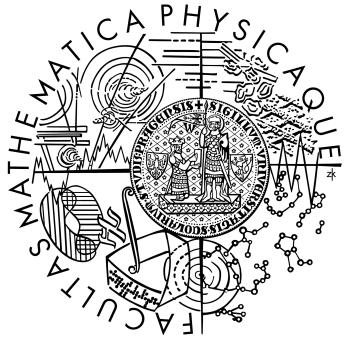
Katedra pravděpodobnosti a matematické statistiky



MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

doc. RNDr. Arnošt Komárek, Ph.D.

NMSA331 Matematická statistika 1

Zimní semestr 2024–25

Osoby a obsazení

Přednášky (Pondělí 12:20 – 15:30 v K3)

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

doc. RNDr. Arnošt Komárek, Ph.D.

komarek@karlin.mff.cuni.cz

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

2. patro vedle schodů

Cvičení (Pondělí 10:40 v K4)

doc. Ing. Marek Omelka, Ph.D.

omelka@karlin.mff.cuni.cz

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

2. patro vedle schodů

Studijní materiály

Webpage přednášky

https://msekce.karlin.mff.cuni.cz/~komarek/vyuka/2024_25/nmsa331-2024.html

Webpage cvičení

https://msekce.karlin.mff.cuni.cz/~omelka/Vyuka_nmsa331_2425.php

Studijní materiály

- **Nejzákladnější 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 2024 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 podpůrné texty.

Znalosti z jiných předmětů

- 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).

Zkouška

- **Písemná část:** 100 minut, obvykle dopoledne.
- **Ústní část:** zadání otázky/problému
  $\rightarrow$ písemná příprava (cca 20 minut)
  $\rightarrow$ 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).

Nepravidelnosti

- Přednáška po Novém roce (6.1.2025) již nebude („povinná“). Místo a čas budou potenciálně využity pro předtermín zkoušky, případně k dokončení témat, jež se nebudou zkoušet nyní, ale mohou se objevit u státní závěrečné zkoušky.

  Požadavky k SZZ se neptají, kolik přednášek skutečně proběhlo, v tomto semestru jich bude pouze 24.

- Náhrada této přednášky: středa 30.10.2024 od 15:40 (do 18:50) v K1.

Matematická statistika

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

- Lze přenést do stylu Definice $\rightarrow$ Věta $\rightarrow$ 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

$\approx$ snaha řešit problémy reálného světa pomocí dat (která jsou prý všude kolem nás) aplikací metod matematické statistiky.

- Je to spíš řemeslo...

  které lze jen obtížně (resp. nejvýše fachidiotsky) 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...

Statistika a analýza dat

- 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 $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ konverguje skoro jistě k náhodnému vektoru $\boldsymbol{X}$ pro $n \to \infty$ tehdy a jen tehdy když

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

Značíme $\boldsymbol{X}_n \xrightarrow{\text{s.j.}} \boldsymbol{X}$, $n \to \infty$.

Definice 1.2 Konvergence v pravděpodobnosti.

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

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

Značíme $\boldsymbol{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{X}$, $n \to \infty$.

Definice 1.3 Konvergence v distribuci.

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

$$\lim_{n\to\infty} F_{\boldsymbol{X}_n}(\boldsymbol{x}) = F_{\boldsymbol{X}}(\boldsymbol{x})$$

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

Tvrzení 1.1 Vztahy mezi konvergencemi.

(i) $\boldsymbol{X}_n \xrightarrow{s.j.} \boldsymbol{X},\ n \to \infty \quad \Longrightarrow \quad \boldsymbol{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{X},\ n \to \infty;$

(ii) $\boldsymbol{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{X},\ n \to \infty \quad \Longrightarrow \quad \boldsymbol{X}_n \xrightarrow{\mathcal{D}} \boldsymbol{X},\ n \to \infty.$

Poznámka.
Opačně platí POUZE implikace $\boldsymbol{X}_n \xrightarrow{\mathcal{D}} \boldsymbol{c},\ \boldsymbol{c} \in \mathbb{R}^k \quad \Longrightarrow \quad \boldsymbol{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{c}.$

Tvrzení 1.2 Věta o spojité transformaci.

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

(i) $\boldsymbol{X}_n \xrightarrow{s.j.} \boldsymbol{X}, n \to \infty \implies g(\boldsymbol{X}_n) \xrightarrow{s.j.} g(\boldsymbol{X}), n \to \infty;$

(ii) $\boldsymbol{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{X}, n \to \infty \implies g(\boldsymbol{X}_n) \xrightarrow{\mathrm{P}} g(\boldsymbol{X}), n \to \infty;$

(iii) $\boldsymbol{X}_n \xrightarrow{\mathcal{D}} \boldsymbol{X}, n \to \infty \implies g(\boldsymbol{X}_n) \xrightarrow{\mathcal{D}} g(\boldsymbol{X}), n \to \infty.$

Tvrzení 1.3 Cramérova-Sluckého věta.

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

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

Tvrzení 1.4 Postačující podmínka pro konzistenci.

Nechť

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

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

$$\boldsymbol{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{\mu}, \quad n \to \infty.$$

Oddíl **1.2**

Zákon velkých čísel

Tvrzení 1.5 Silný zákon velkých čísel.

Nechť $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \overset{i.i.d.}{\sim} \boldsymbol{X}, \quad \mathbb{E}\boldsymbol{X} = \boldsymbol{\mu} \in \mathbb{R}^k$. *Potom*

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

Oddíl **1.3**

Centrální limitní věta

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

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

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

Poznámky.

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

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

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

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

Tvrzení 1.7 Δ-metoda.

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

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

pro nějaký vektor konstant $\boldsymbol{\mu} \in \mathbb{R}^k$ *a* $k \times k$ *reálnou matici* $\boldsymbol{\Sigma}$.
Nechť $g : \mathbb{R}^k \to \mathbb{R}^p$ *je funkce, která je spojitě diferencovatelná na nějakém okolí bodu* $\boldsymbol{\mu}$. *Označme*

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

Potom

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

2

Náhodný výběr

Oddíl **2.1**

Definice náhodného výběru

Definice 2.1 Náhodný výběr.

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

Značíme $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}, \quad \boldsymbol{X} \sim F_{\boldsymbol{X}}$

nebo stručněji $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} F_{\boldsymbol{X}}$.

Definice 2.2 (Statistický) model.

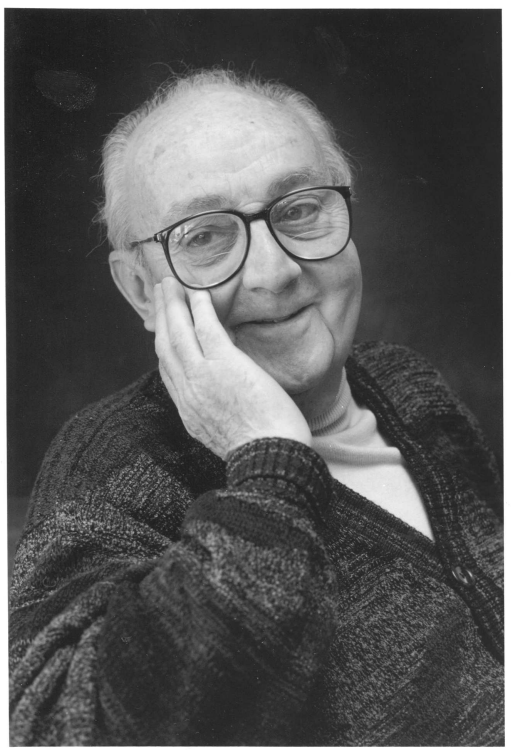
(Statistický) model (*statistical model*) pro náhodný výběr $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ je prespecifikovaná množina rozdělení $\mathcal{F}$, která obsahuje (neznámé) rozdělení $F_{\boldsymbol{X}}$.

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 $\boldsymbol{S}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \boldsymbol{S}(\mathbb{X})$ náhodného výběru $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \equiv \mathbb{X}$ se nazývá statistika (*statistic*).

Definice 2.4 Výběrový průměr a výběrový rozptyl.

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

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

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

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

Vlastnosti výběrového průměru

Lemma 2.1 Výběrový průměr a nejmenší čtverce.

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

Věta 2.2 Statistické vlastnosti výběrového průměru.

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

(i) $\mathbb{E}\overline{X}_n = \mu, \quad \operatorname{var}\overline{X}_n = \dfrac{\sigma^2}{n}$.

(ii) $\overline{X}_n \xrightarrow{\mathrm{P}} \mu, \quad n \to \infty$.

(iii) $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2), \quad n \to \infty$.

Věta 2.3 Statistické vlastnosti (empirické) relativní četnosti.

Nechť $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X, \quad X \sim \mathcal{A}lt(p), \quad p \in (0, 1)$. *Potom*

(i) $\mathbb{E}\overline{X}_n = p, \quad \mathrm{var}\overline{X}_n = \dfrac{p(1-p)}{n}.$

(ii) $\overline{X}_n \xrightarrow{\mathrm{P}} p, \quad n \to \infty.$

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

(iv) $n\overline{X}_n \sim \mathcal{B}i(n, p).$

Lemma 2.4 Alternativní vyjádření výběrového rozptylu.

Pro $n \geq 2$:

(i) $S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right).$

(ii) *Označme*

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

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

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

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

Důkaz.

(i)

$$\frac{n-1}{n} S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i^2 - 2 X_i \overline{X}_n + \overline{X}_n^2)$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i^2 - 2\overline{X}_n^2 + \overline{X}_n^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2.$$

(ii)

$$\boldsymbol{X}^\top \mathbb{A} \boldsymbol{X} = \boldsymbol{X}^\top (\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n{}^\top) \boldsymbol{X} = \boldsymbol{X}^\top \boldsymbol{X} - \frac{1}{n} \boldsymbol{X}^\top \mathbf{1}_n \mathbf{1}_n{}^\top \boldsymbol{X}$$

$$= \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \Big(\sum_{i=1}^{n} X_i \Big)^2 = \sum_{i=1}^{n} X_i^2 - n \overline{X}_n^2 = (n-1) S_n^2.$$

Invariance S_n^2 vůči posunutí plyne z faktu, že
$\mathbb{A}\mathbb{A} = \mathbb{A}$ a $\mathbb{A}\,\mathbf{1}_n = \mathbf{0}_n$, $\mathbf{1}_n{}^\top \mathbb{A} = \mathbf{0}_n{}^\top$.

Lemma 2.5 Střední hodnota kvadratické formy.

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

$$\mathbb{E}\left(\boldsymbol{Z}^\top \mathbb{B}\, \boldsymbol{Z}\right) = \boldsymbol{\mu}^\top \mathbb{B}\, \boldsymbol{\mu} + \mathrm{tr}\left(\mathbb{B}\, \boldsymbol{\Sigma}\right).$$

Věta 2.6 Statistické vlastnosti výběrového rozptylu.

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

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

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

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

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

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

Věta 2.6 Statistické vlastnosti výběrového rozptylu, pokrač.

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

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

kde

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

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

Lemma 2.7 Nezávislost lineární a kvadratické transformace za normality.

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

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

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

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

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

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

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

Věta 2.8 Vlastnosti výběrového rozptylu za normality.

Nechť $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X, \quad X \sim \mathcal{N}(\mu, \sigma^2)$. *Potom*

(i) $\dfrac{(n-1)\,S_n^2}{\sigma^2} \sim \chi^2_{n-1}$.

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

Lemma A.4

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

$$\boldsymbol{X}^\top \mathbb{A}\, \boldsymbol{X} \sim \chi^2_{\mathrm{tr}(\mathbb{A}\,\boldsymbol{\Sigma})}.$$

Věta 2.9 Asymptotické rozdělení T statistiky.

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

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

Věta 2.10 Rozdělení T statistiky za normality.

Nechť $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X, \quad X \sim \mathcal{N}(\mu, \sigma^2)$. *Potom*

$$T_n := \frac{\sqrt{n}\,(\overline{X}_n - \mu)}{S_n} \sim \mathrm{t}_{n-1}.$$

Definice 2.5 F-rozdělení.

Nechť $X \sim \chi^2_n$ a $Y \sim \chi^2_m$ 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 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X, \quad X \sim \mathcal{N}(\mu_X, \sigma_X^2)$

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

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

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

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

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

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

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

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

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

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

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

$$\boldsymbol{X}_{(\bullet)} = \left(X_{(1)}, \ldots, X_{(n)}\right)^\top.$$

Věta 2.12 Sdružená hustota uspořádaného výběru.

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

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

Věta 2.13 Distribuční funkce k-té pořádkové statistiky.

Distribuční funkce k-té pořádkové statistiky jest

$$F_{(k)}(x) = \mathrm{P}(X_{(k)} \leq x) = \sum_{j=k}^{n} \binom{n}{j} F^j(x) \{1 - F(x)\}^{n-j}$$

$$= \frac{1}{\mathrm{B}(k, n-k+1)} \int_0^{F(x)} t^{k-1} (1-t)^{n-k} \, \mathrm{d}t.$$

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

$$\mathrm{P}(\boldsymbol{R} = \boldsymbol{r}) = \frac{1}{n!}, \qquad \boldsymbol{r} \in \mathcal{P}_n.$$

Věta 2.16 Vlastnosti pořadí.

Platí

(i) $\mathrm{P}(R_i = k) = \dfrac{1}{n}$ *pro každé* $i, k \in \{1, \ldots, n\}$.

(ii) $\mathrm{P}(R_i = k, R_j = m) = \dfrac{1}{n(n-1)}$ *pro každé* $i \neq j, k \neq m \in \{1, \ldots, n\}$.

(iii) $\mathbb{E}\, R_i = \dfrac{n+1}{2}, \quad \operatorname{var} R_i = \dfrac{n^2-1}{12}$ *pro každé* $i \in \{1, \ldots, n\}$.

(iv) $\operatorname{cov}(R_i, R_j) = -\dfrac{n+1}{12}$ *pro každé* $i \neq j \in \{1, \ldots, n\}$.

Oddíl **2.4**

Transformace ve statistice

Nastavení

- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_x, \quad$ hustota: $f_X, \quad$ nosič: S_X;
- $g: S_x \longrightarrow \mathbb{R}$, ryze monotónní a diferencovatelná;
- $Y_i := g(X_i), i = 1, \ldots, n \quad \equiv \quad$ transformovaný náhodný výběr.

Dostaneme

- $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Y$, $Y \sim F_Y$, hustota: f_Y;
- Pokud F_X spojitá a f_X známá, věta o transformaci poskutuje předpis f_Y.

Transformace stabilizující (asymptotický) rozptyl

- Nechť $\sqrt{n}\,(T_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(\mu)), \quad n \to \infty.$
- Pokud $g \equiv$ reálná funkce spojitě diferencovatelná na okolí bodu μ, potom (Δ-metoda):

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

- Zvolme $g(x) = c \int \frac{1}{\sigma(x)}\,\mathrm{d}x$

$$\Rightarrow \quad g'(\mu) = \frac{c}{\sigma(\mu)}$$

$$\Rightarrow \quad \sqrt{n}\,(g(T_n) - g(\mu)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c^2), \quad n \to \infty.$$

Standardizace

- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_x$, rozdělení s $0 < \operatorname{var} X < \infty$;
- Z-skóry:

$$Z_i := \frac{X_i - \overline{X}_n}{S_n}, \qquad i = 1, \ldots, n.$$

- Výběrový průměr Z-skórů je 0.
- Výběrový rozptyl Z-skórů je 1.
- Z-skóry $Z_1, \ldots, Z_n$ nejsou nezávislé.
- Protože platí $\overline{X}_n \xrightarrow{\text{P}} \mathbb{E}\, X$,

$$S_n \xrightarrow{\text{P}} \sqrt{\operatorname{var} X}, \qquad n \to \infty.$$

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

3

Odhadování parametrů

Nastavení/předpoklady

- $\underbrace{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}_{\mathbb{X}} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}, \quad \boldsymbol{X} \sim F_X \in \mathcal{F}$ (model);
- $\boldsymbol{\theta} = \boldsymbol{t}(F) \in \mathbb{R}^p$, $F \in \mathcal{F}$
  $\equiv$ parametr, který za platnosti předpokládaného modelu chceme odhadnout.
- Nechť $F_X \in \mathcal{F}$ $\equiv$ skutečné rozdělení náhodného vektoru $\boldsymbol{X}$.
- $\boldsymbol{\theta}_X := \boldsymbol{t}(F_X)$ $\equiv$ skutečná hodnota parametru zájmu.

Oddíl **3.1**

Bodový odhad

Definice 3.1 Bodový odhad.

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

$$\widehat{\theta}_n = \boldsymbol{T}_n(\mathbb{X}) = \boldsymbol{T}_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n),$$

kde $\boldsymbol{T}_n$ je nějaká Borelovsky měřitelná funkce dat.

Definice 3.2 Nestrannost a konzistence.

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

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

$$\mathbb{E}_{F_X} \widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_X$$

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

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

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\theta}_X, \quad n \to \infty$$

pro každé $F_X \in \mathcal{F}$.

Definice 3.3 Vychýlení.

Nechť odhad $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{T}_n(\mathbb{X})$ parametru $\boldsymbol{\theta}_X$ má konečnou střední hodnotu (pro každé $F_X \in \mathcal{F}$). Rozdíl

$$\text{bias}\left(\widehat{\boldsymbol{\theta}}_n\right) := \mathbb{E}_{F_X}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)$$

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

Definice 3.4 Střední čtvercová a směrodatná chyba.

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

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

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

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

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

Věta 3.1 Postačující podmínka konzistence odhadu.

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

$$\lim_{n\to\infty} \mathbb{E}_{F_X}\left(\widehat{\theta}_n\right) = \theta_X \qquad \& \qquad \lim_{n\to\infty} \mathrm{var}_{F_X}\left(\widehat{\theta}_n\right) = 0.$$

Potom je $\widehat{\theta}_n$ (slabě) konzistentním odhadem parametru θ_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 $\mathbb{E}_{F_X}\left|\widehat{\theta}_n\right| = \infty$.
- **Příklad**: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{A}lt(p_X)$, $0 < p_X < 1$.

  Uvažme $\widehat{\theta}_n = \left(\overline{X}_n\right)^{-1}$ jako odhad parametru $\theta_X := p_X^{-1}$.

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

Oddíl **3.2**

Volba parametru zájmu

Typy dat (škály měření)

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:** $\mathbb{E}\,X, \operatorname{var} X, F_X(x), F_X^{-1}(\alpha), \ldots$

KATEGORIÁLNÍ data

- hodnoty „pouze“ kódují příslušnost do určité kategorie, skupiny
- **Matematická reprezentace: diskrétní** náhodná veličina s konečným nosičem $\{\omega_1, \ldots, \omega_K\}$, $\omega_1 < \cdots < \omega_K \equiv$ „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 $\mathrm{P}(X = \omega_k)$, $k = 1, \ldots, K$ má smysl.

     Veličiny typu $\mathbb{E}\,X, \mathrm{var}\,X, F_X(x), F_X^{-1}(\alpha), \ldots$
     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 $\{\omega_1, \ldots, \omega_K\}$, $\omega_1 < \cdots < \omega_K \equiv$ „nálepky“ jednotlivých kategorií
- Další dělení:
  2. **ordinální** veličiny: kategorie lze smyslupně uspořádat

     (známka ve škole: $1, 2, 3, 4, 5$,

     spokojenost s. . . : -2 = *velmi nespokojen*, -1 = *nespokojen*, 0 = *neutrální*, 1 = *spokojen*, 2 = *velmi spokojen*)

     Parametr zájmu: kromě $\mathrm{P}(X = \omega_k)$, $k = 1, \ldots, K$ má smysl též

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

     někdy i další, např. $\mathbb{E}\, X$

BINÁRNÍ data

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

$\Rightarrow$ **Matematická reprezentace:**

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

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

Statistické metody

- Pro kvantitativní data
  - pracují s charakteristikami jako $\mathbb{E}\,X$, $\operatorname{var} X$, $F_X(x)$, $F_X^{-1}(\alpha)$, ...
  - 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)

$$\underbrace{X_1, \ldots, X_n}_{\mathbb{X}} \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_X \in \mathcal{F},$$

$$\mathcal{F} = \big\{\text{rozdělení s hustotou } f(x;\, \theta),\ \theta \in \Theta \subseteq \mathbb{R}^p\big\},$$

hustota $f(\cdot;\, \theta)$ vzhledem k σ-konečné míře, tvar $f(\cdot;\, \cdot)$ známý.

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

CÍL: odhad parametru θ_X.

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

- Máme konzistentní odhady momentů:

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

- Platí: $\mathbb{E}_F\, X^q$, $q = 1, \ldots, p$ je funkcí θ

(konst. funkce je možnou komplikací).

Oddíl **3.4**

Intervalový odhad

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

- $\underbrace{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}_{\mathbb{X}} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}, \quad \boldsymbol{X} \sim F_X \in \mathcal{F}$ (model)

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

- $\theta = t(F) \in \mathbb{R}, F \in \mathcal{F}$

  $\equiv$ parametr, který za platnosti předpokládaného modelu chceme odhadnout.

- Nechť $F_X \in \mathcal{F}$ $\equiv$ skutečné rozdělení náhodného vektoru $\boldsymbol{X}$.

- $\theta_X := t(F_X)$ $\equiv$ 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

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

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

$$\lim_{n\to\infty} \mathsf{P}_{F_X}\left(B_n(\mathbb{X}) \ni \theta_X\right) = 1 - \alpha, \qquad \text{pro každé rozdělení } F_X \in \mathcal{F}.$$

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 α je předem zvolené, nejběžnější volba: $\alpha = 0{,}05 \to$ 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 $F_X \in \mathcal{F}$)

$$\mathsf{P}_{F_X}\Big((\eta_L(\mathbb{X}), \eta_U(\mathbb{X})) \ni \theta_X\Big) = 1 - \alpha,$$

$$\mathsf{P}_{F_X}\big(\eta_L(\mathbb{X}) < \eta_U(\mathbb{X})\big) = \mathsf{P}_{F_X}\big(\eta_L(\mathbb{X}) > -\infty\big) = \mathsf{P}_{F_X}\big(\eta_U(\mathbb{X}) < \infty\big) = 1.$$

Obvykle konstruován, aby platilo (alespoň asymptoticky)

$$\mathsf{P}_{F_X}\big(\eta_L(\mathbb{X}) \geq \theta_X\big) = \frac{\alpha}{2} = \mathsf{P}_{F_X}\big(\eta_U(\mathbb{X}) \leq \theta_X\big).$$

 software: `alternative = "two.sided"`

Jednostranný (*one-sided*) interval spolehlivosti

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

$$\mathsf{P}_{F_X}\left(\eta_L(\mathbb{X}) < \theta_X\right) = 1 - \alpha, \qquad \text{pro každé } F_X \in \mathcal{F}.$$

R software: `alternative = "greater"`

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

$$\mathsf{P}_{F_X}\left(\eta_U(\mathbb{X}) > \theta_X\right) = 1 - \alpha, \qquad \text{pro každé } F_X \in \mathcal{F}.$$

R software: `alternative = "less"`

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

- Oblast spolehlivosti/konfidenční oblast, množina (*confidence region, set*) $B_n(\mathbb{X}) \subset \mathbb{R}^p$:

$$\mathsf{P}_{F_X}\left(B_n(\mathbb{X}) \ni \boldsymbol{\theta}_X\right) = 1 - \alpha, \qquad \text{pro každé } F_X \in \mathcal{F}.$$

Příklad. **Odhad střední hodnoty normálního rozdělení**

Data: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

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

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

$$\left(\overline{X}_n - \mathrm{t}_{n-1}\left(1-\frac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}},\quad \overline{X}_n + \mathrm{t}_{n-1}\left(1-\frac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}}\right)$$

$$\equiv \left(\overline{X}_n \mp \mathrm{t}_{n-1}\left(1-\frac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}}\right)$$

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

Data: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

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

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

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

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

κ	0,95	0,975	
u_κ	1,64	1,96	`qnorm(...)`
$t_{500}(\kappa)$	1,65	1,96	`qt(...)`
$t_{100}(\kappa)$	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: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

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

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

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

Lemma 3.2 Interval spolehlivosti po transformaci parametru.

Je-li (η_L, η_U) *(asymptotický) interval spolehlivosti pro parametr* θ_X *s pravděpodobností pokrytí* $1-\alpha$ *a je-li* ψ *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: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

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

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

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

Příklad. **Odhad parametru Poissonova rozdělení**

Data: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_X \equiv \mathcal{P}o(\lambda_X), \lambda_X > 0$

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

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

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

Příklad. Odhad proporce

Data a model: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim \mathcal{A}lt(p_X), 0 < p_X < 1$

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

(Domácí) úkol. Využijte asymptotiku pro $\widehat{p}_n = \overline{X}_n$

$$\sqrt{n}\,(\widehat{p}_n - p_X) \xrightarrow{\mathcal{D}} \mathcal{N}(0, p_X(1-p_X)), \quad \widehat{p}_n \xrightarrow{\mathrm{P}} p_X, \qquad n \to \infty$$

k odvození asyptotického intervalu spolehlivosti pro p_X.

$$\left(\widehat{p}_n - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}, \quad \widehat{p}_n + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}\right)$$

$$\equiv \left(\widehat{p}_n \mp u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}\right)$$

Oddíl **3.5**

Empirické odhady

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

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

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

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

- Nechť $F_X \in \mathcal{F}$ $\equiv$ skutečné rozdělení náhodné veličiny X.

- $\boldsymbol{\theta}_X := \boldsymbol{t}(F_X)$ $\equiv$ skutečná hodnota parametru zájmu.

Definice 3.6 Empirická distribuční funkce.

Funkci

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

nazýváme empirická distribuční funkce
(*empirical cumulative distribution function – ecdf*)
náhodného výběru $X_1, \ldots, X_n$.

Věta 3.3 Vlastnosti empirické distribuční funkce.

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

(i) $\mathbb{E}_{F_X} \widehat{F}_n(x) = F_X(x)$ *(nestrannost)*, $\quad \operatorname{var}_{F_X} \widehat{F}_n(x) = \dfrac{F_X(x)\{1 - F_X(x)\}}{n}$;

(ii) $\widehat{F}_n(x) \xrightarrow{\mathrm{P}} F_X(x),\ n \to \infty$ *(bodová konzistence)*;

(iii) $\sqrt{n}\{\widehat{F}_n(x) - F_X(x)\} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0,\ F_X(x)\{1 - F_X(x)\}\Big),\ n \to \infty$;

(iv) $n\widehat{F}_n(x) \sim \mathcal{B}i\big(n, F_X(x)\big)$;

(v) $\sup\limits_{x \in \mathbb{R}} \left|\widehat{F}_n(x) - F_X(x)\right| \xrightarrow{\mathrm{P}} 0,\ n \to \infty$ *(stejnoměrná konzistence)*.

Idea empirických odhadů

- $\theta_X = \boldsymbol{t}(F_X)$: parametr zájmu
- $\widehat{\theta}_n = \boldsymbol{t}(\widehat{F}_n)$: empirický odhad parametru θ_X

Příklad. **Empirický odhad střední hodnoty**

Data: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_X$

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

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

$$\widehat{\mu}_n = \mathbb{E}_{\widehat{F}_n} X = \int_{-\infty}^{\infty} x \, \mathrm{d}\widehat{F}_n(x) = \cdots = \overline{X}_n$$

Empirické odhady momentů

- Nechť h je měřitelná reálná funkce taková, že
$$\forall F_X \in \mathcal{F} \quad \mathbb{E}_{F_X} \left| h(X) \right| < \infty.$$
- Snadno: empirický odhad pro $\theta_X = \mathbb{E}_{F_X} h(X)$ jest $\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$.

Příklad. **Empirický odhad rozptylu**

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

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

$$\widehat{\sigma}_n^2 = \cdots = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 = \frac{n-1}{n} S_n^2$$

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

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

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

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

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

- šikmost (*skewness*) $\gamma_3 := \dfrac{\mathbb{E}_{F_X} \left(X - \mathbb{E}\, X\right)^3}{\left(\mathrm{var}_{F_X} X\right)^{3/2}}$

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

- špičatost (*kurtosis*) $\gamma_4 := \dfrac{\mathbb{E}_{F_X} \left(X - \mathbb{E}\, X\right)^4}{\left(\mathrm{var}_{F_X} X\right)^{2}}$

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

Empirické odhady kvantilů

- Kvantilová funkce rozdělení F_X

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

- α-kvantil rozdělení F_X $\qquad u_X(\alpha) := F_X^{-1}(\alpha).$
- Platí: $\qquad F_X(u_X(\alpha)) \geq \alpha, \qquad \forall h > 0 \quad F_X(u_X(\alpha) - h) < \alpha.$

Definice 3.7 Výběrový kvantil.

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

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

Výběrový medián ($\alpha = 0{,}50$)

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

Software, včetně R, `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 R (`median()`, `quantile(, probs = 0.50)`):

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

Lemma 3.4 Charakterizace výběrových kvantilů.

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

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

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 ($\alpha = 0{,}50$)

- $\widehat{m}_n = \underset{c\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \left|X_i - c\right|$,

  srovnej s $\overline{X}_n = \underset{c\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(X_i - c\right)^2$.

- $\min\limits_{c\in\mathbb{R}} \sum_{i=1}^{n} \left|X_i - c\right|$ dosaženo na množině M_n, 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. $\widehat{u}_n(\alpha) = \widehat{F}_n^{-1}(\alpha)$ (jednoznačné).

Věta 3.5 Vlastnosti výběrových kvantilů.

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

(i) *Potom* $\widehat{u}_n(\alpha) \xrightarrow{\mathrm{P}} u_X(\alpha), \quad n \to \infty.$

(ii) *Pokud navíc existuje hustota f_X, která je spojitá a nenulová v bodě $u_X(\alpha)$, pak*

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

Intervalový odhad pro $u_X(\alpha)$

- Lze využít

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

- 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 $\left(X_{(k_L)},\, X_{(k_U)}\right)$, $1 \le k_L < k_U \le n$, tj. požadujeme, pro každé $F_X \in \mathcal{F}$

$$\mathrm{P}_{F_X}\left(\left(X_{(k_L)},\, X_{(k_U)}\right) \ni u_X(\alpha)\right) \ge 1-\beta.$$

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

$$\mathrm{P}\left(\mathcal{B}i(n,\, \alpha) \le k_L - 1\right) \le \frac{\beta}{2}, \qquad \mathrm{P}\left(\mathcal{B}i(n,\, \alpha) \ge k_U\right) \le \frac{\beta}{2}.$$

Normální aproximace a korekce na spojitost

$$\mathrm{P}\Big(\mathcal{B}i(n,\alpha) \le k_L - 1\Big) = \mathrm{P}\Big(\mathcal{B}i(n,\alpha) < k_L\Big) \qquad = \mathrm{P}\Big(\mathcal{B}i(n,\alpha) \le k_L - \tfrac{1}{2}\Big)$$

$$\mathrm{P}\Big(\mathcal{B}i(n,\alpha) \ge k_U\Big) \qquad = \mathrm{P}\Big(\mathcal{B}i(n,\alpha) > k_U - 1\Big) = \mathrm{P}\Big(\mathcal{B}i(n,\alpha) > k_U - \tfrac{1}{2}\Big)$$

$$\mathrm{P}\Big(\mathcal{B}i(n,\alpha) \le k_L - \tfrac{1}{2}\Big) = \mathrm{P}\left(\frac{\mathcal{B}i(n,\alpha) - n\alpha}{\sqrt{n\alpha(1-\alpha)}} \le \frac{k_L - \frac{1}{2} - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right)$$

$$\doteq \Phi\left(\frac{k_L - \frac{1}{2} - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right)$$

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

$$\doteq 1 - \Phi\left(\frac{k_U - \frac{1}{2} - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right)$$

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

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

- $\underbrace{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}_{\mathbb{X}} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}, \quad \boldsymbol{X} \sim F_X \in \mathcal{F}$ (model)

  $\mathcal{F} = \{\text{libovolné rozdělení na } \mathbb{R}^p \text{s konečnou varianční maticí}\}$

  $\equiv$ neparametrický $\mathcal{L}^2$ model

- $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})^\top$, $i = 1, \ldots, n$, $\quad \boldsymbol{X} = (X_1, \ldots, X_p)^\top$

- **Parametry zájmu:**

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

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

Empirické odhady (zřejmé)

$$\widehat{\boldsymbol{\mu}}_n := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i = \overline{\boldsymbol{X}}_n = (\overline{X}_1, \ldots, \overline{X}_p)^\top$$

$$\widehat{\boldsymbol{\Sigma}}_n := \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)\,(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^\top$$

Výběrová varianční matice

$$\mathbb{S}_n^2 := \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)\,(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^\top$$

- diagonála $S_j^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} (X_{i,j} - \overline{X}_j)^2, \quad j = 1, \ldots, p$

  (výběrové rozptyly)

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

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

  (výběrové kovariance)

Tvrzení 3.6 Vlastnosti výběrového průměru a výběrové varianční matice.

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

$$\mathbb{E}_{F_X} \overline{\boldsymbol{X}}_n = \boldsymbol{\mu}_X, \qquad \overline{\boldsymbol{X}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\mu}_X, \quad n \to \infty.$$

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

$$\mathbb{E}_{F_X} \mathbb{S}_n^2 = \boldsymbol{\Sigma}_X, \qquad \mathbb{S}_n^2 \xrightarrow{\mathrm{P}} \boldsymbol{\Sigma}_X, \quad n \to \infty,$$

$$\widehat{\boldsymbol{\Sigma}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\Sigma}_X, \quad n \to \infty.$$

Definice 3.8 Výběrový korelační koeficient.

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

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

4

Principy testování hypotéz

Oddíl **4.1**

Základní pojmy a definice

Nastavení/předpoklady

- $\underbrace{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}_{\mathbb{X}} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}, \quad \boldsymbol{X} \sim F_X \in \mathcal{F}$ (model);
- $\boldsymbol{\theta} := \boldsymbol{t}(F) \in \mathbb{R}^p, F \in \mathcal{F} \equiv$ parametr zájmu
- $\Theta = \{\boldsymbol{t}(F),\ F \in \mathcal{F}\} \equiv$ parametrický prostor
- $\boldsymbol{\theta}_X := \boldsymbol{t}(F_X) \equiv$ skutečná hodnota parametru

---

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

(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: $\theta = \mathbb{E}_F X = \int x \mathrm{d}F(x), \quad \Theta = \mathbb{R}$

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

- Nechť $\Theta_0, \Theta_1 \subset \Theta, \quad \Theta_0 \cap \Theta_1 = \emptyset$, ne nutně $\Theta_0 \cup \Theta_1 = \Theta$.
- Hledáme odpověď na otázku, zda $\boldsymbol{\theta}_X \in \Theta_0$ nebo $\boldsymbol{\theta}_X \in \Theta_1$.

Definice 4.1 Hypotéza a alternativa.

Množinu Θ_0 nazýváme [nulová] hypotéza, množinu Θ_1 nazýváme alternativní hypotéza (alternativa).

Označme

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

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

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

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

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

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

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

H_1: $\theta_X \neq \theta_0$, $\theta_0 \in \mathbb{R}$ 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 $\mathbb{X} = \{\boldsymbol{X}_1, \ldots \boldsymbol{X}_n\}$

$\longrightarrow$ 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 $\left(U_n(\mathbb{X}), C\right)$ a $\left(U_n^*(\mathbb{X}), C^*\right)$ nazveme ekvivalntní, právě když skoro jistě platí

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

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

$$\mathrm{P}_F\big(U_n(\mathbb{X}) \in B\big) = \mathrm{P}_{\boldsymbol{\theta}}\big(U_n(\mathbb{X}) \in B\big) = \int \mathbb{I}\big\{U_n(\mathbb{X}) \in B\big\}\, \mathrm{d}F(\boldsymbol{x}_1) \cdots \mathrm{d}F(\boldsymbol{x}_n).$$

Definice 4.4 Hladina testu.

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

(i) Jestliže kritický obor C splňuje podmínku

$$\sup_{F \in \mathcal{F}_0} \mathrm{P}_F\big(U_n(\mathbb{X}) \in C\big) = \alpha,$$

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

(ii) Jestliže kritický obor C splňuje podmínku

$$\sup_{F \in \mathcal{F}_0} \lim_{n \to \infty} \mathrm{P}_F\big(U_n(\mathbb{X}) \in C\big) = \alpha,$$

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

$\rightarrow$ asymptotický test

Konstrukce statistického testu

1. Stanovíme (regulátor stanoví) požadovanou hladinu α.
2. Navrhneme/najdeme vhodnou testovou statistiku $U_n(\mathbb{X})$.
3. Navrhneme/zvolíme kritický obor $C = C(\alpha)$ tak, aby hladina byla (asymptoticky) rovna α, současně se snažíme o co nejnižší 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 $\alpha \in (0, 1)$.

  $\rightarrow$ Při zadaném α volíme kritický obor tak, že (dosažená) hladina $\alpha' < \alpha$

  $\rightarrow$ konzervativní test.
- Test, kde je skutečná pravděpodobnost chyby I. druhu $>$ požadovaná hladina α

  $\rightarrow$ antikonzervativní test.

Definice 4.5 Silofunkce a síla testu.

Funkce $\beta_n(F) = \mathrm{P}_F\big(U_n(\mathbb{X}) \in C\big)$, funkce $\mathcal{F} \to [0, 1]$, 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

$= 1 - \mathrm{P}_F\big(U_n(\mathbb{X}) \notin C\big)$

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

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

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X,\ X \sim F_X,\ F_X \in \mathcal{F}, \qquad \theta_X := \mathbb{E}_{F_X} X = \int x \,\mathrm{d}F_X(x) \in \mathbb{R} = \Theta$$

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 $F \in \mathcal{F}^A$:

$$\beta_n(F) = \beta_n(\theta) = \mathrm{P}_\theta\Big(\big|U_n(\mathbb{X})\big| \geq u_{1-\alpha/2}\Big) = \mathrm{P}_\theta\left(\left|\frac{\overline{X}_n - \theta_0}{\frac{\sigma_0}{\sqrt{n}}}\right| \geq u_{1-\alpha/2}\right)$$

$$= \cdots = \Phi\Big(-u_{1-\alpha/2} - \big|\nu_n\big|\Big) + 1 - \Phi\Big(u_{1-\alpha/2} - \big|\nu_n\big|\Big),$$

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

Určení rozsahu výběru

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

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

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

Síla testu závisí na

- hladině testu α (dáno regulátorem)
- rozptylu pozorování σ_0^2 (vlastnost „jevu“, který pozoroujeme)
- alternativě (resp. její „vzdálenosti“ δ 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 θ_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 α.
- $C(\alpha)$ 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 α nazveme konzistentním testem, jestliže $\forall F \in \mathcal{F}_1$ platí $\lim_{n\to\infty} \beta_n(F) = 1$.

Definice 4.7 Nestranný test.

Test $(U_n(\mathbb{X}), C)$ na hladině α 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

$\equiv$ 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

$\equiv$ 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!

Volba kritického oboru

Příklady, připomenutí: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X$, 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) $\sim F_X, F_X \in \mathcal{L}_+^2$, $\theta_X := \mathbb{E}_{F_X} X$, $\sigma^2 := \mathrm{var}_{F_X} X$

- 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)	$\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\sigma_0^2}}$	$\sim \mathcal{N}(0, 1)$
(B)	$\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{S_n^2}}$	$\sim \mathrm{t}_{n-1}$
(C)	$\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{S_n^2}}$	$\xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$, $n \to \infty$

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

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

$c_L(\alpha)$, $c_U(\alpha)$: 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 $\big(U_n(\mathbb{X}), C(\alpha)\big)$ na hladině α,

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

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

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

Definice 4.8 P-hodnota.

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

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

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

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

Rozhodování na základě p-hodnoty

- Zamítni H_0, pokud $p(\boldsymbol{x}) \leq \alpha$.

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

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

- Nezamítni H_0, pokud $p(\boldsymbol{x}) > \alpha$.

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

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

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

Např. $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta_X, \sigma_X^2\right)$

H_0: $\theta_X \leq \theta_0$

H_1: $\theta_X > \theta_0$

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

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

$$c_U(\alpha) = \mathsf{t}_{n-1}(1-\alpha)$$

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

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

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

$$\begin{aligned} &\mathrm{H}_0\colon \quad \theta_X \geq \theta_0 \\ &\mathrm{H}_1\colon \quad \theta_X < \theta_0 \end{aligned}$$

$$U_n(\mathbb{X}) = \frac{\sqrt{n}\left(\overline{X}_n - \theta_0\right)}{S_n}, \qquad \text{pokud } \theta_X = \theta_0\text{:} \quad U_n(\mathbb{X}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\, 1),$$

$$c_L(\alpha) = -u_{1-\alpha}$$

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.

Výpočet p-hodnoty pro **oboustranný** kritický obor

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

Např. $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta_X,\, \sigma_X^2\right)$

$$\begin{aligned} \mathrm{H}_0\colon &\quad \theta_X = \theta_0 \\ \mathrm{H}_1\colon &\quad \theta_X \neq \theta_0 \end{aligned}$$

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

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

$$-c_L(\alpha) = c_U(\alpha) = \mathrm{t}_{n-1}\left(1 - \tfrac{\alpha}{2}\right)$$

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*

$$\forall \alpha \in (0, 1): \qquad \sup_{F \in \mathcal{F}_0} \mathsf{P}_F\big(U_n(\mathbb{X}) \in C(\alpha)\big) = \mathsf{P}_{F_X}\big(U_n(\mathbb{X}) \in C(\alpha)\big).$$

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

$$\underbrace{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}_{\mathbb{X}} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}, \quad \boldsymbol{X} \sim F_X \in \mathcal{F}$$

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

Intervalový odhad pro θ_X:

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

$$\text{resp.} \quad \lim_{n\to\infty} \mathsf{P}_{F_X}\Big(\big(\eta_L(\mathbb{X}), \eta_U(\mathbb{X})\big) \ni \theta_X\Big) = 1 - \alpha.$$

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

H_1: $\theta_X \neq \theta_0$

Tvrzení 4.2 Dualita intervalových odhadů a testování.

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

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

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

Tento test má hladinu α *[přesně/asymptoticky].*

Tvrzení 4.2 Dualita intervalových odhadů a testování, pokrač.

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

$$\mathsf{P}_F\Big(U_n(\mathbb{X}, \theta) \in C_\theta(\alpha)\Big) \;=\; \alpha, \quad \textit{resp.} \lim_{n\to\infty} \mathsf{P}_F\Big(U_n(\mathbb{X}, \theta) \in C_\theta(\alpha)\Big) \;=\; \alpha.$$

Sestavme množinu

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

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

$$\mathsf{P}_{F_X}\Big(B_n(\mathbb{X}) \;\ni\; \theta_X\Big) \;=\; 1-\alpha, \quad \textit{resp.} \lim_{n\to\infty} \mathsf{P}_{F_X}\Big(B_n(\mathbb{X}) \;\ni\; \theta_X\Big) \;=\; 1-\alpha.$$

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

Příklad (B).

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

$\mathsf{H}_0\colon \quad \theta_X = \theta$

$\mathsf{H}_1\colon \quad \theta_X \neq \theta$

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

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

$$C_\theta(\alpha) = \left(-\infty, -\mathsf{t}_{n-1}(1-\tfrac{\alpha}{2})\right] \cup \left[\mathsf{t}_{n-1}(1-\tfrac{\alpha}{2}), \infty\right)$$

Příklad (C). Jednostranný interval spolehlivosti

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

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

H_1: $\quad \theta_X < \theta$

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

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

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

Duální k intervalu spolehlivosti $\left(-\infty, \eta_U(\mathbb{X})\right)$, $\qquad \eta_U(\mathbb{X}) = \overline{X}_n + \frac{S_n}{\sqrt{n}}\, u_{1-\alpha}$.

5

Jednovýběrové a párové problémy pro kvantitativní data

Nastavení/předpoklady

- Kvantitativní data

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

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

  $\rightarrow$ testy

  $\rightarrow$ intervaly spolehlivosti

Oddíl **5.1**

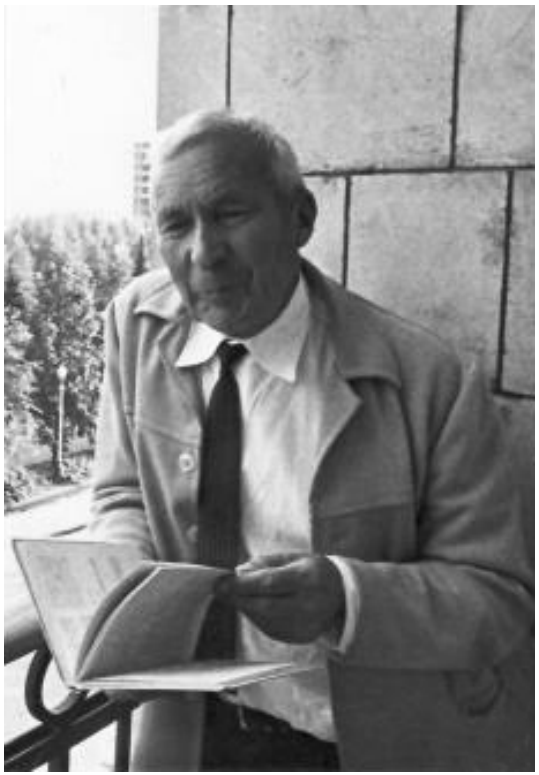
Jednovýběrový Kolmogorovův-Smirnovův test

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

Testovaný parametr: Celá distribuční funkce F_X

Hypotéza a alternativa:

$$\begin{aligned} H_0&: \quad \forall x \in \mathbb{R} \quad F_X(x) = F_0(x) \\ H_1&: \quad \exists x \in \mathbb{R} \quad F_X(x) \neq F_0(x) \end{aligned}$$

F_0: pevně specifikovaná distribuční funkce bez neznámých parametrů

Připomenutí (empirická distribuční funkce):

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

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

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

$K_n^- = \sup_{x \in \mathbb{R}} \left(F_0(x) - \widehat{F}_n(x) \right)$.

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

Lemma 5.1 Kolmogorovova-Smirnovova statistika a pořádkové statistiky.

Je-li F_0 spojitá, pak platí

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

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

Tvrzení 5.2 Asymptotika pro Kolmogorovovu-Smirnovovu statistiku.

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

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

kde Z má distribuční funkci

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

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

Asymptotický test:

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

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

k_n: hodnota statistiky K_n dosažená/realizovaná s daty

R: `ks.test(x, y = F0, exact = FALSE)`

Poznámky.

- $K_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right| \xrightarrow{\mathrm{P}} \sup_{x \in \mathbb{R}} \left| F_X(x) - F_0(x) \right| > 0$ (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

  $$\mathrm{H}_1\colon F_X(x) \geq F_0(x), \quad \exists x \in \mathbb{R}\ F_X(x) > F_0(x).$$

Intervaly spolehlivosti pro F_X

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

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

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

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

$$\forall x \in S_X \quad \forall F_X \in \mathcal{F} \qquad \lim_{n\to\infty} \mathrm{P}_{F_X}\Big(IS_n(x) \ni F_X(x)\Big) = 1 - \alpha$$

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

Intervaly spolehlivosti pro F_X

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

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

$$\forall F_X \in \mathcal{F} \qquad \lim_{n\to\infty} \mathrm{P}_{F_X}\Big(\forall x \in S_X \quad B_n(x) \ni F_X(x)\Big) = 1-\alpha$$

- **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 spojité
   - $K_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|$ lze použít jako testovou statistiku
   - jiné asymptotické rozdělení než uvedené v Tvrzení 5.2
   - použití kvantilů $k_{1-\alpha}$

     $\rightarrow$ (asymptoticky) konzervativní test (hladina OK, ale je slabší)

2. F_0 spojité, ale v datech shody (zaokrouhlování, ...)
   - formálně pozorujeme $\widetilde{X}_1 = \texttt{round}(X_1), \ldots, \widetilde{X}_n = \texttt{round}(X_n)$
     s distribuční funkcí $\widetilde{F}_X(x)$
   - K-S lze použít, pokud $\sup_{x \in \mathbb{R}} \left| \widetilde{F}_X(x) - F_X(x) \right|$ „malé“
     (splněno při rozumném zaokrouhlování)

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

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: $F_X \in \mathcal{F}_0 \overset{\text{např.}}{=} \{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0\}$
  H_1: $F_X \notin \mathcal{F}_0$
- obecně $\mathcal{F}_0 = \{F(x; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}$
- testová statistika $\widetilde{K}_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x; \widehat{\boldsymbol{\theta}}_n) \right|$,
  kde $\widehat{\boldsymbol{\theta}}_n =$ vhodný odhad parametru $\boldsymbol{\theta}_X$
- neplatí Tvrzení 5.2, použití kvantilů $k_{1-\alpha}$
  $\rightarrow$ (asymptoticky) silně konzervativní test (velmi slabý)
- (asymptotické) rozdělení $\widetilde{K}_n$ komplikované a závisí na (neznámém) $\boldsymbol{\theta}_X$
- pro H_0: $F_X \equiv$ normální rozdělení existují speciální testy
  $\rightarrow$ Lilliefors, Shapiro-Wilk, D'Agostino, . . .

Oddíl **5.2**

Jednovýběrový t-test

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: $T_n := \dfrac{\sqrt{n}\,(\overline{X}_n - \mu_0)}{S_n}$

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

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

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

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

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

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

R: `t.test(x, mu = `μ_0`, conf.level = `$1-\alpha$`)`

Lze i jednostranně:

H_0: $\mu_X \leq \mu_0$

H_1: $\mu_X > \mu_0$

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

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

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

R: `t.test(x, mu = μ0, alternative = "greater", conf.level = 1 − α)`

H_0: $\mu_X \geq \mu_0$

H_1: $\mu_X < \mu_0$

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

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

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

R: `t.test(x, mu = μ0, alternative = "less", conf.level = 1 − α)`

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

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

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

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

Přesný test:

Zamítáme $H_0 \iff B_n \le c_L(\alpha) \ \vee \ B_n \ge c_U(\alpha)$

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

$$\begin{aligned} c_L(\alpha) \quad &= \max\left\{k_1 \in \mathbb{N}_0 : \ \mathrm{P}\left(\mathcal{B}i(n, \tfrac{1}{2}) \le k_1\right) \ \le \ \tfrac{\alpha}{2}\right\} \\ &= \max\left\{k_1 \in \mathbb{N}_0 : \ \tfrac{1}{2^n} \sum_{j=0}^{k_1} \binom{n}{j} \ \le \ \tfrac{\alpha}{2}\right\} \\ c_U(\alpha) \quad &= \min\left\{k_2 \in \mathbb{N}_0 : \ \mathrm{P}\left(\mathcal{B}i(n, \tfrac{1}{2}) \ge k_2\right) \ \le \ \tfrac{\alpha}{2}\right\} \\ &= \min\left\{k_2 \in \mathbb{N}_0 : \ \tfrac{1}{2^n} \sum_{j=k_2}^{n} \binom{n}{j} \ \le \ \tfrac{\alpha}{2}\right\} \end{aligned}$$

symetrie: $c_L(\alpha) = k, \quad c_U(\alpha) = n - k$ pro nějaké $k \in \left\{0, \ldots, \tfrac{n}{2}\right\}$

G_0: distribuční funkce rozdělení $\mathcal{B}i(n, \tfrac{1}{2})$

b_n: hodnota statistiky B_n dosažená/realizovaná s daty

Asymptotický test:

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

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

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

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

  ---

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

Porušení předpokladů

- V předpokladech bylo vyžadováno spojité rozdělení. Pro platnost všech odvození stačí vyžadovat $\mathrm{P}(X = m_0) = 0$

  (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

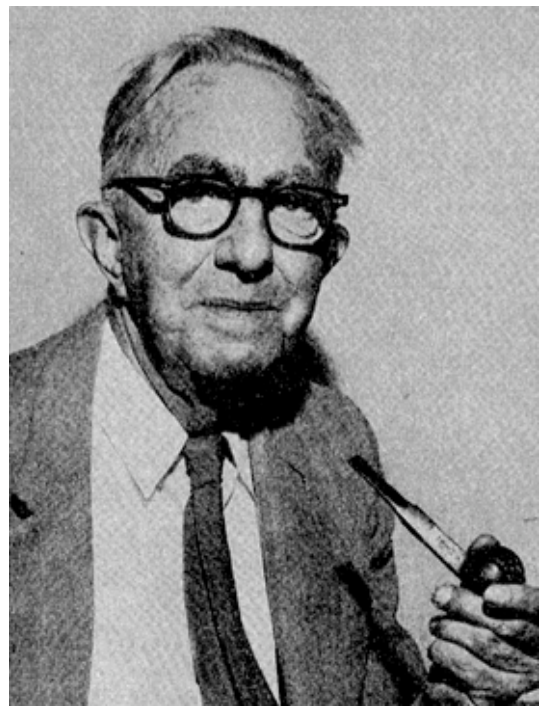
≡ Wilcoxon signed-rank test

Frank Wilcoxon

2.9.1892 (County Cork, IRL)

– 18.11.1965 (Tallahassee, FL, USA)

(fyzikální) chemik (a statistik)



Jednovýběrový Wilcoxonův test

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

Model: $\mathcal{F} = \{$spojitá rozdělení s hustotou f splňující

$$\exists \delta \in \mathbb{R}: \quad f(\delta - x) = f(\delta + x) \quad \text{pro } x \in \mathbb{R}\}$$

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 δ_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:

- $Z_i := X_i - \delta_0, \quad i = 1, \ldots, n$
- Seřaď $|Z_i|$: $0 < |Z|_{(1)} < \cdots < |Z|_{(n)}$
- $R_i =$ pořadí Z_i

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

  tj. $|Z_i| = |Z|_{(R_i)}$
- $W_n := \sum_{i \in \mathcal{I}} R_i$,

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

  $=$ sum of ranks of either sign,

  odsud *Wilcoxon signed-rank test*

Věta 5.4 Vlastnosti testové statistiky Wilcoxonova jednovýběrového testu za hypotézy.

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

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

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

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

Asymptotický test:

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

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

R: `wilcox.test(x, mu = δ0, 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

$$\mathrm{P}_{\delta_0}(W_n = k), \qquad k = 0, \ldots, \frac{n(n+1)}{2} \qquad \sim W_n^0 \text{ s distr. funkcí } G_{n,0}$$

R: `dsignrank()`, `psignrank()`, `qsignrank()`

Přesný test:

Zamítáme $\mathrm{H}_0 \iff W_n \leq c_L(\alpha) \vee W_n \geq c_U(\alpha)$

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

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

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

w_n: hodnota statistiky W_n dosažená/realizovaná s daty

R: `wilcox.test(x, mu = δ0, exact = TRUE)`

Porušení předpokladů

1. Shody kvůli zaokrouhlování
   - $X_i = \delta_0 \longrightarrow$ vyřadit
     (nelze určit, zda před zaokrouhlením bylo $X_i < \delta_0$ nebo $X_i > \delta_0$)
   - $X_i \neq \delta_0 \longrightarrow$ použít průměrná pořadí
     (+ drobná úprava teorie)

2. Nesymetrie rozdělení (hustota f ne symetrická)
   - jednovýběrový Wilcoxonův test se používá na testování o hodnotě
     $$\begin{aligned} \delta_X &= \text{pseudo-medián}(X) \\ &= \text{medián}\left(\frac{X_1 + X_2}{2}\right) \end{aligned}$$
   - δ_X leží mezi 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) α, ale odchylky nejsou velké ani pro značně asymetrická rozdělení jako např. exponenciální

Oddíl **5.5**

Jednovýběrový χ^2 test na rozptyl

Jednovýběrový χ^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: $\sigma_X^2 := \mathrm{var}_{F_X} X$

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: $U_n := \dfrac{(n-1)\,S_n^2}{\sigma_0^2}, \qquad S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)$

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

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

$\chi^2_{n-1}\left(\frac{\alpha}{2}\right)$:	$\frac{\alpha}{2}$-kvantil rozdělení χ^2_{n-1}
$\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)$:	$\left(1-\frac{\alpha}{2}\right)$-kvantil rozdělení χ^2_{n-1}
G_{n-1}:	distribuční funkce rozdělení χ^2_{n-1}
u_n:	hodnota statistiky U_n dosažená/realizovaná s daty

Duální interval spolehlivosti pro σ^2_X

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

$$\forall F_X \in \mathcal{F} \qquad \mathrm{P}_{F_X}\left(IS_n \ni \sigma^2_X\right) = 1-\alpha \qquad \text{přesný IS}$$

Lze i jednostranně:

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

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

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

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

Duální IS pro σ_X^2: $\left(\frac{(n-1)\,S_n^2}{\chi^2_{n-1}(1-\alpha)},\ \infty \right)$

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

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

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

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

Duální IS pro σ_X^2: $\left(0,\ \frac{(n-1)\,S_n^2}{\chi^2_{n-1}(\alpha)} \right)$

Poznámky.

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

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

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

Oddíl **5.6**

Párové testy

Párové testy

- Technicky jednovýběrové testy.
- Odlišná interpretace!

Nastavení/předpoklady

Náhodný výběr z dvourozměrného rozdělení:

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

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

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

Párový t-test

$\equiv$ jednovýběrový t-test s rozdíly $Z_1, \ldots, Z_n$

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

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

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

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

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

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

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

R: `t.test(x, y, paired = TRUE, mu = δ0, conf.level = 1 − α)`

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

$\equiv$ test na významnost změny (zlepšení/zhoršení)

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

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

Přesné rozdělení B_n za H_0: $B_n \sim \mathcal{B}i(n, \frac{1}{2})$

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

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

Poznámky.

- $m_Z = 0$ obecně neznamená $m_X = m_Y$!
- $m_Z = 0 \Leftrightarrow \mathrm{P}(X \leq Y) = \mathrm{P}(X \geq Y) = \frac{1}{2}$
- Jako test hypotézy nulového efektu je test konzistentní (a citlivý) proti alternativám, kde $\mathrm{P}(X \leq Y) \neq \mathrm{P}(X \geq Y)$.
- Zobecnění testu s H_0: $m_Z = m_0$ odpovídá testování H_0: $\mathrm{P}(X - Y \leq m_0) = \mathrm{P}(X - Y \geq m_0) = \frac{1}{2}$.
- 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)

$\equiv$ porovnání středu symetrie rozdělení Z s předem danou konstantou

---

Model: $\mathcal{F} = \{Z$ má spojité rozdělení s hustotou f splňující

$$\exists \delta \in \mathbb{R}: \quad f(\delta - x) = f(\delta + x) \quad \text{pro } x \in \mathbb{R}\}$$

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 δ_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:

- $Z_i^* := X_i - Y_i - \delta_0, \quad i = 1, \ldots, n$
- Seřaď $|Z_i^*|$: $0 < |Z^*|_{(1)} < \cdots < |Z^*|_{(n)}$
- $R_i =$ pořadí Z_i^*

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

  tj. $|Z_i^*| = |Z^*|_{(R_i)}$
- $W_n := \sum_{i \in \mathcal{I}} R_i,$

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

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

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X, \quad X \sim F_X,$$

$$Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} Y, \quad Y \sim F_Y,$$

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

- 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: $\theta_X \neq \theta_Y$ 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: $\mathcal{F} = \{\text{všechna spojitá rozdělení, } F_X \text{ i } F_Y\}$

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: $K_{n,m} = \sup_{x \in \mathbb{R}} \left| \widehat{F}_X(x) - \widehat{F}_Y(x) \right|$

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

Tvrzení 6.1 Asymptotika pro statistiku dvouvýběrového Kolmogorovova-Smirnovova testu.

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

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

kde Z *má distribuční funkci*

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

Poznámka.

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

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

Asymptotický test:

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

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

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

---

```
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 $n, m \to \infty$

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

  - 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 $\longrightarrow$ 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, . . .)

Dvouvýběrový t-test bez předpokladu shody rozptylů

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

Značení: $\sigma_X^2 := \mathrm{var}_{F_X} X$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$,

$\sigma_Y^2 := \mathrm{var}_{F_Y} Y$, $\overline{Y}_m = \frac{1}{m}\sum_{j=1}^m Y_j$, $S_Y^2 = \frac{1}{m-1}\sum_{j=1}^m (Y_j - \overline{Y}_m)^2$.

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

Nechť $X_1, \ldots, X_n$ *a* $Y_1, \ldots, Y_m$ *jsou nezávislé náhodné výběry z rozdělení se středními hodnotami* μ_X *a* μ_Y *a konečnými nenulovými rozptyly* σ_X^2 *a* σ_Y^2. *Pak*

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

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

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

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

---

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

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

$$\forall F_X, F_Y \in \mathcal{F} \qquad \mathrm{P}_{F_X,F_Y}\left(IS_{n,m} \ni \mu_X - \mu_Y\right) \longrightarrow 1 - \alpha$$

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

- 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 σ_X^2/σ_Y^2.
  Lepší zůstat u asymptotiky...

- V praxi se používá (přesnější pro konečné n a m) aproximace rozdělení $T_{n,m}$ za H_0

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

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

  - **Welchův** t-test
  - R: `t.test(x, y, mu = δ0, conf.level = 1 − α)`
    `t.test(z ∼ I, mu = δ0, conf.level = 1 − α)`

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:

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

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

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

Hypotéza a alternativa:

$$\begin{aligned} &\mathrm{H}_0\colon \quad \mu_X - \mu_Y = \delta_0 \\ &\mathrm{H}_1\colon \quad \mu_X - \mu_Y \neq \delta_0 \end{aligned}$$

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

Testová statistika:

$$T_{n,m} := \frac{\overline{X}_n - \overline{Y}_m - \delta_0}{\sqrt{S^2_{n,m}\left(\frac{1}{n} + \frac{1}{m}\right)}} = \sqrt{\frac{n\,m}{n+m}}\,\frac{\overline{X}_n - \overline{Y}_m - \delta_0}{S_{n,m}}$$

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

Věta 6.3 Asymptotika pro statistiku dvouvýběrového t-testu.

Nechť $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ jsou nezávislé náhodné výběry z rozdělení se středními hodnotami μ_X a μ_Y a konečnými nenulovými rozptyly σ^2_X a σ^2_Y. Pak

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

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

kde $\sigma^2_* = \frac{(1-\lambda)\sigma^2_X + \lambda\sigma^2_Y}{\lambda\sigma^2_X + (1-\lambda)\sigma^2_Y}$ $(= 1, \text{ pokud } \sigma^2_X = \sigma^2_Y \text{ nebo } \lambda = \frac{1}{2})$.

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

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

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S^2_{n,m}\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim \mathrm{t}_{n+m-2}.$$

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

Připomenutí: $\mathrm{t}_{n+m-2}\left(1-\frac{\alpha}{2}\right) \longrightarrow u_{1-\frac{\alpha}{2}}, \quad n, m \to \infty.$

Testujeme $\quad \mathrm{H}_0\colon \ \mu_X - \mu_Y = \delta_0$

$\mathrm{H}_1\colon \ \mu_X - \mu_Y \neq \delta_0$

Zamítáme $\mathrm{H}_0 \iff \left|T_{n,m}\right| \geq \mathrm{t}_{n+m-2}\left(1-\frac{\alpha}{2}\right)$

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

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 =` δ_0`, var.equal = TRUE)`

  `t.test(z ~ I, mu =` δ_0`, var.equal = TRUE)`

- Jednostranné testy analogicky.

Duální interval spolehlivosti pro $\mu_X - \mu_Y$

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

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

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

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

-  : `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: $\quad T_{n,m} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_*^2), \qquad \sigma_*^2 = \dfrac{(1-\lambda)\,\sigma_X^2 + \lambda\,\sigma_Y^2}{\lambda\,\sigma_X^2 + (1-\lambda)\,\sigma_Y^2},$

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

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

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 $\Rightarrow$ proveď Welchův t-test

   nezamítnul shodu rozptylů $\Rightarrow$ proveď 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 $\longrightarrow$ 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 $\delta_0 = 0$

  $\approx$ test hypotézy nulového rozdílu

  citlivý na rozdíl ve středních hodnotách.

- Jako test hypotézy nulového rozdílu konzistentní

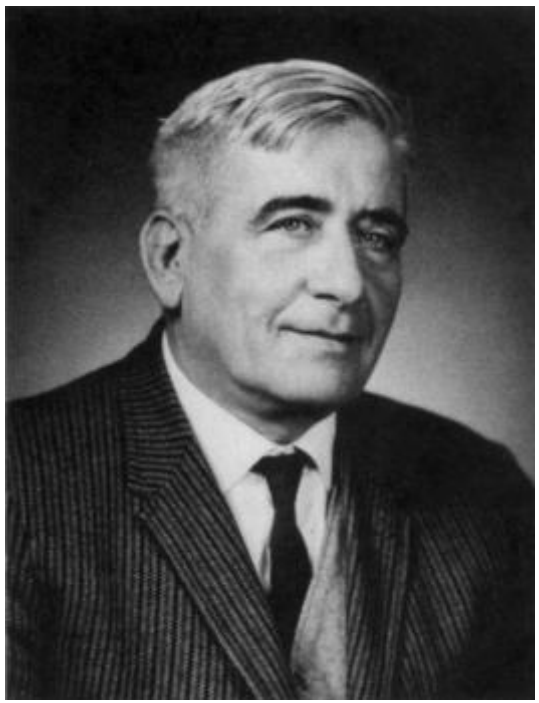
  proti alternativám, kde $\mu_X \neq \mu_Y$.

Oddíl **6.4**

Dvouvýběrový Wilcoxonův test

≡ Wilcoxon rank-sum, Mannův-Whitneyho test

Henry Berthold Mann	**Donald Random Whitney**
27.10.1905 (Vídeň)	27.11.1915 (East Cleveland, Ohio)
– 1.2.2000 (Tucson, Arizona)	– 16.8.2007



Dvouvýběrový Wilcoxonův test

- Neparametrický test založený na pořadích

Model: $\mathcal{F} = \{$ existuje rostoucí funkce g a existuje $\delta \in \mathbb{R}$ tak, že

$g(X) \sim \widetilde{F}_X$, spojitá distribuční funkce,

$g(Y) \sim \widetilde{F}_Y$, spojitá distribuční funkce,

$\forall x \in \mathbb{R} \quad \widetilde{F}_X(x) = \widetilde{F}_Y(x - \delta)\}$

$\mathcal{F} \equiv$ (zobecněný) model posunutí v poloze

Testovaný parametr: posunutí δ_{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:

➠ $\underbrace{(Z_1, \ldots, Z_{n+m})}_{\text{spojený výběr}} \equiv (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$

➠ $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n+m)}$:

pořádkové statistiky ze spojeného výběru

➠ $R_i \equiv$ pořadí X_i mezi $Z_{(1)}, \ldots, Z_{(n+m)}$, tj. $X_i = Z_{(R_i)}$

➠ $W_{n,m} := \sum_{i=1}^{n} R_i =$ 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

$$\mathrm{P}_{\delta_{XY}=0}(W_{n,m} = k), \qquad k = \frac{n(n+1)}{2}, \ldots, nm + \frac{n(n+1)}{2}$$

$$\sim W^0_{n,m} \text{ s distr. funkcí } G_{n,m,0}$$

- $\mathrm{P}\big(\text{libovolné uspořádání } Z_1, \ldots, Z_{n+m}\big) = \dfrac{1}{(n+m)!}$
- $\mathrm{P}\big(R_1 = r_1, \ldots, R_n = r_n\big) = \dfrac{m!}{(n+m)!}$,

  pro každé $r_1 \neq r_2 \neq \cdots \neq r_n \in \{1, \ldots, n+m\}$

R: `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 $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ jsou nezávislé náhodné výběry, pro které platí model $\mathcal{F}$. Nechť dále platí hypotéza H_0: $\delta_{XY} = 0$. Pak

(i) $$\mathbb{E}_{H_0} W_{n,m} = \frac{n(n+m+1)}{2},$$

$$\operatorname{var}_{H_0} W_{n,m} = \frac{nm(n+m+1)}{12}.$$

(ii) *Pokud $n, m \to \infty$,* $$\frac{W_{n,m} - \mathbb{E}_{H_0} W_{n,m}}{\sqrt{\operatorname{var}_{H_0} W_{n,m}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

Asymptotický test:

$$U_{n,m} := \frac{W_{n,m} - \frac{n\,(n+m+1)}{2}}{\sqrt{\frac{n\,m\,(n+m+1)}{12}}}$$

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

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

$u_{n,m}$: hodnota statistiky $U_{n,m}$ dosažená/realizovaná s daty

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

`wilcox.test(z ~ I, exact = FALSE)`

ve výstupu: `W` $= n\,m + \frac{n\,(n+1)}{2} - W_{n,m} =: W^*_{n,m}$

(Mannovo-Whitheyho vyjádření testové statistiky)

Mannova-Whitneyho formulace Wilcoxonova testu

Testová statistika:

- Uvaž všechny dvojice (X_i, Y_j), $\quad i = 1, \ldots, n, j = 1, \ldots, m$

- $W^*_{n,m} := \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}\{X_i < Y_j\}$

  $=$ Mannova-Whitneyho statistika, hodnoty $\in \{0, \ldots, n\,m\}$

Tvrzení 6.6 Vlastnosti Mannovy-Whitneyho statistiky.

(i) $W_{n,m} + W^*_{n,m} = n\,m + \dfrac{n\,(n+1)}{2}$.

(ii) *Pokud* $\min(n, m) \to \infty$, *pak*

$$\frac{W^*_{n,m}}{n\,m} \xrightarrow{\mathrm{P}} \mathrm{P}(X < Y).$$

Důsledky, za platnosti modelu $\mathcal{F}$ a H_0: $\delta_{XY} = 0$:

- $\mathbb{E}_{H_0} W^*_{n,m} = \mathbb{E}_{H_0} (- W_{n,m}) + n\,m + \frac{n(n+1)}{2} = \frac{n\,m}{2}$.
- $\mathrm{var}_{H_0} W^*_{n,m} = \mathrm{var}_{H_0} W_{n,m} = \frac{n\,m(n+m+1)}{12}$.
- Pro $n,\ m \to \infty$:

$$\frac{W^*_{n,m} - \frac{n\,m}{2}}{\sqrt{\frac{n\,m\,(n+m+1)}{12}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\ 1).$$

Důsledky části (ii) Tvrzení 6.6.

- $W^*_{n,m}$ je konzistentním odhadem parametru $\theta_{XY} := \mathrm{P}(X < Y)$.
- Zřejmě platí $F_X = F_Y \implies \theta_{XY} = \frac{1}{2}$.
- Za platnosti zobecněného modelu posunutí (nikoliv obecně) platí též $\theta_{XY} = \frac{1}{2} \implies F_X = F_Y$.

$\Rightarrow$ Wilcoxonův test jako test hypotézy nulového rozdílu H_0: $F_X = F_Y$ je konzistentní vůči alternativám, kde $\theta_{XY} \neq \frac{1}{2}$.

Porušení předpokladů

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 $\mathbb{E}\,X = \mathbb{E}\,Y$,

mnohde (Žižkov/Albertov/Malá Strana/...) se doporučuje dvoukrokový postup

1. Otestuj normalitu výběrů
   (např. Shapirův-Wilkův test, R: `shapiro.test()`).
2. Test zamítnul normalitu $\Rightarrow$ proveď Wilcoxonův test
   nezamítnul normalitu $\Rightarrow$ proveď 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:

$$\mathcal{F} = \left\{F_X \equiv \mathcal{N}(\mu_X, \sigma_X^2),\ F_Y \equiv \mathcal{N}(\mu_Y, \sigma_Y^2),\ \mu_X, \mu_Y \in \mathbb{R}, \sigma_X^2 > 0, \sigma_Y^2 > 0\right\}$$

Testované parametry: $\sigma_X^2 = \mathrm{var}_{F_X} X,\ \sigma_Y^2 = \mathrm{var}_{F_Y} Y$

Hypotéza a alternativa: $\mathrm{H}_0\colon\ \sigma_X^2 = \sigma_Y^2$

$\mathrm{H}_1\colon\ \sigma_X^2 \neq \sigma_Y^2$

Testová statistika: $F := \dfrac{S_X^2}{S_Y^2}$

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

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

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 $\frac{\sigma_X^2}{\sigma_Y^2}$

$$IS = \left(\frac{S_X^2}{S_Y^2}\,\frac{1}{F_{n-1,m-1}\left(1-\frac{\alpha}{2}\right)},\quad \frac{S_X^2}{S_Y^2}\,\frac{1}{F_{n-1,m-1}\left(\frac{\alpha}{2}\right)}\right),$$

$$\forall F_X, F_Y \in \mathcal{F} \qquad \mathrm{P}_{F_X,F_Y}\left(IS \ni \frac{\sigma_X^2}{\sigma_Y^2}\right) = 1-\alpha.$$

- 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

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Y, \quad Y \sim \mathcal{A}lt(p_X) \equiv \text{model } \mathcal{F}, \quad 0 < p_X < 1$$

$$\mathbb{Y} := (Y_1, \ldots, Y_m)^\top$$

- Parametr zájmu: $p_X = \mathsf{P}_{F_X}(Y = 1) = \mathbb{E}_{F_X} Y$
- Značení a již známá fakta:

$$X_n := \sum_{i=1}^n Y_i \sim \mathcal{B}i(n, p_X)$$

$$n - X_n = \sum_{i=1}^n (1 - Y_i) \sim \mathcal{B}i(n, 1 - p_X)$$

$$\widehat{p}_n := \frac{X_n}{n} = \overline{Y}_n \equiv \text{ nestranný a konzistentní odhad } p_X$$

$$\mathbb{E}_{p_X} \widehat{p}_n = p_X$$

$$\mathrm{var}_{p_X} \widehat{p}_n = \frac{\mathrm{var}_{p_X} Y}{n} = \frac{p_X(1 - p_X)}{n}$$

Clopperova-Pearsonova metoda

- Přesný interval spolehlivosti a test o p_X

Test: H_0: $p_X = p_0$

H_1: $p_X \neq p_0$, $\qquad 0 < p_0 < 1$: dáno předem

Testová statistika: $X_n := \sum_{i=1}^{n} Y_i$

Kritický obor: zamítni $H_0 \quad \Leftrightarrow \quad X_n \leq c_L(\alpha)$ 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

$$p(x_n) = 2 \min\left\{ \mathrm{P}_{p_0}(X_n \leq x_n),\ \mathrm{P}_{p_0}(X_n \geq x_n) \right\}$$

$$= 2 \min\left\{ G_0(x_n),\ 1 - G_0(x_n - 1) \right\},$$

$G_0 =$ distribuční funkce $\mathcal{B}i(n, p_0)$

Clopperův-Pearsonův interval spolehlivosti

pro p_X s pokrytím **alespoň** $1-\alpha$

IS_n = množina $p \in (0,1)$, pro něž test na hladině α nezamítá $H_0: p_X = p$

$$G_p(x) := \sum_{j=0}^{x} \binom{n}{j} p^j (1-p)^{n-j} \equiv \text{distribuční funkce } \mathcal{B}i(n,p)$$

$$\ldots$$

Odsud $IS_n = (p_L, p_U)$: $p_L < p_U$ řeší rovnice

$$\sum_{j=0}^{X_n} \binom{n}{j} p^j (1-p)^{n-j} = \frac{\alpha}{2}, \qquad \sum_{j=X_n}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \frac{\alpha}{2}$$

$$p_L = \frac{X_n\, q_L(\alpha)}{X_n\, q_L(\alpha) + n - X_n + 1}, \qquad p_U = \frac{(X_n+1)\, q_U(\alpha)}{(X_n+1)\, q_U(\alpha) + n - X_n},$$

$q_L(\alpha) = \frac{\alpha}{2}$-kvantil rozdělení $F_{2X_n,\, 2(n-X_n+1)}$,

$q_U(\alpha) = (1-\frac{\alpha}{2})$-kvantil rozdělení $F_{2(X_n+1),\, 2(n-X_n)}$,

$p_L = 0$, pokud $X_n = 0$, $\qquad p_U = 1$, pokud $X_n = n$

- Vyšli jsme z testu s pravděpodobností chyby I. druhu $\leq \alpha$.
- To jest,

$$\forall\, p_x \in (0,1) \qquad \mathrm{P}_{p_X}\Big((p_L, p_U) \ni p_X\Big) \quad \geq \quad 1-\alpha$$

$$\Rightarrow \quad \text{konzervativní interval spolehlivosti.}$$

- Pokrytí IS je $\geq \quad 1-\alpha$ pro libovolný rozsah výběru.

Klasická asymptotická metoda

Bylo: $$Z_n := \frac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{\widehat{p}_n\,(1-\widehat{p}_n)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \quad n \to \infty.$$

$\Rightarrow$ asymptotický test H_0: $p_X = p_0$

H_1: $p_X \neq p_0$

Kritický obor: zamítni H_0 $\Leftrightarrow$ $$\left| \frac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{\widehat{p}_n\,(1-\widehat{p}_n)}} \right| \geq u_{1-\frac{\alpha}{2}}$$

p-hodnota: $z_n =$ hodnota $\left| \frac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{\widehat{p}_n\,(1-\widehat{p}_n)}} \right|$ v datech,

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

Duální interval spolehlivosti (asymptotický)

$$\forall\, 0 < p_x < 1 \qquad \lim_{n\to\infty} \mathrm{P}_{p_x}\left(\left(\widehat{p}_n \mp \sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}\, u_{1-\frac{\alpha}{2}} \right) \ni p_x \right) \;=\; 1-\alpha,$$

$\Rightarrow$ (asymptotický) interval spolehlivosti pro p_X s pokrytím $1-\alpha$

$$\left(\max\left\{0,\ \widehat{p}_n - \sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}\, u_{1-\frac{\alpha}{2}} \right\},\quad \min\left\{1,\ \widehat{p}_n + \sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}\, u_{1-\frac{\alpha}{2}} \right\} \right)$$

Nevýhody:

- Pro $p_X \to 0$ a $p_X \to 1$ pomalá asymptotika (potřeba velké n).
- Rule of thumb:

  dostatečně velké $n \quad \Leftrightarrow \quad n\,p_x \geq 5 \;\;\&\;\; n\,(1-p_x) \geq 5.$

Poznámka. Hypotézy H_0: $p_X = p_0$

H_1: $p_X \neq p_0$

lze testovat též pomocí jednovýběrového t-testu (asymptotická verze).

SDC: Ukažte, že příslušná testová statistika má tvar

$$T_n = \frac{\sqrt{n-1}\,(\widehat{p}_n - p_0)}{\sqrt{\widehat{p}_n\,(1-\widehat{p}_n)}}.$$

S kvantily t_{n-1}: mírně konzervativnější test.

Wilsonova metoda

Bylo též: $W_n := \dfrac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{p_0\,(1-p_0)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \quad n \to \infty.$

$\Rightarrow$ asymptotický test H_0: $p_X = p_0$

H_1: $p_X \neq p_0$

Kritický obor: zamítni $\mathrm{H}_0 \Leftrightarrow \left| \dfrac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{p_0\,(1-p_0)}} \right| \geq u_{1-\frac{\alpha}{2}}$

p-hodnota: $w_n =$ hodnota $\left| \dfrac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{p_0\,(1-p_0)}} \right|$ v datech,

$$p(w_n) = 2\left(1 - \Phi\left(|w_n|\right)\right)$$

$\Rightarrow$ **Wilsonův** test

Duální interval spolehlivosti (asymptotický)

$$IS_n = \left\{ p \in (0,1) : \quad \left| \frac{\sqrt{n}\,(\widehat{p}_n - p)}{\sqrt{p\,(1-p)}} \right| < u_{1-\frac{\alpha}{2}} \right\}$$

- potřeba řešit $\sqrt{n}\,\left|\widehat{p}_n - p\right| < \sqrt{p(1-p)}\,u_{1-\frac{\alpha}{2}}$
- kvadratická rovnice $n\,(\widehat{p}_n - p)^2 = p(1-p)\,u^2_{1-\frac{\alpha}{2}}$

Wilsonův interval spolehlivosti pro p_X

$$IS_n = \left(\left(\widehat{p}_n + \frac{u^2_{1-\frac{\alpha}{2}}}{2\,n} \mp \sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n} + \frac{u^2_{1-\frac{\alpha}{2}}}{4\,n^2}} \right) \frac{1}{1 + \frac{u^2_{1-\frac{\alpha}{2}}}{n}} \right)$$

- Lepší vlastnosti (pokrytí pro konečné n) než klasický asymptotický interval spolehlivosti.

Logitová metoda

- Asymptotika v předchozích částech založena na odhadu/statistice $\widehat{p}_n \in [0, 1]$.
- Rozdělení statistiky $\widehat{p}_n$ 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 θ_X: $\widehat{\theta}_n = \log \frac{\widehat{p}_n}{1-\widehat{p}_n}$

$\vdots$

Asymptotický interval spolehlivosti pro θ_X:

$$\left(\theta_{L,n},\ \theta_{U,n}\right), \qquad \begin{aligned} \theta_{L,n} &= \widehat{\theta}_n - u_{1-\frac{\alpha}{2}}\, D_n, \\ \theta_{U,n} &= \widehat{\theta}_n + u_{1-\frac{\alpha}{2}}\, D_n, \end{aligned} \qquad D_n = \frac{1}{\sqrt{n\,\widehat{p}_n\,(1-\widehat{p}_n)}}$$

Asymptotický interval spolehlivosti pro p_X:

$$\forall\, 0 < p_X < 1 \qquad \lim_{n\to\infty} \mathrm{P}_{p_X}\left(\left(\frac{\mathrm{e}^{\theta_{L,n}}}{1+\mathrm{e}^{\theta_{L,n}}},\ \frac{\mathrm{e}^{\theta_{U,n}}}{1+\mathrm{e}^{\theta_{U,n}}}\right) \ni p_X\right) = 1-\alpha$$

Oddíl **7.2**

Dvouvýběrový problém

Nastavení/předpoklady

- Binární data, dva nezávislé náhodné výběry

  $$Y_{1,1}, \ldots, Y_{1,n} \overset{\text{i.i.d.}}{\sim} \mathcal{A}lt(p_1),$$

  $$Y_{2,1}, \ldots, Y_{2,m} \overset{\text{i.i.d.}}{\sim} \mathcal{A}lt(p_2),$$

  $\mathbb{Y}_1 = \left(Y_{1,1}, \ldots, Y_{1,n}\right)^\top$ nezávislé s $\mathbb{Y}_2 = \left(Y_{2,1}, \ldots, Y_{2,m}\right)^\top$

- $X_1 := \sum_{i=1}^{n} Y_{1,i} \sim \mathcal{B}i(n, p_1)$,

  $X_2 := \sum_{i=1}^{m} Y_{2,i} \sim \mathcal{B}i(m, p_2)$, $\qquad X_1$ nezávislé s X_2

- Cíl: porovnat p_1 a p_2

- Pokud $Y = 1 \equiv$ negativní událost

  $\longrightarrow \quad p_1, p_2$: riziko události

- Odhady (viz dříve): $\quad \widehat{p}_1 = \dfrac{X_1}{n}, \quad \widehat{p}_2 = \dfrac{X_2}{n}$

Příklad. Registrační studie genové terapie Comirnaty, děti 6–23 měsíců, 2 dávky terapie

$Y = 1 \quad \equiv$ PCR detekce RNA viru SARS-CoV-2 do jisté doby od terapie

Aktivní ošetření	Placebo
$n = 1\,178$	$m = 598$
$x_1 = 83$	$x_2 = 51$
$\widehat{p}_1 = 0{,}070$	$\widehat{p}_2 = 0{,}085$

Kvantifikace odlišnosti pravděpodobností/rizik

1. **Rozdíl pravděpodobností** (změna rizika)

$$d_X := p_1 - p_2 \in [-1, 1], \qquad \widehat{d} = \widehat{p}_1 - \widehat{p}_2 = -0{,}015$$

2. **Podíl pravděpodobností** (relativní riziko)

$$r_X := \frac{p_1}{p_2} \in [0, \infty], \qquad \widehat{r} = \frac{\widehat{p}_1}{\widehat{p}_2} = 0{,}826$$

3. **Poměr šancí** (odds ratio) $\quad o_X := \dfrac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} \in [0, \infty], \qquad \widehat{o} = \dfrac{\frac{\widehat{p}_1}{1-\widehat{p}_1}}{\frac{\widehat{p}_2}{1-\widehat{p}_2}} = 0{,}813$

Tvrzení 7.1 Asymptotika pro odhady veličin kvantifikujících odlišnost pravděpodobností.

Nechť $p_1, p_2 \in (0, 1)$, $n \to \infty$, $m \to \infty$, $\frac{n}{m} \to q \in (0, \infty)$. *Potom*

(i) $$\frac{\widehat{d} - d_X}{\sqrt{\widehat{V}_d}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{kde} \quad \widehat{V}_d = \frac{\widehat{p}_1 (1 - \widehat{p}_1)}{n} + \frac{\widehat{p}_2 (1 - \widehat{p}_2)}{m}.$$

(ii) $$\frac{\log(\widehat{r}) - \log(r_X)}{\sqrt{\widehat{V}_r}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{kde} \quad \widehat{V}_r = \frac{1 - \widehat{p}_1}{n \widehat{p}_1} + \frac{1 - \widehat{p}_2}{m \widehat{p}_2}.$$

(iii) $$\frac{\log(\widehat{o}) - \log(o_X)}{\sqrt{\widehat{V}_o}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

$$\text{kde} \quad \widehat{V}_o = \frac{1}{n \widehat{p}_1} + \frac{1}{n (1 - \widehat{p}_1)} + \frac{1}{m \widehat{p}_2} + \frac{1}{m (1 - \widehat{p}_2)}.$$

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 $H_0 \Leftrightarrow$

$$\left| \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{V}_d}} \right| \geq u_{1-\frac{\alpha}{2}}$$

$$\left| \frac{\log\left(\frac{\widehat{p}_1}{\widehat{p}_2}\right)}{\sqrt{\widehat{V}_r}} \right| \geq u_{1-\frac{\alpha}{2}}$$

$$\left| \frac{\log\left(\frac{\frac{\widehat{p}_1}{1-\widehat{p}_1}}{\frac{\widehat{p}_2}{1-\widehat{p}_2}}\right)}{\sqrt{\widehat{V}_o}} \right| \geq u_{1-\frac{\alpha}{2}}$$

- p-hodnoty jako vždy...
- Lze i jednostranně...

Duální intervaly spolehlivosti

Jako vždy. . . , např. pro poměr šancí $o_X = \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right)$:

$$\frac{\log(\widehat{o}) - \log(o_X)}{\sqrt{\widehat{V}_o}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \qquad n \to \infty,\ m \to \infty,\ \frac{n}{m} \to q \in (0, \infty).$$

Tedy $\forall p_1, p_2 \in (0, 1)$

$$\mathrm{P}_{p_1, p_2}\left(\left|\frac{\log(\widehat{o}) - \log(o_X)}{\sqrt{\widehat{V}_o}}\right| < u_{1-\frac{\alpha}{2}}\right) \longrightarrow 1 - \alpha$$

$$\mathrm{P}_{p_1, p_2}\left(\left(\log(\widehat{o}) \mp \sqrt{\widehat{V}_o}\, u_{1-\frac{\alpha}{2}}\right) \ni \log(o_X)\right) \longrightarrow 1 - \alpha$$

$$\mathrm{P}_{p_1, p_2}\left(\left(\widehat{o}\exp\left(-\sqrt{\widehat{V}_o}\, u_{1-\frac{\alpha}{2}}\right),\ \widehat{o}\exp\left(\sqrt{\widehat{V}_o}\, u_{1-\frac{\alpha}{2}}\right)\right) \ni o_X\right) \longrightarrow 1 - \alpha$$

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:	$n = 1\,178,$	$x_1 = 83,$	$\widehat{p}_1 = 0{,}070$
Placebo:	$m = 598,$	$x_2 = 51,$	$\widehat{p}_2 = 0{,}085$

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:	$n = 1\,178,$	$x_1 = 3,$	$\widehat{p}_1 = 0{,}003$
Placebo:	$m = 598,$	$x_2 = 0,$	$\widehat{p}_2 = 0$

Použitelnost asymptotických výsledků:

- Pro $p_1 \to 0$, $p_1 \to 1$, $p_2 \to 0$, $p_2 \to 1$ je asymptotika pomalá

  (potřeba velké n, resp. m).

- Rule of thumb:

$$\text{dostatečně velké } n \quad \Leftrightarrow \quad n\,p_1 \geq 5 \;\;\&\;\; n(1-p_1) \geq 5,$$
$$\text{dostatečně velké } m \quad \Leftrightarrow \quad m\,p_2 \geq 5 \;\;\&\;\; m(1-p_2) \geq 5.$$

8

Multinomické rozdělení a kontingenční tabulky

Nastavení/předpoklady

Kategoriální data, náhodná veličina $Z \sim \mathcal{D}\mathrm{iscr}(1, \ldots, K)$

$$\mathrm{P}(Z = k) = p_k, \qquad k = 1, \ldots, K,$$

$$0 < p_k < 1, \quad k = 1, \ldots, K, \qquad \sum_{k=1}^{K} p_k = 1$$

Data: $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} Z$

Četnosti (*counts*):

$$X_1 := \sum_{i=1}^{n} \mathbb{I}\{Z_i = 1\} \sim \mathcal{B}i(n, p_1)$$

$$\vdots$$

$$X_K := \sum_{i=1}^{n} \mathbb{I}\{Z_i = K\} \sim \mathcal{B}i(n, p_K)$$

- $X_1, \ldots, X_K$ ne nezávislé! $\boldsymbol{X} := (X_1, \ldots, X_K)^\top$
- Zřejmě, pro $x_1, \ldots, x_K \in \mathbb{N}_0$, $\sum_{k=1}^{K} x_k = n$

$$\mathrm{P}(X_1 = x_1, \ldots, X_K = x_K) = \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}$$

Oddíl **8.1**

Multinomické rozdělení

Definice 8.1 Multinomické rozdělení.

Nechť $K \geq 2$, $n \geq 1$, $\boldsymbol{p} = (p_1, \ldots, p_K)^\top$ tak, že $0 < p_k < 1$, $k = 1, \ldots, K$, $\sum_{k=1}^{K} p_k = 1$. Náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_K)^\top$ má multinomické rozdělení (*multinomial distribution*) $\mathcal{M}ult_K(n, \boldsymbol{p})$, právě když jeho hustota vzhledem k součinové míře na $\mathbb{Z}^K$ je

$$\begin{aligned} &\mathrm{P}(X_1 = x_1, \ldots, X_K = x_K) \\ &\qquad = \begin{cases} \dfrac{n!}{x_1! \cdots x_K!}\, p_1^{x_1} \cdots p_K^{x_K}, & \text{pokud } \sum_{k=1}^{K} x_k = n,\ x_k \in \mathbb{N}_0, \\ 0, & \text{jinak.} \end{cases} \end{aligned}$$

Věta 8.1 Rozklad multinomického rozdělení.

Nechť $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{i.i.d.}{\sim} \boldsymbol{Y}$, *kde* $\boldsymbol{Y} \sim \mathcal{M}ult_K(1, \boldsymbol{p})$.

Pak $\boldsymbol{X} := \sum_{i=1}^{n} \boldsymbol{Y}_i \sim \mathcal{M}ult_K(n, \boldsymbol{p})$.

Věta 8.2 Vlastnosti multinomického rozdělení.

Nechť $\boldsymbol{X} = (X_1, \ldots, X_K)^\top \sim \mathcal{M}ult_K(n, \boldsymbol{p})$. *Pak*

(i) $X_k \sim \mathcal{B}i(n, p_k), \quad k = 1, \ldots, K.$

(ii) $\mathbb{E}\, X_k = n p_k, \quad \operatorname{var} X_k = n p_k (1 - p_k), \quad k = 1, \ldots, K.$

(iii) $\operatorname{cov}(X_j, X_k) = -n p_j p_k, \quad j \neq k.$

(iv) $\operatorname{var} \boldsymbol{X} = n \{\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\}.$

Věta 8.3 Asymptotické vlastnosti multinomického rozdělení.

Nechť $\boldsymbol{X}_n = (X_{n,1}, \ldots, X_{n,K})^\top \sim \mathcal{M}ult_K(n, \boldsymbol{p})$. *Pak*

(i) $\frac{1}{\sqrt{n}}(\boldsymbol{X}_n - n\boldsymbol{p}) \xrightarrow{\mathcal{D}} \mathcal{N}_K(\boldsymbol{0}, \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top), \quad n \to \infty.$

(ii) $\sum_{k=1}^{K} \frac{(X_{n,k} - n p_k)^2}{n p_k} \xrightarrow{\mathcal{D}} \chi^2_{K-1}, \quad n \to \infty.$

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

$E_k = n p_k$: očekávaná četnost (***expected*** *count)* kategorie k

Odhady parametrů multinomického rozdělení

Data: $\boldsymbol{X}_n = (X_{n,1}, \ldots, X_{n,K})^\top, \quad X_{n,k} \sim \mathcal{B}i(n, p_k), \; k = 1, \ldots, K.$

Bylo (konzistentní a nestranný odhad): $\widehat{p}_k = \dfrac{X_{n,k}}{n}, \; k = 1, \ldots, K.$

Konzistentní a nestranný odhad vektoru $\boldsymbol{p}$: $\widehat{\boldsymbol{p}}_n = \dfrac{\boldsymbol{X}_n}{n}.$

$$\forall \boldsymbol{c} \in \mathbb{R}^K \quad \sqrt{n}\left(\boldsymbol{c}^\top \widehat{\boldsymbol{p}}_n - \boldsymbol{c}^\top \boldsymbol{p}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_c), \quad n \to \infty,$$

$$\text{kde } V_c = \boldsymbol{c}^\top \left\{\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right\} \boldsymbol{c}$$

$$\widehat{V}_c := \boldsymbol{c}^\top \left\{\text{diag}\left(\widehat{\boldsymbol{p}}_n\right) - \widehat{\boldsymbol{p}}_n \widehat{\boldsymbol{p}}_n^\top\right\} \boldsymbol{c} \xrightarrow{\text{P}} V_c, \quad n \to \infty$$

$$\text{pokud } V_c > 0 \quad \frac{\sqrt{n}\left(\boldsymbol{c}^\top \widehat{\boldsymbol{p}}_n - \boldsymbol{c}^\top \boldsymbol{p}\right)}{\sqrt{\widehat{V}_c}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad n \to \infty$$

Asymptotický test pro zadané $\boldsymbol{c} \in \mathbb{R}^K$ a $\gamma_0 \in \mathbb{R}$

$$\begin{aligned} &\mathrm{H}_0\colon \quad \boldsymbol{c}^\top \boldsymbol{p} = \gamma_0 \\ &\mathrm{H}_1\colon \quad \boldsymbol{c}^\top \boldsymbol{p} \neq \gamma_0 \end{aligned}$$

Zamítej H_0 na hladině α $\quad \Longleftrightarrow \quad \left| \dfrac{\sqrt{n}\left(\boldsymbol{c}^\top \widehat{\boldsymbol{p}}_n - \gamma_0\right)}{\sqrt{\widehat{V}_c}} \right| \geq u_{1-\frac{\alpha}{2}}$

(Duální) **asymptotický interval spolehlivosti** pro $\boldsymbol{c}^\top \boldsymbol{p}$ s pokrytím $1 - \alpha$

$$\left(\boldsymbol{c}^\top \widehat{\boldsymbol{p}}_n \mp \sqrt{\frac{\widehat{V}_c}{n}}\, u_{1-\frac{\alpha}{2}} \right)$$

Inference pro $\theta = g(\boldsymbol{p})$, kde g dostatečně hladká

- Δ-metoda
- pokud $V(\boldsymbol{p}) = \left(\dfrac{\partial g}{\partial \boldsymbol{p}}\right)^\top \left\{\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right\} \dfrac{\partial g}{\partial \boldsymbol{p}} > 0$

χ^2-test dobré shody pro multinomické rozdělení

Data (četnosti): $\boldsymbol{X} = (X_1, \ldots, X_K)^\top$, $\boldsymbol{X} \sim \mathcal{M}ult(n, \boldsymbol{p})$

H_0: $\boldsymbol{p} = \boldsymbol{p}^0$, $\qquad \boldsymbol{p}^0 = (p_1^0, \ldots, p_K^0)^\top$

H_1: $\boldsymbol{p} \neq \boldsymbol{p}^0$ $\qquad$ předem zadané pravděpodobnosti

Testová statistika:

$$\chi^2 = \sum_{k=1}^{K} \frac{(X_k - n p_k^0)^2}{n p_k^0} = \sum_{k=1}^{K} \left(\frac{X_k - n p_k^0}{\sqrt{n p_k^0}} \right)^2$$

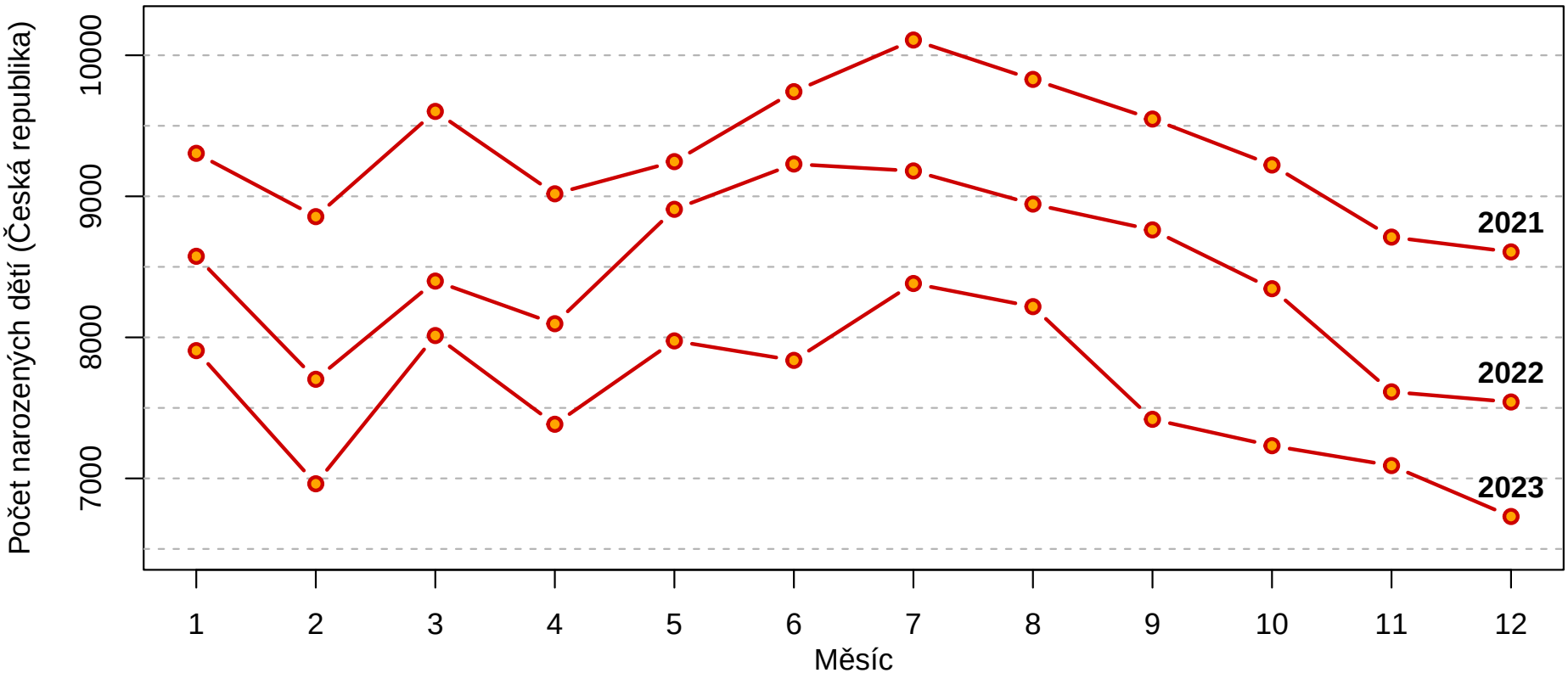
$$= \sum_{k=1}^{K} \left(\frac{O_k - E_k}{\sqrt{E_k}} \right)^2 \overset{\text{as.}}{\sim} \chi^2_{K-1}$$

Zamítej H_0 na hladině α $\iff$ $\chi^2 \geq \chi^2_{K-1}(1-\alpha)$

p-hodnota $= 1 - G_{K-1}(s_X)$,

s_X = napoz. hodn. test. stat., $\quad G_{K-1}$: distr. funkce rozděl. χ^2_{K-1}

Příklad. Rodí se děti během roku rovnoměrně?



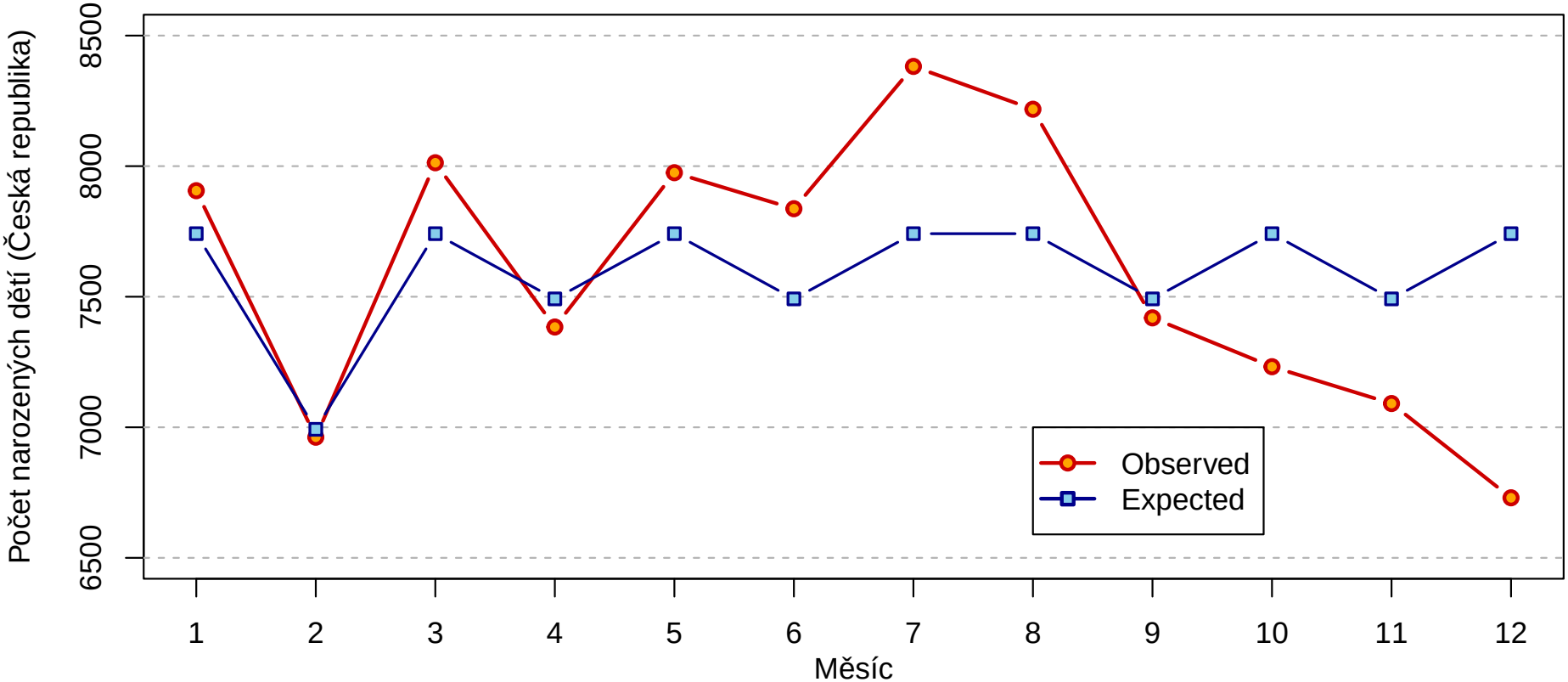
Pro vybraný rok: $X_1, X_2, \ldots, X_{12} \equiv$

počet narozených dětí v lednu, únoru, . . . , prosinci

$p_1^0 = \frac{31}{365}, p_2^0 = \frac{28}{365}, \ldots, p_{12}^0 = \frac{31}{365}$

R: `chisq.test(x, p = p^0)`

Rok **2023**:



$\chi^2 = 307{,}8$

$\chi^2_{11}(0{,}95) = 19{,}7$

p-hodnota $< 0{,}001$

Poznámky.

- Asymptotika (rychlost konvergence k rozdělení χ^2_{K-1})

  Rule of thumb: aproximace rozdělením χ^2_{K-1} uspokojivá,

  pokud $\forall k = 1, \ldots, K \quad E_k = n\,p_k^0 \geq 5$.

- $K = 2$: $\quad p_0 := p_1^0, \quad p_2^0 = 1 - p_0, \quad X_2 = n - X_1$

$$\chi^2 = \frac{(X_1 - n\,p_0)^2}{n\,p_0} + \frac{(n - X_1 - n\,(1 - p_0))^2}{n\,(1 - p_0)} = W_n^2,$$

$$W_n = \frac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{p_0\,(1 - p_0)}}, \quad \widehat{p}_n = \frac{X_1}{n}$$

W_n: Wilsonova testová statistika pro test H_0: $p_X = p_0$ ve výběru z $\mathcal{A}lt(p_X)$

Za platnosti H_0: $\quad W_n \overset{\text{as.}}{\sim} \mathcal{N}(0, 1), \quad \chi^2 = W_n^2 \overset{\text{as.}}{\sim} \chi^2_1$

χ^2-test dobré shody pro multinomické rozdělení s odhadnutými parametry

Motivační příklad

$Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} Z$ (ne nutně kategoriální data)

H_0: $Z \sim F_0(z) = F(z; \theta_0)$, θ_0: známý parametr

Test dobré shody

- intervaly $(a_{k-1}, a_k]$, $k = 1, \ldots, K$, $a_0 := -\infty$, $a_K := \infty$, $K << n$
- $X_k := \sum_{i=1}^{n} \mathbb{I}\{Z_k \in (a_{k-1}, a_k]\}$, $k = 1, \ldots, K$
- H_0: $(X_1, \ldots, X_K)^\top \sim \mathcal{M}ult_K(n, \boldsymbol{p}_0)$, kde $p_k^0 = F(a_k; \theta_0) - F(a_{k-1}; \theta_0)$
- zamítnutí H_0 $\equiv$ výběr $Z_1, \ldots, Z_n$ nepochází z rozdělení s distribuční funkcí $F_0 \equiv F(\cdot; \theta_0)$

Co dělat, pokud θ_0 neznámé, např. $F(z; \theta_0) \equiv \mathcal{N}(\mu_0, \sigma_0^2)$?

Obecně

Model $\mathcal{F}_0$: $\boldsymbol{X} = (X_1, \ldots, X_K)^\top \sim \mathcal{M}ult_K(n, \boldsymbol{p}(\theta_X))$

- $\theta_X \in \Theta \subset \mathbb{R}^d$: neznámý paranmetr, $\quad d < K - 1$
- $\boldsymbol{p} : \Theta \longrightarrow (0, 1)^K$ splňující $\forall \boldsymbol{\theta} \in \Theta \quad \mathbf{1}_K^\top \boldsymbol{p}(\boldsymbol{\theta}) = 1$
- Testujeme H_0: platí model $\mathcal{F}_0$

Krok 1: odhad θ_X

metoda maximální věrohodnosti (ML, *maximum likelihood*)

v **multinomickém modelu**

$\vdots$

MLE $\widehat{\theta}_n$ splňuje $$\sum_{k=1}^{K} \frac{X_k}{p_k(\widehat{\theta}_n)} \cdot \frac{\partial p_k(\widehat{\theta}_n)}{\partial \theta} = \mathbf{0}_d$$

soustava věrohodnostních (odhadovacích) rovnic

Krok 2: testování

H_0: $\exists \theta_X \in \Theta \;\; \boldsymbol{p} = \boldsymbol{p}(\theta_X)$ (platí model $\mathcal{F}_0$)

H_0: $\forall \theta_X \in \Theta \;\; \boldsymbol{p} \neq \boldsymbol{p}(\theta_X)$ (model $\mathcal{F}_0$ neplatí)

Testová statistika

$$\chi^2 = \sum_{k=1}^{K} \frac{\left(X_k - n\,p_k(\widehat{\theta}_n)\right)^2}{n\,p_k(\widehat{\theta}_n)}$$

Tvrzení 8.4 Asymptotické rozdělení χ^2 statistiky při neznámých parametrech.

Platí-li hypotéza H_0, má testová statistika χ^2 (za jistých předpokladů regularity) asymptoticky rozdělení χ^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 $O_k := X_k$, $\widehat{E}_k := n\, p_k(\widehat{\theta}_n)$, $\quad k = 1, \ldots, K$
- Testová statistika:

$$\chi^2 = \sum_{k=1}^{K} \frac{(O_k - \widehat{E}_k)^2}{\widehat{E}_k} \overset{\text{as.}}{\sim} \chi^2_{K-d-1} \qquad \text{za } H_0$$

- H_0 zamítáme $\Leftrightarrow \chi^2 \geq \chi^2_{K-d-1}(1-\alpha)$
- Použitelnost asymptotiky: rule of thumb $\widehat{E}_k \geq 5 \ \ \forall k$

Zpět k **motivačnímu příkladu**

(test dobré shody s parametrickým rozdělením)

$Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} Z$ (ne nutně kategoriální data)

H_0: $Z \sim F_X(z) = F(z; \theta_X)$, $\theta_X \in \Theta$: neznámý parametr

---

dále obdobně jako u testu dobré shody se známými parametry:

- intervaly $(a_{k-1}, a_k]$, $k = 1, \ldots, K$, $a_0 := -\infty$, $a_K := \infty$, $K << n$
- $X_k := \sum_{i=1}^{n} \mathbb{I}\{Z_k \in (a_{k-1}, a_k]\}$, $k = 1, \ldots, K$
- H_0: $(X_1, \ldots, X_K)^\top \sim \mathcal{M}ult_K(n, \boldsymbol{p}(\theta_X))$,

  kde $p_k(\theta_X) = F(a_k; \theta_X) - F(a_{k-1}; \theta_X)$
- $\widehat{\theta}_n$ řeší $\displaystyle\sum_{k=1}^{K} \frac{X_k}{p_k(\widehat{\theta}_n)} \cdot \frac{\partial p_k(\widehat{\theta}_n)}{\partial \theta} = \mathbf{0}_d$
- testová statistika $\displaystyle\chi^2 = \sum_{k=1}^{K} \frac{\left(X_k - n\, p_k(\widehat{\theta}_n)\right)^2}{n\, p_k(\widehat{\theta}_n)} \overset{\text{as.}}{\sim} \chi^2_{K-d-1}$ (za H_0)

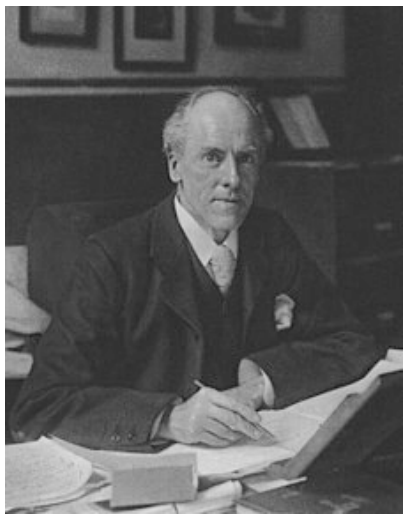
Karl Pearson

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

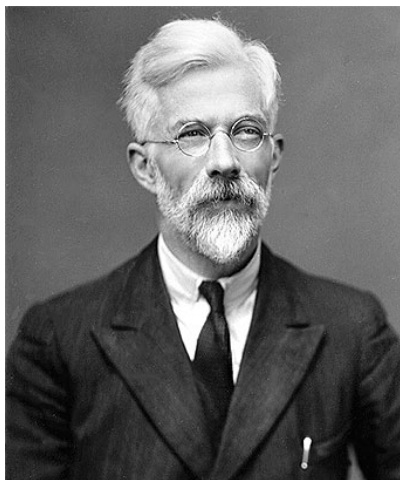
17.2.1890 (Londýn, ENG)

– 29.7.1962 (Adelaide, AUS)

matematik, statistik, biolog, genetik, . . .

(ve věku 23 let)

- zakladatel moderní matematické statistiky
- + zakladatel biostatistiky
  a populační genetiky
- metoda maximální věrohodnosti,
  plánování experimentů
- mnohé další



Pearsonův χ^2 test a spor o (nejenom) stupně volnosti

1900: K. Pearson [43], článek v *Philosophical Magazine*

– koncept χ^2-testu a též principy statistického testování hypotéz
(≈ 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 2×2

– v článku o „klinické studii“ o efektivitě vakcín na choleru 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 Raphaelem 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 (nejenom) 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

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Z \end{pmatrix}, \qquad \begin{aligned} X &\in \{1, \ldots, J\}, \\ Z &\in \{1, \ldots, K\}. \end{aligned}$$

- **Pozorovaná** četnost kategorie (j, k)

$$O_{j,k} = n_{j,k} := \sum_{i=1}^{N} \mathbb{I}\{X_i = j, Z_i = k\}, \quad j = 1, \ldots, J, \; k = 1, \ldots, K$$

- **Sdružené pravděpodobnosti**

$$p_{j,k} := \mathrm{P}(X_i = j, Z_i = k), \quad j = 1, \ldots, J, \; k = 1, \ldots, K,$$

$$\boldsymbol{p} = (p_{1,1}, \ldots, p_{J,K})^\top$$

- Vektor pozorovaných četností

$$\boldsymbol{n} := (n_{1,1}, \ldots, n_{J,K})^\top, \qquad \text{zřejmě: } \boldsymbol{n} \sim \mathcal{M}ult_{JK}(N, \boldsymbol{p})$$

$\Rightarrow$ kontingenční tabulka, **sdruženě multinomický model**

(Pearsonův) χ^2-test nezávislosti (v kontingenční tabulce)

$$X \perp\!\!\!\perp Z \iff \mathrm{P}(X = j, Z = k) = \mathrm{P}(X = j) \cdot \mathrm{P}(Z = k) \quad \forall j, k$$
$$p_{j,k} = p_{j+} \cdot p_{+k}$$

- Při nezávislosti: $\boldsymbol{n} \sim \mathcal{M}ult_{JK}(N, \boldsymbol{p})$,

  kde $\boldsymbol{p} = (p_{1,1}, \ldots, p_{JK})^\top$ = funkce $d = J + K - 2$ parametrů

  $$\theta_X = (p_{1+}, \ldots, p_{(J-1)+}, p_{+1}, \ldots, p_{+(K-1)})^\top$$

- Test H_0: $X \perp\!\!\!\perp Z$

  $\equiv$ χ^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

$$\ell_N(\boldsymbol{\theta}) = \log\left(\frac{N!}{\prod_{j,k} n_{j,k}!} \prod_{j,k} (p_{j+}\, p_{+k})^{n_{j,k}}\right)$$

$$= \sum_{j=1}^{J}\sum_{k=1}^{K} n_{j,k} \log(\underbrace{p_{j+}\, p_{+k}}_{p_{j,k}(\boldsymbol{\theta})}) + \log\left(\frac{N!}{\prod_{j,k} n_{j,k}!}\right)$$

Skórový vektor

$$\frac{\partial \ell_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{n_{j,k}}{p_{j+}\, p_{+k}} \cdot \frac{\partial p_{j,k}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

$$\boldsymbol{\theta} = \left(p_{1+}, \ldots, p_{(J-1)+}, p_{+1}, \ldots, p_{+(K-1)}\right)^\top$$

Pro $k = 1, \ldots, K$:

$$\frac{\partial p_{j,k}(\boldsymbol{\theta})}{\partial p_{j+}} = \frac{\partial \left(p_{j+}\, p_{+k}\right)}{\partial p_{j+}} = p_{+k}, \qquad j = 1, \ldots, J-1,$$

$$\frac{\partial p_{J,k}(\boldsymbol{\theta})}{\partial p_{j+}} = \frac{\partial \left\{ \left(1 - p_{1+} - \cdots - p_{(J-1)+}\right) p_{+k} \right\}}{\partial p_{j+}} = -\, p_{+k}, \qquad j = 1, \ldots, J-1.$$

Odsud

$$\frac{\partial \ell_N}{\partial p_{j+}} = \sum_{k=1}^{K} \frac{n_{j,k}}{p_{j+}\, p_{+k}}\, p_{+k} - \sum_{k=1}^{K} \frac{n_{J,k}}{p_{J+}\, p_{+k}}$$

$$= \sum_{k=1}^{K} \frac{n_{j,k}}{p_{j+}} - \sum_{k=1}^{K} \frac{n_{J,k}}{p_{J+}}, \qquad j = 1, \ldots, J-1.$$

Podobně

$$\frac{\partial \ell_N}{\partial p_{+k}} = \sum_{j=1}^{J} \frac{n_{j,k}}{p_{+k}} - \sum_{j=1}^{J} \frac{n_{j,K}}{p_{+K}}, \qquad k = 1, \ldots, K-1.$$

Věrohodnostní rovnice $(J + K - 2$ rovnic)

$$\frac{n_{j+}}{p_{j+}} = \frac{n_{J+}}{p_{J+}}, \qquad j = 1, \ldots, J-1,$$

$$\frac{n_{+k}}{p_{+k}} = \frac{n_{+K}}{p_{+K}}, \qquad k = 1, \ldots, K-1.$$

Maximálně věrohodný odhad $\widehat{\theta}_N = \left(\widehat{p}_{1+}, \ldots, \widehat{p}_{(J-1)+}, \widehat{p}_{+1}, \ldots, \widehat{p}_{+(K-1)}\right)^\top$

$$\widehat{p}_{j+} = \frac{n_{j+}}{N}, \qquad j = 1, \ldots, J-1,$$

$$\widehat{p}_{J+} = \frac{n_{J+}}{N},$$

$$\widehat{p}_{+k} = \frac{n_{+k}}{N}, \qquad k = 1, \ldots, K-1,$$

$$\widehat{p}_{+K} = \frac{n_{+K}}{N}.$$

Maximálně věrohodné odhady pro $p_{1,1}, \ldots, p_{J,K}$ za nezávislosti

$$\widehat{p}_{j,k} \;=\; p_{j,k}\big(\widehat{\theta}_N\big) \;=\; \widehat{p}_{j+}\,\widehat{p}_{+k} \;=\; \frac{n_{j+}\,n_{+k}}{N^2}, \qquad j = 1, \ldots, J,\; k = 1, \ldots, K.$$

(Odhadnuté) očekávané četnosti

$$\widehat{E}_{j,k} := N\,\widehat{p}_{j,k} \;=\; \frac{n_{j+}\,n_{+k}}{N}, \qquad j = 1, \ldots, J,\; k = 1, \ldots, K.$$

χ^2-statistika

$$\chi^2 \;=\; \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\big(O_{j,k} - \widehat{E}_{j,k}\big)^2}{\widehat{E}_{j,k}}$$

$$= \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\Big(n_{j,k} \;-\; \frac{n_{j+}\,n_{+k}}{N}\Big)^2}{\frac{n_{j+}\,n_{+k}}{N}} \overset{\text{as.}}{\sim} \chi^2_{JK-(J+K-2)-1} \quad \text{(pokud } X \perp\!\!\!\perp Z\text{)}$$

Poznámka: $JK - (J + K - 2) - 1 \;=\; (J-1)\,(K-1)$.

(Pearsonův) χ^2-test nezávislosti

Data a model:

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Z \end{pmatrix}, \qquad p_{j,k} = \mathrm{P}(X = j, Z = k), \quad \begin{matrix} j = 1, \ldots, J, \\ k = 1, \ldots, K. \end{matrix}$$

Pozorované četnosti: $\boldsymbol{n} \sim \mathcal{M}ult_{J\,K}\left(N, \left(p_{1,1}, \ldots, p_{J,k}\right)^\top\right)$

Testované hypotézy:

H_0: $X \perp\!\!\!\perp Z$, tj. $\forall(j, k)$ $p_{j,k} = p_{j+}\, p_{+k}$

H_1: $X \not\perp\!\!\!\perp Z$, tj. $\exists(j, k)$ $p_{j,k} \neq p_{j+}\, p_{+k}$

Testová statistika:

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(O_{j,k} - \widehat{E}_{j,k}\right)^2}{\widehat{E}_{j,k}}$$

$$= \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n_{j,k} - \frac{n_{j+}\, n_{+k}}{N}\right)^2}{\frac{n_{j+}\, n_{+k}}{N}} \overset{\text{as.}}{\sim} \chi^2_{(J-1)\,(K-1)} \quad \text{(pokud } X \perp\!\!\!\perp Z\text{)}$$

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 χ^2 dosažená/realizovaná s daty

R: `chisq.test(table(x, z))`

Použitelnost asymptotické aproximace rozdělením $\chi^2_{(J-1)(K-1)}$

Pravidlo palce: použitelná asymptotika $\iff$ $\widehat{E}_{j,k} \geq 5 \quad \forall\,(j,k)$

$$\frac{n_{j+}\, n_{+k}}{N} \geq 5 \quad \forall\,(j,k)$$

$$\frac{\min_j n_{j+}\ \min_k n_{+k}}{N} \geq 5$$

Společenské (a jim podobné) vědy: pozor na příliš ambiciozní dotazníky!

Test homogenity multinomických rozdělení

$\equiv$ **dvou/vícevýběrový** test s kategoriálními daty

$X \equiv$ příslušnost do jedné z J skupin (dáno předem)

$Z \equiv$ kategoriální odezva $\in \{1, \ldots, K\}$ (dotazník, ...)

Cíl testovat $\quad H_0\colon \mathcal{L}(Z \mid X = 1) = \cdots = \mathcal{L}(Z \mid X = J)$

Označme:

$$\begin{aligned} \boldsymbol{p}_{(1)} &= \big(\mathrm{P}(Z = 1 \mid X = 1), \ldots, \mathrm{P}(Z = K \mid X = 1)\big)^\top \\ &\vdots \\ \boldsymbol{p}_{(J)} &= \big(\mathrm{P}(Z = 1 \mid X = J), \ldots, \mathrm{P}(Z = K \mid X = K)\big)^\top \end{aligned}$$

Testované hypotézy:

$$H_0\colon \boldsymbol{p}_{(1)} = \cdots = \boldsymbol{p}_{(J)}$$

$$H_1\colon \exists\, j \neq \ell \in \{1, \ldots, J\} \quad \boldsymbol{p}_{(j)} \neq \boldsymbol{p}_{(\ell)}$$

Řádkové marginální četnosti: $n_{1+}, \ldots, n_{J+} \equiv$ rozsahy jednotlivých výběrů

Pozorované četnosti:

$$\boldsymbol{n}_{(1)} := \left(n_{1,1}, \ldots, n_{1,K}\right)^\top \sim \mathcal{M}ult_K\left(n_{1+}, \boldsymbol{p}_{(1)}\right)$$

$$\vdots$$

$$\boldsymbol{n}_{(J)} := \left(n_{J,1}, \ldots, n_{J,K}\right)^\top \sim \mathcal{M}ult_K\left(n_{J+}, \boldsymbol{p}_{(J)}\right)$$

Řádkově multinomický model

Pozorování (pohlížejme na X jako na náhodné), pro všechna (j, k):

$$\mathrm{P}\left(Z = k \,\middle|\, X = j\right) = \frac{\mathrm{P}(Z = k, X = j)}{\mathrm{P}(X = j)}$$

$$\text{iff } X \perp\!\!\!\perp Z \quad \frac{\mathrm{P}(Z = k)\,\mathrm{P}(X = j)}{\mathrm{P}(X = j)} = \mathrm{P}(Z = k)$$

To jest, $X \perp\!\!\!\perp Z$

$$\iff \quad \forall\, k \quad \mathrm{P}\big(Z = k \,\big|\, X = 1\big) = \cdots = \mathrm{P}\big(Z = k \,\big|\, X = J\big) = \mathrm{P}(Z = k)$$

$$\iff \quad \boldsymbol{p}_{(1)} = \cdots = \boldsymbol{p}_{(J)}$$

Nezávislost ve **sdruženě multinomickém** modelu

$\iff$

Homogenita řádkových rozdělení v **řádkově multinomickém** modelu

Testová statistika:

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(O_{j,k} - \widehat{E}_{j,k}\right)^2}{\widehat{E}_{j,k}}$$

$$= \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n_{j,k} - \frac{n_{j+}\, n_{+k}}{N}\right)^2}{\frac{n_{j+}\, n_{+k}}{N}} \overset{\text{as.}}{\sim} \chi^2_{(J-1)\,(K-1)} \quad \text{(pokud } \boldsymbol{p}_{(1)} = \cdots = \boldsymbol{p}_{(J)}\text{)}$$

$O_{j,k}$ = observed count = $n_{j,k}$, $\qquad j = 1, \ldots, J,\ k = 1, \ldots, K$

$\widehat{E}_{j,k}$ = expected count (za homogenity)

$= n_{j+}\, \widehat{\mathrm{P}}(Z = k \mid X = j) \overset{\text{homogenita}}{=} n_{j+}\, \widehat{\mathrm{P}}(Z = k)$

$= n_{j+}\, \dfrac{n_{+k}}{N}, \qquad j = 1, \ldots, J,\ k = 1, \ldots, K$

Poznámky

- Technické provedení testu homogenity řádkových rozdělení je shodné s χ^2-testem nezávislosti, ale **interpretace** je jiná!
- Zdůvodnění asymptotiky by nyní bylo odlišné. Nikoliv $N \to \infty$, ale
  - $n_{1+} \to \infty, \ldots, n_{J+} \to \infty$
  - vyjádření, že žádný z výběrů limitně „nemizí“

Oddíl **8.3**

Kontingenční tabulky 2×2 (a $J \times 2$)

Nastavení/předpoklady

Dvourozměrná binární data

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Z \end{pmatrix}, \qquad X \in \{1, 2\}, \quad Z \in \{0, 1\}.$$

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 $(X, Z)^\top$ 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

$$\chi^2 = \sum_{j=1}^{2}\sum_{k=0}^{1} \frac{(n_{j,k} - \frac{n_{j+}\, n_{+k}}{N})^2}{\frac{n_{j+}\, n_{+k}}{N}} \overset{\text{as.}}{\sim} \chi^2_1$$

(pokud $X \perp\!\!\!\perp Z$)

Test homogenity 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

$$p_1 := \mathrm{P}(Z = 1;\ \text{sk. } 1)$$

$$\equiv \mathrm{P}(Z = 1 \,|\, X = 1) = \frac{\mathrm{P}(X = 1, Z = 1)}{\mathrm{P}(X = 1)} \overset{\text{iff } X \perp\!\!\!\perp Z}{=} \mathrm{P}(Z = 1)$$

$$p_2 := \mathrm{P}(Z = 1;\ \text{sk. } 2)$$

$$\equiv \mathrm{P}(Z = 1 \,|\, X = 2) = \frac{\mathrm{P}(X = 2, Z = 1)}{\mathrm{P}(X = 2)} \overset{\text{iff } X \perp\!\!\!\perp Z}{=} \mathrm{P}(Z = 1)$$

Při podmínění hodnotou X, resp. při pevných n_{1+} a n_{2+}:

$$n_{1,1} \sim \mathcal{B}i(n_{1+}, p_1), \qquad n_{2,1} \sim \mathcal{B}i(n_{2+}, p_2)$$

viz předchozí část

$$\chi^2\text{-test nezávislosti } X \text{ a } Z \quad \equiv \quad \text{test } \mathrm{H}_0\colon p_1 = p_2$$

Opakování.

$$\widehat{p}_1 = \frac{n_{1,1}}{n_{1+}}, \qquad \operatorname{var}\widehat{p}_1 = \frac{p_1\,(1-p_1)}{n_{1+}}$$

$$\widehat{p}_2 = \frac{n_{2,1}}{n_{2+}}, \qquad \operatorname{var}\widehat{p}_2 = \frac{p_2\,(1-p_2)}{n_{2+}}$$

$$d_x = p_1 - p_2, \qquad \widehat{d}_X = \frac{n_{1,1}}{n_{1+}} - \frac{n_{2,1}}{n_{2+}},$$

$$r_x = \frac{p_1}{p_2}, \qquad \widehat{r}_X = \frac{n_{1,1}\,n_{2+}}{n_{2,1}\,n_{1+}},$$

$$o_x = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}, \qquad \widehat{o}_X = \frac{n_{1,1}\,n_{2,0}}{n_{1,0}\,n_{2,1}}$$

(křížový poměr)

Pokud X náhodné a $p_1 \equiv \mathrm{P}\big(Z = 1 \,\big|\, X = 1\big)$, $\quad p_2 \equiv \mathrm{P}\big(Z = 1 \,\big|\, X = 2\big)$, 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.

$$\widehat{p}_1 = \frac{n_{1,1}}{n_{1+}}, \qquad \widehat{\operatorname{var}}\,\widehat{p}_1 = \frac{\widehat{p}_1\,(1-\widehat{p}_1)}{n_{1+}}$$

$$\widehat{p}_2 = \frac{n_{2,1}}{n_{2+}}, \qquad \widehat{\operatorname{var}}\,\widehat{p}_2 = \frac{\widehat{p}_2\,(1-\widehat{p}_2)}{n_{2+}}$$

Testová statistika pro testování H_0: $d_X = 0$:

$$T_d = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1\,(1-\widehat{p}_1)}{n_{1+}} + \frac{\widehat{p}_2\,(1-\widehat{p}_2)}{n_{2+}}}} \overset{\text{as.}}{\sim} \mathcal{N}(0,\,1) \qquad \text{za platnosti } \mathrm{H}_0$$

Za platnosti H_0: $p_1 = p_2 =: p$ je $\operatorname{var}\left(\widehat{p}_1 - \widehat{p}_2\right) = p\,(1-p)\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)$

$\longrightarrow$ konzistentní odhad p: $\widetilde{p} := \frac{n_{1,1}+n_{2,1}}{N}$

a testová statistika pro testování H_0: $d_X = 0$:

$$T_d^* = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widetilde{p}\,(1-\widetilde{p})\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)}} \overset{\text{as.}}{\sim} \mathcal{N}(0,\,1) \qquad \text{za platnosti } \mathrm{H}_0$$

Test homogenity několika binomických rozdělení

$X \equiv$ označení skupiny $\in \{1, \ldots, J\}$

$Z \equiv$ 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

$n_{1+}, \ldots, n_{J+}$: (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

$$p_1 := \mathrm{P}(Z = 1;\ \text{sk. } 1)$$

$$\equiv \mathrm{P}\big(Z = 1 \,\big|\, X = 1\big) = \frac{\mathrm{P}(X = 1, Z = 1)}{\mathrm{P}(X = 1)} \overset{\text{iff } X \perp\!\!\!\perp Z}{=} \mathrm{P}(Z = 1)$$

$$\vdots$$

$$p_J := \mathrm{P}(Z = 1;\ \text{sk. J})$$

$$\equiv \mathrm{P}\big(Z = 1 \,\big|\, X = J\big) = \frac{\mathrm{P}(X = J, Z = 1)}{\mathrm{P}(X = J)} \overset{\text{iff } X \perp\!\!\!\perp Z}{=} \mathrm{P}(Z = 1)$$

Při podmínění hodnotou X, resp. při pevných $n_{1+}, \ldots, n_{J+}$:

$$n_{1,1} \sim \mathcal{B}i(n_{1+}, p_1),\ \ldots,\ n_{J,1} \sim \mathcal{B}i(n_{J+}, p_J)$$

analogicky jako při $J = 2$

$$\chi^2\text{-test nezávislosti } X \text{ a } Z \quad \equiv \quad \text{test } \mathrm{H}_0\text{: } p_1 = \cdots = p_J$$

Testová statistika: $$\chi^2 = \sum_{j=1}^{J}\sum_{k=0}^{1} \frac{(n_{j,k} - \frac{n_{j+}\, n_{+k}}{N})^2}{\frac{n_{j+}\, n_{+k}}{N}} \overset{\text{as.}}{\sim} \chi^2_{J-1}$$

(pokud $p_1 = \cdots = p_J$)

trocha algebraických úprav:

$$\chi^2 = \frac{1}{\widetilde{p}\,(1-\widetilde{p})} \sum_{j=1}^{J} n_{j+}\,(\widehat{p}_j - \widetilde{p})^2,$$

$$\widehat{p}_j = \frac{n_{j,1}}{n_{j+}} = \widehat{\mathrm{P}}(Z = 1 \,|\, X = j), \qquad j = 1, \ldots, J,$$

$$\widetilde{p} = \frac{n_{1,1} + \cdots + n_{J,1}}{N} = \widehat{\mathrm{P}}(Z = 1)$$

Fisherův faktoriálový test

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

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \sim \begin{pmatrix} X \\ Z \end{pmatrix}, \qquad \begin{aligned} X &\in \{0, 1\}, \\ Z &\in \{0, 1\}. \end{aligned}$$

Kontingenční tabulka:

		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

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:

$$n_{0,0} \,\big|\, n_{0+},\, n_{1+},\, n_{+0},\, n_{+1} \;\sim\; \text{Hypergeom}\big(n_{0+};\; n_{+0},\, n_{+1}\big)$$

$$\text{P(tabulka)} \;=\; \text{P}\big(n_{0,0} \,\big|\, n_{0+},\, n_{1+},\, n_{+0},\, n_{+1}\big) \;=\; \frac{\binom{n_{+0}}{n_{0,0}}\,\binom{n_{+1}}{n_{0,1}}}{\binom{N}{n_{0+}}}$$

$$= \;\frac{n_{0+}!\; n_{1+}!\; n_{+0}!\; n_{+1}!}{n_{0,0}!\; n_{0,1}!\; n_{1,0}!\; n_{1,1}!\; N!}$$

$$\text{p-hodnota} \;=\; \sum_{\text{TAB}\in\mathcal{T}} \text{P}\big(\text{TAB}\big),$$

$$\mathcal{T} \;=\; \Big\{\text{TAB},\quad \text{pro kterou}\quad \text{P}\big(\text{TAB}\big) \;\le\; \text{P}\big(\text{pozorovaná TAB}\big)\Big\}$$

R: `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:

	$\begin{matrix}0 & 4\\ 4 & 0\end{matrix}$	$\begin{matrix}1 & 3\\ 3 & 1\end{matrix}$	$\begin{matrix}2 & 2\\ 2 & 2\end{matrix}$	$\begin{matrix}3 & 1\\ 1 & 3\end{matrix}$	$\begin{matrix}4 & 0\\ 0 & 4\end{matrix}$
$P(\text{TAB})$	0,014	0,229	0,514	0,229	0,014

p-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

$$\begin{pmatrix} X_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_N \\ Z_N \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \begin{pmatrix} X \\ Z \end{pmatrix}, \qquad \begin{aligned} X &\in \{0, 1\}, \\ Z &\in \{0, 1\}. \end{aligned}$$

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:

$$\begin{aligned} & \text{H}_0\colon p_{1+} = p_{+1} \quad (\text{P}(X = 1) = \text{P}(Z = 1)) \\ \iff\ & \text{H}_0\colon p_{0,1} = p_{1,0} \quad (\text{P}(X = 0, Z = 1) = \text{P}(X = 1, Z = 0)) \end{aligned}$$

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:

$$\widehat{p}_{0,0} = \frac{n_{0,0}}{N}, \qquad \widehat{p}_{0,1} = \frac{n_{0,1} + n_{1,0}}{2\,N} = \widehat{p}_{1,0}, \qquad \widehat{p}_{1,1} = \frac{n_{1,1}}{N}$$

χ^2-statistika (test dobré shody při $d = 2$ neznámých parametrech):

$$\chi^2 = \sum_{j=0}^{1}\sum_{k=0}^{1} \frac{(n_{j,k} - N\widehat{p}_{j,k})^2}{N\widehat{p}_{j,k}} = \frac{(n_{0,1} - n_{1,0})^2}{(n_{0,1} + n_{1,0})} \overset{\text{as.}}{\sim} \chi^2_{4-1-2} = \chi^2_1$$

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