
- ▶ Lze do značné míry přenést do učebního textu, naučit se sám četbou tohoto textu a očekávat obdobný výsledek jako v jiných matematických předmětech.

## (Statistická) analýza dat

≈ snaha řešit problémy reálného světa pomocí dat (která jsou pryč všude kolem nás) aplikací metod matematické statistiky.

- ▶ Je to spíš řemeslo...  
které lze jen obtížně (resp. nejvýše fachiotsky) provozovat  
(s pomocí počítače) bez znalosti základů a hlavně pochopení principů.
- ▶ Dovednost nelze popsat učebním textem, lze ji pouze více či méně úspěšně předat.
- ▶ Dovednost získat lze pouze (opakovaným) vykonáváním příslušné činnosti...



- ▶ Teorii a řemeslo nelze oddělit  
(vykládat nejprve jedno a potom předávat druhé).
- ▶ Zkouška se bude snažit ověřit jak znalost teorie,  
tak (věku a zkušenostem přiměřenou) dovednost.

# 1

**Vybrané asymptotické výsledky**

## Oddíl 1.1

# Konvergence náhodných vektorů

## Definice 1.1 Konvergence skoro jistě.

Posloupnost  $\{\mathbf{X}_n\}_{n=1}^{\infty}$  konverguje **skoro jistě** k náhodnému vektoru  $\mathbf{X}$  pro  $n \rightarrow \infty$  tehdy a jen tehdy když

$$\mathbf{P}\left(\omega : \lim_{n \rightarrow \infty} \|\mathbf{X}_n(\omega) - \mathbf{X}(\omega)\| = 0\right) = 1.$$

Značíme  $\mathbf{X}_n \xrightarrow{\text{s.j.}} \mathbf{X}, n \rightarrow \infty$ .

## Definice 1.2 Konvergence v pravděpodobnosti.

Posloupnost  $\{\mathbf{X}_n\}_{n=1}^{\infty}$  konverguje **v pravděpodobnosti** k náhodnému vektoru  $\mathbf{X}$  pro  $n \rightarrow \infty$  tehdy a jen tehdy když

$$\forall \varepsilon > 0 \quad \lim_{n \rightarrow \infty} \mathbf{P}\left(\omega : \|\mathbf{X}_n(\omega) - \mathbf{X}(\omega)\| > \varepsilon\right) = 0.$$

Značíme  $\mathbf{X}_n \xrightarrow{\mathbf{P}} \mathbf{X}, n \rightarrow \infty$ .

### Definice 1.3 Konvergence v distribuci.

Posloupnost  $\{\mathbf{X}_n\}_{n=1}^{\infty}$  konverguje v distribuci k náhodnému vektoru  $\mathbf{X}$  pro  $n \rightarrow \infty$  tehdy a jen tehdy když

$$\lim_{n \rightarrow \infty} F_{\mathbf{X}_n}(\mathbf{x}) = F_{\mathbf{X}}(\mathbf{x})$$

v každém bodě  $\mathbf{x}$ , v němž je  $F_{\mathbf{X}}$  spojitá. Značíme  $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}, n \rightarrow \infty$ .

## Tvrzení 1.1 Vztahy mezi konvergencemi.

$$(i) \mathbf{X}_n \xrightarrow{s.j.} \mathbf{X}, n \rightarrow \infty \implies \mathbf{X}_n \xrightarrow{P} \mathbf{X}, n \rightarrow \infty;$$

$$(ii) \mathbf{X}_n \xrightarrow{P} \mathbf{X}, n \rightarrow \infty \implies \mathbf{X}_n \xrightarrow{D} \mathbf{X}, n \rightarrow \infty.$$

### **Poznámka.**

Opačně platí **POUZE** implikace  $\mathbf{X}_n \xrightarrow{D} \mathbf{c}, \mathbf{c} \in \mathbb{R}^k \implies \mathbf{X}_n \xrightarrow{P} \mathbf{c}.$

## Tvrzení 1.2 Věta o spojitě transformaci.

Nechť  $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \dots$  jsou náhodné vektory  
a funkce  $g : \mathbb{R}^k \rightarrow \mathbb{R}^m$  je spojitá na množině  $C$  takové, že  $P(\mathbf{X} \in C) = 1$ .  
Potom:

$$(i) \quad \mathbf{X}_n \xrightarrow{s.j.} \mathbf{X}, n \rightarrow \infty \implies g(\mathbf{X}_n) \xrightarrow{s.j.} g(\mathbf{X}), n \rightarrow \infty;$$

$$(ii) \quad \mathbf{X}_n \xrightarrow{P} \mathbf{X}, n \rightarrow \infty \implies g(\mathbf{X}_n) \xrightarrow{P} g(\mathbf{X}), n \rightarrow \infty;$$

$$(iii) \quad \mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}, n \rightarrow \infty \implies g(\mathbf{X}_n) \xrightarrow{\mathcal{D}} g(\mathbf{X}), n \rightarrow \infty.$$

### Tvrzení 1.3 Cramérova-Sluckého věta.

Nechť  $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ ,  $\mathbb{A}_n \xrightarrow{\mathcal{P}} \mathbb{A}$  a  $\mathbf{B}_n \xrightarrow{\mathcal{P}} \mathbf{b}$ ,  $n \rightarrow \infty$ ,  
kde  $\mathbf{X}_n$  a  $\mathbf{X}$  jsou  $k$ -rozměrné náhodné vektory,  $\mathbb{A}_n$  je náhodná matice o dimenzích  $m \times k$ ,  $\mathbb{A}$  je matice konstant o dimenzích  $m \times k$ ,  $\mathbf{B}_n$  jsou  $m$ -rozměrné náhodné vektory a  $\mathbf{b}$  je  $m$ -rozměrný vektor konstant, pak

$$\mathbb{A}_n \mathbf{X}_n + \mathbf{B}_n \xrightarrow{\mathcal{D}} \mathbb{A} \mathbf{X} + \mathbf{b}, \quad n \rightarrow \infty.$$



## Tvrzení 1.4 Postačující podmínka pro konzistenci.

Nechť

$$a_n(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \mathbf{X}, \quad n \rightarrow \infty,$$

kde  $a_n > 0$  je reálná posloupnost splňující  $a_n \rightarrow \infty, n \rightarrow \infty$  a  $\boldsymbol{\mu} \in \mathbb{R}^k$  je vektor konstant. Potom

$$\mathbf{X}_n \xrightarrow{\mathcal{P}} \boldsymbol{\mu}, \quad n \rightarrow \infty.$$

## Oddíl 1.2

# Zákon velkých čísel

## Tvrzení 1.5 Silný zákon velkých čísel.

Nechť  $\mathbf{X}_1, \mathbf{X}_2, \dots \stackrel{i.i.d.}{\sim} \mathbf{X}$ ,  $\mathbb{E}\mathbf{X} = \boldsymbol{\mu} \in \mathbb{R}^k$ . Potom

$$\overline{\mathbf{X}}_n \xrightarrow{s.j.} \boldsymbol{\mu}, \quad n \rightarrow \infty.$$

## Oddíl 1.3

# Centrální limitní věta

## Tvrzení 1.6 Centrální limitní věta pro i.i.d. náhodné vektory.

Nechť  $\mathbf{X}_1, \mathbf{X}_2, \dots \stackrel{i.i.d.}{\sim} \mathbf{X}$ ,  $\mathbb{E}\mathbf{X} = \boldsymbol{\mu} \in \mathbb{R}^k$ ,  $\text{var}\mathbf{X} = \boldsymbol{\Sigma}$ , kde  $\boldsymbol{\Sigma}$  je  $k \times k$  matice s konečnými prvky. Potom

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}) = \sqrt{n} (\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma}), \quad n \rightarrow \infty.$$

### Poznámky.

(i) Pokud  $\boldsymbol{\Sigma} > 0$  (pozitivně definitní matice), můžeme též psát

$$\frac{1}{\sqrt{n}} \boldsymbol{\Sigma}^{-1/2} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}) = \sqrt{n} \boldsymbol{\Sigma}^{-1/2} (\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k), \quad n \rightarrow \infty.$$

(ii) V jednorozměrném případě s  $\boldsymbol{\Sigma} = \sigma^2 > 0$ :

$$\frac{\sqrt{n} (\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty.$$

## Tvrzení 1.7 $\Delta$ -metoda.

Nechť náhodná posloupnost  $\{\mathbf{T}_n\}_{n=1}^{\infty}$  splňuje

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma}), \quad n \rightarrow \infty$$

pro nějaký vektor konstant  $\boldsymbol{\mu} \in \mathbb{R}^k$  a  $k \times k$  reálnou matici  $\boldsymbol{\Sigma}$ .

Nechť  $g : \mathbb{R}^k \rightarrow \mathbb{R}^p$  je funkce, která je *spojitě diferencovatelná* na nějakém okolí bodu  $\boldsymbol{\mu}$ . Označme

$$\mathbb{D}(\mathbf{x}) = \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}, \quad \mathbf{x} \in \mathbb{R}^k.$$

Potom

$$\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\mu})) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \mathbb{D}(\boldsymbol{\mu}) \boldsymbol{\Sigma} \mathbb{D}^{\top}(\boldsymbol{\mu})), \quad n \rightarrow \infty.$$

# 2

## Náhodný výběr

## Oddíl **2.1**

# **Definice náhodného výběru**



## Definice 2.1 Náhodný výběr.

Náhodná posloupnost  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  **nezávislých stejně rozdělených** (*independent identically distributed – i.i.d.*) náhodných vektorů definovaných na pravděpodobnostním prostoru  $(\Omega, \mathcal{A}, \mathbf{P})$ , kde všechny z nich mají distribuční funkci  $F_{\mathbf{X}}$  se nazývá **náhodný výběr z rozdělení  $F_{\mathbf{X}}$**  (*random sample from the distribution  $F_{\mathbf{X}}$* ). Konstanta  $n$  se nazývá **rozsah výběru** (*sample size*).

Značíme  $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{i.i.d.}}{\sim} \mathbf{X}, \quad \mathbf{X} \sim F_{\mathbf{X}}$

nebo stručněji  $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{i.i.d.}}{\sim} F_{\mathbf{X}}$ .

## Definice 2.2 (Statistický) model.

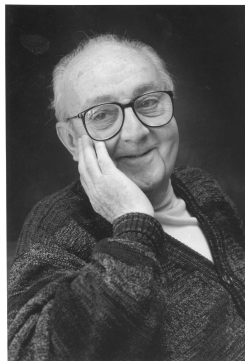
(Statistický) model (*statistical model*) pro náhodný výběr  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  je **prespecifikovaná** množina rozdělení  $\mathcal{F}$ , která obsahuje (neznámé) rozdělení  $F_{\mathbf{X}}$ .

**Essentially, all models are wrong, but some are useful. The practical question is how wrong do they have to be to not be useful.**

George E. P. Box

October 18, 1919, Gravesend,  
Kent, England

– March 28, 2013, Madison,  
Wisconsin, USA



## Oddíl **2.2**

# **Statistiky**

## Definice 2.3 Statistika.

Libovolná měřitelná funkce  $\mathbf{S}(\mathbf{X}_1, \dots, \mathbf{X}_n) = \mathbf{S}(\mathbb{X})$  náhodného výběru  $\mathbf{X}_1, \dots, \mathbf{X}_n \equiv \mathbb{X}$  se nazývá **statistika** (*statistic*).

## Definice 2.4 Výběrový průměr a výběrový rozptyl.

(i) Veličina  $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$

se nazývá **výběrový průměr** (*sample mean*) náhodného výběru  $\mathbb{X} \equiv X_1, \dots, X_n$ .

(ii) Pro  $n \geq 2$ , se veličina  $S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

nazývá **výběrový rozptyl** (*sample variance*) náhodného výběru  $\mathbb{X} \equiv X_1, \dots, X_n$ .

## Lemma 2.1 Výběrový průměr a nejmenší čtverce.

Platí:  $\bar{X}_n = \operatorname{argmin}_{c \in \mathbb{R}} \sum_{i=1}^n (X_i - c)^2$ .

## Věta 2.2 Statistické vlastnosti výběrového průměru.

Nechť  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X$ ,  $X \sim F_X \in \mathcal{F} = \mathcal{L}^2$   
s  $\mu := \mathbb{E} X \in \mathbb{R}$  a  $\sigma^2 := \operatorname{var} X \in (0, \infty)$ . Potom

- (i)  $\mathbb{E} \bar{X}_n = \mu$ ,  $\operatorname{var} \bar{X}_n = \frac{\sigma^2}{n}$ .
- (ii)  $\bar{X}_n \xrightarrow{P} \mu$ ,  $n \rightarrow \infty$ .
- (iii)  $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$ ,  $n \rightarrow \infty$ .

## Věta 2.3 Statistické vlastnosti (empirické) relativní četnosti.

Nechť  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X$ ,  $X \sim \mathcal{Alt}(p)$ ,  $p \in (0, 1)$ . Potom

$$(i) \mathbb{E}\bar{X}_n = p, \quad \text{var}\bar{X}_n = \frac{p(1-p)}{n}.$$

$$(ii) \bar{X}_n \xrightarrow{P} p, \quad n \rightarrow \infty.$$

$$(iii) \sqrt{n}(\bar{X}_n - p) \xrightarrow{\mathcal{D}} \mathcal{N}(0, p(1-p)), \quad n \rightarrow \infty.$$

$$(iv) n\bar{X}_n \sim \mathcal{Bi}(n, p).$$

### Lemma 2.4 Alternativní vyjádření výběrového rozptylu.

Pro  $n \geq 2$ :

$$(i) \ S_n^2 = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_n^2 \right).$$

(ii) Označme

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad \mathbb{A} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top = \begin{pmatrix} 1 - \frac{1}{n} & & -\frac{1}{n} \\ & \ddots & \\ -\frac{1}{n} & & 1 - \frac{1}{n} \end{pmatrix}.$$

Potom můžeme pro libovolné  $c \in \mathbb{R}$  psát

$$S_n^2 = \frac{1}{n-1} \mathbf{X}^\top \mathbb{A} \mathbf{X} = \frac{1}{n-1} \mathbf{Y}^\top \mathbb{A} \mathbf{Y},$$

kde  $\mathbf{Y} = \mathbf{X} - c \mathbf{1}_n$ .

Důkaz.

(i)

$$\begin{aligned}\frac{n-1}{n} S_n^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2X_i \bar{X}_n + \bar{X}_n^2) \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - 2\bar{X}_n^2 + \bar{X}_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2.\end{aligned}$$

(ii)

$$\begin{aligned}\mathbf{X}^\top \mathbb{A} \mathbf{X} &= \mathbf{X}^\top (\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top) \mathbf{X} = \mathbf{X}^\top \mathbf{X} - \frac{1}{n} \mathbf{X}^\top \mathbf{1}_n \mathbf{1}_n^\top \mathbf{X} \\ &= \sum_{i=1}^n X_i^2 - \frac{1}{n} \left( \sum_{i=1}^n X_i \right)^2 = \sum_{i=1}^n X_i^2 - n \bar{X}_n^2 = (n-1) S_n^2.\end{aligned}$$

Invariance  $S_n^2$  vůči posunutí plyne z faktu, že

$$\mathbb{A} \mathbb{A} = \mathbb{A} \quad \text{a} \quad \mathbb{A} \mathbf{1}_n = \mathbf{0}_n, \quad \mathbf{1}_n^\top \mathbb{A} = \mathbf{0}_n^\top.$$





## Lemma 2.5 Střední hodnota kvadratické formy.

Nechť  $\mathbf{Z}$  je  $n$ -rozměrný náhodný vektor s konečnou střední hodnotou  $\boldsymbol{\mu} = \mathbb{E}\mathbf{Z}$  a konečnou varianční maticí  $\boldsymbol{\Sigma} = \text{var}\mathbf{Z}$ . Nechť  $\mathbb{B}$  je libovolná  $n \times n$  matice. Potom platí:

$$\mathbb{E}(\mathbf{Z}^\top \mathbb{B} \mathbf{Z}) = \boldsymbol{\mu}^\top \mathbb{B} \boldsymbol{\mu} + \text{tr}(\mathbb{B} \boldsymbol{\Sigma}).$$

## Věta 2.6 Statistické vlastnosti výběrového rozptylu.

Nechť  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X$ ,  $X \sim F_X \in \mathcal{F} = \mathcal{L}^2$   
s  $\mu := \mathbb{E} X \in \mathbb{R}$  a  $\sigma^2 := \text{var} X \in (0, \infty)$ . Potom

(i)  $S_n^2 \xrightarrow{P} \sigma^2$ ,  $n \rightarrow \infty$ .

(ii)  $\mathbb{E} S_n^2 = \sigma^2$ .

(iii) Pokud navíc  $\mathcal{F} = \mathcal{L}^4$ , tj. existuje  $\mathbb{E} X^4 < \infty$ , potom

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^4(\gamma_4 - 1)), \quad n \rightarrow \infty,$$

kde  $\gamma_4 = \frac{\mathbb{E}(X - \mu)^4}{\sigma^4}$  je **špičatost (kurtosis)** rozdělení  $X$ .

## Věta 2.6 Statistické vlastnosti výběrového rozptylu, pokrač.

(iv) Pokud  $\mathcal{F} = \mathcal{L}^4$ , potom také

$$\sqrt{n} \left[ \begin{pmatrix} \bar{X}_n \\ S_n^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \right] \xrightarrow{\mathcal{D}} \mathcal{N}_2(\mathbf{0}_2, \mathbf{\Sigma}), \quad n \rightarrow \infty,$$

kde

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & \sigma^3 \gamma_3 \\ \sigma^3 \gamma_3 & \sigma^4 (\gamma_4 - 1) \end{pmatrix},$$

$$\gamma_3 = \frac{\mathbb{E}(X - \mu)^3}{\sigma^3} \quad \text{je šikmost (skewness) rozdělení } X.$$

## Lemma 2.7 Nezávislost lineární a kvadratické transformace za normality.

Nechť  $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  a  $\mathbf{A}$  je pozitivně semidefinitní matice velikosti  $n \times n$ .

(i) Nechť  $\mathbf{B}$  je libovolná matice velikosti  $m \times n$  splňující rovnost

$$\mathbf{B} \boldsymbol{\Sigma} \mathbf{A} = \mathbf{0}_{m \times n}.$$

Potom jsou náhodná veličina  $\mathbf{X}^\top \mathbf{A} \mathbf{X}$  a náhodný vektor  $\mathbf{B} \mathbf{X}$  *nezávislé*.

(ii) Nechť  $\mathbf{B}$  je libovolná pozitivně semidefinitní matice velikosti  $n \times n$  splňující rovnost

$$\mathbf{B} \boldsymbol{\Sigma} \mathbf{A} = \mathbf{0}_{n \times n}.$$

Potom jsou náhodné veličiny  $\mathbf{X}^\top \mathbf{A} \mathbf{X}$  a  $\mathbf{X}^\top \mathbf{B} \mathbf{X}$  *nezávislé*.

## Věta 2.8 Vlastnosti výběrového rozptylu za normality.

Nechť  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X$ ,  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Potom

(i)  $\frac{(n-1) S_n^2}{\sigma^2} \sim \chi_{n-1}^2$ .

(ii)  $\bar{X}_n$  a  $S_n^2$  jsou *nezávislé* náhodné veličiny.

## Lemma A.4

Nechť  $\mathbf{X} \sim \mathcal{N}_n(\mathbf{0}_n, \mathbf{\Sigma})$  a  $\mathbb{A}$  je pozitivně semidefinitní matice velikosti  $n \times n$  taková, že  $\mathbb{A} \mathbf{\Sigma}$  je *nenulová* a *idempotentní*. Potom

$$\mathbf{X}^\top \mathbb{A} \mathbf{X} \sim \chi_{\text{tr}(\mathbb{A} \mathbf{\Sigma})}^2.$$

## Věta 2.9 Asymptotické rozdělení $T$ statistiky.

Nechť  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X$ ,  $X \sim F_X \in \mathcal{F} = \mathcal{L}^2$

s  $\mu := \mathbb{E}X \in \mathbb{R}$  a  $\sigma^2 := \text{var}X \in (0, \infty)$ . Potom

$$T_n := \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty.$$

## Věta 2.10 Rozdělení $T$ statistiky za normality.

Nechť  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X$ ,  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Potom

$$T_n := \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t_{n-1}.$$

## Definice 2.5 F-rozdělení.

Necht'  $X \sim \chi_n^2$  a  $Y \sim \chi_m^2$  jsou **nezávislé**. Rozdělení náhodné veličiny

$$Z := \frac{X/n}{Y/m}$$

se nazývá [Fisherovo-Snedecorovo] **F-rozdělení s  $n$  a  $m$  stupni volnosti** (*degrees of freedom*).

Značíme  $Z \sim \mathcal{F}_{n,m}$ .

## Věta 2.11 O $F$ statistice.

Nechť  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X$ ,  $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$

a  $Y_1, \dots, Y_m \stackrel{i.i.d.}{\sim} Y$ ,  $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ .

Nechť dále jsou náhodné vektory  $(X_1, \dots, X_n)^\top$  a  $(Y_1, \dots, Y_m)^\top$   
*nezávislé*. Nechť

$$\bar{X}_n = \sum_{i=1}^n X_i, \quad S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

$$\bar{Y}_m = \sum_{j=1}^m Y_j, \quad S_Y^2 = \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y}_m)^2.$$

Potom platí

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim \mathcal{F}_{n-1, m-1}.$$



## Oddíl **2.3**

# **Uspořádaný náhodný výběr**

## Definice 2.6 Uspořádaný náhodný výběr.

Nechť  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X$ ,  $X \sim F_X$ ,  $F_X \in \mathcal{F} = \{\text{jednorozměrná spojitá rozdělení s distribuční funkcí } F \text{ a hustotou } f\}$ .

- (i) Seřazením náhodných veličin  $X_1, \dots, X_n$  (jejich realizací) od nejmenší po největší získáme **uspořádaný náhodný výběr** (*ordered random sample*):

$$X_{(1)} < X_{(2)} < \dots < X_{(n-1)} < X_{(n)}.$$

Symbolem  $X_{(k)}$  rozumíme  $k$ -tou nejmenší hodnotu mezi  $X_1, \dots, X_n$ . Náhodná veličina  $X_{(k)}$  se nazývá  **$k$ -tá pořádková statistika** (*order statistic*).

- (ii) **Pořadím** (*rank*) náhodné veličiny  $X_i$  ve výběru  $X_1, \dots, X_n$  rozumíme přirozené číslo  $R_i \in \{1, \dots, n\}$  takové, že

$$X_i = X_{(R_i)}.$$

Vektor pořádkových statistik (*vector of order statistics*, celý uspořádaný výběr) budeme značit  $\mathbf{X}_{(\bullet)}$ , to jest,

$$\mathbf{X}_{(\bullet)} = (X_{(1)}, \dots, X_{(n)})^\top.$$

## Věta 2.12 Sdružená hustota uspořádaného výběru.

Náhodný vektor  $\mathbf{X}_{(\bullet)} = (X_{(1)}, \dots, X_{(n)})^\top$  má následující hustotu (vzhledem k Lebesgueově míře)

$$p(y_1, \dots, y_n) = \begin{cases} n! f(y_1) \cdot f(y_2) \cdots f(y_n), & \text{pokud } y_1 < \cdots < y_n, \\ 0, & \text{jinak.} \end{cases}$$

### Věta 2.13 Distribuční funkce $k$ -té pořádkové statistiky.

*Distribuční funkce  $k$ -té pořádkové statistiky jest*

$$\begin{aligned} F_{(k)}(x) &= P(X_{(k)} \leq x) = \sum_{j=k}^n \binom{n}{j} F^j(x) \{1 - F(x)\}^{n-j} \\ &= \frac{1}{B(k, n - k + 1)} \int_0^{F(x)} t^{k-1} (1 - t)^{n-k} dt. \end{aligned}$$

### Věta 2.14 Hustota $k$ -té pořádkové statistiky.

*Hustota (vzhledem k Lebesgueově míře)  $k$ -té pořádkové statistiky jest*

$$f_{(k)}(x) = n \binom{n-1}{k-1} f(x) F^{k-1}(x) \{1 - F(x)\}^{n-k}.$$

## Věta 2.15 Rozdělení vektoru pořadí.

Náhodný vektor  $\mathbf{R} = (R_1, \dots, R_n)^\top$  pořadí má rovnoměrné rozdělení na množině  $\mathcal{P}_n$  vech permutací posloupnosti  $(1, \dots, n)$ . To jest,

$$P(\mathbf{R} = \mathbf{r}) = \frac{1}{n!}, \quad \mathbf{r} \in \mathcal{P}_n.$$

## Věta 2.16 Vlastnosti pořadí.

Platí

- (i)  $P(R_i = k) = \frac{1}{n}$  pro každé  $i, k \in \{1, \dots, n\}$ .
- (ii)  $P(R_i = k, R_j = m) = \frac{1}{n(n-1)}$  pro každé  $i \neq j, k \neq m \in \{1, \dots, n\}$ .
- (iii)  $\mathbb{E} R_i = \frac{n+1}{2}, \quad \text{var } R_i = \frac{n^2 - 1}{12}$  pro každé  $i \in \{1, \dots, n\}$ .
- (iv)  $\text{cov}(R_i, R_j) = -\frac{n+1}{12}$  pro každé  $i \neq j \in \{1, \dots, n\}$ .

## Oddíl **2.4**

# **Transformace ve statistice**

## Nastavení

- ▶  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X, \quad \text{hustota: } f_X, \quad \text{nosič: } S_X;$
- ▶  $g: S_X \rightarrow \mathbb{R},$  **ryze monotónní** a **diferencovatelná**;
- ▶  $Y_i := g(X_i), i = 1, \dots, n \quad \equiv \quad \text{transformovaný náhodný výběr.}$

## Dostaneme

- ▶  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} Y, \quad Y \sim F_Y, \quad \text{hustota: } f_Y;$
- ▶ Pokud  $F_X$  spojitá a  $f_X$  známá, **věta o transformaci** poskytuje předpis  $f_Y$ .

## Transformace stabilizující (asymptotický) rozptyl

- ▶ Necht'  $\sqrt{n}(T_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(\mu))$ ,  $n \rightarrow \infty$ .
- ▶ Pokud  $g \equiv$  reálná funkce spojitě diferencovatelná na okolí bodu  $\mu$ , potom ( $\Delta$ -metoda):

$$\sqrt{n}(g(T_n) - g(\mu)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \{g'(\mu)\}^2 \sigma^2(\mu)\right).$$

- ▶ Zvolme  $g(x) = c \int \frac{1}{\sigma(x)} dx$   
 $\Rightarrow g'(\mu) = \frac{c}{\sigma(\mu)}$   
 $\Rightarrow \sqrt{n}(g(T_n) - g(\mu)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c^2), \quad n \rightarrow \infty.$



## Standardizace

►  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X$ ,  $X \sim F_X$ , rozdělení s  $0 < \text{var}X < \infty$ ;

► Z-skóry:

$$Z_i := \frac{X_i - \bar{X}_n}{S_n}, \quad i = 1, \dots, n.$$

► Výběrový průměr Z-skórů je 0.

► Výběrový rozptyl Z-skórů je 1.

► Z-skóry  $Z_1, \dots, Z_n$  nejsou nezávislé.

► Protože platí  $\bar{X}_n \xrightarrow{P} \mathbb{E} X$ ,

$$S_n \xrightarrow{P} \sqrt{\text{var} X}, \quad n \rightarrow \infty.$$

chovají se náhodné veličiny  $Z_1, \dots, Z_n$ , pro velké  $n$ , skoro jako náhodný výběr z rozdělení s nulovou střední hodnotou a jednotkovým rozptylem.

# 3

## **Odhadování parametrů**

## Nastavení/předpoklady

- ▶  $\underbrace{X_1, \dots, X_n}_{\mathbf{X}} \stackrel{\text{i.i.d.}}{\sim} \mathbf{X}, \quad \mathbf{X} \sim F_X \in \mathcal{F} \text{ (model);}$
- ▶  $\theta = \mathbf{t}(F) \in \mathbb{R}^p, F \in \mathcal{F}$   
 $\equiv$  parametr, který za platnosti předpokládaného modelu chceme odhadnout.
- ▶ Necht'  $F_X \in \mathcal{F} \quad \equiv$  skutečné rozdělení náhodného vektoru  $\mathbf{X}$ .
- ▶  $\theta_X := \mathbf{t}(F_X) \quad \equiv$  skutečná hodnota parametru zájmu.

## Oddíl **3.1**

# **Bodový odhad**

### Definice 3.1 Bodový odhad.

Bodovým odhadem (*point estimate*) parametru  $\theta_X = t(F_X) \in \mathbb{R}^p$  rozumíme  $p$ -rozměrný náhodný vektor  $\hat{\theta}_n$ , který spočteme jako

$$\hat{\theta}_n = T_n(\mathbb{X}) = T_n(\mathbf{X}_1, \dots, \mathbf{X}_n),$$

kde  $T_n$  je nějaká Borelovsky měřitelná funkce dat.

### Definice 3.2 Nestrannost a konzistence.

Mějme náhodný výběr  $\mathbb{X} \equiv (\mathbf{X}_1, \dots, \mathbf{X}_n)$  z rozdělení  $F_X \in \mathcal{F}$  a (bodový) odhad  $\hat{\theta}_n = \mathbf{T}_n(\mathbb{X})$  parametru  $\theta_X = \mathbf{t}(F_X)$ .

- (i) Řekneme, že odhad  $\hat{\theta}_n$  je **nestranný** (*unbiased*) odhad parametru  $\theta_X$  v modelu  $\mathcal{F}$ , právě když

$$\mathbb{E}_{F_X} \hat{\theta}_n = \theta_X$$

pro každé  $n$  (pro něž je odhad definován) a pro každé  $F_X \in \mathcal{F}$ .

- (ii) Řekneme, že odhad  $\hat{\theta}_n$  je **(slabě) konzistentní** (*(weakly) consistent*) odhad parametru  $\theta_X$  v modelu  $\mathcal{F}$ , právě když

$$\hat{\theta}_n \xrightarrow{\text{P}} \theta_X, \quad n \rightarrow \infty$$

pro každé  $F_X \in \mathcal{F}$ .

### Definice 3.3 Vychýlení.

Nechť odhad  $\hat{\theta}_n = T_n(\mathbb{X})$  parametru  $\theta_X$  má konečnou střední hodnotu (pro každé  $F_X \in \mathcal{F}$ ). Rozdíl

$$\text{bias}(\hat{\theta}_n) := \mathbb{E}_{F_X}(\hat{\theta}_n - \theta_X)$$

nazýváme **vychýlením** (*bias*) odhadu  $\hat{\theta}_n$ .

### Definice 3.4 Střední čtvercová a směrodatná chyba.

Nechť odhad  $\hat{\theta}_n = T_n(\mathbb{X})$  jednorozměrného parametru  $\theta_X$  má konečný rozptyl (pro každé  $F_X \in \mathcal{F}$ ).

(i) Výraz  $\text{MSE}(\hat{\theta}_n) := \mathbb{E}_{F_X}(\hat{\theta}_n - \theta_X)^2$

nazýváme **střední čtvercovou chybou** (*mean square error*) odhadu  $\hat{\theta}_n$ .

(ii) Výraz  $\text{S.E.}(\hat{\theta}_n) := \sqrt{\text{var}_{F_X}(\hat{\theta}_n)}$

nazýváme **směrodatnou chybou** (*standard error*) odhadu  $\hat{\theta}_n$ .

### Věta 3.1 Postačující podmínka konzistence odhadu.

Nechť  $\hat{\theta}_n$  je odhad jednorozměrného parametru  $\theta_X \in \mathbb{R}$ , pro nějž platí, pro všechna  $F_X \in \mathcal{F}$ :

$$\lim_{n \rightarrow \infty} \mathbb{E}_{F_X}(\hat{\theta}_n) = \theta_X \quad \& \quad \lim_{n \rightarrow \infty} \text{var}_{F_X}(\hat{\theta}_n) = 0.$$

Potom je  $\hat{\theta}_n$  (slabě) konzistentním odhadem parametru  $\theta_X$ .

#### Poznámka.

- ▶ Opačná implikace neplatí!
- ▶ Existují odhady, které jsou konzistentní  
a současně pro každé  $F_X \in \mathcal{F}$  a každé  $n \geq 1$  je  $\mathbb{E}_{F_X}|\hat{\theta}_n| = \infty$ .

- ▶ **Příklad:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Alt}(p_X)$ ,  $0 < p_X < 1$ .

Uvažme  $\hat{\theta}_n = (\bar{X}_n)^{-1}$  jako odhad parametru  $\theta_X := p_X^{-1}$ .

Ukažte (pro každé  $p_X \in (0, 1)$ )  $\mathbb{E}_{p_X} \hat{\theta}_n = \infty$ ,  $\hat{\theta}_n \xrightarrow{P} \theta_X$ ,  $n \rightarrow \infty$ .



## Oddíl **3.2**

# **Volba parametru zájmu**

## KVANTITATIVNÍ data

- ▶ hodnoty mají konkrétní **numerický** význam a často též dané jednotky (fyzikální, peněžní, ...)  
(počet, procento, délka, objem, hmotnost, úrok. míra, koncentrace, energie, teplota, doba trvání, velikost úhlu, kalendářní rok, ...)
- ▶ existuje smysluplné **uspořádání** jejich hodnot
- ▶ **rozdíly** jejich hodnot mají reálnou interpretaci
- ▶ Další (méně podstatné) dělení:
  1. **poměrové** veličiny: typicky nezáporné s jasně definovanou nulovou hodnotou a interpretovatelnými podíly
  2. **intervalové** veličiny: nemají pevně definovanou nulu a nemají interpretovatelné podíly
- ▶ **Matematická reprezentace:** **spojité** i **diskrétní** náhodné veličiny
- ▶ **Parametr zájmu:**  $\mathbb{E} X$ ,  $\text{var } X$ ,  $F_X(x)$ ,  $F_X^{-1}(\alpha)$ , ...

## KATEGORIÁLNÍ data

- ▶ hodnoty „pouze“ **kódují** příslušnost do určité kategorie, skupiny
- ▶ **Matematická reprezentace:** **diskrétní** náhodná veličina s **konečným** nosičem  $\{\omega_1, \dots, \omega_K\}$ ,  $\omega_1 < \dots < \omega_K \equiv$  „nálepky“ jednotlivých kategorií
- ▶ Další dělení:
  1. **nominální** veličiny: neexistuje smysluplné uspořádání jejich kategorií  
(kraj bydliště: 1 = *Praha*, 2 = *SČ kraj*, ..., 14 = *Zlínský kraj*)

**Parametr zájmu:** pouze  $P(X = \omega_k)$ ,  $k = 1, \dots, K$  má smysl.

Veličiny typu  $\mathbb{E} X$ ,  $\text{var } X$ ,  $F_X(x)$ ,  $F_X^{-1}(\alpha)$ , ...

nejsou prakticky interpretovatelné,

i když matematicky jsou správně definovány.

## KATEGORIÁLNÍ data

- ▶ hodnoty „pouze“ **kódují** příslušnost do určité kategorie, skupiny
- ▶ **Matematická reprezentace:** **diskrétní** náhodná veličina s **konečným** nosičem  $\{\omega_1, \dots, \omega_K\}$ ,  $\omega_1 < \dots < \omega_K \equiv$  „nálepky“ jednotlivých kategorií
- ▶ Další dělení:
  - 2. **ordinální** veličiny: kategorie lze smyslupně **uspořádat**

(známka ve škole: 1, 2, 3, 4, 5,  
spokojenost s...:  $-2 = \text{velmi nespokojen}$ ,  $-1 = \text{nespokojen}$ ,  $0 = \text{neutrální}$ ,  $1 = \text{spokojen}$ ,  $2 = \text{velmi spokojen}$ )

**Parametr zájmu:** kromě  $P(X = \omega_k)$ ,  $k = 1, \dots, K$  má smysl též

$$F_X(x) = P(X \leq x), x \in \{\omega_1, \dots, \omega_K\},$$

někdy i další, např.  $\mathbb{E} X$

## BINÁRNÍ data

- ▶ pouze dvě možné hodnoty, speciální případ kategoriálních dat
- ▶ hodnoty výhodné kódovat jako 0 a 1

⇒ **Matematická reprezentace:**

náhodná veličina s **alternativním** rozdělením  $\mathcal{Alt}(p_X)$ ,  $0 < p_X < 1$

- ▶ **Parametr zájmu:**  $p_X = P(X = 1) = \mathbb{E} X$

## Statistické metody

- ▶ Pro **kvantitativní** data
  - pracují s charakteristikami jako  $\mathbb{E} X$ ,  $\text{var } X$ ,  $F_X(x)$ ,  $F_X^{-1}(\alpha)$ , ...
  - částečně využitelné též s kategoriálními **ordinálními** daty
- ▶ Pro **kategoriální** data
  - pracují s pravděpodobnostmi jednotlivých kategorií

## Bodový odhad parametru, základní obecné metody

- ▶ Metoda maximální věrohodnosti (*maximum likelihood – ML*)  
viz NMSA332 *Matematická statistika 2* v LS
- ▶ Momentová metoda

## Oddíl 3.3

# Momentová metoda



## Nastavení/předpoklady (parametrický model, jednorozměrné rozdělení)

$$\underbrace{X_1, \dots, X_n}_X \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X \in \mathcal{F},$$

$\mathcal{F} = \{\text{rozdělení s hustotou } f(x; \theta), \theta \in \Theta \subseteq \mathbb{R}^p\},$   
hustota  $f(\cdot; \theta)$  vzhledem k  $\sigma$ -konečné míře, tvar  $f(\cdot; \cdot)$  známý.

- Rozdělení  $F_X$  má hustotu  $f(x; \theta_X)$ .

**CÍL:** odhad parametru  $\theta_X$ .

**PŘEDPOKLAD:**  $\mathbb{E}_F |X|^p < \infty$ , pro každé  $F \in \mathcal{F}$ .

- 
- Máme **konzistentní** odhady momentů:

$$\text{pro každé } F \in \mathcal{F} \quad \frac{1}{n} \sum_{i=1}^n X_i^q \xrightarrow{P} \mathbb{E}_F X^q, \quad n \rightarrow \infty, \quad q = 1, \dots, p.$$

- Platí:  $\mathbb{E}_F X^q, q = 1, \dots, p$  je funkcí  $\theta$   
(konst. funkce je možnou komplikací).

## Oddíl **3.4**

# **Intervalový odhad**

## Nastavení/předpoklady (obecný (neparametrický) model, jednorozměrný parametr)

$$\triangleright \underbrace{\mathbf{X}_1, \dots, \mathbf{X}_n}_{\mathbb{X}} \stackrel{\text{i.i.d.}}{\sim} \mathbf{X}, \quad \mathbf{X} \sim F_X \in \mathcal{F} \text{ (model)}$$

$\mathcal{F} = \{\text{libovolné rozdělení na } \mathbb{R}^p\} \equiv \text{neparametrický model}$

$$\triangleright \theta = t(F) \in \mathbb{R}, F \in \mathcal{F} \\ \equiv \text{parametr, který za platnosti předpokládaného modelu} \\ \text{chceme odhadnout.}$$

$$\triangleright \text{Necht' } F_X \in \mathcal{F} \equiv \text{skutečné rozdělení náhodného vektoru } \mathbf{X}.$$

$$\triangleright \theta_X := t(F_X) \equiv \text{skutečná hodnota parametru zájmu.}$$

### Definice 3.5 Intervalový odhad.

Interval  $B_n = B_n(\mathbb{X}) \subset \mathbb{R}$  se nazývá **intervalový odhad** (*interval estimate*) parametru  $\theta_X \in \mathbb{R}$  o **spolehlivosti** (*confidence*)  $1 - \alpha$  v modelu  $\mathcal{F}$ , právě když

$$\underbrace{P_{F_X} \left[ \omega \in \Omega : B_n(\mathbb{X}(\omega)) \ni \theta_X \right]}_{P_{F_X}(B_n(\mathbb{X}) \ni \theta_X)} = 1 - \alpha, \quad \text{pro každé rozdělení } F_X \in \mathcal{F}.$$

Interval  $B_n = B_n(\mathbb{X}) \subset \mathbb{R}$  se nazývá **asymptotický intervalový odhad** parametru  $\theta_X \in \mathbb{R}$  o **(přibližné) spolehlivosti**  $1 - \alpha$  v modelu  $\mathcal{F}$ , právě když

$$\lim_{n \rightarrow \infty} P_{F_X}(B_n(\mathbb{X}) \ni \theta_X) = 1 - \alpha, \quad \text{pro každé rozdělení } F_X \in \mathcal{F}.$$

#### Alternativní názvy a poznámky

- ▶ Interval spolehlivosti/konfidenční interval s pravděpodobností pokrytí/o spolehlivosti  $1 - \alpha$  (*confidence interval with a coverage probability/confidence level  $1 - \alpha$* )
- ▶  $(1 - \alpha)$  100% interval spolehlivosti/konfidenční interval
- ▶ Číslo  $\alpha$  je předem zvolené, nejběžnější volba:  $\alpha = 0,05 \rightarrow 95\%$  intervaly spolehlivosti

## Oboustranný (*two-sided*) interval spolehlivosti

Interval tvaru  $(\eta_L(\mathbb{X}), \eta_U(\mathbb{X}))$ ,

$\eta_L(\mathbb{X}), \eta_U(\mathbb{X})$  dvě náhodné veličiny (statistiky) splňující (pro každé  $F_X \in \mathcal{F}$ )

$$P_{F_X} \left( (\eta_L(\mathbb{X}), \eta_U(\mathbb{X})) \ni \theta_X \right) = 1 - \alpha,$$

$$P_{F_X}(\eta_L(\mathbb{X}) < \eta_U(\mathbb{X})) = P_{F_X}(\eta_L(\mathbb{X}) > -\infty) = P_{F_X}(\eta_U(\mathbb{X}) < \infty) = 1.$$

Obvykle konstruován, aby platilo (alespoň asymptoticky)

$$P_{F_X}(\eta_L(\mathbb{X}) \geq \theta_X) = \frac{\alpha}{2} = P_{F_X}(\eta_U(\mathbb{X}) \leq \theta_X).$$

 software: `alternative = "two.sided"`

## Jednostranný (*one-sided*) interval spolehlivosti

- ▶ Levostranný (dolní) interval spolehlivosti:  $(\eta_L(\mathbb{X}), \infty)$

$$P_{F_X}(\eta_L(\mathbb{X}) < \theta_X) = 1 - \alpha, \quad \text{pro každé } F_X \in \mathcal{F}.$$

 software: `alternative = "greater"`

- ▶ Pravostranný (horní) interval spolehlivosti:  $(-\infty, \eta_U(\mathbb{X}))$

$$P_{F_X}(\eta_U(\mathbb{X}) > \theta_X) = 1 - \alpha, \quad \text{pro každé } F_X \in \mathcal{F}.$$

 software: `alternative = "less"`

## Vícerozměrně pro $\theta_X \in \mathbb{R}^p$

- ▶ Oblast spolehlivosti/konfidenční oblast, množina (*confidence region, set*)

$$B_n(\mathbb{X}) \subset \mathbb{R}^p:$$

$$P_{F_X}(B_n(\mathbb{X}) \ni \theta_X) = 1 - \alpha, \quad \text{pro každé } F_X \in \mathcal{F}.$$

## Příklad. Odhad střední hodnoty normálního rozdělení

**Data:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

**Model:**  $F_X \in \mathcal{F} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$

**Odhadovaný parametr:**  $\theta_X = \mathbb{E}_{F_X} X =: \mu_X$

$$\begin{aligned} \left( \bar{X}_n - t_{n-1} \left( 1 - \frac{\alpha}{2} \right) \frac{S_n}{\sqrt{n}}, \quad \bar{X}_n + t_{n-1} \left( 1 - \frac{\alpha}{2} \right) \frac{S_n}{\sqrt{n}} \right) \\ \equiv \left( \bar{X}_n \mp t_{n-1} \left( 1 - \frac{\alpha}{2} \right) \frac{S_n}{\sqrt{n}} \right) \end{aligned}$$

## Příklad. Odhad střední hodnoty v libovolném $\mathcal{L}^2$ rozdělení

**Data:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

**Model:**  $F_X \in \mathcal{F} = \mathcal{L}^2$  (rozdělení s konečným rozptylem)

**Odhadovaný parametr:**  $\theta_X = \mathbb{E}_{F_X} X =: \mu_X$

$$\left( \bar{X}_n - u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}, \quad \bar{X}_n + u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \right) \equiv \left( \bar{X}_n \mp u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \right)$$



## Vybrané kvantily (pro oboustranné IS s pokrytím 90 %, resp. 95 %)

| $\kappa$          | 0,95 | 0,975 |                         |
|-------------------|------|-------|-------------------------|
| $u_{\kappa}$      | 1,64 | 1,96  | <code>qnorm(...)</code> |
| $t_{500}(\kappa)$ | 1,65 | 1,96  | <code>qt(...)</code>    |
| $t_{100}(\kappa)$ | 1,66 | 1,98  |                         |
| $t_{50}(\kappa)$  | 1,68 | 2,01  |                         |
| $t_{10}(\kappa)$  | 1,81 | 2,23  |                         |
| $t_5(\kappa)$     | 2,02 | 2,57  |                         |
| $t_2(\kappa)$     | 2,92 | 4,30  |                         |
| $t_1(\kappa)$     | 6,31 | 12,71 |                         |

## Příklad. Odhad rozptylu normálního rozdělení

**Data:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

**Model:**  $F_X \in \mathcal{F} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$

**Odhadovaný parametr:**  $\theta_X = \text{var}_{F_X} X =: \sigma_X^2$

$$\left( \frac{(n-1) S_n^2}{\chi_{n-1}^2 (1 - \frac{\alpha}{2})}, \frac{(n-1) S_n^2}{\chi_{n-1}^2 (\frac{\alpha}{2})} \right)$$

### Lemma 3.2 Interval spolehlivosti po transformaci parametru.

Je-li  $(\eta_L, \eta_U)$  (asymptotický) interval spolehlivosti pro parametr  $\theta_X$  s pravděpodobností pokrytí  $1 - \alpha$  a je-li  $\psi$  **rostoucí** reálná funkce na parametrickém prostoru  $\Theta = \{t(F), F \in \mathcal{F}\} \subseteq \mathbb{R}$ , pak  $(\psi(\eta_L), \psi(\eta_U))$  je (asymptotický) interval spolehlivosti pro parametr  $\psi(\theta_X)$  s pravděpodobností pokrytí  $1 - \alpha$ .

#### Příklad. Odhad směrodatné odchylky normálního rozdělení

**Data:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

**Model:**  $F_X \in \mathcal{F} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$

**Odhadovaný parametr:**  $\theta_X = \sqrt{\text{var}_{F_X} X} =: \sigma_X$

$$\left( \sqrt{\frac{(n-1) S_n^2}{\chi_{n-1}^2(1 - \frac{\alpha}{2})}}, \sqrt{\frac{(n-1) S_n^2}{\chi_{n-1}^2(\frac{\alpha}{2})}} \right)$$

## Příklad. Odhad parametru Poissonova rozdělení

**Data:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X \equiv \mathcal{Po}(\lambda_X), \lambda_X > 0$

**Model:**  $F_X \in \mathcal{F} = \{\mathcal{Po}(\lambda), \lambda > 0\}$

**Odhadovaný parametr:**  $\theta_X = \lambda_X = \mathbb{E}_{F_X} X = \text{var}_{F_X} X$

$$\left( \left[ \max \left\{ 0, \sqrt{\bar{X}_n} - \frac{u_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \right\} \right]^2, \left( \sqrt{\bar{X}_n} + \frac{u_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \right)^2 \right)$$

## Příklad. Odhad proporce

**Data a model:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim \text{Alt}(p_X), 0 < p_X < 1$

**Odhadovaný parametr:**  $\theta_X = p_X = \mathbb{E}_{F_X} X = \mathbb{P}_{F_X}(X = 1)$

**(Domácí) úkol.** Využijte asymptotiku pro  $\hat{p}_n = \bar{X}_n$

$$\sqrt{n}(\hat{p}_n - p_X) \xrightarrow{\mathcal{D}} \mathcal{N}(0, p_X(1 - p_X)), \quad \hat{p}_n \xrightarrow{\mathbb{P}} p_X, \quad n \rightarrow \infty$$

k odvození asymptotického intervalu spolehlivosti pro  $p_X$ .

$$\begin{aligned} \left( \hat{p}_n - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \quad \hat{p}_n + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right) \\ \equiv \left( \hat{p}_n \mp u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right) \end{aligned}$$

## Oddíl **3.5**

# **Empirické odhady**

## Nastavení/předpoklady (obecný (neparametrický) model, jednorozměrné rozdělení)

$$\text{▶ } \underbrace{X_1, \dots, X_n}_{\mathbb{X}} \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X \in \mathcal{F} \text{ (model)}$$

$\mathcal{F} = \{\text{libovolné rozdělení na } \mathbb{R}\} \equiv \text{neparametrický model}$

$$\text{▶ } \theta = \mathbf{t}(F) \in \mathbb{R}^q, F \in \mathcal{F} \\ \equiv \text{parametr, který chceme odhadnout} \\ \text{(vybrané charakteristiky rozdělení } F).$$

▶ Necht'  $F_X \in \mathcal{F} \equiv$  skutečné rozdělení náhodné veličiny  $X$ .

▶  $\theta_X := \mathbf{t}(F_X) \equiv$  skutečná hodnota parametru zájmu.

## Definice 3.6 Empirická distribuční funkce.

Funkci

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \leq x), \quad x \in \mathbb{R}$$

nazýváme **empirická distribuční funkce**

(*empirical cumulative distribution function – ecdf*)

náhodného výběru  $X_1, \dots, X_n$ .



### Věta 3.3 Vlastnosti empirické distribuční funkce.

Pro libovolné  $x \in \mathbb{R}$  a pro všechna  $F_X \in \mathcal{F}$  platí

- (i)  $\mathbb{E}_{F_X} \hat{F}_n(x) = F_X(x)$  (*neustrannost*),  $\text{var}_{F_X} \hat{F}_n(x) = \frac{F_X(x) \{1 - F_X(x)\}}{n}$ ;
- (ii)  $\hat{F}_n(x) \xrightarrow{P} F_X(x)$ ,  $n \rightarrow \infty$  (*bodová konzistence*);
- (iii)  $\sqrt{n} \{ \hat{F}_n(x) - F_X(x) \} \xrightarrow{\mathcal{D}} \mathcal{N}(0, F_X(x) \{1 - F_X(x)\})$ ,  $n \rightarrow \infty$ ;
- (iv)  $n \hat{F}_n(x) \sim \text{Bi}(n, F_X(x))$ ;
- (v)  $\sup_{x \in \mathbb{R}} | \hat{F}_n(x) - F_X(x) | \xrightarrow{P} 0$ ,  $n \rightarrow \infty$  (*stejnoměrná konzistence*).

- ▶  $\theta_X = t(F_X)$ : parametr zájmu
- ▶  $\hat{\theta}_n = t(\hat{F}_n)$ : **empirický odhad** parametru  $\theta_X$

## Příklad. **Empirický odhad střední hodnoty**

**Data:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

**Model:**  $F_X \in \mathcal{F} = \{\text{libovolné rozdělení na } \mathbb{R}\}$

**Odhadovaný parametr:**  $\theta_X = \mathbb{E}_{F_X} X = \int_{-\infty}^{\infty} x dF_X(x) =: \mu_X$

$$\hat{\mu}_n = \mathbb{E}_{\hat{F}_n} X = \int_{-\infty}^{\infty} x d\hat{F}_n(x) = \dots = \bar{X}_n$$

- Nechť  $h$  je měřitelná reálná funkce taková, že

$$\forall F_X \in \mathcal{F} \quad \mathbb{E}_{F_X} |h(X)| < \infty.$$

- Snadno: empirický odhad pro  $\theta_X = \mathbb{E}_{F_X} h(X)$  jest  $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n h(X_i)$ .

### Příklad. Empirický odhad rozptylu

**Data a model:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X \in \mathcal{F} = \mathcal{L}^2$

**Odhadovaný parametr:**  $\theta_X = \text{var}_{F_X} X =: \sigma_X^2$

$$\hat{\sigma}_n^2 = \dots = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S_n^2$$

Analogicky (za předpokladu existence příslušných **absolutních** momentů), empirické odhady pro

- **necentrální momenty**  $\mu'_k := \mathbb{E}_{F_X} X^k$ ,  $k = 1, 2, \dots$

$$\hat{\mu}'_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad \text{konzistentní, nestranné}$$

- **centrální momenty**  $\mu_k := \mathbb{E}_{F_X} (X - \mathbb{E} X)^k$ ,  $k = 1, 2, \dots$

$$\hat{\mu}'_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^k \quad \text{konzistentní, ne nutně nestranné}$$

- **šikmost (skewness)**  $\gamma_3 := \frac{\mathbb{E}_{F_X} (X - \mathbb{E} X)^3}{(\text{var}_{F_X} X)^{3/2}}$

$$\hat{\gamma}_3 = \frac{\hat{\mu}_3}{(\hat{\sigma}_n^2)^{3/2}} \quad \text{konzistentní}$$

- **špičatost (kurtosis)**  $\gamma_4 := \frac{\mathbb{E}_{F_X} (X - \mathbb{E} X)^4}{(\text{var}_{F_X} X)^2}$

$$\hat{\gamma}_4 = \frac{\hat{\mu}_4}{(\hat{\sigma}_n^2)^2} \quad \text{konzistentní}$$

- ▶ **Kvantilová funkce** rozdělení  $F_X$

$$F_X^{-1}(\alpha) := \inf\{x : F_X(x) \geq \alpha\}, \quad 0 < \alpha < 1.$$

- ▶  **$\alpha$ -kvantil** rozdělení  $F_X$        $u_X(\alpha) := F_X^{-1}(\alpha).$

- ▶ Platí:       $F_X(u_X(\alpha)) \geq \alpha, \quad \forall h > 0 \quad F_X(u_X(\alpha) - h) < \alpha.$

### Definice 3.7   Výběrový kvantil.

Pro  $\alpha \in (0, 1)$  definujeme **výběrový (empirický)  $\alpha$ -kvantil** (*empirical quantile*) jako

$$\hat{u}_n(\alpha) = \hat{F}_n^{-1}(\alpha).$$

## Výběrový medián ( $\alpha = 0,50$ )

$$\hat{m}_n := \hat{u}_n(0,50) = \begin{cases} X_{(\frac{n+1}{2})}, & n \text{ liché,} \\ X_{(\frac{n}{2})}, & n \text{ sudé.} \end{cases}$$

**Software**, včetně , `quantile()`

- ▶ i jiné definice výběrových kvantilů;
- ▶ obvykle nějaká lineární interpolace mezi  $X_{(k_\alpha-1)}$ ,  $X_{(k_\alpha)}$ ,  $X_{(k_\alpha+1)}$
- ▶ např. výběrový medián v  (`median()`, `quantile(, probs = 0.50)`):

$$\tilde{m}_n = \begin{cases} X_{(\frac{n+1}{2})}, & n \text{ liché,} \\ 0,5 (X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}), & n \text{ sudé.} \end{cases}$$

### Lemma 3.4 Charakterizace výběrových kvantilů.

Nechť  $\alpha \in (0, 1)$ . Pro výběrový  $\alpha$ -kvantil  $\hat{u}_n(\alpha)$  platí

$$\hat{u}_n(\alpha) = \operatorname{argmin}_{c \in \mathbb{R}} \sum_{i=1}^n \varrho_\alpha(X_i - c),$$

kde  $\varrho_\alpha(u) = \alpha u \mathbb{I}(u \geq 0) + (1 - \alpha)(-u) \mathbb{I}(u < 0)$

$$= \begin{cases} \alpha |u|, & u \geq 0, \\ (1 - \alpha) |u|, & u < 0. \end{cases}$$

## Výběrový medián ( $\alpha = 0,50$ )

$$\blacktriangleright \hat{m}_n = \operatorname{argmin}_{c \in \mathbb{R}} \sum_{i=1}^n |X_i - c|,$$

$$\text{srovnej s } \bar{X}_n = \operatorname{argmin}_{c \in \mathbb{R}} \sum_{i=1}^n (X_i - c)^2.$$

$$\blacktriangleright \min_{c \in \mathbb{R}} \sum_{i=1}^n |X_i - c| \text{ dosaženo na množině } M_n, \text{ kde}$$

$$M_n = \begin{cases} \left\{ X_{(\frac{n+1}{2})} \right\}, & n \text{ liché,} \\ \left[ X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)} \right], & n \text{ sudé.} \end{cases}$$



## Statistické vlastnosti výběrových kvantilů

- ▶ Omezíme se na **spojitá** rozdělení s **rostoucí** (alespoň lokálně) distribuční funkcí  $F_X$  a hustotou  $f_X$ .
- ▶ Stále **neparametrický** model, ale zúžení množiny uvažovaných rozdělení.
- ▶ Výběrový kvantil dle definice 3.7, tj.  $\hat{u}_n(\alpha) = \hat{F}_n^{-1}(\alpha)$  (jednoznačné).

### Věta 3.5 Vlastnosti výběrových kvantilů.

Nechť  $\alpha \in (0, 1)$ . Nechť  $X_1, \dots, X_n$  je náhodný výběr z rozdělení, které má distribuční funkci  $F_X$  **spojitou** a **rostoucí** na nějakém okolí bodu  $u_X(\alpha)$ .

- (i) Potom  $\hat{u}_n(\alpha) \xrightarrow{P} u_X(\alpha)$ ,  $n \rightarrow \infty$ .
- (ii) Pokud navíc existuje hustota  $f_X$ , která je spojitá a nenulová v bodě  $u_X(\alpha)$ , pak

$$\sqrt{n}\{\hat{u}_n(\alpha) - u_X(\alpha)\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V(\alpha)), \quad n \rightarrow \infty, \quad \text{kde } V(\alpha) = \frac{\alpha(1-\alpha)}{f_X^2(u_X(\alpha))}.$$

- ▶ Lze využít

$$\sqrt{n}\{\hat{u}_n(\alpha) - u_X(\alpha)\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V(\alpha)), \quad n \rightarrow \infty, \quad \text{kde } V(\alpha) = \frac{\alpha(1-\alpha)}{f_X^2(u_X(\alpha))}?$$

- ▶ Je-li  $F_X$  **spojitá v bodě  $u_X(\alpha)$** , lze postupovat s využitím **pořádkových statistik**.
- ▶ Hledáme **oboustranný** interval spolehlivosti pro  $u_X(\alpha)$  s pokrytím  $1 - \beta$  ve tvaru  $(X_{(k_L)}, X_{(k_U)})$ ,  $1 \leq k_L < k_U \leq n$ , tj. požadujeme, pro každé  $F_X \in \mathcal{F}$

$$P_{F_X} \left( (X_{(k_L)}, X_{(k_U)}) \ni u_X(\alpha) \right) \geq 1 - \beta.$$

- ▶ Hledej největší a nejmenší přirozená čísla  $k_L < k_U$  tak, aby

$$P\left(Bi(n, \alpha) \leq k_L - 1\right) \leq \frac{\beta}{2}, \quad P\left(Bi(n, \alpha) \geq k_U\right) \leq \frac{\beta}{2}.$$

## Normální aproximace a korekce na spojitost

$$P\left(\mathcal{B}i(n, \alpha) \leq k_L - 1\right) = P\left(\mathcal{B}i(n, \alpha) < k_L\right) = P\left(\mathcal{B}i(n, \alpha) \leq k_L - \frac{1}{2}\right)$$

$$P\left(\mathcal{B}i(n, \alpha) \geq k_U\right) = P\left(\mathcal{B}i(n, \alpha) > k_U - 1\right) = P\left(\mathcal{B}i(n, \alpha) > k_U - \frac{1}{2}\right)$$

---

$$\begin{aligned} P\left(\mathcal{B}i(n, \alpha) \leq k_L - \frac{1}{2}\right) &= P\left(\frac{\mathcal{B}i(n, \alpha) - n\alpha}{\sqrt{n\alpha(1-\alpha)}} \leq \frac{k_L - \frac{1}{2} - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right) \\ &\doteq \Phi\left(\frac{k_L - \frac{1}{2} - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right) \end{aligned}$$

$$\begin{aligned} P\left(\mathcal{B}i(n, \alpha) > k_U - \frac{1}{2}\right) &= P\left(\frac{\mathcal{B}i(n, \alpha) - n\alpha}{\sqrt{n\alpha(1-\alpha)}} > \frac{k_U - \frac{1}{2} - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right) \\ &\doteq 1 - \Phi\left(\frac{k_U - \frac{1}{2} - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right) \end{aligned}$$

---

$$k_L = \left\lfloor \frac{1}{2} + n\alpha - u_{1-\frac{\beta}{2}} \sqrt{n\alpha(1-\alpha)} \right\rfloor, \quad k_U = \left\lceil \frac{1}{2} + n\alpha + u_{1-\frac{\beta}{2}} \sqrt{n\alpha(1-\alpha)} \right\rceil.$$

## Nastavení/předpoklady (obecný (neparametrický) $\mathcal{L}^2$ model, vícerozměrné rozdělení)

$$\blacktriangleright \underbrace{\mathbf{X}_1, \dots, \mathbf{X}_n}_{\mathbb{X}} \stackrel{\text{i.i.d.}}{\sim} \mathbf{X}, \quad \mathbf{X} \sim F_X \in \mathcal{F} \text{ (model)}$$

$\mathcal{F} = \{\text{libovolné rozdělení na } \mathbb{R}^p \text{ s konečnou varianční maticí}\}$   
 $\equiv$  **neparametrický  $\mathcal{L}^2$  model**

$$\blacktriangleright \mathbf{X}_i = (X_{i,1}, \dots, X_{i,p})^\top, \quad i = 1, \dots, n, \quad \mathbf{X} = (X_1, \dots, X_p)^\top$$

**Parametry zájmu:**

$$\boldsymbol{\mu}_X := \mathbb{E}_{F_X} \mathbf{X},$$

$$\boldsymbol{\Sigma}_X := \text{var}_{F_X} \mathbf{X} = \mathbb{E}_{F_X} (\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_X)^\top = \mathbb{E}_{F_X} \mathbf{X} \mathbf{X}^\top - \boldsymbol{\mu}_X \boldsymbol{\mu}_X^\top$$

## Empirické odhady (zřejmé)

$$\hat{\boldsymbol{\mu}}_n := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \bar{\mathbf{X}}_n = (\bar{X}_1, \dots, \bar{X}_p)^\top$$

$$\hat{\boldsymbol{\Sigma}}_n := \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{X}}_n) (\mathbf{x}_i - \bar{\mathbf{X}}_n)^\top$$

## Výběrová varianční matice

$$\mathbb{S}_n^2 := \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{X}}_n) (\mathbf{x}_i - \bar{\mathbf{X}}_n)^\top$$

► diagonála  $\mathbb{S}_j^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{i,j} - \bar{X}_j)^2, \quad j = 1, \dots, p$

(výběrové rozptyly)

► prvek  $(j, m)$  matice  $\mathbb{S}_n^2$ :

$$\mathbb{S}_{j,m}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{i,j} - \bar{X}_j) (X_{i,m} - \bar{X}_m), \quad j \neq m$$

(výběrové kovariance)

### Tvrzení 3.6 Vlastnosti výběrového průměru a výběrové varianční matice.

(i) Je-li  $\mathbb{E}_{F_X} |X_j| < \infty$  pro každé  $j = 1, \dots, p$ , potom

$$\mathbb{E}_{F_X} \bar{\mathbf{X}}_n = \boldsymbol{\mu}_X, \quad \bar{\mathbf{X}}_n \xrightarrow{P} \boldsymbol{\mu}_X, \quad n \rightarrow \infty.$$

(ii) Je-li  $\text{var}_{F_X} X_j < \infty$  pro každé  $j = 1, \dots, p$ , potom

$$\begin{aligned} \mathbb{E}_{F_X} \mathbb{S}_n^2 &= \boldsymbol{\Sigma}_X, & \mathbb{S}_n^2 &\xrightarrow{P} \boldsymbol{\Sigma}_X, \quad n \rightarrow \infty, \\ \hat{\boldsymbol{\Sigma}}_n &\xrightarrow{P} \boldsymbol{\Sigma}_X, \quad n \rightarrow \infty. \end{aligned}$$

### Definice 3.8 Výběrový korelační koeficient.

Výběrový korelační koeficient (*sample correlation coefficient*)  $\hat{\varrho}_{j,m}$  veličin  $X_j$  a  $X_m$ ,  $j, m = 1, \dots, p, j \neq m$ , definujeme jako

$$\hat{\varrho}_{j,m} = \frac{S_{j,m}}{S_j S_m} = \frac{\sum_{i=1}^n (X_{i,j} - \bar{X}_j) (X_{i,m} - \bar{X}_m)}{\sqrt{\sum_{i=1}^n (X_{i,j} - \bar{X}_j)^2 \sum_{i=1}^n (X_{i,m} - \bar{X}_m)^2}}.$$

# 4

## **Principy testování hypotéz**



## Oddíl **4.1**

# **Základní pojmy a definice**

## Nastavení/předpoklady

- ▶  $\underbrace{X_1, \dots, X_n}_{\mathbf{X}} \stackrel{\text{i.i.d.}}{\sim} \mathbf{X}, \quad \mathbf{X} \sim F_X \in \mathcal{F} \text{ (model);}$
- ▶  $\theta := \mathbf{t}(F) \in \mathbb{R}^p, F \in \mathcal{F} \equiv$  parametr zájmu
- ▶  $\Theta = \{\mathbf{t}(F), F \in \mathcal{F}\} \equiv$  parametrický prostor
- ▶  $\theta_X := \mathbf{t}(F_X) \equiv$  skutečná hodnota parametru

---

**Ilustrační příklady** pro  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, X \sim F_X, F_X \in \mathcal{F}$

- (A)  $X \sim \mathcal{N}(\theta_X, \sigma_0^2), \sigma_0^2 > 0$  známé,  
model  $\mathcal{F}^A = \{\mathcal{N}(\theta, \sigma_0^2), \theta \in \mathbb{R}\}$
- (B)  $X \sim \mathcal{N}(\theta_X, \sigma_X^2), \sigma_X^2 > 0$  neznámé,  
model  $\mathcal{F}^B = \{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}, \sigma^2 > 0\}$
- (C)  $X \sim F_X, F_X \in \mathcal{L}_+^2$  (rozdělení s konečným a nenulovým rozptylem),  
model  $\mathcal{F}^C = \mathcal{L}_+^2$  (neparametrický model)

Testovaný parametr:  $\theta = \mathbb{E}_F X = \int x dF(x), \quad \Theta = \mathbb{R}$

$\theta_X = \mathbb{E}_{F_X} X = \int x dF_X(x)$  (skutečná hodnota parametru)

- ▶ Necht'  $\Theta_0, \Theta_1 \subset \Theta$ ,  $\Theta_0 \cap \Theta_1 = \emptyset$ , ne nutně  $\Theta_0 \cup \Theta_1 = \Theta$ .
- ▶ Hledáme odpověď na otázku, zda  $\theta_X \in \Theta_0$  nebo  $\theta_X \in \Theta_1$ .

### Definice 4.1 Hypotéza a alternativa.

Množinu  $\Theta_0$  nazýváme [nulová] **hypotéza**, množinu  $\Theta_1$  nazýváme **alternativní hypotéza (alternativa)**.

### Označme

$$\mathcal{F}_0 := \{F \in \mathcal{F} : t(F) \in \Theta_0\}$$

- ▶ pokud  $\mathcal{F}_0 = \{F_0\}$ , mluvíme o **jednoduché hypotéze**,
- ▶ jinak **složená hypotéza**

$$\mathcal{F}_1 := \{F \in \mathcal{F} : t(F) \in \Theta_1\}$$

- ▶ pokud  $\mathcal{F}_1 = \{F_1\}$ , mluvíme o **jednoduché alternativě**,
- ▶ jinak **složená alternativa**

**Ilustrační příklady** pro  $X_1, \dots, X_n \sim X, X \sim F_X, F_X \in \mathcal{F}$

**Oboustranný** test parametru  $\theta = t(F) = \int x \, dF(x) \in \mathbb{R}$

(test o střední hodnotě) –  $H_0: \theta_X = \theta_0$

$H_1: \theta_X \neq \theta_0, \quad \theta_0 \in \mathbb{R}$  zvoleno předem

V modelech

**(A)**  $X \sim \mathcal{N}(\theta_X, \sigma_0^2), \sigma_0^2 > 0$  známé,

model  $\mathcal{F}^A = \{\mathcal{N}(\theta, \sigma_0^2), \theta \in \mathbb{R}\}$

**(B)**  $X \sim \mathcal{N}(\theta_X, \sigma_X^2), \sigma_X^2 > 0$  neznámé,

model  $\mathcal{F}^B = \{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}, \sigma^2 > 0\}$

**(C)**  $X \sim F_X, F_X \in \mathcal{L}_+^2$  (rozdělení s konečným a nenulovým rozptylem),

model  $\mathcal{F}^C = \mathcal{L}_+^2$  (neparametrický model)

## Rozhodování

Podklady pro  $\equiv$  Data  $\equiv$  Náhodný výběr  $\mathbb{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$

→ statistika  $U_n(\mathbb{X}) \equiv$  **testová** statistika (obvykle jednorozměrná)  
+ **kritický obor**  $C \subset \mathbb{R}$

## Rozhodovací pravidlo

- ▶  $U_n(\mathbb{X}) \in C \rightarrow$  **zamítáme** hypotézu  $H_0$  **ve prospěch** alternativy  $H_1$
- ▶  $U_n(\mathbb{X}) \notin C \rightarrow$  hypotézu  $H_0$  **nelze zamítnout (nezamítáme)**  
ve prospěch alternativy  $H_1$

## Definice 4.2 Statistický test.

**Statistický test** je definován pomocí testové statistiky  $U_n(\mathbb{X})$ , kritického oboru  $C$  a výše uvedeného pravidla pro zamítání hypotézy.

Dva testy  $(U_n(\mathbb{X}), C)$  a  $(U_n^*(\mathbb{X}), C^*)$  nazveme ekvivaltní, právě když **skoro jistě** platí

$$U_n(\mathbb{X}) \in C \iff U_n^*(\mathbb{X}) \in C^*.$$

## Oddíl 4.2

# Hladina a síla testu

### Definice 4.3 Chyba I. a II. druhu.

- (i) Jestliže test **zamítnul** platnou (nulovou) hypotézu, říkáme, že nastala **chyba I. druhu** (*type I error*).
- (ii) Jestliže test **nezamítnul** neplatnou (nulovou) hypotézu, říkáme, že nastala **chyba II. druhu** (*type II error*).

Pro  $F \in \mathcal{F}$ ,  $\theta = t(F)$  a  $B \in \mathcal{B}$  značíme

$$P_F(U_n(\mathbb{X}) \in B) = P_\theta(U_n(\mathbb{X}) \in B) = \int \mathbb{I}\{U_n(\mathbb{X}) \in B\} dF(\mathbf{x}_1) \cdots dF(\mathbf{x}_n).$$

#### Definice 4.4 Hladina testu.

Nechť  $\alpha \in (0, 1)$  je předem stanovené číslo.

(i) Jestliže kritický obor  $C$  splňuje podmínku

$$\sup_{F \in \mathcal{F}_0} P_F(U_n(\mathbb{X}) \in C) = \alpha,$$

říkáme, že test  $(U_n(\mathbb{X}), C)$  má **hladinu významnosti (significance level)** přesně  $\alpha$ .  
→ **přesný test**

(ii) Jestliže kritický obor  $C$  splňuje podmínku

$$\sup_{F \in \mathcal{F}_0} \lim_{n \rightarrow \infty} P_F(U_n(\mathbb{X}) \in C) = \alpha,$$

říkáme, že test  $(U_n(\mathbb{X}), C)$  má hladinu  $\alpha$  **asymptoticky**.

→ **asymptotický test**



## Konstrukce statistického testu

1. Stanovíme (regulátor stanoví) požadovanou hladinu  $\alpha$ .
  2. Navrhujeme/najdeme vhodnou testovou statistiku  $U_n(\mathbb{X})$ .
  3. Navrhujeme/zvolíme kritický obor  $C = C(\alpha)$  tak, aby hladina byla (asymptoticky) rovna  $\alpha$ , současně se snažíme o co nejnížší pravděpodobnost chyby II. druhu.
- 

### Poznámky.

- ▶ Hladina testu bývá obvykle malá, nejčastěji  $\alpha = 0,05$ .
- ▶ Má-li testová statistika  $U_n(\mathbb{X})$  **diskrétní** rozdělení, není možné dosáhnout všech hladin  $\alpha \in (0, 1)$ .
  - Při zadaném  $\alpha$  volíme kritický obor tak, že (dosažená) hladina  $\alpha' < \alpha$ 
    - **konzervativní** test.
- ▶ Test, kde je skutečná pravděpodobnost chyby I. druhu  $>$  požadovaná hladina  $\alpha$ 
  - **antikonzervativní** test.

## Definice 4.5 Silofunkce a síla testu.

Funkce  $\beta_n(F) = P_F(U_n(\mathbb{X}) \in C)$ , funkce  $\mathcal{F} \rightarrow [0, 1]$ , se nazývá **silofunkce testu** (*power function*).

Pro  $F \in \mathcal{F}_1$  se číslo  $\beta_n(F)$  nazývá **síla** (*power*) testu proti alternativě  $F$ .

Síla = pravděpodobnost (správného) zamítnutí neplatné  $H_0$ , pokud ve skutečnosti platí alternativa  $F$

$$= 1 - P_F(U_n(\mathbb{X}) \notin C)$$

= 1 mínus pravděpodobnost chyby II. druhu, pokud platí alternativa  $F$

## Příklad. Test o střední hodnotě v normálním rozdělení se známým rozptylem

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, X \sim F_X, F_X \in \mathcal{F}, \quad \theta_X := \mathbb{E}_{F_X} X = \int x \, dF_X(x) \in \mathbb{R} = \Theta$$

Model A:  $\mathcal{F} = \mathcal{F}^A = \{\mathcal{N}(\theta, \sigma_0^2), \theta \in \mathbb{R}\}$ ,  $\sigma_0^2 > 0$  známé

Testujeme  $H_0: \theta_X = \theta_0$

$H_1: \theta_X \neq \theta_0$ ,  $\theta_0 \in \mathbb{R}$  zvoleno předem

---

Označme  $\delta := \theta - \theta_0$ , pro  $F \in \mathcal{F}^A$ :

$$\begin{aligned} \beta_n(F) &= \beta_n(\theta) = \mathbf{P}_\theta \left( |U_n(\mathbb{X})| \geq u_{1-\alpha/2} \right) = \mathbf{P}_\theta \left( \left| \frac{\bar{X}_n - \theta_0}{\frac{\sigma_0}{\sqrt{n}}} \right| \geq u_{1-\alpha/2} \right) \\ &= \dots = \Phi \left( -u_{1-\alpha/2} - |\nu_n| \right) + 1 - \Phi \left( u_{1-\alpha/2} - |\nu_n| \right), \end{aligned}$$

$$\nu_n = \frac{\delta}{\frac{\sigma_0}{\sqrt{n}}}$$

## Určení rozsahu výběru

Jaké  $n$  je potřeba, aby test s pravděpodobností alespoň  $\beta$  (např. 0,80) zamítnul, pokud je  $|\theta_X - \theta_0|$  alespoň zadané  $\delta$ ?

$$\text{Řešíme } \Phi(-u_{1-\alpha/2} - |\nu_n|) + 1 - \Phi(u_{1-\alpha/2} - |\nu_n|) \geq \beta.$$

$$\text{Přibližné řešení: } n \geq (u_{1-\alpha/2} + u_\beta)^2 \frac{\sigma_0^2}{\delta^2}.$$

---

**Síla testu** závisí na

- ▶ hladině testu  $\alpha$  (dáno regulátorem)
- ▶ rozptylu pozorování  $\sigma_0^2$  (vlastnost „jevu“, který pozorujeme)
- ▶ alternativě (resp. její „vzdálenosti“  $\delta$  od hypotézy)
- ▶ počtu pozorování (můžeme ovlivnit)

## Poznámky.

### Volba testové statistiky $U_n(\mathbb{X})$

- ▶ Kvantifikuje (ne)shodu dat s  $H_0$   
 $\equiv$  rozdělení  $U_n(\mathbb{X})$  citlivé na skutečnou hodnotu testovaného parametru  $\theta_X$ .
- ▶ Za platnosti  $H_0$  rozdělení  $U_n(\mathbb{X})$  alespoň asymptoticky nezávisí na neznámých (rušivých) parametrech a příslušné rozdělení je známé.

### Volba kritického oboru $C(\alpha)$

- ▶ Musí dodržovat hladinu testu  $\alpha$ .
- ▶  $C(\alpha)$  odpovídá množinám hodnot  $U_n(\mathbb{X})$ , které jsou za platnosti  $H_0$  méně pravděpodobné než za  $H_1$   
 $\equiv \forall F \in \mathcal{F}_1 \quad \beta_n(F) \geq \alpha.$

#### Definice 4.6 Konzistentní test.

Test  $(U_n(\mathbb{X}), C)$  na hladině  $\alpha$  nazveme **konzistentním** testem, jestliže  $\forall F \in \mathcal{F}_1$  platí  $\lim_{n \rightarrow \infty} \beta_n(F) = 1$ .

#### Definice 4.7 Nestranný test.

Test  $(U_n(\mathbb{X}), C)$  na hladině  $\alpha$  nazveme **nestranným** testem, jestliže  $\forall F \in \mathcal{F}_1$  platí  $\beta_n(F) \geq \alpha$ .

- ▶ Nestrannost a konzistence testu příliš nesouvisí s nestranností a konzistencí odhadu.
- ▶ Nestrannost testu  $\equiv$  síla proti každé alternativě je  $\geq \alpha$ .

## Poznámky k interpretaci výsledku testu

### Zamítnutí nulové hypotézy

- ≡ rozdělení dat průkazně neodpovídá  $H_0$
- ▶ pravděpodobnost chybného zamítnutí  $H_0$  VŽDY omezena shora hladinou testu (je  $\leq \alpha$ )

**Hypotézu  $H_0$  vyvracíme, prokazujeme platnost alternativy  $H_1$ .**

### Nezamítnutí nulové hypotézy

- ≡ rozdělení dat není dostatečně odlišné od rozdělení dat, které předpokládá  $H_0$
- ▶ nelze potvrdit platnost  $H_1$ , ale ani  $H_0$
- ▶  $P_{H_1}(\text{nezamítám } H_0) = P(\text{chyba II. druhu}) = 1 - \text{síla}$  může být vysoká!

**Hypotézu  $H_0$  nemůžeme vyvrátit ve prospěch alternativy  $H_1$ , ale nemůžeme ji ani potvrdit!**

**Příklady, připomenutí:**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X$ , uvažujeme tři modely:

**(A)**  $X \sim \mathcal{N}(\theta_X, \sigma_0^2)$ ,  $\theta_X \in \mathbb{R}$ ,  $\sigma_0^2 > 0$  známé

**(B)**  $\sim \mathcal{N}(\theta_X, \sigma^2)$ ,  $\theta_X \in \mathbb{R}$ ,  $\sigma^2 > 0$  neznámé

**(C)**  $\sim F_X$ ,  $F_X \in \mathcal{L}_+^2$ ,  $\theta_X := \mathbb{E}_{F_X} X$ ,  $\sigma^2 := \text{var}_{F_X} X$

► Parametr zájmu:  $\theta_X = \mathbb{E}_{F_X} X$

►  $\theta_0 \in \mathbb{R}$ : „významná“ předem zvolená hodnota (objeví se v  $H_0$ )

---

Test. statistika  $U_n(\mathbb{X})$

Rozdělení, pokud  $\theta_X = \theta_0$

**(A)**  $\frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sqrt{\sigma_0^2}}$

$\sim \mathcal{N}(0, 1)$

**(B)**  $\frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sqrt{S_n^2}}$

$\sim t_{n-1}$

**(C)**  $\frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sqrt{S_n^2}}$

$\xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty$



$C(\alpha) \equiv$  kritický obor testu na hladině  $\alpha$  obvykle tvaru:

| $C(\alpha)$                                                | $U_n(\mathbb{X}) \in C(\alpha) \Leftrightarrow$ |
|------------------------------------------------------------|-------------------------------------------------|
| (i) $[c_U(\alpha), \infty)$                                | $U_n(\mathbb{X}) \geq c_U(\alpha)$              |
| (ii) $(-\infty, c_L(\alpha)]$                              | $U_n(\mathbb{X}) \leq c_L(\alpha)$              |
| (iii) $(-\infty, -c_U(\alpha)] \cup [c_U(\alpha), \infty)$ | $ U_n(\mathbb{X})  \geq c_U(\alpha) \quad > 0$  |
| (iv) $(-\infty, c_L(\alpha)] \cup [c_U(\alpha), \infty)$   | $c_L(\alpha) < c_U(\alpha)$                     |

$c_L(\alpha), c_U(\alpha)$ : **kritické hodnoty**

## Oddíl 4.3

# P-hodnota

Uvažujme testování  $H_0: \theta_X \in \Theta_0$

$H_1: \theta_X \in \Theta_1$

Pro  $\alpha \in (0, 1)$  mějme test  $(U_n(\mathbb{X}), C(\alpha))$  na hladině  $\alpha$ ,

tj. pro  $\alpha \in (0, 1)$  je dáno pravidlo, jak vypadá  $C(\alpha)$ .

Dodefinujeme  $C(1) = \mathbb{R}$ .

Pro testy s **diskrétně** rozdělenou testovou statistikou, nechť  $C(\alpha) =$  kritický obor testu na hladině  $\alpha' < \alpha$ , kde  $\alpha'$  je nejbližší dosažitelná hladina.

### **Definice 4.8** P-hodnota.

Nechť  $u_x = U_n(\mathbf{x})$  je realizovaná hodnota testové statistiky. Pak **p-hodnotu** (*p-value*) neboli dosaženou hladinu testu definujeme jako

$$p(\mathbf{x}) = \inf\{\alpha \in (0, 1] : u_x \in C(\alpha)\}.$$

*přesný* test  $\rightarrow$  *přesná* p-hodnota

*asymptotický* test  $\rightarrow$  *asymptotická* p-hodnota

## Rozhodování na základě p-hodnoty

- Zamítní  $H_0$ , pokud  $p(\mathbf{x}) \leq \alpha$ .

Zamítáme na všech hladinách  $\alpha' \geq p(\mathbf{x})$ , proto též termín **dosažená hladina testu**.

Nelze stanovovat hladinu testu poté, co je spočtena p-hodnota!

- Nezamítní  $H_0$ , pokud  $p(\mathbf{x}) > \alpha$ .

p-hodnota **není**  $P(\text{platí } H_0)$ !

## Výpočet p-hodnoty pro **jednostranný** kritický obor

Předpokládejme, že  $C(\alpha) = [c_U(\alpha), \infty)$  a **spojité** rozdělení testové statistiky.

Např.  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_X, \sigma_X^2)$

$$H_0: \theta_X \leq \theta_0$$

$$H_1: \theta_X > \theta_0$$

---

$$U_n(\mathbb{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{S_n},$$

pokud  $\theta_X = \theta_0$ :  $U_n(\mathbb{X}) \sim t_{n-1}$ ,

$$c_U(\alpha) = t_{n-1}(1 - \alpha)$$

## Výpočet p-hodnoty pro **jednostranný** kritický obor

Předpokládejme, že  $C(\alpha) = (-\infty, c_L(\alpha)]$  a **asymptotické spojité** rozdělení testové statistiky.

Např.  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, X \sim F_X \in \mathcal{L}_+^2, \quad \theta_X := \mathbb{E}_{F_X} X \quad \begin{array}{l} H_0: \theta_X \geq \theta_0 \\ H_1: \theta_X < \theta_0 \end{array}$

$$U_n(\mathbb{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{S_n}, \quad \text{pokud } \theta_X = \theta_0: \quad U_n(\mathbb{X}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$
$$c_L(\alpha) = -u_{1-\alpha}$$

P-hodnotu lze chápat jako **maximální** možnou pravděpodobnost, že bychom za platnosti  $H_0$  při opakování studie/experimentu napozorovali data, která by byla s  $H_0$  ve stejném nebo větším rozporu **(vedla by k extrémnější hodnotě testové statistiky)** než data, která analyzujeme.

Předpokládejme, že  $C(\alpha) = (-\infty, c_L(\alpha)] \cup [c_U(\alpha), \infty)$  a **spojité** rozdělení testové statistiky.

Např.  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_X, \sigma_X^2)$

$$H_0: \theta_X = \theta_0$$

$$H_1: \theta_X \neq \theta_0$$

---

$$U_n(\mathbb{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{S_n},$$

$$\text{pokud } \theta_X = \theta_0: \quad U_n(\mathbb{X}) \sim t_{n-1},$$

$$-c_L(\alpha) = c_U(\alpha) = t_{n-1}(1 - \frac{\alpha}{2})$$



## Rozdělení p-hodnoty za nulové hypotézy

Uvažme nyní  $p(\mathbb{X}) \equiv$  statistika (funkce dat).

### Tvrzení 4.1 Rozdělení p-hodnoty za nulové hypotézy.

*Nechť platí nulová hypotéza, tj.  $F_X \in \mathcal{F}_0$ , nechť navíc platí*

$$\forall \alpha \in (0, 1) : \quad \sup_{F \in \mathcal{F}_0} P_F(U_n(\mathbb{X}) \in C(\alpha)) = P_{F_X}(U_n(\mathbb{X}) \in C(\alpha)).$$

*Nechť má statistika  $U_n(\mathbb{X})$  **spojité** rozdělení. Pak  $p(\mathbb{X}) \sim \mathcal{U}(0, 1)$ .*

## Oddíl 4.4

# Dualita intervalových odhadů a testování hypotéz

$$\underbrace{\mathbf{X}_1, \dots, \mathbf{X}_n}_{\mathbb{X}} \stackrel{\text{i.i.d.}}{\sim} \mathbf{X}, \quad \mathbf{X} \sim F_X \in \mathcal{F}$$

- ▶  $\theta = t(F) \in \mathbb{R}$ : parametr zájmu;
- ▶  $\theta_X = t(F_X)$ : skutečná hodnota parametru zájmu.

Intervalový odhad pro  $\theta_X$ :

$$\forall F_X \in \mathcal{F}: \quad \mathbf{P}_{F_X} \left( (\eta_L(\mathbb{X}), \eta_U(\mathbb{X})) \ni \theta_X \right) = 1 - \alpha,$$

$$\text{resp.} \quad \lim_{n \rightarrow \infty} \mathbf{P}_{F_X} \left( (\eta_L(\mathbb{X}), \eta_U(\mathbb{X})) \ni \theta_X \right) = 1 - \alpha.$$

souvisí s testováním  $H_0: \theta_X = \theta_0$

$$\underline{H_1: \theta_X \neq \theta_0}$$

## Tvrzení 4.2 Dualita intervalových odhadů a testování.

- (i) Nechť je dán *oboustranný* interval spolehlivosti pro parametr  $\theta_X$  s pokrytím  $1 - \alpha$  [přesným/asymptotickým] tvaru  $(\eta_L(\mathbb{X}), \eta_U(\mathbb{X}))$ . Uvažujme test hypotézy  $H_0: \theta_X = \theta_0$  proti alternativě  $H_1: \theta_X \neq \theta_0$  založený na *rozhodovacím pravidle*

$$H_0 \text{ zamítneme} \quad \Longleftrightarrow \quad \theta_0 \notin (\eta_L(\mathbb{X}), \eta_U(\mathbb{X})),$$

$$H_0 \text{ nezamítneme} \quad \Longleftrightarrow \quad \theta_0 \in (\eta_L(\mathbb{X}), \eta_U(\mathbb{X})).$$

Tento test má hladinu  $\alpha$  [přesně/asymptoticky].

## Tvrzení 4.2 Dualita intervalových odhadů a testování, pokrač.

- (ii) *Nechť pro všechna  $\theta \in \Theta$  je dán test  $(U_n(\mathbb{X}, \theta), C_\theta(\alpha))$  hypotézy  $H_0: \theta_X = \theta$  proti alternativě  $H_1: \theta_X \neq \theta$  takový, že pro všechna  $F \in \mathcal{F}$ , pro která  $\theta = t(F)$  je*

$$P_F(U_n(\mathbb{X}, \theta) \in C_\theta(\alpha)) = \alpha, \quad \text{resp.} \quad \lim_{n \rightarrow \infty} P_F(U_n(\mathbb{X}, \theta) \in C_\theta(\alpha)) = \alpha.$$

*Sestavme množinu*

$$B_n(\mathbb{X}) = \{\theta : U_n(\mathbb{X}, \theta) \notin C_\theta(\alpha)\}$$

*obsahující všechny parametry  $\theta \in \Theta$ , pro něž se při pozorovaných datech nezamítá hypotéza  $H_0: \theta_X = \theta$ . Pak pro všechna  $F_X \in \mathcal{F}$*

$$P_{F_X}(B_n(\mathbb{X}) \ni \theta_X) = 1 - \alpha, \quad \text{resp.} \quad \lim_{n \rightarrow \infty} P_{F_X}(B_n(\mathbb{X}) \ni \theta_X) = 1 - \alpha.$$

*Je-li  $B_n(\mathbb{X})$  interval, pak se jedná o interval spolehlivosti pro parametr  $\theta_X$  s pravděpodobností pokrytí  $1 - \alpha$  [přesnou/asymptotickou].*

### Příklad (B).

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_X, \sigma_X^2) \quad H_0: \theta_X = \theta$$

$$H_1: \theta_X \neq \theta$$

---

$$U_n(\mathbb{X}, \theta) = \frac{\sqrt{n}(\bar{X}_n - \theta)}{S_n},$$

$$\text{pokud } \theta_X = \theta: \quad U_n(\mathbb{X}, \theta) \sim t_{n-1},$$

$$C_\theta(\alpha) = (-\infty, -t_{n-1}(1 - \frac{\alpha}{2})] \cup [t_{n-1}(1 - \frac{\alpha}{2}), \infty)$$

### Příklad (C). **Jednostranný interval spolehlivosti**

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, X \sim F_X \in \mathcal{L}_+^2, \quad \theta_X := \mathbb{E}_{F_X} X$$

$$H_0: \theta_X \geq \theta$$

$$H_1: \theta_X < \theta$$

---

$$U_n(\mathbb{X}, \theta) = \frac{\sqrt{n}(\bar{X}_n - \theta)}{S_n},$$

$$\text{pokud } \theta_X = \theta: \quad U_n(\mathbb{X}, \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), n \rightarrow \infty$$

$$C_\theta(\alpha) = (-\infty, -u_{1-\alpha}]$$

$$\text{Duální k intervalu spolehlivosti } (-\infty, \eta_U(\mathbb{X})), \quad \eta_U(\mathbb{X}) = \bar{X}_n + \frac{S_n}{\sqrt{n}} u_{1-\alpha}.$$

# 5

**Jednovýběrové a párové problémy  
pro kvantitativní data**



## Nastavení/předpoklady

- ▶ Kvantitativní data

$$\underbrace{X_1, \dots, X_n}_{\mathbf{X}} \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X \in \mathcal{F} \text{ (model);}$$

- ▶  $\theta_X = t(F_X) \in \mathbb{R}$ : parametr zájmu
  - testy
  - intervaly spolehlivosti

## Oddíl **5.1**

# **Jednovýběrový Kolmogorovův-Smirnovův test**

# Jednovýběrový Kolmogorovův-Smirnovův test

---

## Andrej Nikolajevič Kolmogorov

25.4.1903 (Tambov)

– 20.10.1987 (Moskva)



## Nikolaj Vasiljevič Smirnov

17.10.1900 (Moskva)

– 2.6.1966 (Moskva)



# Jednovýběrový Kolmogorovův-Smirnovův test

---

- ▶ Test shody distribuční funkce dat s pevně danou distribuční funkcí  
– test dobré shody (*goodness-of-fit test*)
  - ▶ **Neparametrický** test
- 

Model:  $\mathcal{F} = \{\text{všechna } \underline{\text{spojitá}} \text{ rozdělení}\}$

Testovaný parametr: Celá distribuční funkce  $F_X$

Hypotéza a alternativa:  $H_0: \forall x \in \mathbb{R} \quad F_X(x) = F_0(x)$

$H_1: \exists x \in \mathbb{R} \quad F_X(x) \neq F_0(x)$

---

$F_0$ : pevně specifikovaná distribuční funkce **bez neznámých** parametrů

**Připomenutí** (empirická distribuční funkce):

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \leq x\}, \quad x \in \mathbb{R}.$$

Testová statistika:  $K_n := \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$

---

Označme  $K_n^+ = \sup_{x \in \mathbb{R}} (\hat{F}_n(x) - F_0(x))$ ,  
 $K_n^- = \sup_{x \in \mathbb{R}} (F_0(x) - \hat{F}_n(x))$ .

To jest  $K_n = \max\{K_n^+, K_n^-\}$ .

**Lemma 5.1** Kolmogorovova-Smirnovova statistika a pořádkové statistiky.

---

Je-li  $F_0$  *spojitá*, pak platí

$$K_n^+ = \max_{1 \leq i \leq n} \left( \frac{i}{n} - F_0(X_{(i)}) \right),$$

$$K_n^- = \max_{1 \leq i \leq n} \left( F_0(X_{(i)}) - \frac{i-1}{n} \right).$$

## **Tvrzení 5.2** Asymptotika pro Kolmogorovovu-Smirnovovu statistiku.

Nechť  $X_1, \dots, X_n$  je náhodný výběr ze **spojitého** rozdělení s distribuční funkcí  $F_X$ . Potom

$$\sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_X(x)| \xrightarrow{\mathcal{D}} Z, \quad n \rightarrow \infty,$$

kde  $Z$  má distribuční funkci

$$G(z) = \begin{cases} 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 z^2}, & z > 0, \\ 0, & z \leq 0. \end{cases}$$

**Značení:**  $k_{1-\alpha} = G^{-1}(1 - \alpha), \quad 0 < \alpha < 1.$

## Asymptotický test:

Zamítáme  $H_0 \iff \sqrt{n} K_n \geq k_{1-\alpha}$

$$p = 1 - G(\sqrt{n} k_n)$$

$k_n$ : hodnota statistiky  $K_n$  dosažená/realizovaná s daty

---

: `ks.test(x, y = F_0, exact = FALSE)`

## Poznámky.

- ▶  $K_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)| \xrightarrow{P} \sup_{x \in \mathbb{R}} |F_X(x) - F_0(x)| > 0$  (za platnosti  $H_1$ ),  
odsud **konzistence** testu
- ▶ Jde o univerzální test, málo citlivý na konkrétní porušení  $H_0$  (např. změnu střední hodnoty).
- ▶ Test lze uvažovat též **jednostranně** s alternativou  
 $H_1: F_X(x) \geq F_0(x), \quad \exists x \in \mathbb{R} \ F_X(x) > F_0(x).$



## Intervaly spolehlivosti pro $F_X$

(a) pro pevně dané (jedno konkrétní)  $x \in S_X = \{x^* : F_X(x^*) \in (0, 1)\}$ :

$$IS_n(x) = \left( \hat{F}_n(x) \mp \sqrt{\frac{\hat{F}_n(x)(1 - \hat{F}_n(x))}{n}} u_{1-\alpha/2} \right),$$

$$\text{resp. dolní mez} = \max \left\{ 0, \hat{F}_n(x) - \sqrt{\frac{\hat{F}_n(x)(1 - \hat{F}_n(x))}{n}} u_{1-\alpha/2} \right\}$$

$$\text{horní mez} = \min \left\{ 1, \hat{F}_n(x) + \sqrt{\frac{\hat{F}_n(x)(1 - \hat{F}_n(x))}{n}} u_{1-\alpha/2} \right\}$$

$$\forall x \in S_X \quad \forall F_X \in \mathcal{F} \quad \lim_{n \rightarrow \infty} P_{F_X} \left( IS_n(x) \ni F_X(x) \right) = 1 - \alpha$$

▀ **bodový** (asymptotický) interval spolehlivosti pro  $F_X(x)$

(b) pro všechna  $x \in S_X$  najednou (sdruženě)

$$B_n(x) = \left( \max \left\{ 0, \hat{F}_n(x) - \frac{k_{1-\alpha}}{\sqrt{n}} \right\}, \min \left\{ 1, \hat{F}_n(x) + \frac{k_{1-\alpha}}{\sqrt{n}} \right\} \right),$$

$$\forall F_X \in \mathcal{F} \quad \lim_{n \rightarrow \infty} \mathbf{P}_{F_X} \left( \forall x \in S_X \quad B_n(x) \ni F_X(x) \right) = 1 - \alpha$$

▀ **simultánní** (asymptotické) intervaly spolehlivosti pro  $F_X(x)$

▀ **pás spolehlivosti** (*confidence band*) pro  $F_X$

# Komplikace spojené s použitím K-S testu (porušení předpokladů)

## 1. $F_0$ není spojitě

- ▶  $K_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$  lze použít jako testovou statistiku
- ▶ jiné asymptotické rozdělení než uvedené v Tvzení 5.2
- ▶ použití kvantilů  $k_{1-\alpha}$   
→ (asymptoticky) konzervativní test (hladina OK, ale je slabší)

## 2. $F_0$ spojitě, ale v datech shody (zaokrouhlování, ...)

- ▶ formálně pozorujeme  $\tilde{X}_1 = \text{round}(X_1), \dots, \tilde{X}_n = \text{round}(X_n)$   
s distribuční funkcí  $\tilde{F}_X(x)$
- ▶ K-S lze použít, pokud  $\sup_{x \in \mathbb{R}} |\tilde{F}_X(x) - F_X(x)|$  „malé“  
(splněno při rozumném zaokrouhlování)

## 3. Hypotéza není jednoduchá (nejčastější problém, ...)

▶ např. **test normality**:  $H_0: F_X \equiv$  normální rozdělení

$H_1: F_X$  není normální (gaussovské)

---

▶ tj.  $H_0: F_X \in \mathcal{F}_0 \stackrel{\text{např.}}{=} \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$

$H_1: F_X \notin \mathcal{F}_0$

---

▶ obecně  $\mathcal{F}_0 = \{F(x; \theta), \theta \in \Theta\}$

▶ testová statistika  $\tilde{K}_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x; \hat{\theta}_n)|$ ,

kde  $\hat{\theta}_n =$  vhodný odhad parametru  $\theta_X$

▶ neplatí Tvzení 5.2, použití kvantilů  $k_{1-\alpha}$

→ (asymptoticky) **silně konzervativní** test (velmi slabý)

▶ (asymptotické) rozdělení  $\tilde{K}_n$  komplikované a závisí na (neznámém)  $\theta_X$

▶ pro  $H_0: F_X \equiv$  normální rozdělení existují speciální testy

→ Lilliefors, Shapiro-Wilk, D'Agostino, ...

## Oddíl **5.2**

# **Jednovýběrový t-test**

- ▶ Test o střední hodnotě v normálním výběru, resp. výběru z rozdělení s konečným rozptylem
- 

Model:  $\mathcal{F}_N = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$

$$\mathcal{F}_{as} = \mathcal{L}_+^2$$

Testovaný parametr:  $\mu_X := \mathbb{E}_{F_X} X$

Hypotéza a alternativa:  $H_0: \mu_X = \mu_0$

$$H_1: \mu_X \neq \mu_0$$

---

$\mu_0 \in \mathbb{R}$ : dáno předem

Testová statistika:  $T_n := \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n}$

$$\text{Zamítáme } H_0 \iff |T_n| \geq t_{n-1} \left(1 - \frac{\alpha}{2}\right)$$

$$p = 2(1 - F_{n-1}(|t_n|))$$

$t_{n-1} \left(1 - \frac{\alpha}{2}\right)$ :  $(1 - \frac{\alpha}{2})$ -kvantil rozdělení  $t_{n-1}$

$F_{n-1}$ : distribuční funkce rozdělení  $t_{n-1}$

$t_n$ : hodnota statistiky  $T_n$  dosažená/realizovaná s daty

## Duální interval spolehlivosti pro $\mu_X$

$$IS_n = \left( \bar{X}_n \mp t_{n-1} \left(1 - \frac{\alpha}{2}\right) \frac{S_n}{\sqrt{n}} \right),$$

$$\forall F_X \in \mathcal{F}_N \quad P_{F_X}(IS_n \ni \mu_X) = 1 - \alpha \quad \text{přesný IS}$$

$$\forall F_X \in \mathcal{F}_{as} \quad \lim_{n \rightarrow \infty} P_{F_X}(IS_n \ni \mu_X) = 1 - \alpha \quad \text{asymptotický IS}$$

**R**: `t.test(x, mu =  $\mu_0$ , conf.level =  $1 - \alpha$ )`

Lze i jednostranně:

$$H_0: \mu_X \leq \mu_0$$

$$\text{Zamítáme } H_0 \iff T_n \geq t_{n-1}(1 - \alpha)$$

$$\underline{H_1: \mu_X > \mu_0}$$

$$p = 1 - F_{n-1}(t_n)$$

$$\text{Duální IS pro } \mu_X: \left( \bar{X}_n - t_{n-1}(1 - \alpha) \frac{S_n}{\sqrt{n}}, \infty \right)$$

 `t.test(x, mu =  $\mu_0$ , alternative = "greater", conf.level = 1 -  $\alpha$ )`

---

$$H_0: \mu_X \geq \mu_0$$

$$\text{Zamítáme } H_0 \iff T_n \leq -t_{n-1}(1 - \alpha)$$

$$\underline{H_1: \mu_X < \mu_0}$$

$$p = F_{n-1}(t_n)$$

$$\text{Duální IS pro } \mu_X: \left( -\infty, \bar{X}_n + t_{n-1}(1 - \alpha) \frac{S_n}{\sqrt{n}} \right)$$

 `t.test(x, mu =  $\mu_0$ , alternative = "less", conf.level = 1 -  $\alpha$ )`



## Poznámky.

- ▶ t-test nevyžaduje normalitu, pouze konečný druhý moment (a dostatek pozorování).
- ▶ Ověřování normality před použitím t-testu je (kromě jiného) ztrátou času.

## Oddíl **5.3**

# **Jednovýběrový znaménkový test**

# Jednovýběrový znaménkový test

---

- ▶ Porovnání mediánu s pevně danou hodnotou
  - ▶ **Neparametrický** test
- 

Model:  $\mathcal{F} = \{\text{všechna } \underline{\text{spojitá}} \text{ rozdělení}\}$

Testovaný parametr: medián  $m_X := F_X^{-1}(0,5)$

Hypotéza a alternativa:  $H_0: m_X = m_0$

$H_1: m_X \neq m_0$

---

$m_0 \in \mathbb{R}$ : dáno předem

Testová statistika:  $B_n := \sum_{i=1}^n \mathbb{I}\{X_i > m_0\}$

### Věta 5.3 Vlastnosti testové statistiky znaménkového testu.

Nechť  $X_1, \dots, X_n$  je náhodný výběr z libovolného **spojitého** rozdělení s mediánem  $m_X$ . Pak

$$(i) \sum_{i=1}^n \mathbb{I}\{X_i > m_X\} \sim \text{Bi}(n, 1/2).$$

$$(ii) \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \mathbb{I}\{X_i > m_X\} - \frac{1}{2} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1/4), \quad n \rightarrow \infty.$$

### Přesný test:

Zamítáme  $H_0 \iff B_n \leq c_L(\alpha) \vee B_n \geq c_U(\alpha)$

$$p = 2 \min\{G_0(b_n), 1 - G_0(b_n - 1)\}$$

$$\begin{aligned} c_L(\alpha) &= \max\left\{k_1 \in \mathbb{N}_0 : P\left(\text{Bi}\left(n, \frac{1}{2}\right) \leq k_1\right) \leq \frac{\alpha}{2}\right\} \\ &= \max\left\{k_1 \in \mathbb{N}_0 : \frac{1}{2^n} \sum_{j=0}^{k_1} \binom{n}{j} \leq \frac{\alpha}{2}\right\} \end{aligned}$$

$$\begin{aligned} c_U(\alpha) &= \min\left\{k_2 \in \mathbb{N}_0 : P\left(\text{Bi}\left(n, \frac{1}{2}\right) \geq k_2\right) \leq \frac{\alpha}{2}\right\} \\ &= \min\left\{k_2 \in \mathbb{N}_0 : \frac{1}{2^n} \sum_{j=k_2}^n \binom{n}{j} \leq \frac{\alpha}{2}\right\} \end{aligned}$$

**symetrie:**  $c_L(\alpha) = k, \quad c_U(\alpha) = n - k \quad \text{pro nějaké } k \in \{0, \dots, \frac{n}{2}\}$

$G_0$ :            distribuční funkce rozdělení  $\text{Bi}(n, \frac{1}{2})$

$b_n$ :            hodnota statistiky  $B_n$  dosažená/realizovaná s daty

### Asymptotický test:

$$Z_n := \frac{B_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{2}{\sqrt{n}} \left( B_n - \frac{n}{2} \right)$$

$$\text{Zamítáme } H_0 \iff |Z_n| \geq u_{1-\frac{\alpha}{2}}$$

$$p = 2(1 - \Phi(|z_n|))$$

$z_n$ : hodnota statistiky  $Z_n$  dosažená/realizovaná s daty

## Interval spolehlivosti pro $m_X$

Viz empirické odhady.

---

### Poznámky.

►  $B_n = \sum_{i=1}^n \mathbb{I}\{X_i > m_0\}$ :

k výpočtu netřeba znát konkrétní hodnoty  $X_i$ , stačí počet  $> m_0$

► test i intervaly spolehlivosti lze **jednostranně**

► snadno test o libovolném kvantilu  $u_X(\beta)$ ,  $0 < \beta < 1$

$$H_0: u_X(\beta) = u_0$$

$$H_1: u_X(\beta) \neq u_0, \quad u_0 \in \mathbb{R} \text{ dáno}$$

---

$$\text{testová statistika } B_n := \sum_{i=1}^n \mathbb{I}\{X_i > u_0\} \stackrel{H_0}{\sim} Bi(n, 1 - \beta)$$

- ▶ V předpokladech bylo vyžadováno **spojité** rozdělení. Pro platnost všech odvození stačí vyžadovat  $P(X = m_0) = 0$   
(stačí schopnost jednoznačně určit počet pozorování nad/pod  $m_0$ ).
- ▶ Pokud v datech shoda  $X_i$  s  $m_0$  (kvůli zaokrouhlování), vyloučit příslušná pozorování a provést test s menším výběrem.



## Oddíl **5.4**

# **Jednovýběrový Wilcoxonův test**

## ≡ Wilcoxon signed-rank test

---

### Frank Wilcoxon

2.9.1892 (County Cork, IRL)

– 18.11.1965 (Tallahassee, FL, USA)

(fyzikální) chemik (a statistik)



- ▶ V *jistém smyslu* opět test o mediánu, resp. střední hodnotě
  - ▶ **Neparametrický** test
- 

Model:  $\mathcal{F} = \{ \text{spojitá} \text{ rozdělení s hustotou } f \text{ splňující} \\ \exists \delta \in \mathbb{R} : f(\delta - x) = f(\delta + x) \text{ pro } x \in \mathbb{R} \}$

V modelu  $\mathcal{F}$ :  $\delta = F^{-1}(0,5) = \text{medián}_F$   
 $= \mathbb{E}_F X$ ,      pokud  $\mathbb{E}_F X$  existuje

Testovaný parametr:    střed symetrie  $\delta_X$

Hypotéza a alternativa:     $H_0: \delta_X = \delta_0$

$H_1: \delta_X \neq \delta_0$

---

$\delta_0 \in \mathbb{R}$ : dáno předem

## Testová statistika:

$$\blacksquare Z_i := X_i - \delta_0, \quad i = 1, \dots, n$$

$$\blacksquare \text{Seřad' } |Z_i|: 0 < |Z|_{(1)} < \dots < |Z|_{(n)}$$

$$\blacksquare R_i = \text{pořadí } Z_i$$

v uspořádaném výběru  $|Z|_{(1)}, \dots, |Z|_{(n)}$ ,

$$\text{tj. } |Z_i| = |Z|_{(R_i)}$$

$$\blacksquare W_n := \sum_{i \in \mathcal{I}} R_i,$$

$$\mathcal{I} = \{i \in \{1, \dots, n\} : Z_i > 0\}$$

= sum of **rank**s of either **sign**,

odtud *Wilcoxon signed-rank test*

## Věta 5.4 Vlastnosti testové statistiky Wilcoxonova jednovýběrového testu za hypotézy.

Nechť  $X_1, \dots, X_n$  je náhodný výběr z libovolného **spojitého** rozdělení splňujícího model  $\mathcal{F}$  pro jednovýběrový Wilcoxonův test. Nechť platí hypotéza  $\delta_X = \delta_0$ . Pak

$$(i) \quad \mathbb{E}_{\delta_0} W_n = \frac{n(n+1)}{4},$$

$$\text{var}_{\delta_0} W_n = \frac{n(n+1)(2n+1)}{24}.$$

$$(ii) \quad \frac{W_n - \mathbb{E}_{\delta_0} W_n}{\sqrt{\text{var}_{\delta_0} W_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty.$$

### Asymptotický test:

$$U_n := \frac{W_n - \frac{n(n-1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

$$\text{Zamítáme } H_0 \iff |U_n| \geq u_{1-\frac{\alpha}{2}}$$

$$p = 2(1 - \Phi(|u_n|))$$

$u_n$ : hodnota statistiky  $U_n$  dosažená/realizovaná s daty

 `wilcox.test(x, mu =  $\delta_0$ , exact = FALSE)`

ve výstupu:  $V = W_n$

## Přesné rozdělení statistiky $W_n$ za hypotézy

„Snadno“ (trocha kombinatoriky) lze pro konečné  $n$  odvodit přesné rozdělení statistiky  $W_n$  (při  $\delta_X = \delta_0$ ), tj. hodnoty pravděpodobností

$$P_{\delta_0}(W_n = k), \quad k = 0, \dots, \frac{n(n+1)}{2} \quad \sim W_n^0 \text{ s distr. funkcí } G_{n,0}$$

 `dsignrank()`, `psignrank()`, `qsignrank()`

### Přesný test:

Zamítáme  $H_0 \iff W_n \leq c_L(\alpha) \vee W_n \geq c_U(\alpha)$

$$p = 2 \min\{G_{n,0}(w_n), 1 - G_{n,0}(w_n - 1)\}$$

$$c_L(\alpha) = \max\left\{k_1 \in \mathbb{N}_0 : P(W_n^0 \leq k_1) \leq \frac{\alpha}{2}\right\}$$

$$c_U(\alpha) = \min\left\{k_2 \in \mathbb{N}_0 : P(W_n^0 \geq k_2) \leq \frac{\alpha}{2}\right\}$$

$w_n$ : hodnota statistiky  $W_n$  dosažená/realizovaná s daty

 `wilcox.test(x, mu =  $\delta_0$ , exact = TRUE)`

## 1. Shody kvůli zaokrouhlování

- ▶  $X_i = \delta_0 \rightarrow$  vyřadit  
(nelze určit, zda před zaokrouhlením bylo  $X_i < \delta_0$  nebo  $X_i > \delta_0$ )
- ▶  $X_i \neq \delta_0 \rightarrow$  použít **průměrná** pořadí  
(+ drobná úprava teorie)

## 2. Nesymetrie rozdělení (hustota $f$ ne symetrická)

- ▶ jednovýběrový Wilcoxonův test se používá na testování o hodnotě  
$$\delta_X = \text{pseudo-medián}(X)$$
$$= \text{medián}\left(\frac{X_1 + X_2}{2}\right)$$
- ▶  $\delta_X$  leží mezi  $\text{medián}_{F_X}$  a  $\mathbb{E}_{F_X} X$  (existuje-li)
- ▶  $\text{medián}_{F_X} < \mathbb{E}_{F_X} X$ , resp.  $\text{medián}_{F_X} > \mathbb{E}_{F_X} X$  v závislosti na šikmosti rozdělení
- ▶ **Problémy:**
  - ▶ interpretace pseudo-mediánu
  - ▶ skutečná hladina testu není (asymptoticky)  $\alpha$ , ale odchylky nejsou velké ani pro značně asymetrická rozdělení jako např. exponenciální



## Oddíl 5.5

# Jednovýběrový $\chi^2$ test na rozptyl

► Test o rozptylu v normálním výběru

---

Model:  $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$

Testovaný parametr:  $\sigma_X^2 := \text{var}_{F_X} X$

Hypotéza a alternativa:  $H_0: \sigma_X^2 = \sigma_0^2$

$H_1: \sigma_X^2 \neq \sigma_0^2$

---

$\sigma_0^2 \in (0, \infty)$ : dáno předem

Testová statistika:  $U_n := \frac{(n-1) S_n^2}{\sigma_0^2}, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

$$\text{Zamítáme } H_0 \iff U_n \leq \chi_{n-1}^2\left(\frac{\alpha}{2}\right) \vee U_n \geq \chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right)$$

$$p = 2 \min\{G_{n-1}(u_n), 1 - G_{n-1}(u_n)\}$$

$$\chi_{n-1}^2\left(\frac{\alpha}{2}\right): \quad \frac{\alpha}{2}\text{-kvantil rozdělení } \chi_{n-1}^2$$

$$\chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right): \quad \left(1 - \frac{\alpha}{2}\right)\text{-kvantil rozdělení } \chi_{n-1}^2$$

$$G_{n-1}: \quad \text{distribuční funkce rozdělení } \chi_{n-1}^2$$

$$u_n: \quad \text{hodnota statistiky } U_n \text{ dosažená/realizovaná s daty}$$


---

### Duální interval spolehlivosti pro $\sigma_X^2$

$$IS_n = \left( \frac{(n-1) S_n^2}{\chi_{n-1}^2(1 - \alpha/2)}, \frac{(n-1) S_n^2}{\chi_{n-1}^2(\alpha/2)} \right),$$

$$\forall F_X \in \mathcal{F} \quad P_{F_X}(IS_n \ni \sigma_X^2) = 1 - \alpha \quad \text{přesný IS}$$

Lze i jednostranně:

$$H_0: \sigma_X^2 \leq \sigma_0^2$$

$$H_1: \sigma_X^2 > \sigma_0^2$$

---

$$\text{Zamítáme } H_0 \Leftrightarrow U_n \geq \chi_{n-1}^2(1 - \alpha)$$

$$p = 1 - G_{n-1}(u_n)$$

$$\text{Duální IS pro } \sigma_X^2: \left( \frac{(n-1) S_n^2}{\chi_{n-1}^2(1 - \alpha)}, \infty \right)$$

---

$$H_0: \sigma_X^2 \geq \sigma_0^2$$

$$H_1: \sigma_X^2 < \sigma_0^2$$

---

$$\text{Zamítáme } H_0 \Leftrightarrow U_n \leq \chi_{n-1}^2(\alpha)$$

$$p = G_{n-1}(u_n)$$

$$\text{Duální IS pro } \sigma_X^2: \left( 0, \frac{(n-1) S_n^2}{\chi_{n-1}^2(\alpha)} \right)$$

## Poznámky.

- ▶ Při porušení normality výběru nedodrží test hladinu ani asymptoticky!
- ▶ Asymptoticky validní test na rozptyl by bylo možné zkonstruovat na základě (Věta 2.6):

$$\sqrt{n}(S_n^2 - \sigma_X^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_X^4 (\gamma_{X,4} - 1)), \quad n \rightarrow \infty,$$

kde  $\gamma_{X,4} = \frac{\mathbb{E}_{F_X}(X - \mu_X)^4}{\sigma_X^4}$  je špičatost rozdělení  $X$ .

## Oddíl **5.6**

# **Párové testy**

# Párové testy

---

- ▶ Technicky **jednovýběrové** testy.
  - ▶ Odlišná interpretace!
- 

## Nastavení/předpoklady

Náhodný výběr z **dvourozměrného** rozdělení:

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \stackrel{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \begin{pmatrix} X \\ Y \end{pmatrix} \sim F_{XY} \in \mathcal{F}$$

- ▶  $F_X, F_Y$ : marginální rozdělení  $X$ , resp.  $Y$ .
- ▶  $X$  a  $Y$  ne nutně nezávislé.
- ▶ **CÍL**: porovnat  $\theta_X = t(F_X)$  a  $\theta_Y = t(F_Y)$ .
- ▶  $Z_i := X_i - Y_i, i = 1, \dots, n, \quad Z := X - Y.$
- ▶  $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} Z, \quad Z \sim F_Z.$

## Hypotéza nulového efektu

- ▶ Nejčastější použití párových testů.
- ▶  $X \equiv$  měření **před** intervencí/ošetřením/...
- ▶  $Y \equiv$  měření **po** intervenci/ošetření/...
- ▶ Obecně:  $H_0: \forall x \in \mathbb{R} \quad F_X(x) = F_Y(x)$   
 $H_1: \exists x \in \mathbb{R} \quad F_X(x) \neq F_Y(x)$ 

---
- ▶ Konkrétní test obvykle volíme tak, aby byl citlivý na **změnu charakteristiky**, která nás hlavně zajímá.



## Oddíl **5.7**

# **Párový t-test**

≡ jednovýběrový t-test s rozdíly  $Z_1, \dots, Z_n$

Model:  $\mathcal{F}_N = \{Z = X - Y \sim \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$

$\mathcal{F}_{as} = \{Z = X - Y \sim F_X \in \mathcal{L}_+^2\}$

Testované parametry:  $\mu_X := \mathbb{E}_{F_X} X, \quad \mu_Y := \mathbb{E}_{F_Y} Y$

Hypotéza a alternativa:  $H_0: \mu_X - \mu_Y = \delta_0$

$H_1: \mu_X - \mu_Y \neq \delta_0$

---

$\delta_0 \in \mathbb{R}$ : dáno předem

Testová statistika:  $T_n := \frac{\sqrt{n}(\bar{Z}_n - \delta_0)}{S_Z} = \frac{\bar{X}_n - \bar{Y}_n - \delta_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{n} - 2 \frac{S_{XY}}{n}}}$

$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n) (Y_i - \bar{Y}_n)$  (výběrová kovariance)

$$\text{Zamítáme } H_0 \iff |T_n| \geq t_{n-1} \left(1 - \frac{\alpha}{2}\right)$$

$$p = 2(1 - F_{n-1}(|t_n|))$$

$t_{n-1} \left(1 - \frac{\alpha}{2}\right)$ :  $(1 - \frac{\alpha}{2})$ -kvantil rozdělení  $t_{n-1}$

$F_{n-1}$ : distribuční funkce rozdělení  $t_{n-1}$

$t_n$ : hodnota statistiky  $T_n$  dosažená/realizovaná s daty

**Duální interval spolehlivosti pro  $\mu_X - \mu_Y$**

$$IS_n = \left( (\bar{X}_n - \bar{Y}_n) \mp t_{n-1} \left(1 - \frac{\alpha}{2}\right) \frac{S_Z}{\sqrt{n}} \right),$$

$\forall F_{XY} \in \mathcal{F}_N$        $P_{F_{XY}}(IS_n \ni \mu_X - \mu_Y) = 1 - \alpha$       přesný IS

$\forall F_{XY} \in \mathcal{F}_{as}$        $\lim_{n \rightarrow \infty} P_{F_{XY}}(IS_n \ni \mu_X - \mu_Y) = 1 - \alpha$       asymptotický IS

**R**: `t.test(x, y, paired = TRUE, mu =  $\delta_0$ , conf.level =  $1 - \alpha$ )`

## Hypotéza nulového efektu

►  $H_0: \forall x \in \mathbb{R} \quad F_X(x) = F_Y(x)$

$H_1: \exists x \in \mathbb{R} \quad F_X(x) \neq F_Y(x)$

---

- Párový t-test s  $\delta_0 = 0$   $\approx$  test hypotézy nulového efektu  
citlivý na rozdíl ve středních hodnotách.
- Jako test hypotézy nulového efektu **konzistentní**  
proti alternativám, kde  $\mathbb{E}_{F_X} X \neq \mathbb{E}_{F_Y} Y$ .
- Test **ne konzistentní**, pokud  $H_0$  sice neplatí, ale zůstává  $\mathbb{E}_{F_X} X = \mathbb{E}_{F_Y} Y$ .

## Oddíl **5.8**

# **Párový znaménkový test**

# Párový znaménkový test

---

≡ test na významnost **změny** (zlepšení/zhoršení)

---

Model:  $\mathcal{F} = \{Z = X - Y \text{ má jakékoliv } \underline{\text{spojité}} \text{ rozdělení}\}$

Testovaný parametr: medián  $m_Z$  rozdílu  $Z = X - Y$

Hypotéza a alternativa:  $H_0: m_Z = 0$

$H_1: m_Z \neq 0$

---

Testová statistika:  $B_n := \sum_{i=1}^n \mathbb{I}\{Z_i > 0\} = \text{počet párů, kde } X_i > Y_i$

Přesné rozdělení  $B_n$  za  $H_0$ :  $B_n \sim \text{Bi}(n, \frac{1}{2})$

Asymptotické rozdělení  $B_n$  za  $H_0$ :  $\frac{B_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty,$

tj.  $B_n \overset{\text{as}}{\approx} \mathcal{N}\left(\frac{n}{2}, \frac{n}{4}\right)$

## Poznámky.

- ▶  $m_Z = 0$  obecně neznamená  $m_X = m_Y$ !
- ▶  $m_Z = 0 \Leftrightarrow P(X \leq Y) = P(X \geq Y) = \frac{1}{2}$
- ▶ Jako test hypotézy **nulového efektu** je test konzistentní (a citlivý) proti alternativám, kde  $P(X \leq Y) \neq P(X \geq Y)$ .
- ▶ Zobecnění testu s  $H_0: m_Z = m_0$  odpovídá testování  $H_0: P(X - Y \leq m_0) = P(X - Y \geq m_0) = \frac{1}{2}$ .
- ▶ Pokud  $\mathbb{E}|Z| < \infty$  a  $Z$  má hustotu **symetrickou** okolo 0, potom  $m_Z = 0 \Leftrightarrow \mathbb{E}Z = \mathbb{E}X - \mathbb{E}Y = 0$ .
- ▶ K provedení testu není potřeba znát přesné hodnoty  $X_i$  a  $Y_i$ . Stačí vědět, kolikrát je  $X_i > Y_i$  (kolikrát došlo ke zlepšení/zhoršení).

## Oddíl **5.9**

# **Párový Wilcoxonův test**



## Párový Wilcoxonův test (Wilcoxon signed-rank test)

---

≡ porovnání středu symetrie rozdělení  $Z$  s předem danou konstantou

---

Model:  $\mathcal{F} = \{Z \text{ má } \textcolor{red}{\text{spojité}} \text{ rozdělení s hustotou } f \text{ splňující}$   
 $\exists \delta \in \mathbb{R} : f(\delta - x) = f(\delta + x) \text{ pro } x \in \mathbb{R}\}$

Model  $\mathcal{F}$ : požadavek na symetrii hustoty  $Z$ , nikoliv  $X$  a  $Y$ .

Pokud existují stř. hodnoty a  $Z$  má symetrickou hustotu,  
potom  $\delta = \mathbb{E} Z = \mathbb{E} X - \mathbb{E} Y$ .

Testovaný parametr: střed symetrie  $\delta_Z$

Hypotéza a alternativa:  $H_0: \delta_Z = \delta_0$

$H_1: \delta_Z \neq \delta_0$

---

$\delta_0 \in \mathbb{R}$ : dáno předem

### Testová statistika:

$$\Rightarrow Z_i^* := X_i - Y_i - \delta_0, \quad i = 1, \dots, n$$

$$\Rightarrow \text{Seřad' } |Z_i^*|: 0 < |Z^*|_{(1)} < \dots < |Z^*|_{(n)}$$

$$\Rightarrow R_i = \text{pořadí } Z_i^*$$

v uspořádaném výběru  $|Z^*|_{(1)}, \dots, |Z^*|_{(n)}$ ,

$$\text{tj. } |Z_i^*| = |Z^*|_{(R_i)}$$

$$\Rightarrow W_n := \sum_{i \in \mathcal{I}} R_i,$$

$$\mathcal{I} = \{i \in \{1, \dots, n\} : Z_i^* > 0\}$$

Další postup: shodné s jednovýběrovým Wilcoxonovým testem

## Poznámky.

- ▶ Při symetrii rozdělení  $Z$ :

Párový Wilcoxonův test  $\equiv$  test o  $\mathbb{E} X - \mathbb{E} Y$ ,

ale párový t-test je vhodnější (nevyžaduje symetrii rozdělení rozdílů. . . ).

- ▶ Pro  $\delta_0 = 0$  lze test použít k testování hypotézy **nulového efektu**.

# 6

**Dvouvýběrové problémy pro  
kvantitativní data**

## Nastavení/předpoklady

- Kvantitativní data, dva nezávislé náhodné výběry

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X, \quad X \sim F_X,$$

$$Y_1, \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} Y, \quad Y \sim F_Y,$$

$$\mathbb{X} = (X_1, \dots, X_n)^\top \text{ nezávislé s } \mathbb{Y} = (Y_1, \dots, Y_m)^\top$$

- Model  $\mathcal{F} \equiv$  model (předpoklady) pro  $F_X$  a  $F_Y$

- Parametry zájmu:  $\theta_X = t(F_X)$ ,  $\theta_Y = t(F_Y)$ ,

obvykle stejného typu, např.  $\theta_X = \mathbb{E}_{F_X} X$ ,  $\theta_Y = \mathbb{E}_{F_Y} Y$

- Základní problémy:  $H_0: \theta_X = \theta_Y$       resp.  $H_0: \theta_X - \theta_Y = \delta_0$

$$H_1: \theta_X \neq \theta_Y \qquad H_1: \theta_X - \theta_Y \neq \delta_0$$

---

$\delta_0 \in \mathbb{R}$  pevně dané

## Oddíl **6.1**

# **Dvouvýběrový Kolmogorovův-Smirnovův test**

# Dvouvýběrový Kolmogorovův-Smirnovův test

---

- ▶ Globální test shody dvou rozdělení
  - ▶ **Neparametrický** test
- 

Model:  $\mathcal{F} = \{\text{všechna } \underline{\text{spojitá}} \text{ rozdělení, } F_X \text{ i } F_Y\}$

Testovaný parametr: Celé distribuční funkce  $F_X$  a  $F_Y$

Hypotéza a alternativa:  $H_0: \forall x \in \mathbb{R} \quad F_X(x) = F_Y(x)$

$H_1: \exists x \in \mathbb{R} \quad F_X(x) \neq F_Y(x)$

---

$H_0 \equiv$  hypotéza **nulového rozdílu**

Testová statistika:  $K_{n,m} = \sup_{x \in \mathbb{R}} |\hat{F}_X(x) - \hat{F}_Y(x)|$

$\hat{F}_X, \hat{F}_Y$ : empirické distribuční funkce pro  $X$  a  $Y$  výběr

## **Tvrzení 6.1** Asymptotika pro statistiku dvouvýběrového Kolmogorovova-Smirnovova testu.

Nechť  $X_1, \dots, X_n, Y_1, \dots, Y_m$  jsou **nezávislé** náhodné výběry ze **spojitého** rozdělení s distribuční funkcí  $F_0$  ( $= F_X = F_Y$ ). Potom

$$\sqrt{\frac{nm}{n+m}} K_{n,m} \xrightarrow{\mathcal{D}} Z, \quad n, m \rightarrow \infty,$$

kde  $Z$  má distribuční funkci

$$G(z) = \begin{cases} 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 z^2}, & z > 0, \\ 0, & z \leq 0. \end{cases}$$

### **Poznámka.**

Asymptotické rozdělení  $K_{n,m}$  nezávisí na **neznámém** rozdělení  $F_0$ .

**Značení:**  $k_{1-\alpha} = G^{-1}(1 - \alpha)$ ,  $0 < \alpha < 1$ .



## Asymptotický test:

$$\text{Zamítáme } H_0 \iff \sqrt{\frac{nm}{n+m}} K_{n,m} \geq k_{1-\alpha}$$

$$p = 1 - G\left(\sqrt{\frac{nm}{n+m}} k_{n,m}\right)$$

$k_{n,m}$ : hodnota statistiky  $K_{n,m}$  dosažená/realizovaná s daty

---

**R:** `ks.test(x, y, exact = FALSE)`  
`ks.test(z ~ I, exact = FALSE)`

## Poznámky.

- ▶ Pokud  $\exists x \in \mathbb{R} \quad F_X(x) \neq F_Y(x)$ , potom pro  $n, m \rightarrow \infty$

$$\underbrace{\sup_{x \in \mathbb{R}} |\hat{F}_X(x) - \hat{F}_Y(x)|}_{K_{n,m}} \xrightarrow{P} \sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| > 0$$

- ▶ konzistence testu vůči libovolné alternativě
- ▶ test reaguje na libovolné porušení  $H_0$
- ▶ malá síla vůči specifickým porušením  $H_0$  (např.  $\mathbb{E} X \neq \mathbb{E} Y$ )
- ▶ Statistika  $K_{n,m}$  je invariantní vůči prostým transformacím dat, test lze zformulovat též jako pořadový test (*rank test*)
- ▶ **Porušení předpokladů:**
  - ▶ **diskrétní** rozdělení  $\rightarrow$  konzervativní test
  - ▶ **zaokrouhlování:** potřeba předpokládat, že zaokrouhlování probíhá stejným způsobem v obou výběrech

## Oddíl 6.2

# Dvouvýběrový t-test bez předpokladu shody rozptylů

# Dvouvýběrový (Studentův) t-test

## William Sealy Gosset

13.6.1876 (Canterbury, ENG)

– 16.10.1937 (Beaconsfield, ENG)

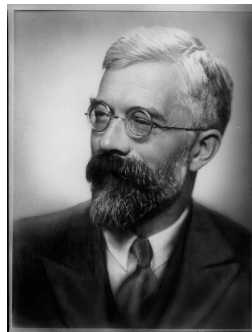
(chemik, pivovarník, statistik,  
hlavní sládek pivovaru Guinness)

## Sir Ronald Aylmer Fisher

17.2.1890 (Londýn, ENG)

– 29.7.1962 (Adelaide, AUS)

(matematik, statistik, biolog,  
genetik, ...)



## Bernard Lewis Welch

1911 (Sunderland, ENG) – 29.12.1989

...

*Welch then attended University College London to study statistics. [Pearson](#) and [Fisher](#) were creating a centre at the College for studies in statistical inference and the use of statistical methods in biological science.*

...

*From 1939 to 1946 Welch served as a Scientific Officer on the Ordnance Board of the Ministry of Supply. He then returned to academic life by way of an appointment to a Readership in Statistics in the then Department of Mathematics in the University of Leeds. [Leeds was then one of the few universities that had a statistician on its mathematical staff.](#) Welch was appointed to the Chair in Statistics in 1968. Following the establishment of the School of Mathematics, he was appointed Head of the newly created Department of Statistics where he remained until his retirement (1976).*

...

*Doctoral student: [Sir David Roxbee Cox \(1924–2022\)](#)*

# Dvouvýběrový t-test bez předpokladu shody rozptylů

- Porovnání **střední hodnoty** dvou výběrů

Model:  $\mathcal{F} = \{F_X \in \mathcal{L}_+^2, F_Y \in \mathcal{L}_+^2\}$

Testované parametry:  $\mu_X := \mathbb{E}_{F_X} X, \mu_Y := \mathbb{E}_{F_Y} Y$

Hypotéza a alternativa:  $H_0: \mu_X - \mu_Y = \delta_0$

$H_1: \mu_X - \mu_Y \neq \delta_0$

---

$\delta_0 \in \mathbb{R}$ : dáno předem

Testová statistika:  $T_{n,m} := \frac{\bar{X}_n - \bar{Y}_m - \delta_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$

Značení:  $\sigma_X^2 := \text{var}_{F_X} X, \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$   
 $\sigma_Y^2 := \text{var}_{F_Y} Y, \bar{Y}_m = \frac{1}{m} \sum_{j=1}^m Y_j, S_Y^2 = \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y}_m)^2.$

## Věta 6.2 Asymptotika pro statistiku dvouvýběrového t-testu bez předpokladu shody rozptylů.

Nechť  $X_1, \dots, X_n$  a  $Y_1, \dots, Y_m$  jsou *nezávislé* náhodné výběry z rozdělení se středními hodnotami  $\mu_X$  a  $\mu_Y$  a konečnými nenulovými rozptyly  $\sigma_X^2$  a  $\sigma_Y^2$ . Pak

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{pro } n, m \rightarrow \infty, \quad \frac{n}{m} \rightarrow q \in (0, \infty).$$

Zamítáme  $H_0 \iff |T_{n,m}| \geq u_{1-\frac{\alpha}{2}}$

$$p = 2(1 - \Phi(|t_{n,m}|))$$

$t_{n,m}$ : hodnota statistiky  $T_{n,m}$  dosažená/realizovaná s daty

---

**Duální interval spolehlivosti** (asymptotický) pro  $\mu_X - \mu_Y$

$$IS_{n,m} = \left( \bar{X}_n - \bar{Y}_m \mp u_{1-\frac{\alpha}{2}} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \right),$$

$$\forall F_X, F_Y \in \mathcal{F} \quad P_{F_X, F_Y}(IS_{n,m} \ni \mu_X - \mu_Y) \longrightarrow 1 - \alpha$$

$$\text{pro } n, m \rightarrow \infty, \quad \frac{n}{m} \rightarrow q \in (0, \infty).$$

► Jednostranné testy a intervaly spolehlivosti analogicky.



## Poznámky.

- ▶ I s předpokladem normality, tj.  $F_X \equiv \mathcal{N}(\mu_X, \sigma_X^2)$ ,  $F_Y \equiv \mathcal{N}(\mu_Y, \sigma_Y^2)$  závisí **přesné** rozdělení  $T_{n,m}$  na **neznámém** poměru  $\sigma_X^2/\sigma_Y^2$ .  
Lepší zůstat u asymptotiky...
- ▶ V praxi se používá (přesnější pro konečné  $n$  a  $m$ ) aproximace rozdělení  $T_{n,m}$  za  $H_0$

$$T_{n,m} \dot{\sim} t_\nu, \quad \nu = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{(S_X^2)^2}{n^2(n-1)} + \frac{(S_Y^2)^2}{m^2(m-1)}},$$

tj. zamítáme  $H_0 \Leftrightarrow |T_{n,m}| \geq t_\nu(1 - \frac{\alpha}{2})$  atd.

### ▶ **Welchův** t-test

- ▶  **R**: `t.test(x, y, mu =  $\delta_0$ , conf.level =  $1 - \alpha$ )`  
`t.test(z ~ I, mu =  $\delta_0$ , conf.level =  $1 - \alpha$ )`

## Oddíl 6.3

# Dvouvýběrový t-test za předpokladu shody rozptylů

# Dvouvýběrový t-test za předpokladu shody rozptylů

- Porovnání **střední hodnoty** dvou výběrů se shodnou variabilitou (homoskedastické výběry)

## Model:

$$\mathcal{F}_N = \{F_X \equiv \mathcal{N}(\mu_X, \sigma^2), F_Y \equiv \mathcal{N}(\mu_Y, \sigma^2), \mu_X, \mu_Y \in \mathbb{R}, \sigma^2 > 0\}$$

$$\mathcal{F}_{as} = \{F_X \in \mathcal{L}_+^2, F_Y \in \mathcal{L}_+^2, \text{var}_{F_X} X = \text{var}_{F_Y} Y =: \sigma^2\}$$

Testované parametry:  $\mu_X := \mathbb{E}_{F_X} X, \mu_Y := \mathbb{E}_{F_Y} Y$

Hypotéza a alternativa:  $H_0: \mu_X - \mu_Y = \delta_0$

$$H_1: \mu_X - \mu_Y \neq \delta_0$$

---

$$\delta_0 \in \mathbb{R}: \text{dáno předem}$$

Testová statistika:  $T_{n,m} := \frac{\bar{X}_n - \bar{Y}_m - \delta_0}{\sqrt{S_{n,m}^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} = \sqrt{\frac{nm}{n+m}} \frac{\bar{X}_n - \bar{Y}_m - \delta_0}{S_{n,m}}$

Značení:  $S_{n,m}^2 = \frac{1}{n+m-2} \left\{ \sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{j=1}^m (Y_j - \bar{Y}_m)^2 \right\}.$

### Věta 6.3 Asymptotika pro statistiku dvouvýběrového t-testu.

Nechť  $X_1, \dots, X_n$  a  $Y_1, \dots, Y_m$  jsou **nezávislé** náhodné výběry z rozdělení se středními hodnotami  $\mu_X$  a  $\mu_Y$  a konečnými nenulovými rozptyly  $\sigma_X^2$  a  $\sigma_Y^2$ . Pak

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_{n,m}^2 \left( \frac{1}{n} + \frac{1}{m} \right)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_*^2),$$

$$\text{pro } n, m \rightarrow \infty, \quad \frac{n}{n+m} \rightarrow \lambda \in (0, \infty),$$

kde  $\sigma_*^2 = \frac{(1-\lambda)\sigma_X^2 + \lambda\sigma_Y^2}{\lambda\sigma_X^2 + (1-\lambda)\sigma_Y^2} \quad (= 1, \text{ pokud } \sigma_X^2 = \sigma_Y^2 \text{ nebo } \lambda = \frac{1}{2}).$

### Věta 6.4 Přesné rozdělení statistiky dvouvýběrového t-testu za normality.

Nechť  $X_1, \dots, X_n$  a  $Y_1, \dots, Y_m$  jsou *homoskedastické nezávislé* náhodné výběry z *normálních* rozdělení  $\mathcal{N}(\mu_X, \sigma^2)$ , resp.  $\mathcal{N}(\mu_Y, \sigma^2)$ . Pak

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_{n,m}^2 \left( \frac{1}{n} + \frac{1}{m} \right)}} \sim t_{n+m-2}.$$

## Testování pro model $\mathcal{F}_{as}$ i $\mathcal{F}_N$

Připomenutí:  $t_{n+m-2}(1 - \frac{\alpha}{2}) \rightarrow u_{1-\frac{\alpha}{2}}, \quad n, m \rightarrow \infty.$

Testujeme  $H_0: \mu_X - \mu_Y = \delta_0$

$H_1: \mu_X - \mu_Y \neq \delta_0$

Zamítáme  $H_0 \iff |T_{n,m}| \geq t_{n+m-2}(1 - \frac{\alpha}{2})$

$p = 2(1 - F_{n+m-2}(|t_{n,m}|))$

$F_{n+m-2}$ : distribuční funkce rozdělení  $t_{n+m-2}$

$t_{n,m}$ : hodnota statistiky  $T_{n,m}$  dosažená/realizovaná s daty

►  `t.test(x, y, mu =  $\delta_0$ , var.equal = TRUE)`  
`t.test(z ~ I, mu =  $\delta_0$ , var.equal = TRUE)`

► Jednostranné testy analogicky.

## Duální interval spolehlivosti pro $\mu_X - \mu_Y$

$$IS_{n,m} = \left( \bar{X}_n - \bar{Y}_m \mp t_{n+m-2} \left(1 - \frac{\alpha}{2}\right) \sqrt{S_{n,m}^2 \left(\frac{1}{n} + \frac{1}{m}\right)} \right).$$

$$\forall F_X, F_Y \in \mathcal{F}_N \quad P_{F_X, F_Y}(IS_{n,m} \ni \mu_X - \mu_Y) = 1 - \alpha,$$

$$\forall F_X, F_Y \in \mathcal{F}_{as} \quad P_{F_X, F_Y}(IS_{n,m} \ni \mu_X - \mu_Y) \longrightarrow 1 - \alpha,$$

$$\text{pro } n, m \rightarrow \infty, \quad \frac{n}{n+m} \rightarrow \lambda \in (0, \infty).$$

►  `t.test(x, y, var.equal = TRUE, conf.level = 1 -  $\alpha$ )`  
`t.test(z ~ I, var.equal = TRUE, conf.level = 1 -  $\alpha$ )`

► Jednostranné intervaly spolehlivosti analogicky.

## Porušení předpokladu shody rozptylů

Za platnosti  $H_0$ :  $T_{n,m} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_*^2)$ ,  $\sigma_*^2 = \frac{(1-\lambda)\sigma_X^2 + \lambda\sigma_Y^2}{\lambda\sigma_X^2 + (1-\lambda)\sigma_Y^2}$ ,

$$n, m \rightarrow \infty, \quad \frac{n}{n+m} \rightarrow \lambda \in (0, 1).$$

$\lambda = \frac{1}{2}$ , tj.  $n \approx m$

$\sigma_*^2 = 1$ , tedy  $T_{n,m} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$

$\Rightarrow$  hladina testu (asymptoticky) OK

$\sigma_*^2 > 1$ , např.  $\sigma_X^2 > \sigma_Y^2$  a  $\lambda < \frac{1}{2}$  ( $n < m$ )

pravděpodobnost chyby I. druhu  $> \alpha$ , anti-konzervativní test

$\sigma_*^2 < 1$

konzervativní test

pokud  $n = m$ :  $\frac{S_X^2}{n} + \frac{S_Y^2}{m} = \dots = S_{n,m}^2 \left( \frac{1}{n} + \frac{1}{m} \right)$

statistiky t-testu **s/bez** předpokladu shodných rozptylů jsou **shodné**  
a mají stejné asymptotické rozdělení za  $H_0$



## Někdo doporučuje dvoukrokový postup

1. Otestuj shodu rozptylů (viz F-test zanedlouho).
2. F-test zamítnul shodu rozptylů  $\Rightarrow$  proved' Welchův t-test  
nezamítnul shodu rozptylů  $\Rightarrow$  proved' t-test předpokládající shodu rozptylů

## NEDĚLAT !!!

- ▶ V každém kroku možnost chyby I. druhu.
- ▶ Chybí kontrola nad pravděpodobností chyby I. druhu pro hlavní testovací problém  $H_0: \mu_X - \mu_Y = \delta_0$ .
- ▶ Pokud různě velké rozsahy výběru a nelze předpokládat shodu rozptylů  $\rightarrow$  rovnou Welchův t-test.

## Hypotéza nulového rozdílu

►  $H_0: \forall x \in \mathbb{R} \quad F_X(x) = F_Y(x)$

$H_1: \exists x \in \mathbb{R} \quad F_X(x) \neq F_Y(x)$

---

- Dvouvýběrový t-test s  $\delta_0 = 0$

≈ test hypotézy nulového rozdílu

citlivý na rozdíl ve středních hodnotách.

- Jako test hypotézy nulového rozdílu konzistentní  
proti alternativám, kde  $\mu_X \neq \mu_Y$ .

## Oddíl **6.4**

# **Dvouvýběrový Wilcoxonův test**

## ≡ Wilcoxon rank-sum, Mannův-Whitneyho test

---

### Henry Berthold Mann

27.10.1905 (Vídeň)

– 1.2.2000 (Tucson, Arizona)



### Donald Random Whitney

27.11.1915 (East Cleveland, Ohio)

– 16.8.2007



► **Neparametrický** test založený na **pořadích**

---

Model:  $\mathcal{F} = \{ \text{existuje rostoucí funkce } g \text{ a existuje } \delta \in \mathbb{R} \text{ tak, že}$   
 $g(X) \sim \tilde{F}_X, \text{ spojitá distribuční funkce,}$   
 $g(Y) \sim \tilde{F}_Y, \text{ spojitá distribuční funkce,}$   
 $\forall x \in \mathbb{R} \quad \tilde{F}_X(x) = \tilde{F}_Y(x - \delta) \}$

$\mathcal{F} \equiv$  (zobecněný) model posunutí v poloze

Testovaný parametr: posunutí  $\delta_{XY}$

Hypotéza a alternativa:  $H_0: \delta_{XY} = 0$

$H_1: \delta_{XY} \neq 0$

---

Lze zobecnit na testování  $H_0: \delta_{XY} = \delta_0$ ,  
kde  $\delta_0 \in \mathbb{R}$  je předem dané.

## Poznámky.

- ▶ Žádné z rozdělení nemusí být symetrické.
- ▶ Za platnosti modelu  $\mathcal{F}$  a nulové hypotézy je  $\mathcal{L}(X) = \mathcal{L}(Y)$  a ted též  $m_X = m_Y$ ,  $\mathbb{E} X = \mathbb{E} Y$  (pokud existují).
- ▶ Dvouvýběrový Wilcoxonův test se obvykle interpretuje jako test **shody mediánů**.

---

### Testová statistika:

$$\Rightarrow \underbrace{(Z_1, \dots, Z_{n+m})}_{\text{spojený výběr}} \equiv (X_1, \dots, X_n, Y_1, \dots, Y_m)$$

$$\Rightarrow Z_{(1)} < Z_{(2)} < \dots < Z_{(n+m)}:$$

pořádkové statistiky ze spojeného výběru

$$\Rightarrow R_i \equiv \text{pořadí } X_i \text{ mezi } Z_{(1)}, \dots, Z_{(n+m)}, \quad \text{tj. } X_i = Z_{(R_i)}$$

$$\Rightarrow W_{n,m} := \sum_{i=1}^n R_i = \text{sum of ranks}$$

## Přesné rozdělení statistiky $W_{n,m}$ za hypotézy

„Snadno“ (trocha kombinatoriky) lze pro konečné  $n$  a  $m$  odvodit přesné rozdělení statistiky  $W_{n,m}$ , tj. hodnoty pravděpodobností

$$P_{\delta_{XY}=0}(W_{n,m} = k), \quad k = \frac{n(n+1)}{2}, \dots, nm + \frac{n(n+1)}{2} \\ \sim W_{n,m}^0 \text{ s distr. funkcí } G_{n,m,0}$$

$$\blacktriangleright P(\text{libovolné uspořádání } Z_1, \dots, Z_{n+m}) = \frac{1}{(n+m)!}$$

$$\blacktriangleright P(R_1 = r_1, \dots, R_n = r_n) = \frac{m!}{(n+m)!},$$

pro každé  $r_1 \neq r_2 \neq \dots \neq r_n \in \{1, \dots, n+m\}$

 `dwilcox()`, `pwilcox()`, `qwilcox()`

Přesný test: `wilcox.test(x, y, exact = TRUE)`

`wilcox.test(z ~ I, exact = TRUE)`

## **Tvrzení 6.5** Vlastnosti statistiky dvouvýběrového Wilcoxonova testu za hypotézy.

Nechť  $X_1, \dots, X_n$  a  $Y_1, \dots, Y_m$  jsou nezávislé náhodné výběry, pro které platí model  $\mathcal{F}$ . Nechť dále platí hypotéza  $H_0: \delta_{XY} = 0$ . Pak

$$(i) \quad \mathbb{E}_{H_0} W_{n,m} = \frac{n(n+m+1)}{2},$$

$$\text{var}_{H_0} W_{n,m} = \frac{nm(n+m+1)}{12}.$$

$$(ii) \quad \text{Pokud } n, m \rightarrow \infty, \quad \frac{W_{n,m} - \mathbb{E}_{H_0} W_{n,m}}{\sqrt{\text{var}_{H_0} W_{n,m}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$



## Asymptotický test:

$$U_{n,m} := \frac{W_{n,m} - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

Zamítáme  $H_0 \iff |U_{n,m}| \geq u_{1-\frac{\alpha}{2}}$

$$p = 2(1 - \Phi(|u_{n,m}|))$$

$u_{n,m}$ : hodnota statistiky  $U_{n,m}$  dosažená/realizovaná s daty

 `wilcox.test(x, y, exact = FALSE)`

`wilcox.test(z ~ I, exact = FALSE)`

ve výstupu:  $W = nm + \frac{n(n+1)}{2} - W_{n,m} =: W_{n,m}^*$   
(Mannovo-Whitheyho vyjádření testové statistiky)

## Testová statistika:

Uvaž všechny dvojice  $(X_i, Y_j)$ ,  $i = 1, \dots, n, j = 1, \dots, m$

$$W_{n,m}^* := \sum_{i=1}^n \sum_{j=1}^m \mathbb{I}\{X_i < Y_j\}$$

= Mannova-Whitneyho statistika, hodnoty  $\in \{0, \dots, nm\}$

## Tvrzení 6.6 Vlastnosti Mannovy-Whitneyho statistiky.

$$(i) \quad W_{n,m} + W_{n,m}^* = nm + \frac{n(n+1)}{2}.$$

(ii) Pokud  $\min(n, m) \rightarrow \infty$ , pak

$$\frac{W_{n,m}^*}{nm} \xrightarrow{P} P(X < Y).$$

**Důsledky**, za platnosti modelu  $\mathcal{F}$  a  $H_0$ :  $\delta_{XY} = 0$ :

$$\blacktriangleright \mathbb{E}_{H_0} W_{n,m}^* = \mathbb{E}_{H_0} (-W_{n,m}) + nm + \frac{n(n+1)}{2} = \frac{nm}{2}.$$

$$\blacktriangleright \text{var}_{H_0} W_{n,m}^* = \text{var}_{H_0} W_{n,m} = \frac{nm(n+m+1)}{12}.$$

$\blacktriangleright$  Pro  $n, m \rightarrow \infty$ :

$$\frac{W_{n,m}^* - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

## Důsledky části (ii) Tvrzení 6.6.

- ▶  $W_{n,m}^*$  je **konzistentním** odhadem parametru  $\theta_{XY} := P(X < Y)$ .
  - ▶ Zřejmě platí  $F_X = F_Y \implies \theta_{XY} = \frac{1}{2}$ .
  - ▶ Za platnosti **zobecněného modelu posunutí** (nikoliv obecně) platí též  $\theta_{XY} = \frac{1}{2} \implies F_X = F_Y$ .
- $\implies$  Wilcoxonův test jako test **hypotézy nulového rozdílu**  $H_0: F_X = F_Y$  je konzistentní vůči alternativám, kde  $\theta_{XY} \neq \frac{1}{2}$ .

## 1. Shody kvůli zaokrouhlování

- ▶ pro  $W_{n,m}$ : použít průměrná pořadí
- ▶ pro  $W_{n,m}^*$ : použít  $W_{n,m}^* = \sum_{i=1}^n \sum_{j=1}^m \left[ \mathbb{I}\{X_i < Y_j\} + \frac{1}{2} \mathbb{I}\{X_i = Y_j\} \right]$
- ▶ korekce asymptotického rozptylu testové statistiky

## 2. Neplatí zobecněný model posunutí

- ▶ (asymptotická) hladina testu OK  
(za  $H_0$  je  $F_X = F_Y$ , tj. zde platí model posunutí)
- ▶ vliv na interpretaci výsledku testu,  
zamítnutí  $H_0: \delta_{XY} = 0$  neznamená nutně, že  $m_X \neq m_Y$ , resp.  $\mathbb{E} X \neq \mathbb{E} Y$
- ▶ vliv na sílu  
(test je konzistentní při platnosti zobecněného modelu posunutí, jinak ne nutně)

**Chceme testovat  $E X = E Y$ ,**

**mnohde** (Žižkov/Albertov/Malá Strana/...) **se doporučuje dvoukrokový postup**

1. Otestuj normalitu výběrů

(např. Shapirův-Wilkův test, : `shapiro.test()`).

2. Test zamítnul normalitu  $\Rightarrow$  proved' Wilcoxonův test  
nezamítnul normalitu  $\Rightarrow$  proved' t-test

**NEDĚLAT !!!**

- ▶ V každém kroku možnost chyby I. druhu.
- ▶ Chybí kontrola nad pravděpodobností chyby I. druhu pro hlavní testovací problém  $H_0: \mathbb{E} X = \mathbb{E} Y$ .
- ▶ **Wilcoxonův** test srovnává střední hodnoty pouze za předpokladu zobecněného modelu posunutí.
- ▶ **Welchův** t-test
  - ▶ Testuje hypotézu, kterou chceme testovat.
  - ▶ Asymptoticky OK
    - i bez normality,
    - i bez shody rozptylů (tj. bez platnosti modelu posunutí)

## Oddíl **6.5**

# **Dvouvýběrový F-test shody rozptylů**



# Dvouvýběrový F-test shody rozptylů

- Porovnání rozptylu dvou nezávislých normálně rozdělených náhodných výběrů

## Model:

$$\mathcal{F} = \{F_X \equiv \mathcal{N}(\mu_X, \sigma_X^2), F_Y \equiv \mathcal{N}(\mu_Y, \sigma_Y^2), \mu_X, \mu_Y \in \mathbb{R}, \sigma_X^2 > 0, \sigma_Y^2 > 0\}$$

Testované parametry:  $\sigma_X^2 = \text{var}_{F_X} X, \sigma_Y^2 = \text{var}_{F_Y} Y$

Hypotéza a alternativa:  $H_0: \sigma_X^2 = \sigma_Y^2$

$$H_1: \sigma_X^2 \neq \sigma_Y^2$$

Testová statistika:  $F := \frac{S_X^2}{S_Y^2}$

Značení:  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$   
 $\bar{Y}_m = \frac{1}{m} \sum_{j=1}^m Y_j, \quad S_Y^2 = \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y}_m)^2.$

Zamítáme  $H_0 \iff F \leq F_{n-1,m-1}\left(\frac{\alpha}{2}\right)$  nebo  $F \geq F_{n-1,m-1}\left(1 - \frac{\alpha}{2}\right)$

$$p = 2 \min\{G(f), 1 - G(f)\}$$

$G$ : distribuční funkce rozdělení  $F_{n-1,m-1}$

$f$ : hodnota statistiky  $F$  dosažená/realizovaná s daty

---

**Duální interval spolehlivosti pro  $\frac{\sigma_X^2}{\sigma_Y^2}$**

$$IS = \left( \frac{S_X^2}{S_Y^2} \frac{1}{F_{n-1,m-1}\left(1 - \frac{\alpha}{2}\right)}, \frac{S_X^2}{S_Y^2} \frac{1}{F_{n-1,m-1}\left(\frac{\alpha}{2}\right)} \right),$$

$$\forall F_X, F_Y \in \mathcal{F} \quad P_{F_X, F_Y} \left( IS \ni \frac{\sigma_X^2}{\sigma_Y^2} \right) = 1 - \alpha.$$

►  `R`: `var.test(x, y, ratio = 1, conf.level = 1 - alpha)`  
`var.test(z ~ I, ratio = 1, conf.level = 1 - alpha)`

► Jednostranné testy a intervaly spolehlivosti analogicky.

## Poznámky.

- ▶ Při porušení předpokladu normality není hladina dodržena ani asymptoticky!
- ▶ Oblíbený alternativní test: **Leveneův** test
  - ▶ testuje trochu jiné parametry variability, nikoliv rozptyly
  - ▶ lze použít k otestování shody rozptylů i ve více než dvou výběrech
  - ▶ : `levene.test()`
- ▶ Beztak není potřeba moc často porovnávat rozptyly...

# 7

**Jednovýběrové a dvouvýběrové  
problémy pro binární data**

## Oddíl **7.1**

# **Jednovýběrový problém**

## Nastavení/předpoklady

### ► Binární data

$$Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} Y, \quad Y \sim \text{Alt}(p_X) \equiv \text{model } \mathcal{F}, \quad 0 < p_X < 1$$

$$\mathbb{Y} := (Y_1, \dots, Y_n)^\top$$

### ► Parametr zájmu: $p_X = P_{F_X}(Y = 1) = \mathbb{E}_{F_X} Y$

### ► Značení a již známá fakta:

$$X_n := \sum_{i=1}^n Y_i \quad \sim \text{Bi}(n, p_X)$$

$$n - X_n = \sum_{i=1}^n (1 - Y_i) \quad \sim \text{Bi}(n, 1 - p_X)$$

$$\hat{p}_n := \frac{X_n}{n} = \bar{Y}_n \equiv \text{nestranný a konzistentní odhad } p_X$$

$$\mathbb{E}_{p_X} \hat{p}_n = p_X$$

$$\text{var}_{p_X} \hat{p}_n = \frac{\text{var}_{p_X} Y}{n} = \frac{p_X(1 - p_X)}{n}$$

- Přesný interval spolehlivosti a test o  $p_X$

Test:  $H_0: p_X = p_0$

$H_1: p_X \neq p_0, \quad 0 < p_0 < 1: \text{dáno předem}$

Testová statistika:  $X_n := \sum_{i=1}^n Y_i$

Kritický obor: zamítní  $H_0 \Leftrightarrow X_n \leq c_L(\alpha)$  nebo  $X_n \geq c_U(\alpha)$

$c_L(\alpha) =$  největší celé číslo splňující

$$P_{p_0}(X_n \leq c_L(\alpha)) = \sum_{j=0}^{c_L(\alpha)} \binom{n}{j} p_0^j (1 - p_0)^{n-j} \leq \frac{\alpha}{2}$$

$c_U(\alpha) =$  nejmenší celé číslo splňující

$$P_{p_0}(X_n \geq c_U(\alpha)) = \sum_{j=c_U(\alpha)}^n \binom{n}{j} p_0^j (1 - p_0)^{n-j} \leq \frac{\alpha}{2}$$

p-hodnota:  $x_n =$  hodnota  $X_n$  v datech

$$p(x_n) = 2 \min \left\{ P_{p_0}(X_n \leq x_n), P_{p_0}(X_n \geq x_n) \right\}$$

$$= 2 \min \left\{ G_0(x_n), 1 - G_0(x_n - 1) \right\},$$

$G_0 =$  distribuční funkce  $Bi(n, p_0)$



## Clopperův-Pearsonův interval spolehlivosti

pro  $p_X$  s pokrytím alespoň  $1 - \alpha$

$IS_n =$  množina  $p \in (0, 1)$ , pro něž test na hladině  $\alpha$  **nezamítá**  $H_0 : p_X = p$

$$G_p(x) := \sum_{j=0}^x \binom{n}{j} p^j (1-p)^{n-j} \equiv \text{distribuční funkce } Bi(n, p)$$

...

Odsud  $IS_n = (p_L, p_U)$ :  $p_L < p_U$  řeší rovnice

$$\sum_{j=0}^{X_n} \binom{n}{j} p^j (1-p)^{n-j} = \frac{\alpha}{2}, \quad \sum_{j=X_n}^n \binom{n}{j} p^j (1-p)^{n-j} = \frac{\alpha}{2}$$

$$p_L = \frac{X_n q_L(\alpha)}{X_n q_L(\alpha) + n - X_n + 1}, \quad p_U = \frac{(X_n + 1) q_U(\alpha)}{(X_n + 1) q_U(\alpha) + n - X_n},$$

$$q_L(\alpha) = \frac{\alpha}{2}\text{-kvantil rozdělení } F_{2X_n, 2(n-X_n+1)},$$

$$q_U(\alpha) = (1 - \frac{\alpha}{2})\text{-kvantil rozdělení } F_{2(X_n+1), 2(n-X_n)},$$

$$p_L = 0, \text{ pokud } X_n = 0, \quad p_U = 1, \text{ pokud } X_n = n$$

► Vyšli jsme z testu s pravděpodobností chyby I. druhu  $\leq \alpha$ .

► To jest,

$$\forall p_x \in (0, 1) \quad P_{p_x}((p_L, p_U) \ni p_x) \geq 1 - \alpha$$

$\Rightarrow$  konzervativní interval spolehlivosti.

► Pokrytí IS je  $\geq 1 - \alpha$  pro libovolný rozsah výběru.

Bylo: 
$$Z_n := \frac{\sqrt{n}(\hat{p}_n - p_X)}{\sqrt{\hat{p}_n(1 - \hat{p}_n)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty.$$

$\Rightarrow$  **asymptotický** test  $H_0: p_X = p_0$   
 $\underline{H_1: p_X \neq p_0}$

Kritický obor: zamítní  $H_0 \Leftrightarrow \left| \frac{\sqrt{n}(\hat{p}_n - p_0)}{\sqrt{\hat{p}_n(1 - \hat{p}_n)}} \right| \geq u_{1 - \frac{\alpha}{2}}$

p-hodnota:  $z_n =$  hodnota  $\left| \frac{\sqrt{n}(\hat{p}_n - p_0)}{\sqrt{\hat{p}_n(1 - \hat{p}_n)}} \right|$  v datech,

$$p(z_n) = 2 \left( 1 - \Phi(|z_n|) \right)$$

## Duální interval spolehlivosti (asymptotický)

$$\forall 0 < p_x < 1 \quad \lim_{n \rightarrow \infty} P_{p_x} \left( \left( \hat{p}_n \mp \sqrt{\frac{\hat{p}_n (1 - \hat{p}_n)}{n}} u_{1-\frac{\alpha}{2}} \right) \ni p_x \right) = 1 - \alpha,$$

⇒ (asymptotický) interval spolehlivosti pro  $p_x$  s pokrytím  $1 - \alpha$

$$\left( \max \left\{ 0, \hat{p}_n - \sqrt{\frac{\hat{p}_n (1 - \hat{p}_n)}{n}} u_{1-\frac{\alpha}{2}} \right\}, \min \left\{ 1, \hat{p}_n + \sqrt{\frac{\hat{p}_n (1 - \hat{p}_n)}{n}} u_{1-\frac{\alpha}{2}} \right\} \right)$$

### Nevýhody:

► Pro  $p_x \rightarrow 0$  a  $p_x \rightarrow 1$  pomalá asymptotika (potřeba velké  $n$ ).

► Rule of thumb:

$$\text{dostatečně velké } n \Leftrightarrow np_x \geq 5 \text{ \& } n(1 - p_x) \geq 5.$$

**Poznámka.** Hypotézy  $H_0: p_X = p_0$

$$H_1: p_X \neq p_0$$

Ize testovat též pomocí **jednovýběrového t-testu** (asymptotická verze).

**SDC:** Ukažte, že příslušná testová statistika má tvar

$$T_n = \frac{\sqrt{n-1} (\hat{p}_n - p_0)}{\sqrt{\hat{p}_n (1 - \hat{p}_n)}}.$$

S kvantily  $t_{n-1}$ : mírně konzervativnější test.

Bylo též: 
$$W_n := \frac{\sqrt{n}(\hat{p}_n - p_X)}{\sqrt{p_0(1-p_0)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty.$$

$\Rightarrow$  **asymptotický** test  $H_0: p_X = p_0$   
 $H_1: p_X \neq p_0$

Kritický obor: zamítní  $H_0 \Leftrightarrow \left| \frac{\sqrt{n}(\hat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}} \right| \geq u_{1-\frac{\alpha}{2}}$

p-hodnota:  $w_n = \text{hodnota } \left| \frac{\sqrt{n}(\hat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}} \right|$  v datech,

$$p(w_n) = 2 \left( 1 - \Phi(|w_n|) \right)$$

$\Rightarrow$  **Wilsonův** test

## Duální interval spolehlivosti (asymptotický)

$$IS_n = \left\{ p \in (0, 1) : \left| \frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \right| < u_{1-\frac{\alpha}{2}} \right\}$$

- ▶ potřeba řešit  $\sqrt{n}|\hat{p}_n - p| < \sqrt{p(1-p)} u_{1-\frac{\alpha}{2}}$
- ▶ kvadratická rovnice  $n(\hat{p}_n - p)^2 = p(1-p) u_{1-\frac{\alpha}{2}}^2$

## Wilsonův interval spolehlivosti pro $p_X$

$$IS_n = \left( \left( \hat{p}_n + \frac{u_{1-\frac{\alpha}{2}}^2}{2n} \mp \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{u_{1-\frac{\alpha}{2}}^2}{4n^2}} \right) \frac{1}{1 + \frac{u_{1-\frac{\alpha}{2}}^2}{n}} \right)$$

- ▶ Lepší vlastnosti (pokrytí pro konečné  $n$ ) než klasický asymptotický interval spolehlivosti.

- ▶ **Asymptotika** v předchozích částech založena na odhadu/statistice  $\hat{p}_n \in [0, 1]$ .
- ▶ Rozdělení statistiky  $\hat{p}_n$  aproximujeme rozdělením  $\mathcal{N}(\cdot, \cdot)$  s nosičem  $\mathbb{R}$ .
- ▶ Vylepšení:
  - ▶ Odhaduj  $\theta_X = t(p_X) \in \mathbb{R}$ ,
  - ▶ odhadem/statistikou s nosičem  $\mathbb{R}$ ,
  - ▶ jehož rozdělení je aproximováno rozdělením na  $\mathbb{R}$ .



Odhadovaný parametr:  $\theta_X = \log \frac{p_X}{1 - p_X}$  log-šance (*log-odds*)

Konzistentní odhad  $\theta_X$ :  $\hat{\theta}_n = \log \frac{\hat{p}_n}{1 - \hat{p}_n}$

$\vdots$

**Asymptotický interval spolehlivosti pro  $\theta_X$ :**

$$\begin{aligned} (\theta_{L,n}, \theta_{U,n}), \quad \theta_{L,n} &= \hat{\theta}_n - u_{1-\frac{\alpha}{2}} D_n, & D_n &= \frac{1}{\sqrt{n \hat{p}_n (1 - \hat{p}_n)}} \\ \theta_{U,n} &= \hat{\theta}_n + u_{1-\frac{\alpha}{2}} D_n, \end{aligned}$$

**Asymptotický interval spolehlivosti pro  $p_X$ :**

$$\forall 0 < p_X < 1 \quad \lim_{n \rightarrow \infty} P_{p_X} \left( \left( \frac{e^{\theta_{L,n}}}{1 + e^{\theta_{L,n}}}, \frac{e^{\theta_{U,n}}}{1 + e^{\theta_{U,n}}} \right) \ni p_X \right) = 1 - \alpha$$

## Oddíl **7.2**

# **Dvouvýběrový problém**

## Nastavení/předpoklady

- ▶ Binární data, dva nezávislé náhodné výběry

$$Y_{1,1}, \dots, Y_{1,n} \stackrel{\text{i.i.d.}}{\sim} \mathcal{Alt}(p_1),$$

$$Y_{2,1}, \dots, Y_{2,m} \stackrel{\text{i.i.d.}}{\sim} \mathcal{Alt}(p_2),$$

$$\mathbb{Y}_1 = (Y_{1,1}, \dots, Y_{1,n})^\top \text{ nezávislé s } \mathbb{Y}_2 = (Y_{2,1}, \dots, Y_{2,m})^\top$$

- ▶  $X_1 := \sum_{i=1}^n Y_{1,i} \sim \mathcal{Bi}(n, p_1),$

$$X_2 := \sum_{i=1}^m Y_{2,i} \sim \mathcal{Bi}(m, p_2), \quad X_1 \text{ nezávislé s } X_2$$

- ▶ Cíl: porovnat  $p_1$  a  $p_2$

- ▶ Pokud  $Y = 1 \equiv$  negativní událost

$$\longrightarrow p_1, p_2: \text{riziko události}$$

- ▶ Odhady (viz dříve):  $\hat{p}_1 = \frac{X_1}{n}, \quad \hat{p}_2 = \frac{X_2}{n}$

**Příklad.** Registrační studie genové terapie Comirnaty, děti 6–23 měsíců,  
2 dávky terapie

$Y = 1 \quad \equiv \quad$  PCR detekce RNA viru SARS-CoV-2 do jisté doby od terapie

| Aktivní ošetření    | Placebo             |
|---------------------|---------------------|
| $n = 1\,178$        | $m = 598$           |
| $x_1 = 83$          | $x_2 = 51$          |
| $\hat{p}_1 = 0,070$ | $\hat{p}_2 = 0,085$ |

## Kvantifikace odlišnosti pravděpodobností/rizik

### 1. Rozdíl pravděpodobností (změna rizika)

$$d_X := p_1 - p_2 \in [-1, 1],$$

$$\hat{d} = \hat{p}_1 - \hat{p}_2 = -0,015$$

### 2. Podíl pravděpodobností (relativní riziko)

$$r_X := \frac{p_1}{p_2} \in [0, \infty],$$

$$\hat{r} = \frac{\hat{p}_1}{\hat{p}_2} = 0,826$$

### 3. Poměr šancí (odds ratio) $o_X := \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} \in [0, \infty],$

$$\hat{o} = \frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}} = 0,813$$

## **Tvrzení 7.1** Asymptotika pro odhady veličin kvantifikujících odlišnost pravděpodobností.

Nechť  $p_1, p_2 \in (0, 1)$ ,  $n \rightarrow \infty$ ,  $m \rightarrow \infty$ ,  $\frac{n}{m} \rightarrow q \in (0, \infty)$ . Potom

$$(i) \quad \frac{\hat{d} - d_X}{\sqrt{\hat{V}_d}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{kde} \quad \hat{V}_d = \frac{\hat{p}_1(1 - \hat{p}_1)}{n} + \frac{\hat{p}_2(1 - \hat{p}_2)}{m}.$$

$$(ii) \quad \frac{\log(\hat{r}) - \log(r_X)}{\sqrt{\hat{V}_r}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{kde} \quad \hat{V}_r = \frac{1 - \hat{p}_1}{n\hat{p}_1} + \frac{1 - \hat{p}_2}{m\hat{p}_2}.$$

$$(iii) \quad \frac{\log(\hat{o}) - \log(o_X)}{\sqrt{\hat{V}_o}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

$$\text{kde} \quad \hat{V}_o = \frac{1}{n\hat{p}_1} + \frac{1}{n(1 - \hat{p}_1)} + \frac{1}{m\hat{p}_2} + \frac{1}{m(1 - \hat{p}_2)}.$$

## Testování shody pravděpodobností

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

$$H_0: d_X = 0$$

$$H_1: d_X \neq 0$$

$$H_0: r_X = 1$$

$$H_1: r_X \neq 1$$

$$H_0: o_X = 1$$

$$H_1: o_X \neq 1$$

Zamítej  $H_0 \Leftrightarrow$

$$\left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{V}_d}} \right| \geq u_{1-\frac{\alpha}{2}}$$

$$\left| \frac{\log\left(\frac{\hat{p}_1}{\hat{p}_2}\right)}{\sqrt{\hat{V}_r}} \right| \geq u_{1-\frac{\alpha}{2}}$$

$$\left| \frac{\log\left(\frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}}\right)}{\sqrt{\hat{V}_o}} \right| \geq u_{1-\frac{\alpha}{2}}$$

- ▶ p-hodnoty jako vždy...
- ▶ Lze i jednostranně...

## Duální intervaly spolehlivosti

Jako vždy..., např. pro **poměr šancí**  $o_X = \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right)$ :

$$\frac{\log(\hat{o}) - \log(o_X)}{\sqrt{\hat{V}_o}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty, m \rightarrow \infty, \frac{n}{m} \rightarrow q \in (0, \infty).$$

Tedy  $\forall p_1, p_2 \in (0, 1)$

$$P_{p_1, p_2} \left( \left| \frac{\log(\hat{o}) - \log(o_X)}{\sqrt{\hat{V}_o}} \right| < u_{1-\frac{\alpha}{2}} \right) \rightarrow 1 - \alpha$$

$$P_{p_1, p_2} \left( \left( \log(\hat{o}) \mp \sqrt{\hat{V}_o} u_{1-\frac{\alpha}{2}} \right) \ni \log(o_X) \right) \rightarrow 1 - \alpha$$

$$P_{p_1, p_2} \left( \left( \hat{o} \exp\left(-\sqrt{\hat{V}_o} u_{1-\frac{\alpha}{2}}\right), \hat{o} \exp\left(\sqrt{\hat{V}_o} u_{1-\frac{\alpha}{2}}\right) \right) \ni o_X \right) \rightarrow 1 - \alpha$$

**Příklad.** Registrační studie genové terapie Comirnaty, děti 6–23 měsíců,  
2 dávky terapie

$Y = 1 \quad \equiv \quad$  PCR detekce RNA viru SARS-CoV-2 do jisté doby od terapie

Aktivní ošetření:  $n = 1\,178$ ,  $x_1 = 83$ ,  $\hat{p}_1 = 0,070$

Placebo:  $m = 598$ ,  $x_2 = 51$ ,  $\hat{p}_2 = 0,085$

| Parametr | Odhad  | 95% Interval spolehlivosti | p-hodnota |
|----------|--------|----------------------------|-----------|
| $d_X$    | -0,015 | (-0,042, 0,012)            | 0,277     |
| $r_X$    | 0,826  | (0,591, 1,154)             | 0,263     |
| $o_X$    | 0,813  | (0,565, 1,169)             | 0,264     |

Jednostranně s alternativou  $H_1: p_1 < p_2$

| Parametr | Odhad  | 95% Interval spolehlivosti | p-hodnota |
|----------|--------|----------------------------|-----------|
| $d_X$    | -0,015 | (-1, 0,008)                | 0,139     |
| $r_X$    | 0,826  | (0, 1,094)                 | 0,132     |
| $o_X$    | 0,813  | (0, 1,103)                 | 0,132     |



**Příklad.** Registrační studie genové terapie Comirnaty, děti 6–23 měsíců,  
2 dávky terapie

$Y = 1 \quad \equiv \quad$  Hospitalizace do jisté doby od terapie

---

Aktivní ošetření:  $n = 1\,178$ ,  $x_1 = 3$ ,  $\hat{p}_1 = 0,003$

Placebo:  $m = 598$ ,  $x_2 = 0$ ,  $\hat{p}_2 = 0$

---

### Použitelnost asymptotických výsledků:

- ▶ Pro  $p_1 \rightarrow 0$ ,  $p_1 \rightarrow 1$ ,  $p_2 \rightarrow 0$ ,  $p_2 \rightarrow 1$  je asymptotika pomalá  
(potřeba velké  $n$ , resp.  $m$ ).

- ▶ Rule of thumb:

dostatečně velké  $n \quad \Leftrightarrow \quad np_1 \geq 5 \text{ \& } n(1 - p_1) \geq 5$ ,

dostatečně velké  $m \quad \Leftrightarrow \quad mp_2 \geq 5 \text{ \& } m(1 - p_2) \geq 5$ .

# 8

## **Multinomické rozdělení a kontingenční tabulky**

## Nastavení/předpoklady

Kategoriální data, náhodná veličina  $Z \sim \text{Discr}(1, \dots, K)$

$$P(Z = k) = p_k, \quad k = 1, \dots, K,$$

$$0 < p_k < 1, \quad k = 1, \dots, K, \quad \sum_{k=1}^K p_k = 1$$

---

Data:  $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} Z$

Četnosti (*counts*):  $X_1 := \sum_{i=1}^n \mathbb{I}\{Z_i = 1\} \sim \text{Bi}(n, p_1)$

$\vdots$

$X_K := \sum_{i=1}^n \mathbb{I}\{Z_i = K\} \sim \text{Bi}(n, p_K)$

►  $X_1, \dots, X_K$  nezávislé!      $\mathbf{X} := (X_1, \dots, X_K)^\top$

► Zřejmě, pro  $x_1, \dots, x_K \in \mathbb{N}_0$ ,  $\sum_{k=1}^K x_k = n$

$$P(X_1 = x_1, \dots, X_K = x_K) = \frac{n!}{x_1! \dots x_K!} p_1^{x_1} \dots p_K^{x_K}$$

## Oddíl 8.1

# Multinomické rozdělení

## Definice 8.1 Multinomické rozdělení.

Nechť  $K \geq 2$ ,  $n \geq 1$ ,  $\mathbf{p} = (p_1, \dots, p_K)^\top$  tak, že  $0 < p_k < 1$ ,  $k = 1, \dots, K$ ,  $\sum_{k=1}^K p_k = 1$ . Náhodný vektor  $\mathbf{X} = (X_1, \dots, X_K)^\top$  má **multinomické rozdělení** (*multinomial distribution*)  $\text{Mult}_K(n, \mathbf{p})$ , právě když jeho hustota vzhledem k součinové míře na  $\mathbb{Z}^K$  je

$$P(X_1 = x_1, \dots, X_K = x_K) = \begin{cases} \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}, & \text{pokud } \sum_{k=1}^K x_k = n, x_k \in \mathbb{N}_0, \\ 0, & \text{jinak.} \end{cases}$$

## Věta 8.1 Rozklad multinomického rozdělení.

Nechť  $\mathbf{Y}_1, \dots, \mathbf{Y}_n \stackrel{i.i.d.}{\sim} \mathbf{Y}$ , kde  $\mathbf{Y} \sim \text{Mult}_K(1, \mathbf{p})$ .

Pak  $\mathbf{X} := \sum_{i=1}^n \mathbf{Y}_i \sim \text{Mult}_K(n, \mathbf{p})$ .

## Věta 8.2 Vlastnosti multinomického rozdělení.

Nechť  $\mathbf{X} = (X_1, \dots, X_K)^\top \sim \text{Mult}_K(n, \mathbf{p})$ . Pak

- (i)  $X_k \sim \text{Bi}(n, p_k), \quad k = 1, \dots, K.$
- (ii)  $\mathbb{E} X_k = n p_k, \quad \text{var } X_k = n p_k (1 - p_k), \quad k = 1, \dots, K.$
- (iii)  $\text{cov}(X_j, X_k) = -n p_j p_k, \quad j \neq k.$
- (iv)  $\text{var } \mathbf{X} = n \{\text{diag}(\mathbf{p}) - \mathbf{p} \mathbf{p}^\top\}.$

### Věta 8.3 Asymptotické vlastnosti multinomického rozdělení.

Nechť  $\mathbf{X}_n = (X_{n,1}, \dots, X_{n,K})^\top \sim \text{Mult}_K(n, \mathbf{p})$ . Pak

$$(i) \frac{1}{\sqrt{n}} (\mathbf{X}_n - n\mathbf{p}) \xrightarrow{\mathcal{D}} \mathcal{N}_K(\mathbf{0}, \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top), \quad n \rightarrow \infty.$$

$$(ii) \sum_{k=1}^K \frac{(X_{n,k} - np_k)^2}{np_k} \xrightarrow{\mathcal{D}} \chi_{K-1}^2, \quad n \rightarrow \infty.$$

Levá strana z bodu (ii) napsaná (pro dané  $n \equiv$  analyzovaná data) jinak:

$$\sum_{k=1}^K \frac{(O_k - E_k)^2}{E_k}, \text{ kde}$$

$O_k = X_{n,k}$ : pozorovaná četnost (**observed count**) kategorie  $k$

$E_k = np_k$ : očekávaná četnost (**expected count**) kategorie  $k$

# Odhady parametrů multinomického rozdělení

Data:  $\mathbf{X}_n = (X_{n,1}, \dots, X_{n,K})^\top$ ,  $X_{n,k} \sim \text{Bi}(n, p_k)$ ,  $k = 1, \dots, K$ .

Bylo (konzistentní a nestranný odhad):  $\hat{p}_k = \frac{X_{n,k}}{n}$ ,  $k = 1, \dots, K$ .

Konzistentní a nestranný odhad vektoru  $\mathbf{p}$ :  $\hat{\mathbf{p}}_n = \frac{\mathbf{X}_n}{n}$ .

⋮

$$\forall \mathbf{c} \in \mathbb{R}^K \quad \sqrt{n}(\mathbf{c}^\top \hat{\mathbf{p}}_n - \mathbf{c}^\top \mathbf{p}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_c), \quad n \rightarrow \infty,$$

$$\text{kde } V_c = \mathbf{c}^\top \{\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top\} \mathbf{c}$$

$$\hat{V}_c := \mathbf{c}^\top \{\text{diag}(\hat{\mathbf{p}}_n) - \hat{\mathbf{p}}_n \hat{\mathbf{p}}_n^\top\} \mathbf{c} \xrightarrow{\text{P}} V_c, \quad n \rightarrow \infty$$

$$\text{pokud } V_c > 0 \quad \frac{\sqrt{n}(\mathbf{c}^\top \hat{\mathbf{p}}_n - \mathbf{c}^\top \mathbf{p})}{\sqrt{\hat{V}_c}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty$$



## Asymptotický test pro zadané $\mathbf{c} \in \mathbb{R}^K$ a $\gamma_0 \in \mathbb{R}$

$$H_0: \mathbf{c}^\top \mathbf{p} = \gamma_0$$

$$H_1: \mathbf{c}^\top \mathbf{p} \neq \gamma_0$$

$$\text{Zamítej } H_0 \text{ na hladině } \alpha \iff \left| \frac{\sqrt{n}(\mathbf{c}^\top \hat{\mathbf{p}}_n - \gamma_0)}{\sqrt{\hat{V}_c}} \right| \geq u_{1-\frac{\alpha}{2}}$$

(Duální) **asymptotický interval spolehlivosti** pro  $\mathbf{c}^\top \mathbf{p}$  s pokrytím  $1 - \alpha$

$$\left( \mathbf{c}^\top \hat{\mathbf{p}}_n \mp \sqrt{\frac{\hat{V}_c}{n}} u_{1-\frac{\alpha}{2}} \right)$$

Inference pro  $\theta = g(\mathbf{p})$ , kde  $g$  dostatečně hladká

►  $\Delta$ -metoda

► pokud  $V(\mathbf{p}) = \left( \frac{\partial g}{\partial \mathbf{p}} \right)^\top \left\{ \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top \right\} \frac{\partial g}{\partial \mathbf{p}} > 0$

## $\chi^2$ -test dobré shody pro multinomické rozdělení

Data (četnosti):  $\mathbf{X} = (X_1, \dots, X_K)^\top$ ,  $\mathbf{X} \sim \mathcal{Mult}(n, \mathbf{p})$

$$H_0: \mathbf{p} = \mathbf{p}^0, \quad \mathbf{p}^0 = (p_1^0, \dots, p_K^0)^\top$$

$$H_1: \mathbf{p} \neq \mathbf{p}^0 \quad \text{předem zadané pravděpodobnosti}$$

Testová statistika:

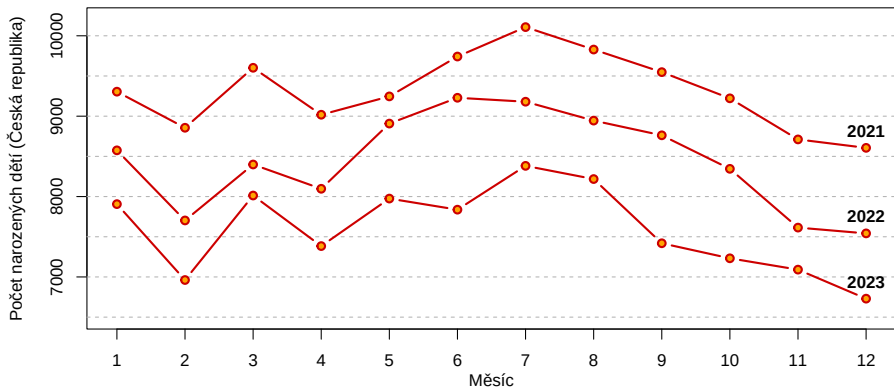
$$\begin{aligned} \chi^2 &= \sum_{k=1}^K \frac{(X_k - np_k^0)^2}{np_k^0} = \sum_{k=1}^K \left( \frac{X_k - np_k^0}{\sqrt{np_k^0}} \right)^2 \\ &= \sum_{k=1}^K \left( \frac{O_k - E_k}{\sqrt{E_k}} \right)^2 \quad \underset{\text{as.}}{\sim} \chi_{K-1}^2 \end{aligned}$$

Zamítej  $H_0$  na hladině  $\alpha \iff \chi^2 \geq \chi_{K-1}^2(1 - \alpha)$

p-hodnota =  $1 - G_{K-1}(s_X)$ ,

$s_X$  = napoz. hodn. test. stat.,  $G_{K-1}$ : distr. funkce rozděl.  $\chi_{K-1}^2$

## Příklad. Rodí se děti během roku rovnoměrně?



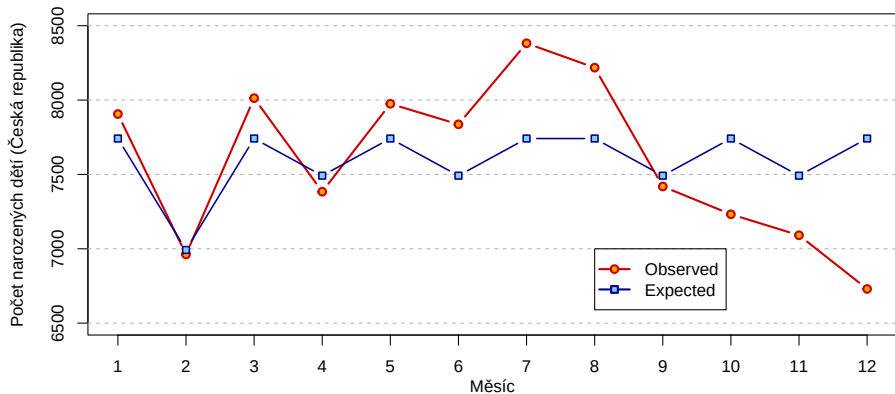
Pro vybraný rok:  $X_1, X_2, \dots, X_{12} \equiv$

počet narozených dětí v lednu, únoru,  $\dots$ , prosinci

$$p_1^0 = \frac{31}{365}, p_2^0 = \frac{28}{365}, \dots, p_{12}^0 = \frac{31}{365}$$

 `chisq.test(x, p =  $p^0$ )`

Rok 2023:



$$\chi^2 = 307,8$$

$$\chi^2_{11}(0,95) = 19,7$$

$$p\text{-hodnota} < 0,001$$

## Poznámky.

- **Asymptotika** (rychlost konvergence k rozdělení  $\chi_{K-1}^2$ )

Rule of thumb: aproximace rozdělením  $\chi_{K-1}^2$  uspokojivá,  
pokud  $\forall k = 1, \dots, K \quad E_k = n p_k^0 \geq 5$ .

- $K = 2$ :  $p_0 := p_1^0, \quad p_2^0 = 1 - p_0, \quad X_2 = n - X_1$

$$\chi^2 = \frac{(X_1 - n p_0)^2}{n p_0} + \frac{(n - X_1 - n(1 - p_0))^2}{n(1 - p_0)} = W_n^2,$$

$$W_n = \frac{\sqrt{n}(\hat{p}_n - p_0)}{\sqrt{p_0(1 - p_0)}}, \quad \hat{p}_n = \frac{X_1}{n}$$

$W_n$ : **Wilsonova** testová statistika pro test  $H_0: p_X = p_0$  ve výběru z  $\mathcal{Alt}(p_X)$

Za platnosti  $H_0$ :  $W_n \stackrel{\text{as.}}{\approx} \mathcal{N}(0, 1), \quad \chi^2 = W_n^2 \stackrel{\text{as.}}{\approx} \chi_1^2$

# $\chi^2$ -test dobré shody pro multinomické rozdělení s odhadnutými parametry

---

## Motivační příklad

$Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} Z$  (ne nutně kategoriální data)

$H_0: Z \sim F_0(z) = F(z; \theta_0), \quad \theta_0: \text{známý parametr}$

---

## Test dobré shody

- ▶ intervaly  $(a_{k-1}, a_k]$ ,  $k = 1, \dots, K$ ,  $a_0 := -\infty$ ,  $a_K := \infty$ ,  $K \ll n$
- ▶  $X_k := \sum_{i=1}^n \mathbb{I}\{Z_i \in (a_{k-1}, a_k]\}$ ,  $k = 1, \dots, K$
- ▶  $H_0: (X_1, \dots, X_K)^\top \sim \text{Mult}_K(n, \mathbf{p}_0)$ , kde  $p_k^0 = F(a_k; \theta_0) - F(a_{k-1}; \theta_0)$
- ▶ zamítnutí  $H_0 \equiv$  výběr  $Z_1, \dots, Z_n$  nepochází z rozdělení s distribuční funkcí  $F_0 \equiv F(\cdot; \theta_0)$

Co dělat, pokud  $\theta_0$  neznámé, např.  $F(z; \theta_0) \equiv \mathcal{N}(\mu_0, \sigma_0^2)$ ?

## Obecně

Model  $\mathcal{F}_0$ :  $\mathbf{X} = (X_1, \dots, X_K)^\top \sim \text{Mult}_K(n, \mathbf{p}(\theta_X))$

- ▶  $\theta_X \in \Theta \subset \mathbb{R}^d$ : neznámý parametr,  $d < K - 1$
- ▶  $\mathbf{p}: \Theta \rightarrow (0, 1)^K$  splňující  $\forall \theta \in \Theta \quad \mathbf{1}_K^\top \mathbf{p}(\theta) = 1$
- ▶ Testujeme  $H_0$ : platí model  $\mathcal{F}_0$

Krok 1: odhad  $\theta_X$

metoda maximální věrohodnosti (ML, *maximum likelihood*)

v **multinomickém modelu**

$$\text{MLE } \hat{\theta}_n \text{ splňuje} \quad \begin{matrix} \vdots \\ \sum_{k=1}^K \frac{X_k}{p_k(\hat{\theta}_n)} \cdot \frac{\partial p_k(\hat{\theta}_n)}{\partial \theta} = \mathbf{0}_d \end{matrix}$$

soustava věrohodnostních (odhadovacích) rovnic

Krok 2: testování

$H_0: \exists \theta_X \in \Theta \quad \mathbf{p} = \mathbf{p}(\theta_X) \quad (\text{platí model } \mathcal{F}_0)$

$H_0: \forall \theta_X \in \Theta \quad \mathbf{p} \neq \mathbf{p}(\theta_X) \quad (\text{model } \mathcal{F}_0 \text{ neplatí})$

---

Testová statistika

$$\chi^2 = \sum_{k=1}^K \frac{\left(X_k - n p_k(\hat{\theta}_n)\right)^2}{n p_k(\hat{\theta}_n)}$$



**Tvrzení 8.4** Asymptotické rozdělení  $\chi^2$  statistiky při neznámých parametrech.

---

*Platí-li hypotéza  $H_0$ , má testová statistika  $\chi^2$  (za jistých předpokladů regularity) asymptoticky rozdělení  $\chi^2_{K-d-1}$ .*

Viz skórový test v teorii maximální věrohodnosti (*Matematická statistika 2*).

## Poznámky

- ▶ Za  $H_0$ :  $\mathbb{E} X_k = n p_k(\theta_X)$  (pro nějaké  $\theta_X \in \Theta$ )
- ▶ Označme  $O_k := X_k$ ,  $\hat{E}_k := n p_k(\hat{\theta}_n)$ ,  $k = 1, \dots, K$
- ▶ Testová statistika:

$$\chi^2 = \sum_{k=1}^K \frac{(O_k - \hat{E}_k)^2}{\hat{E}_k} \stackrel{\text{as.}}{\sim} \chi_{K-d-1}^2 \quad \text{za } H_0$$

- ▶  $H_0$  zamítáme  $\Leftrightarrow \chi^2 \geq \chi_{K-d-1}^2(1 - \alpha)$
- ▶ Použitelnost asymptotiky: rule of thumb  $\hat{E}_k \geq 5 \quad \forall k$

**(test dobré shody s parametrickým rozdělením)**

$Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} Z$  (ne nutně kategoriální data)

$H_0: Z \sim F_X(z) = F(z; \theta_X), \quad \theta_X \in \Theta$ : **neznámý** parametr

dále obdobně jako u testu dobré shody se známými parametry:

► intervaly  $(a_{k-1}, a_k]$ ,  $k = 1, \dots, K$ ,  $a_0 := -\infty$ ,  $a_K := \infty$ ,  $K \ll n$

►  $X_k := \sum_{i=1}^n \mathbb{I}\{Z_k \in (a_{k-1}, a_k]\}$ ,  $k = 1, \dots, K$

►  $H_0: (X_1, \dots, X_K)^\top \sim \text{Mult}_K(n, \mathbf{p}(\theta_X))$ ,

kde  $p_k(\theta_X) = F(a_k; \theta_X) - F(a_{k-1}; \theta_X)$

►  $\hat{\theta}_n$  řeší  $\sum_{k=1}^K \frac{X_k}{p_k(\hat{\theta}_n)} \cdot \frac{\partial p_k(\hat{\theta}_n)}{\partial \theta} = \mathbf{0}_d$

► testová statistika  $\chi^2 = \sum_{k=1}^K \frac{(X_k - n p_k(\hat{\theta}_n))^2}{n p_k(\hat{\theta}_n)} \stackrel{\text{as.}}{\sim} \chi_{K-d-1}^2$  (za  $H_0$ )

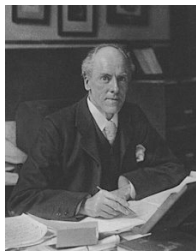
## Karl Pearson

27.3.1857 (Londýn, ENG)

– 27.4.1936 (Coldharbour, Surrey, ENG)

matematik a filozof

- 
- ▶ duchovní otec **matematické statistiky**
  - ▶ + zakladatel oboru **biometrie**  
(aplikace matematické statistiky v biologii)
  - ▶ principy statistického **testování hypotéz**  
a koncept **p-hodnoty**
  - ▶ mnohé další



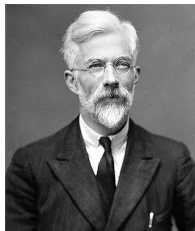
# Sir Ronald Aylmer Fisher

17.2.1890 (Londýn, ENG)

– 29.7.1962 (Adelaide, AUS)

matematik, statistik, biolog, genetik, ...

- ▶ zakladatel **moderní** matematické statistiky
- ▶ + zakladatel **biostatistiky**  
a **populační genetiky**
- ▶ metoda **maximální věrohodnosti**,  
**plánování experimentů**
- ▶ mnohé další



(ve věku 23 let)

## Pearsonův $\chi^2$ test a spor o (ne)jenom) stupně volnosti

1900: K. Pearson [43], článek v *Philosophical Magazine*

- koncept  $\chi^2$ -testu a též principy statistického testování hypotéz  
( $\approx$  vznik oboru **matematická statistika**)
- stupně volnosti (ne takto nazývané)  $K - 1$   
bez ohledu na počet neznámých parametrů

1915: George Udny Yule, Major Greenwood

- upozornili na (opakující se) protichůdné závěry  
dvou „standardních“ testů nezávislost v tabulkách  $2 \times 2$
- v článku o „klinické studii“ o efektivitě vakcín na cholera a tyfus

1922: R. A. Fisher [32], článek v *Journal of the Royal Statistical Society*

- zavedl pojem **stupně volnosti** (*degrees of freedom*),  
matematické zdůvodnění, že stupně volnosti mají být  $K - d - 1$
- upozornil na některé další problémy Pearsonova konceptu  
statistické inference

1922: K. Pearson [65] na stránkách (dodnes prestižního) časopisu *Biometrics*, který v roce 1901 (spolu s Francisem Galtonem a Raphaellem Weldonem) založil:

*The above re-description of what seems to me very elementary considerations would be unnecessary had not a recent writer in the Journal of the Royal Statistical Society [Fisher] appeared to have wholly ignored them. He considers that I have made serious blunders in not linking my degrees of freedom by the number of moments I have taken; ...*

*I hold that such a view is entirely erroneous and that the writer has done no service to the science of statistics by giving it broad cast circulation in the pages of the JRSS. ...*

*I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself or the whole of the theory of probable errors, ...*

---

Více o sporu Pearsona s Fisherem o (ne)jenom) stupně volnosti:

Davis Baird (1983). The Fisher/Pearson chi-squared controversy: A turning point for inductive inference. *The British Journal for the Philosophy of Science*, **34**(2), 105–118.

## Oddíl 8.2

# Kontingenční tabulky



## Nastavení/předpoklady

### Dvourozměrná kategoriální data

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \dots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \stackrel{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Z \end{pmatrix}, \quad \begin{aligned} X &\in \{1, \dots, J\}, \\ Z &\in \{1, \dots, K\}. \end{aligned}$$

- **Pozorovaná četnost** kategorie  $(j, k)$

$$O_{j,k} = n_{j,k} := \sum_{i=1}^N \mathbb{I}\{X_i = j, Z_i = k\}, \quad j = 1, \dots, J, k = 1, \dots, K$$

- **Sdružené pravděpodobnosti**

$$p_{j,k} := P(X_i = j, Z_i = k), \quad j = 1, \dots, J, k = 1, \dots, K,$$

$$\mathbf{p} = (p_{1,1}, \dots, p_{J,K})^\top$$

- **Vektor pozorovaných četností**

$$\mathbf{n} := (n_{1,1}, \dots, n_{J,K})^\top, \quad \text{zřejmě: } \mathbf{n} \sim \text{Mult}_{JK}(N, \mathbf{p})$$

⇒ kontingenční tabulka, **sdruženě multinomický model**

## (Pearsonův) $\chi^2$ -test nezávislosti (v kontingenční tabulce)

$$\begin{aligned} X \perp\!\!\!\perp Z &\iff P(X=j, Z=k) = P(X=j) \cdot P(Z=k) \quad \forall j, k \\ p_{j,k} &= p_{j+} \cdot p_{+k} \end{aligned}$$

► Při nezávislosti:  $\mathbf{n} \sim \text{Mult}_{JK}(N, \mathbf{p})$ ,

kde  $\mathbf{p} = (p_{1,1}, \dots, p_{JK})^\top$  = funkce  $d = J + K - 2$  parametrů

$$\boldsymbol{\theta}_X = (p_{1+}, \dots, p_{(J-1)+}, p_{+1}, \dots, p_{+(K-1)})^\top$$

► Test  $H_0: X \perp\!\!\!\perp Z$

$\equiv \chi^2$ -test dobré shody s odhadnutými parametry

**Maximální věrohodnost, odhad**  $p_{j,k}(\theta_X) = p_{j+}(\theta_X) p_{+k}(\theta_X)$

Logaritmická věrohodnost

$$\begin{aligned}\ell_N(\boldsymbol{\theta}) &= \log\left(\frac{N!}{\prod_{j,k} n_{j,k}!} \prod_{j,k} (p_{j+} p_{+k})^{n_{j,k}}\right) \\ &= \sum_{j=1}^J \sum_{k=1}^K n_{j,k} \log(\underbrace{p_{j+} p_{+k}}_{p_{j,k}(\boldsymbol{\theta})}) + \log\left(\frac{N!}{\prod_{j,k} n_{j,k}!}\right)\end{aligned}$$

Skórový vektor

$$\frac{\partial \ell_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{j=1}^J \sum_{k=1}^K \frac{n_{j,k}}{p_{j+} p_{+k}} \cdot \frac{\partial p_{j,k}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

$$\boldsymbol{\theta} = (p_{1+}, \dots, p_{(J-1)+}, p_{+1}, \dots, p_{+(K-1)})^\top$$

Pro  $k = 1, \dots, K$ :

$$\frac{\partial p_{j,k}(\theta)}{\partial p_{j+}} = \frac{\partial (p_{j+} p_{+k})}{\partial p_{j+}} = p_{+k}, \quad j = 1, \dots, J-1,$$

$$\frac{\partial p_{J,k}(\theta)}{\partial p_{j+}} = \frac{\partial \left\{ (1 - p_{1+} - \dots - p_{(J-1)+}) p_{+k} \right\}}{\partial p_{j+}} = -p_{+k}, \quad j = 1, \dots, J-1.$$

Odsud

$$\begin{aligned} \frac{\partial \ell_N}{\partial p_{j+}} &= \sum_{k=1}^K \frac{n_{j,k}}{p_{j+} p_{+k}} p_{+k} - \sum_{k=1}^K \frac{n_{J,k}}{p_{J+} p_{+k}} \\ &= \sum_{k=1}^K \frac{n_{j,k}}{p_{j+}} - \sum_{k=1}^K \frac{n_{J,k}}{p_{J+}}, \end{aligned} \quad j = 1, \dots, J-1.$$

Podobně

$$\frac{\partial \ell_N}{\partial p_{+k}} = \sum_{j=1}^J \frac{n_{j,k}}{p_{+k}} - \sum_{j=1}^J \frac{n_{j,K}}{p_{+K}}, \quad k = 1, \dots, K-1.$$

## Věrohodnostní rovnice ( $J + K - 2$ rovnic)

$$\frac{n_{j+}}{p_{j+}} = \frac{n_{J+}}{p_{J+}}, \quad j = 1, \dots, J-1,$$

$$\frac{n_{+k}}{p_{+k}} = \frac{n_{+K}}{p_{+K}}, \quad k = 1, \dots, K-1.$$

Maximálně věrohodný odhad  $\hat{\theta}_N = (\hat{p}_{1+}, \dots, \hat{p}_{(J-1)+}, \hat{p}_{+1}, \dots, \hat{p}_{+(K-1)})^\top$

$$\hat{p}_{j+} = \frac{n_{j+}}{N}, \quad j = 1, \dots, J-1,$$

$$\hat{p}_{J+} = \frac{n_{J+}}{N},$$

$$\hat{p}_{+k} = \frac{n_{+k}}{N}, \quad k = 1, \dots, K-1,$$

$$\hat{p}_{+K} = \frac{n_{+K}}{N}.$$

## Maximálně věrohodné odhady pro $p_{1,1}, \dots, p_{J,K}$ za nezávislosti

$$\hat{p}_{j,k} = p_{j,k}(\hat{\theta}_N) = \hat{p}_{j+} \hat{p}_{+k} = \frac{n_{j+} n_{+k}}{N^2}, \quad j = 1, \dots, J, k = 1, \dots, K.$$

## (Odhadnuté) očekávané četnosti

$$\hat{E}_{j,k} := N \hat{p}_{j,k} = \frac{n_{j+} n_{+k}}{N}, \quad j = 1, \dots, J, k = 1, \dots, K.$$

## $\chi^2$ -statistika

$$\begin{aligned} \chi^2 &= \sum_{j=1}^J \sum_{k=1}^K \frac{(O_{j,k} - \hat{E}_{j,k})^2}{\hat{E}_{j,k}} \\ &= \sum_{j=1}^J \sum_{k=1}^K \frac{\left(n_{j,k} - \frac{n_{j+} n_{+k}}{N}\right)^2}{\frac{n_{j+} n_{+k}}{N}} \stackrel{\text{as.}}{\sim} \chi_{JK - (J+K-2) - 1}^2 \quad (\text{pokud } X \perp\!\!\!\perp Z) \end{aligned}$$

Poznámka:  $JK - (J + K - 2) - 1 = (J - 1)(K - 1)$ .

## (Pearsonův) $\chi^2$ -test nezávislosti

Data a model:

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \dots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \stackrel{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Z \end{pmatrix}, \quad p_{j,k} = P(X = j, Z = k), \quad \begin{matrix} j = 1, \dots, J, \\ k = 1, \dots, K. \end{matrix}$$

Pozorované četnosti:  $\mathbf{n} \sim \text{Mult}_{JK}(N, (p_{1,1}, \dots, p_{J,K})^\top)$

Testované hypotézy:

$$\begin{aligned} H_0: & \quad X \perp\!\!\!\perp Z, & \text{ tj. } \forall(j, k) \quad p_{j,k} &= p_{j+} p_{+k} \\ H_1: & \quad X \not\perp\!\!\!\perp Z, & \text{ tj. } \exists(j, k) \quad p_{j,k} &\neq p_{j+} p_{+k} \end{aligned}$$

Testová statistika:

$$\begin{aligned} \chi^2 &= \sum_{j=1}^J \sum_{k=1}^K \frac{(O_{j,k} - \hat{E}_{j,k})^2}{\hat{E}_{j,k}} \\ &= \sum_{j=1}^J \sum_{k=1}^K \frac{\left( n_{j,k} - \frac{n_{j+} n_{+k}}{N} \right)^2}{\frac{n_{j+} n_{+k}}{N}} \stackrel{\text{as.}}{\sim} \chi_{(J-1)(K-1)}^2 \quad (\text{pokud } X \perp\!\!\!\perp Z) \end{aligned}$$

Zamítáme  $H_0 \iff \chi^2 \geq \chi^2_{(J-1)(K-1)}(1 - \alpha)$

$p = 1 - G_{(J-1)(K-1)}(u_N)$

$G_{(J-1)(K-1)}$ : distribuční funkce rozdělení  $\chi^2_{(J-1)(K-1)}$

$u_N$ : hodnota statistiky  $\chi^2$  dosažená/realizovaná s daty

 `chisq.test(table(x, z))`

---

**Použitelnost asymptotické aproximace rozdělením  $\chi^2_{(J-1)(K-1)}$**

Pravidlo palce: použitelná asymptotika  $\iff \hat{E}_{j,k} \geq 5 \quad \forall (j, k)$

$$\frac{n_{j+} n_{+k}}{N} \geq 5 \quad \forall (j, k)$$

$$\frac{\min_j n_{j+} \min_k n_{+k}}{N} \geq 5$$

---

Společenské (a jim podobné) vědy: **pozor na příliš ambiciozní dotazníky!**



# Test homogeneity multinomických rozdělání

≡ **dvou/vícevýběrový** test s kategoriálními daty

$X$  ≡ příslušnost do jedné z  $J$  skupin (dáno předem)

$Z$  ≡ kategoriální odezva  $\in \{1, \dots, K\}$  (dotazník, ...)

Cíl testovat     $H_0: \mathcal{L}(Z | X = 1) = \dots = \mathcal{L}(Z | X = J)$

---

Označme:  $\mathbf{p}_{(1)} = (P(Z = 1 | X = 1), \dots, P(Z = K | X = 1))^T$   
                   $\vdots$   
 $\mathbf{p}_{(J)} = (P(Z = 1 | X = J), \dots, P(Z = K | X = J))^T$

Testované hypotézy:     $H_0: \mathbf{p}_{(1)} = \dots = \mathbf{p}_{(J)}$   
                                   $H_1: \exists j \neq \ell \in \{1, \dots, J\} \quad \mathbf{p}_{(j)} \neq \mathbf{p}_{(\ell)}$

Řádkové marginální četnosti:  $n_{1+}, \dots, n_{J+} \equiv$  rozsahy jednotlivých výběrů

Pozorované četnosti:  $\mathbf{n}_{(1)} := (n_{1,1}, \dots, n_{1,K})^\top \sim \text{Mult}_K(n_{1+}, \mathbf{p}_{(1)})$

$\vdots$

$\mathbf{n}_{(J)} := (n_{J,1}, \dots, n_{J,K})^\top \sim \text{Mult}_K(n_{J+}, \mathbf{p}_{(J)})$

### Řádkově multinomický model

---

**Pozorování** (pohlížejte na  $X$  jako na náhodné), pro všechna  $(j, k)$ :

$$P(Z = k \mid X = j) = \frac{P(Z = k, X = j)}{P(X = j)}$$

$$\text{iff } \underline{X \perp\!\!\!\perp Z} \quad \frac{P(Z = k) P(X = j)}{P(X = j)} = P(Z = k)$$

To jest,  $X \perp\!\!\!\perp Z$

$$\iff \forall k \quad P(Z = k \mid X = 1) = \dots = P(Z = k \mid X = J) = P(Z = k)$$

$$\iff \mathbf{p}_{(1)} = \dots = \mathbf{p}_{(J)}$$

**Nezávislost** ve **sdíleně multinomickém** modelu



**Homogenita řádkových rozdělení** v **řádkově multinomickém** modelu

## Testová statistika:

$$\begin{aligned}\chi^2 &= \sum_{j=1}^J \sum_{k=1}^K \frac{(O_{j,k} - \hat{E}_{j,k})^2}{\hat{E}_{j,k}} \\ &= \sum_{j=1}^J \sum_{k=1}^K \frac{\left(n_{j,k} - \frac{n_{j+} n_{+k}}{N}\right)^2}{\frac{n_{j+} n_{+k}}{N}} \stackrel{\text{as.}}{\sim} \chi^2_{(J-1)(K-1)} \quad (\text{pokud } \mathbf{p}_{(1)} = \dots = \mathbf{p}_{(J)})\end{aligned}$$

---

$O_{j,k}$  = observed count =  $n_{j,k}$ ,  $j = 1, \dots, J, k = 1, \dots, K$

$\hat{E}_{j,k}$  = expected count (za homogeneity)

$$= n_{j+} \hat{P}(Z = k \mid X = j) \stackrel{\text{homogenita}}{=} n_{j+} \hat{P}(Z = k)$$

$$= n_{j+} \frac{n_{+k}}{N}, \quad j = 1, \dots, J, k = 1, \dots, K$$

## Poznámky

- ▶ Technické provedení testu homogenity řádkových rozdělání je shodné s  $\chi^2$ -testem nezávislosti, ale **interpretace** je jiná!
- ▶ Zdůvodnění asymptotiky by nyní bylo odlišné. Nikoliv  $N \rightarrow \infty$ , ale
  - ▶  $n_{1+} \rightarrow \infty, \dots, n_{J+} \rightarrow \infty$
  - ▶ vyjádření, že žádný z výběrů limitně „nemizí“

## Oddíl 8.3

# Kontingenční tabulky $2 \times 2$ (a $J \times 2$ )

## Nastavení/předpoklady

Dvourozměrná binární data

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \dots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \stackrel{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Z \end{pmatrix}, \quad \begin{aligned} X &\in \{1, 2\}, \\ Z &\in \{0, 1\}. \end{aligned}$$

**Kontingenční tabulka:**

|     |   | $Z$       |           |          |
|-----|---|-----------|-----------|----------|
|     |   | 0         | 1         |          |
| $X$ | 1 | $n_{1,0}$ | $n_{1,1}$ | $n_{1+}$ |
|     | 2 | $n_{2,0}$ | $n_{2,1}$ | $n_{2+}$ |
|     |   | $n_{+0}$  | $n_{+1}$  | $N$      |

**Sdružené pravděpodobnosti**

(pokud  $(X, Z)^T$  náhodný vektor):

|     |   | $Z$       |           |          |
|-----|---|-----------|-----------|----------|
|     |   | 0         | 1         |          |
| $X$ | 1 | $p_{1,0}$ | $p_{1,1}$ | $p_{1+}$ |
|     | 2 | $p_{2,0}$ | $p_{2,1}$ | $p_{2+}$ |
|     |   | $p_{+0}$  | $p_{+1}$  | 1        |

## Test nezávislosti $X$ a $Z$

$$\chi^2 = \sum_{j=1}^2 \sum_{k=0}^1 \frac{(n_{j,k} - \frac{n_{j+} n_{+k}}{N})^2}{\frac{n_{j+} n_{+k}}{N}} \stackrel{\text{as.}}{\sim} \chi^2_1$$

(pokud  $X \perp\!\!\!\perp Z$ )



# Test homogeneity dvou binomických rozdělení

$X \equiv$  označení skupiny (1, 2)

$Z \equiv$  0/1 odezva, 1 = „úspěch“

**Četnosti** v tabulce:

|     |       | $Z$       |           |          |
|-----|-------|-----------|-----------|----------|
|     |       | 0         | 1         |          |
| $X$ | sk. 1 | $n_{1,0}$ | $n_{1,1}$ | $n_{1+}$ |
|     | sk. 2 | $n_{2,0}$ | $n_{2,1}$ | $n_{2+}$ |
|     |       |           |           | $N$      |

$n_{1+}, n_{2+}$ : (pevné) rozsahy výběru v jednotlivých skupinách

**Pravděpodobosti:**

|     |       | $Z$       |       |   |
|-----|-------|-----------|-------|---|
|     |       | 0         | 1     |   |
| $X$ | sk. 1 | $1 - p_1$ | $p_1$ | 1 |
|     | sk. 2 | $1 - p_2$ | $p_2$ | 1 |

$$p_1 := P(Z = 1; \text{sk. 1})$$

$$\equiv P(Z = 1 \mid X = 1) = \frac{P(X = 1, Z = 1)}{P(X = 1)} \text{ iff } X \stackrel{\text{red}}{=} Z \quad P(Z = 1)$$

$$p_2 := P(Z = 1; \text{sk. 2})$$

$$\equiv P(Z = 1 \mid X = 2) = \frac{P(X = 2, Z = 1)}{P(X = 2)} \text{ iff } X \stackrel{\text{red}}{=} Z \quad P(Z = 1)$$

Při podmínění hodnotou  $X$ , resp. při pevných  $n_{1+}$  a  $n_{2+}$ :

$$n_{1,1} \sim Bi(n_{1+}, p_1), \quad n_{2,1} \sim Bi(n_{2+}, p_2)$$

viz předchozí část

$$\chi^2\text{-test nezávislosti } X \text{ a } Z \quad \equiv \quad \text{test } H_0: p_1 = p_2$$

## Opakování.

$$\hat{p}_1 = \frac{n_{1,1}}{n_{1+}}, \quad \text{var } \hat{p}_1 = \frac{p_1(1-p_1)}{n_{1+}}$$

$$\hat{p}_2 = \frac{n_{2,1}}{n_{2+}}, \quad \text{var } \hat{p}_2 = \frac{p_2(1-p_2)}{n_{2+}}$$

$$d_x = p_1 - p_2, \quad \hat{d}_x = \frac{n_{1,1}}{n_{1+}} - \frac{n_{2,1}}{n_{2+}},$$

$$r_x = \frac{p_1}{p_2}, \quad \hat{r}_x = \frac{n_{1,1} n_{2+}}{n_{2,1} n_{1+}},$$

$$o_x = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}, \quad \hat{o}_x = \frac{n_{1,1} n_{2,0}}{n_{1,0} n_{2,1}} \quad (\text{křížový poměr})$$

---

Pokud  $X$  náhodné a  $p_1 \equiv P(Z = 1 \mid X = 1)$ ,  $p_2 \equiv P(Z = 1 \mid X = 2)$ , potom

$$d_x = 0 \quad \Leftrightarrow \quad r_x = 1 \quad \Leftrightarrow \quad o_x = 1 \quad \Leftrightarrow \quad X \perp\!\!\!\perp Z$$

několik způsobů, jak otestovat. . .

## Opakování, pokrač.

$$\hat{p}_1 = \frac{n_{1,1}}{n_{1+}}, \quad \widehat{\text{var}} \hat{p}_1 = \frac{\hat{p}_1 (1 - \hat{p}_1)}{n_{1+}}$$

$$\hat{p}_2 = \frac{n_{2,1}}{n_{2+}}, \quad \widehat{\text{var}} \hat{p}_2 = \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_{2+}}$$

Testová statistika pro testování  $H_0: d_X = 0$ :

$$T_d = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_{1+}} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_{2+}}}} \stackrel{\text{as.}}{\sim} \mathcal{N}(0, 1) \quad \text{za platnosti } H_0$$

---

Za platnosti  $H_0: p_1 = p_2 =: p$  je  $\text{var}(\hat{p}_1 - \hat{p}_2) = p(1 - p) \left( \frac{1}{n_{1+}} + \frac{1}{n_{2+}} \right)$

→ konzistentní odhad  $p$ :  $\tilde{p} := \frac{n_{1,1} + n_{2,1}}{N}$

a testová statistika pro testování  $H_0: d_X = 0$ :

$$T_d^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\tilde{p}(1 - \tilde{p}) \left( \frac{1}{n_{1+}} + \frac{1}{n_{2+}} \right)}} \stackrel{\text{as.}}{\sim} \mathcal{N}(0, 1) \quad \text{za platnosti } H_0$$

# Test homogeneity několika binomických rozdělení

$X \equiv$  označení skupiny  $\in \{1, \dots, J\}$

$Z \equiv$  0/1 odezva, 1 = „úspěch“

**Četnosti** v tabulce  $J \times 2$ :

|     |          | $Z$       |           |          |
|-----|----------|-----------|-----------|----------|
|     |          | 0         | 1         |          |
| $X$ | sk. 1    | $n_{1,0}$ | $n_{1,1}$ | $n_{1+}$ |
|     | $\vdots$ | $\vdots$  | $\vdots$  | $\vdots$ |
|     | sk. $J$  | $n_{J,0}$ | $n_{J,1}$ | $n_{J+}$ |
|     |          |           |           | $N$      |

$n_{1+}, \dots, n_{J+}$ : (pevné) rozsahy výběru v jednotlivých skupinách

**Pravděpodobosti:**

|     |          | $Z$       |          |          |
|-----|----------|-----------|----------|----------|
|     |          | 0         | 1        |          |
| $X$ | sk. 1    | $1 - p_1$ | $p_1$    | 1        |
|     | $\vdots$ | $\vdots$  | $\vdots$ | $\vdots$ |
|     | sk. $J$  | $1 - p_J$ | $p_J$    | 1        |

$$p_1 := P(Z = 1; \text{sk. } 1)$$

$$\equiv P(Z = 1 \mid X = 1) = \frac{P(X = 1, Z = 1)}{P(X = 1)} \stackrel{\text{iff } X \perp\!\!\!\perp Z}{=} P(Z = 1)$$

$\vdots$

$$p_J := P(Z = 1; \text{sk. } J)$$

$$\equiv P(Z = 1 \mid X = J) = \frac{P(X = J, Z = 1)}{P(X = J)} \stackrel{\text{iff } X \perp\!\!\!\perp Z}{=} P(Z = 1)$$

Při podmínění hodnotou  $X$ , resp. při pevných  $n_{1+}, \dots, n_{J+}$ :

$$n_{1,1} \sim Bi(n_{1+}, p_1), \dots, n_{J,1} \sim Bi(n_{J+}, p_J)$$

analogicky jako při  $J = 2$

$$\chi^2\text{-test nezávislosti } X \text{ a } Z \quad \equiv \quad \text{test } H_0: p_1 = \dots = p_J$$

**Testová statistika:** 
$$\chi^2 = \sum_{j=1}^J \sum_{k=0}^1 \frac{(n_{j,k} - \frac{n_{j+} n_{+k}}{N})^2}{\frac{n_{j+} n_{+k}}{N}} \stackrel{\text{as.}}{\sim} \chi_{J-1}^2$$

(pokud  $p_1 = \dots = p_J$ )

trocha algebraických úprav:

$$\chi^2 = \frac{1}{\tilde{p}(1-\tilde{p})} \sum_{j=1}^J n_{j+} (\hat{p}_j - \tilde{p})^2,$$

$$\hat{p}_j = \frac{n_{j,1}}{n_{j+}} = \hat{P}(Z = 1 \mid X = j), \quad j = 1, \dots, J,$$

$$\tilde{p} = \frac{n_{1,1} + \dots + n_{J,1}}{N} = \hat{P}(Z = 1)$$

## Motivační problém (*Fisher tea experiment with lady Muriel Bristol*)

- ▶ Lady Bristol tvrdí, že pozná podle chuti, zda při přípravě čaje s mlékem byl v hrnku první čaj nebo mléko.
- ▶ Fisher nechal připravit 4 hrnky prvním způsobem, 4 hrnky druhým způsobem, lady ochutnávala a určovala způsob přípravy.
- ▶ Lady ví, že dostala 4 hrnky, kde mléko bylo první a 4 hrnky, kde mléko bylo druhé.
- ▶  $X \in \{0, 1\}$ , tip lady, 0 = mléko první, 1 = mléko druhé,  
 $Z \in \{0, 1\}$ , způsob přípravy, 0 = mléko první, 1 = mléko druhé

▶ Data:

|     |   | $Z$ |   |   |
|-----|---|-----|---|---|
|     |   | 0   | 1 |   |
| $X$ | 0 | 3   | 1 | 4 |
|     | 1 | 1   | 3 | 4 |
|     |   | 4   | 4 | 8 |



## Dvourozměrná binární data

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \dots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \sim \begin{pmatrix} X \\ Z \end{pmatrix}, \quad \begin{aligned} X &\in \{0, 1\}, \\ Z &\in \{0, 1\}. \end{aligned}$$

### Kontingenční tabulka:

|          |   | <b>Z</b>  |           |          |
|----------|---|-----------|-----------|----------|
|          |   | 0         | 1         |          |
| <b>X</b> | 0 | $n_{0,0}$ | $n_{0,1}$ | $n_{0+}$ |
|          | 1 | $n_{1,0}$ | $n_{1,1}$ | $n_{1+}$ |
|          |   | $n_{+0}$  | $n_{+1}$  | <b>N</b> |

$n_{0+}$ ,  $n_{1+}$ ,  $n_{+0}$ ,  $n_{+1}$ : **pevné** (dopředu známé)

Cíl testovat:  $H_0: X \perp\!\!\!\perp Z$

Za podmínky **pevných marginálních četností** a při nezávislosti  $X$  a  $Z$ :

$$n_{0,0} \mid n_{0+}, n_{1+}, n_{+0}, n_{+1} \sim \text{Hypergeom}(n_{0+}; n_{+0}, n_{+1})$$

$$\begin{aligned} P(\text{tabulka}) &= P(n_{0,0} \mid n_{0+}, n_{1+}, n_{+0}, n_{+1}) = \frac{\binom{n_{+0}}{n_{0,0}} \binom{n_{+1}}{n_{0,1}}}{\binom{N}{n_{0+}}} \\ &= \frac{n_{0+}! n_{1+}! n_{+0}! n_{+1}!}{n_{0,0}! n_{0,1}! n_{1,0}! n_{1,1}! N!} \end{aligned}$$

$$\text{p-hodnota} = \sum_{\text{TAB} \in \mathcal{T}} P(\text{TAB}),$$

$$\mathcal{T} = \left\{ \text{TAB}, \quad \text{pro kterou} \quad P(\text{TAB}) \leq P(\text{pozorovaná TAB}) \right\}$$

 `fisher.test(tab)`

## Poznámky.

- ▶ Diskrétní rozdělení testové statistiky.
- ▶ Skutečná hladina testu  $\leq \alpha$ .
- ▶ Používá se též/spíš jednostranně.
- ▶ Lady Bristol a pití čaje, možné tabulky a jejich hypergeometrické pravděpodobnosti:

|        |       |   |       |   |       |   |       |   |       |   |
|--------|-------|---|-------|---|-------|---|-------|---|-------|---|
|        | 0     | 4 | 1     | 3 | 2     | 2 | 3     | 1 | 4     | 0 |
|        | 4     | 0 | 3     | 1 | 2     | 2 | 1     | 3 | 0     | 4 |
| P(TAB) | 0,014 |   | 0,229 |   | 0,514 |   | 0,229 |   | 0,014 |   |

$$p\text{-hodnota} = 2(0,229 + 0,014) = 0,486$$

- ▶ Kolik hrnků by musela lady Bristol otipovat správně, aby sira Fishera přesvědčila, že pozná, v jakém pořadí byl čaj s mlékem připraven?

# McNemarův test

≡ párový test s 0/1 daty (Quinn McNemar, 1947)

Dvourozměrná binární data

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \dots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \stackrel{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Z \end{pmatrix}, \quad \begin{aligned} X &\in \{0, 1\}, \\ Z &\in \{0, 1\}. \end{aligned}$$

**Kontingenční tabulka a pravděpodobnosti:**

|     |   | $Z$       |           |          |
|-----|---|-----------|-----------|----------|
|     |   | 0         | 1         |          |
| $X$ | 0 | $n_{0,0}$ | $n_{0,1}$ | $n_{0+}$ |
|     | 1 | $n_{1,0}$ | $n_{1,1}$ | $n_{1+}$ |
|     |   | $n_{+0}$  | $n_{+1}$  | $N$      |

|     |   | $Z$       |           |          |
|-----|---|-----------|-----------|----------|
|     |   | 0         | 1         |          |
| $X$ | 0 | $p_{0,0}$ | $p_{0,1}$ | $p_{0+}$ |
|     | 1 | $p_{1,0}$ | $p_{1,1}$ | $p_{1+}$ |
|     |   | $p_{+0}$  | $p_{+1}$  | 1        |

Cíl testovat:

$$H_0: p_{1+} = p_{+1} \quad (P(X = 1) = P(Z = 1))$$

$$\iff H_0: p_{0,1} = p_{1,0} \quad (P(X = 0, Z = 1) = P(X = 1, Z = 0))$$

Typicky:  $X \equiv$  „před“

hodnotitel 1

$Z \equiv$  „po“

hodnotitel 2

$H_0$ : intervence neměla vliv

hodnotitelé se neliší

Tabulka sdružených pravděpodobností za platnosti  $H_0$ :

|           |           |   |
|-----------|-----------|---|
| $p_{0,0}$ | $p_{0,1}$ |   |
| $p_{0,1}$ | $p_{1,1}$ |   |
|           |           | 1 |

Neznámé parametry:

$$p_{0,0}, p_{0,1} = p_{1,0}, \quad d = 2$$

$$p_{1,1} = 1 - p_{0,0} - 2 p_{0,1}$$

Maximálně věrohodné odhady v multinomickém modelu:

$$\hat{p}_{0,0} = \frac{n_{0,0}}{N}, \quad \hat{p}_{0,1} = \frac{n_{0,1} + n_{1,0}}{2N} = \hat{p}_{1,0}, \quad \hat{p}_{1,1} = \frac{n_{1,1}}{N}$$

$\chi^2$ -statistika (test dobré shody při  $d = 2$  neznámých parametrech):

$$\chi^2 = \sum_{j=0}^1 \sum_{k=0}^1 \frac{(n_{j,k} - N\hat{p}_{j,k})^2}{N\hat{p}_{j,k}} = \frac{(n_{0,1} - n_{1,0})^2}{(n_{0,1} + n_{1,0})} \stackrel{\text{as.}}{\sim} \chi_{4-1-2}^2 = \chi_1^2$$

### Poznámky.

- ▶ Existují zobecnění na tabulky  $J \times J$ ,  $J > 2$ 
  - ▶ Homogenita marginálních pravděpodobností ( $\forall j \quad p_{j+} = p_{+j}$ )  
již **není ekvivalentní** se symetrií sdružených pravděpodobností  
( $\forall j, k \quad p_{j,k} = p_{k,j}$ ).
- ▶ Test symetrie: A. H. Bowker (1948).
- ▶ Testy homogenity: A. Stuart (1955), V. P. Bhapkar (1966).

# 9

***K*-výběrový problém pro  
kvantitativní data**

# 10

**Korelační analýza**