NMSA331 Matematická statistika 1

Poznámky k přednášce

Michal Kulich

Naposledy upraveno dne 12. ledna 2017.

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fysikální fakulta University Karlovy

Tento učební text obsahuje přehled všech vět, definic, tvrzení a poznámek probíraných v přednášce „NMSA331 Matematická statistika 1“ v rámci bakalářského studia oboru „Obecná matematika“ na MFF UK. Nejedná se o plnohodnotnou učebnici ani skripta, protože zde nejsou uvedeny důkazy vět a tvrzení, chybí některé příklady a není zde obsažena látka probíraná na cvičení. Při přípravě na zkoušku je nutné tento text doplnit poznámkami z přednášek a cvičení.

Odkazy na potřebné definice, věty a tvrzení z teorie pravděpodobnosti (začínající písmenem P) se týkají příručky „Základy teorie pravděpodobnosti pro předmět Matematická statistika 1“, která je k dispozici na webových stránkách předmětu NMSA331. Např. tvrzení P.2.2 nebo definici P.6.1 lze najít ve 2., resp. 6. kapitole zmíněné příručky.

Autor děkuje prof. RNDr. Jiřímu Andělovi, DrSc., doc. RNDr. Karlu Zvárovi, CSc., a Ing. Marku Omelkovi, Ph.D., za pečlivé pročtení poznámek a pomoc s odstraněním řady drobných chyb a nepřesností.

Michal Kulich
kulich@karlin.mff.cuni.cz

Dáno v Karlíně dne 12. ledna 2017

Obsah

Značení

$\boldsymbol{a}^\mathsf{T}$	transpozice vektoru $\boldsymbol{a}$
$\boldsymbol{a}^{\otimes 2}$	$\boldsymbol{a}\boldsymbol{a}^\mathsf{T}$
$\|\boldsymbol{a}\|$	eukleidovská norma vektoru $\boldsymbol{a}$
$\xrightarrow{\mathsf{P}}$	konvergence v pravděpodobnosti
$\xrightarrow{\mathsf{sj}}$	konvergence skoro jistě
$\xrightarrow{\mathsf{D}}$	konvergence v distribuci
$X \sim \mathcal{L}$	X má přesné rozdělení $\mathcal{L}$
$X \overset{\mathsf{as.}}{\sim} \mathcal{L}$	X má přibližně (asymptoticky) rozdělení $\mathcal{L}$
α	hladina testu
$\beta(\theta)$	síla testu, silofunkce
γ_3	šikmost náhodné veličiny
$\widehat{\gamma}_3$	empirická šikmost
γ_4	špičatost náhodné veličiny
$\widehat{\gamma}_4$	empirická špičatost
Θ	parametrický prostor
Θ_0	nulová hypotéza
Θ_1	alternativa
λ	Lebesgueova míra na $\mathbb{R}$
μ_S	čítací míra na nejvýše spočetné množině S
μ_k	k-tý centrální moment náhodné veličiny
$\widehat{\mu}_k$	empirický odhad k-tého centrálního momentu
μ'_k	k-tý moment náhodné veličiny
$\widehat{\mu}'_k$	empirický odhad k-tého momentu
$\varrho(X, Y)$	korelační koeficient náhodných veličin X a Y
$\widehat{\varrho}_{jm}$	výběrový korelační koeficient j-té a m-té složky náh. vektoru
σ_X	směrodatná odchylka náhodné veličiny X
σ_X^2	rozptyl náhodné veličiny X
$\widehat{\sigma}_n^2$	empirický odhad rozptylu
$\widehat{\Sigma}_n$	výběrová rozptylová matice
φ	hustota normovaného normálního rozdělení
Φ	distribuční funkce normovaného normálního rozdělení

$\chi^2_f(\alpha)$	α-kvantil rozdělení χ^2_f
Ω	prostor elementárních jevů
$\mathbb{1}_B$	indikátor množiny B
$\mathbf{1}_n$	sloupcový vektor jedniček délky n
$\mathcal{A}$	σ-algebra náhodných jevů na Ω
$\mathcal{B}_0$	borelovská σ-algebra na $\mathbb{R}$
$\mathcal{B}_0^n$	borelovská σ-algebra na $\mathbb{R}^n$
$C,\ C(\alpha)$	kritický obor testu
$c_L(\alpha),\ c_U(\alpha)$	kritické hodnoty
$\mathsf{cor}\,(X, Y)$	korelační koeficient náhodných veličin X a Y
$\mathsf{cor}\,(\boldsymbol{X}, \boldsymbol{Y})$	korelační matice náhodných vektorů $\boldsymbol{X}$ a $\boldsymbol{Y}$
$\mathsf{cov}\,(X_1, X_2)$	kovariance náhodných veličin X_1 a X_2
$\mathsf{cov}\,(\boldsymbol{X}_1, \boldsymbol{X}_2)$	kovarianční matice náhodných vektorů $\boldsymbol{X}_1$ a $\boldsymbol{X}_2$
$\mathsf{diag}\,(\boldsymbol{a})$	diagonální matice obsahující složky vektoru $\boldsymbol{a}$ na diagonále
$\mathsf{E}\,X$	střední hodnota náhodné veličiny (vektoru) X
$\mathsf{E}\,(\boldsymbol{U} \mid \boldsymbol{Z} = \boldsymbol{z})$	podmíněná střední hodnota náhodného vektoru $\boldsymbol{U}$, je-li dáno $\boldsymbol{Z} = \boldsymbol{z}$
$\mathsf{E}\,(\boldsymbol{U} \mid \boldsymbol{Z})$	podmíněná střední hodnota náhodného vektoru $\boldsymbol{U}$, je-li dáno $\boldsymbol{Z}$
$\mathcal{F}$	pravděpodobnostní model pro pozorovaná data
$\mathcal{F}_0$	rozdělení splňující nulovou hypotézu
$\mathcal{F}_1$	rozdělení splňující alternativu
f_X	hustota náhodné veličiny (vektoru) X
$f(\boldsymbol{y} \mid \boldsymbol{z})$	podmíněná hustota náhodného vektoru $\boldsymbol{Y}$, je-li dáno $\boldsymbol{Z} = \boldsymbol{z}$
F_X	distribuční funkce náhodné veličiny (vektoru) X
F_X^{-1}	kvantilová funkce náhodné veličiny X
$\widehat{F}_n$	empirická distribuční funkce
$F_{m,n}(\alpha)$	α-kvantil rozdělení $F_{m,n}$
H_0	nulová hypotéza
H_1	alternativa
$\mathbb{I}_n$	jednotková matice $n \times n$
$\mathcal{L}^p$	množina náhodných veličin na $(\Omega, \mathcal{A}, \mathsf{P})$ s konečným p-tým absolutním momentem
$\mathcal{L}(X)$	rozdělení náhodné veličiny (vektoru) X
m_X	medián náhodné veličiny X
$\widehat{m}_n$	výběrový medián
MSE	střední čtvercová odchylka odhadu
P	pravděpodobnost
P_X	rozdělení náhodné veličiny X, její indukovaná míra na výběrovém prostoru
P_θ	rozdělení dat při hodnotě parametru θ

$\mathsf{r}(\mathbb{A})$	hodnost matice $\mathbb{A}$
$\mathbb{R}$	množina reálných čísel
R_i	pořadí i-tého pozorování
SE	směrodatná chyba odhadu
S_n^2	výběrový rozptyl
S_{jm}	výběrová kovariance j-té a m-té složky náh. vektoru
S_X	nosič rozdělení náhodné veličiny X
$t_f(\alpha)$	α-kvantil rozdělení t_f
$\mathsf{tr}(\mathbb{A})$	stopa matice $\mathbb{A}$
$u_X(\alpha)$	α-kvantil náhodné veličiny X
u_α	α-kvantil rozdělení $\mathsf{N}(0, 1)$
$\widehat{u}_n(\alpha)$	výběrový α-kvantil
$\mathsf{var}\, X$	rozptyl náhodné veličiny X
$\mathsf{var}\, \boldsymbol{X}$	rozptylová matice náhodného vektoru $\boldsymbol{X}$
$\mathsf{var}\,(\boldsymbol{U} \mid \boldsymbol{Z} = \boldsymbol{z})$	podmíněný rozptyl náhodného vektoru $\boldsymbol{U}$, je-li dáno $\boldsymbol{Z} = \boldsymbol{z}$
$\mathsf{var}\,(\boldsymbol{U} \mid \boldsymbol{Z})$	podmíněný rozptyl náhodného vektoru $\boldsymbol{U}$, je-li dáno $\boldsymbol{Z}$
$\mathcal{X}$	výběrový prostor
$X_{(k)}$	k-tá pořádková statistika
$\overline{X}_n$	výběrový průměr náhodného výběru $X_1, \ldots, X_n$

1 Náhodný výběr

1.1 Definice náhodného výběru

Nechť je dán pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$.

Definice 1.1 Posloupnost $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ nezávislých stejně rozdělených náhodných vektorů definovaných na $(\Omega, \mathcal{A}, P)$, z nichž každý má distribuční funkci F_0, nazýváme *náhodný výběr z rozdělení* F_0.[*] Konstantu n nazýváme *rozsah výběru.*[†]

Prvky náhodného výběru mohou být buď reálné náhodné veličiny nebo náhodné vektory (matice apod.). Můžeme je nazývat „pozorování“ nebo „data“. Pro označení náhodného výběru jako celku budeme občas používat značení $\boldsymbol{X}$.

Poznámka. Distribuční funkci F_0, z níž pozorování $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ pocházejí, neznáme. Chceme použít pozorování k tomu, abychom se o F_0 něco potřebného dozvěděli. O distribuční funkci F_0 předpokládáme, že patří do nějaké množiny rozdělení $\mathcal{F}$, které říkáme *model.*

Definice 1.2 *Modelem* pro pozorování $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ rozumíme předem stanovenou množinu rozdělení $\mathcal{F}$, do níž patří neznámé rozdělení F_0.

Poznámka. Rozdělení F_0 je neznámé. Rádi bychom použili pozorovaná data $\boldsymbol{X}$, abychom určili jeho jisté charakteristiky, které nazýváme *parametry*. Formálně jde o nějakou konstantu (nebo vektor konstant) $\boldsymbol{\theta}_0 \in \mathbb{R}^k$, kterou bychom uměli zjistit, kdybychom F_0 znali. Hledaný parametr tedy můžeme obecně zapsat ve tvaru $\boldsymbol{\theta}_0 \equiv t(F_0)$, kde t je nějaký funkcionál.

Příklady (Typy modelů pro reálné náhodné veličiny)**.**

1. Za model $\mathcal{F}$ můžeme např. vzít množinu všech [diskrétních, spojitých] rozdělení na $\mathbb{R}$ s konečnou střední hodnotou [s konečným rozptylem]. Hledané parametry mohou být např. $\mathsf{E}\, X_i$, $\mathsf{var}\, X_i$, $\mathsf{P}[X \le x] \equiv F_0(x)$ nebo kvantil $F_0^{-1}(\alpha)$. Takový model nazýváme *neparametrický*[‡], neboť není možné popsat všechna rozdělení v $\mathcal{F}$ pomocí konečně mnoha parametrů. Symbolem Θ označujeme množinu všech přípustných hodnot parametru $\boldsymbol{\theta} \equiv t(F)$ pro všechna $F \in \mathcal{F}$.
2. Za model $\mathcal{F}$ můžeme vzít množinu všech rozdělení s hustotami tvaru $f(x; \boldsymbol{\theta})$ pro $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$, kde $f(\cdot;\cdot)$ je známá funkce a $\boldsymbol{\theta}$ je neznámá konstanta (např. všechna exponenciální, normální, geometrická rozdělení). Tyto modely nazýváme *parametrické*[§]. V parametrickém modelu lze jakékoli jiné parametry vždy vyjádřit jako funkce $\boldsymbol{\theta}$.

Příklady (Parametrické modely)**.**

[*] Angl. *random sample from distribution* F_0 [†] Angl. *sample size* [‡] Angl. *non-parametric model* [§] Angl. *parametric model*

- $\mathcal{F} = \{\mathsf{N}(\mu, \sigma_0^2),\ \mu \in \mathbb{R},\ \sigma_0^2$ pevně dáno$\}$; $\theta = \mu$, $\Theta = \mathbb{R}$.
- $\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\}$; $\boldsymbol{\theta} = (\mu, \sigma^2)^\mathsf{T}$, $\Theta = \mathbb{R} \times \mathbb{R}^+$.
- $\mathcal{F} = \{\mathsf{Exp}(\lambda),\ \lambda \in \mathbb{R}^+\}$; $\theta = \lambda$, $\Theta = \mathbb{R}^+$.
- $\mathcal{F} = \{\mathsf{Alt}(p),\ p \in (0, 1)\}$; $\theta = p$, $\Theta = (0, 1)$.

Poznámka. Model $\mathcal{F}$ a parametr $\boldsymbol{\theta}$, který nás zajímá, volíme sami. Model vyjadřuje naši apriorní (na datech nezávislou) představu o rozdělení pozorovaných veličin. Volba parametru závisí na otázce, kterou se snažíme zodpovědět pomocí statistické analýzy. Volba modelu a parametru ovlivňuje výběr metody pro analýzu dat (a její výsledky).

1.2 Statistiky

Statistická analýza postupuje tak, že se z náhodného výběru počítají veličiny, které obsahují informaci o požadovaných parametrech, a s nimi se dále pracuje. Těmto veličinám se říká statistiky. Uvažujme náhodný výběr $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$.

Definice 1.3 Pojmem *statistika*[*] nazýváme libovolnou měřitelnou funkci $\boldsymbol{S}(\boldsymbol{X})$ pozorování z náhodného výběru $\boldsymbol{X}$. Statistika je náhodná veličina (náhodný vektor, je-li vícerozměrná).

Statistika nesmí záviset na hodnotách, které neznáme a nepozorujeme. Smí to být pouze funkce dat a známých konstant. Mezi nejčastěji používané statistiky patří výběrový průměr a výběrový rozptyl. Uvažujme nyní výběr reálných náhodných veličin $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^\mathsf{T}$ a zaveďme dvě nejčastěji používané statistiky.

Definice 1.4

(i) Veličina $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ se nazývá *výběrový průměr*[†] náhodného výběru $\boldsymbol{X}$.

(ii) Veličina $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ se nazývá *výběrový rozptyl*[‡] náhodného výběru $\boldsymbol{X}$.

Výběrový rozptyl nemá smysl počítat z jediného pozorování ($n = 1$); uvažujeme-li výběrový rozptyl, automaticky předpokládáme, že $n \geq 2$.

1.2.1 Vlastnosti výběrového průměru

Uvažujme obecný model $\mathcal{F} = \mathcal{L}^2$. Pracujeme tedy s náhodným výběrem $\boldsymbol{X}$, jehož složky X_i jsou nezávislé náhodné veličiny s libovolným rozdělením, které má konečné druhé momenty. Označme $\mu \equiv \mathsf{E}\, X_i$ a $\sigma^2 = \mathsf{var}\, X_i$.

Lemma 1.1

$$\overline{X}_n = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^n (X_i - c)^2.$$

Výběrový průměr tedy minimalizuje součet čtverců odchylek jednotlivých pozorování od libovolného reálného čísla.

Zde končí předn. 1 (5.10.)

Snadno spočítáme první dva momenty výběrového průměru a prozkoumáme jeho limitní chování při $n \to \infty$.

[*] Angl. *statistic* [†] Angl. *sample mean* [‡] Angl. *sample variance*

Věta 1.2 (Vlastnosti průměru)
(i) $\mathsf{E}\,\overline{X}_n = \mu$, $\mathsf{var}\,\overline{X}_n = \frac{\sigma^2}{n}$;
(ii) $\overline{X}_n \xrightarrow{\mathsf{P}} \mu$ pro $n \to \infty$;
(iii) $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \sigma^2)$ pro $n \to \infty$.

Poznámka. Platí-li předpoklad normálního rozdělení, tj. $\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\}$, lze body (i) a (iii) předchozí věty zesílit na

$$\sqrt{n}(\overline{X}_n - \mu) \sim \mathsf{N}(0, \sigma^2) \quad \text{neboli} \quad \overline{X}_n \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{1.1}$$

1.2.2 Relativní četnost

Zvolme nějaký náhodný jev $B \in \mathcal{A}$, označme $p \equiv P(B)$. Nechť $p \in (0, 1)$. Nechť existuje posloupnost n nezávislých pozorování jevu B — označme $X_i = 1$, pokud jev B při i-tém pozorování nastal, a $X_i = 0$, pokud jev B při i-tém pozorování nenastal ($i = 1, \ldots, n$). Pak náhodné veličiny $X_1, \ldots, X_n$ představují náhodný výběr z rozdělení $\mathsf{Alt}(p)$.

Výběrový průměr $\overline{X}_n$ je podílem počtu pozorování, při nichž jev B nastal, a celkového počtu pozorování n. Nazýváme jej *(empirická) relativní četnost*[*] jevu B. Pro relativní četnost $\overline{X}_n$ pochopitelně platí Věta 1.2. Uveďme si ji znovu v podobě specializované na tento případ a přidejme ještě jedno nové tvrzení.

Věta 1.3 (Vlastnosti relativní četnosti)
(i) $\mathsf{E}\,\overline{X}_n = p$, $\mathsf{var}\,\overline{X}_n = \frac{p(1-p)}{n}$;
(ii) $\overline{X}_n \xrightarrow{\mathsf{P}} p$ pro $n \to \infty$;
(iii) $\sqrt{n}(\overline{X}_n - p) \xrightarrow{\mathsf{D}} \mathsf{N}(0, p(1-p))$ pro $n \to \infty$;
(iv) $n\overline{X}_n \sim \mathsf{Bi}(n, p)$.

Podle bodu (ii) můžeme pravděpodobnost jevu B zjistit s libovolnou přesností pomocí relativní četnosti, stačí jen mít dostatek pozorování výskytu tohoto jevu.

1.2.3 Vlastnosti výběrového rozptylu

Nejprve uvažujme obecný model $\mathcal{F} = \mathcal{L}^2$. Označme opět $\mu \equiv \mathsf{E}\,X_i$ a $\sigma^2 = \mathsf{var}\,X_i$. Výběrový rozptyl lze přepsat do různých podob, které se k určitým účelům hodí lépe než původní definice.

Věta 1.4

(i)
$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right). \tag{1.2}$$

(ii) Nechť $\mathbf{1}_n$ je sloupcový vektor n jedniček. Označme $\mathbb{A} = \mathbb{I}_n - \frac{1}{n}\mathbf{1}_n^{\otimes 2}$ (matice $n \times n$). Pak
$$S_n^2 = \frac{1}{n-1}\boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X} = \frac{1}{n-1}\boldsymbol{Y}^\mathsf{T}\mathbb{A}\boldsymbol{Y}, \tag{1.3}$$
kde $\boldsymbol{Y} = \boldsymbol{X} - c\mathbf{1}_n$ pro nějaké $c \in \mathbb{R}$.

[*] Angl. *empirical frequency*

Poznámka. Vzorec (1.2) se používá mj. pro numerický výpočet S_n^2. Vzorec (1.3) přepisuje S_n^2 v podobě kvadratické formy a ukazuje, že S_n^2 je invariantní vůči posunutí pozorování X_i o libovolnou konstantu c.

Povšimněte si, že $\mathbf{1}_n^\mathsf{T}\mathbb{A} = \mathbf{0}^\mathsf{T}$ a matice $\mathbb{A}$ je idempotentní, neboli $\mathbb{A}\mathbb{A} = \mathbb{A}$. Dále máme $\mathsf{r}(\mathbb{A}) = \mathsf{tr}(\mathbb{A}) = n - 1$. U kvadratických forem máme k dispozici šikovný vzorec pro výpočet střední hodnoty.

Lemma 1.5 Nechť $\boldsymbol{Z}$ je náhodný vektor délky n se střední hodnotou $\boldsymbol{\mu}$ a konečnou rozptylovou maticí Σ. Nechť $\mathbb{A}$ je libovolná matice $n \times n$. Pak platí

$$\mathsf{E}\,\boldsymbol{Z}^\mathsf{T}\mathbb{A}\boldsymbol{Z} = \boldsymbol{\mu}^\mathsf{T}\mathbb{A}\boldsymbol{\mu} + \mathsf{tr}\,\mathbb{A}\Sigma.$$

Věta 1.6 (Vlastnosti výběrového rozptylu)

(i) $S_n^2 \xrightarrow{\mathsf{P}} \sigma^2$.

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

(iii) Jestliže $\mathcal{F} = \mathcal{L}^4$ (existuje konečný čtvrtý moment X_i), pak

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

kde γ_4 je špičatost rozdělení X_i.

(iv) Jestliže $\mathcal{F} = \mathcal{L}^4$, pak

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

kde $\Sigma = \begin{pmatrix}\sigma^2 & \sigma^3\gamma_3 \\ \sigma^3\gamma_3 & \sigma^4(\gamma_4 - 1)\end{pmatrix}$ a γ_3 je šikmost rozdělení X_i.

Zde končí předn. 2 (6.10.)

Poznámka. Věta 1.6(iii) říká, že variabilita výběrového rozptylu asymptoticky závisí na špičatosti pozorování. Věta 1.6(iv) říká, že výběrový průměr a výběrový rozptyl mají asymptoticky sdružené normální rozdělení. Jejich kovariance asymptoticky závisí na šikmosti pozorování. Je-li šikmost nulová, výběrový průměr a výběrový rozptyl jsou asymptoticky nezávislé.

Nyní přidáme předpoklad normálního rozdělení, tj. budeme pracovat v menším modelu $\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\}$. Pracujeme tedy s náhodným výběrem $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^\mathsf{T}$, kde X_i jsou nezávislé s rozdělením $\mathsf{N}(\mu, \sigma^2)$. Díky jejich nezávislosti platí $\boldsymbol{X} \sim \mathsf{N}_n(\mu\mathbf{1}_n, \sigma^2\mathbb{I}_n)$.

Nejprve uvedeme dva výsledky, které platí pro libovolné normálně rozdělené náhodné vektory.

Lemma 1.7 Nechť $\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \Sigma)$ a $\mathbb{A}$ je positivně semidefinitní matice typu $n \times n$.

(i) Nechť $\mathbb{B}$ je libovolná matice typu $m \times n$ splňující rovnost $\mathbb{B}\Sigma\mathbb{A} = 0$. Pak náhodná veličina $\boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X}$ a náhodný vektor $\mathbb{B}\boldsymbol{X}$ jsou nezávislé.

(ii) Nechť $\mathbb{B}$ je libovolná positivně semidefinitní matice typu $n \times n$ splňující rovnost $\mathbb{B}\Sigma\mathbb{A} = 0$. Pak jsou náhodné veličiny $\boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X}$ a $\boldsymbol{X}^\mathsf{T}\mathbb{B}\boldsymbol{X}$ nezávislé.

Zde končí předn. 3 (12.10.)

Věta 1.8 (Vlastnosti výběrového rozptylu za normality) Nechť $X_i \sim \mathsf{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ jsou nezávislé. Pak platí

(i)
$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{1.4}$$

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

Poznámka. Z definice χ^2 rozdělení a z centrální limitní věty plyne, že pro velké n lze rozdělení χ^2_{n-1} aproximovat rozdělením $\mathsf{N}(n-1, 2(n-1))$. Odtud a z (1.4) dostaneme pro $n \to \infty$

$$\frac{\frac{(n-1)S_n^2}{\sigma^2} - (n-1)}{\sqrt{n-1}} \overset{\text{as.}}{\sim} \mathsf{N}(0, 2)$$

a nakonec

$$\sqrt{\frac{n-1}{n}}\sqrt{n}(S_n^2 - \sigma^2) \overset{\text{as.}}{\sim} \mathsf{N}(0, 2\sigma^4).$$

Uvědomíme-li si, že špičatost normálního rozdělení je 3, vidíme, že tvrzení (i) z věty 1.8 je v souladu s asymptotickým výsledkem věty 1.6(iii). Věta 1.8(i) udává přesné rozdělení S_n^2 pro normální data, zatímco věta 1.6(iii) udává asymptotické rozdělení S_n^2 pro libovolná data s konečným čtvrtým momentem.

Poznámka. Věta 1.8(ii) říká, že jsou-li data normální, $\overline{X}_n$ a S_n^2 jsou nezávislé pro každé konečné $n > 1$.

Věta 1.9 (limitní věta o T statistice) Nechť $X_1, \ldots, X_n$ je náhodný výběr z libovolného rozdělení se střední hodnotou μ a s konečným rozptylem σ^2. Pak

$$T \equiv \sqrt{n}\,\frac{\overline{X}_n - \mu}{S_n} \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1).$$

Nyní opět přidáme předpoklad normálního rozdělení.

Věta 1.10 (věta o T statistice) Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{N}(\mu, \sigma^2)$. Pak

$$T \equiv \sqrt{n}\,\frac{\overline{X}_n - \mu}{S_n} \sim t_{n-1}.$$

Poznámka. Věta 1.10 udává přesné rozdělení statistiky T pro normální data, zatímco věta 1.9 udává asymptotické rozdělení téže statistiky pro libovolná data s konečným rozptylem. Uvědomte si, že pro $n \to \infty$ hustota t_{n-1} konverguje k hustotě $\mathsf{N}(0, 1)$.

Nyní budeme uvažovat dva nezávislé výběry ze dvou různých normálních rozdělení.

Věta 1.11 (věta o F statistice) Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení $\mathsf{N}(\mu_X, \sigma_X^2)$ a $Y_1, \dots, Y_m$ je náhodný výběr z rozdělení $\mathsf{N}(\mu_Y, \sigma_Y^2)$. Nechť jsou vektory $(X_1, \dots, X_n)^\mathsf{T}$ a $(Y_1, \dots, Y_m)^\mathsf{T}$ nezávislé. Označme výběrové průměry obou výběrů $\overline{X}_n$ a $\overline{Y}_m$ a výběrové rozptyly

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{a} \quad S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2.$$

Pak platí

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

1.3 Uspořádaný náhodný výběr

Mějme náhodný výběr $X_1, \dots, X_n$ z jednorozměrného spojitého rozdělení s distribuční funkcí F a hustotou f vzhledem k Lebesgueově míře. Nechť $n \geq 2$. Jelikož $X_1, \dots, X_n$ jsou nezávislé a mají spojité rozdělení, $\mathsf{P}[X_i = X_j] = 0$ pro každé $i \neq j$.

Definice 1.5 (Uspořádaný náhodný výběr a pořadí)

(i) Seřadíme-li všechny náhodné veličiny $X_1, \dots, X_n$ od nejmenší do největší, získáme *uspořádaný náhodný výběr*[*]

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

Symbolem $X_{(k)}$ rozumíme k-tou nejmenší hodnotu mezi pozorováními $X_1, \dots, X_n$; nazýváme ji k-tá *pořádková statistika*[†].

(ii) *Pořadím*[‡] náhodné veličiny X_i ve výběru $X_1, \dots, X_n$ rozumíme přirozené číslo $R_i \in \{1, \dots, n\}$ takové, že $X_i = X_{(R_i)}$.

Poznámka.

1. Hodnoty $X_1, \dots, X_n$ lze jednoznačně určit z n-tice pořádkových statistik a n-tice pořadí.
2. První pořádková statistika je minimum, n-tá pořádková statistika je maximum všech veličin náhodného výběru.
3. Platí $R_i = \sum_{j=1}^{n} \mathbb{1}_{\langle 0,\infty)}(X_i - X_j)$.
4. Pořádkové statistiky a pořadí jsou náhodné veličiny a též statistiky ve smyslu definice 1.3.

Označme symbolem $\mathcal{P}_n$ množinu všech permutací posloupnosti $(1, \dots, n)$. Tato množina má $n!$ prvků.

Věta 1.12 Sdružená hustota náhodného vektoru $(X_{(1)}, \dots, X_{(n)})^\mathsf{T}$ vzhledem k Lebesgueově míře jest

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

[*] Angl. *ordered random sample* [†] Angl. *order statistic* [‡] Angl. *rank*

Poznámka. Náhodné veličiny $X_{(1)}, \dots, X_{(n)}$ nejsou nezávislé. Náhodné veličiny $R_1, \dots, R_n$ nejsou nezávislé.

Věta 1.13 Distribuční funkce k-té pořádkové statistiky jest

$$F_{(k)}(x) = \mathsf{P}[X_{(k)} \le x] = \sum_{j=k}^{n} \binom{n}{j} F^j(x)[1-F(x)]^{n-j} =$$

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

Zde končí předn. 4 (13.10.)

Důsledky.

1. Mají-li X_i rovnoměrné rozdělení na intervalu $(0,\, 1)$, pak $X_{(k)}$ má beta rozdělení $\mathsf{B}(k,\, n-k+1)$. Z toho plyne
$$\mathsf{E}\, X_{(k)} = \frac{k}{n+1}, \qquad \mathsf{var}\, X_{(k)} = \frac{k(n-k+1)}{(n+2)(n+1)^2}.$$
2. Nechť mají X_i jakékoli spojité rozdělení s ryze rostoucí distribuční funkcí F. Nechť $Z \sim \mathsf{B}(k,\, n-k+1)$. Pak $\mathsf{P}[X_{(k)} \le x] = \mathsf{P}[Z \le F(x)] = \mathsf{P}[F^{-1}(Z) \le x]$, tj. $X_{(k)}$ má stejné rozdělení jako $F^{-1}(Z)$.

Věta 1.14 Hustota k-té pořádkové statistiky vzhledem k Lebesgueově míře jest

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

Věta 1.15 Náhodný vektor $(R_1, \dots, R_n)^{\mathsf{T}}$ nabývá všech hodnot na množině $\mathcal{P}_n$, přičemž každá z nich má pravděpodobnost $1/n!$.

Věta 1.16 Platí

(i) $\mathsf{P}[R_i = k] = \frac{1}{n}$ pro všechna $i,\, k \in \{1, \dots, n\}$.
(ii) $\mathsf{P}[R_i = k,\, R_j = m] = \frac{1}{n(n-1)}$ pro všechna $i \ne j,\, k \ne m \in \{1, \dots, n\}$.
(iii) $\mathsf{E}\, R_i = \frac{n+1}{2}$, $\mathsf{var}\, R_i = \frac{n^2-1}{12}$ pro všechna $i \in \{1, \dots, n\}$.
(iv) $\mathsf{cov}\,(R_i,\, R_j) = -\frac{n+1}{12}$ pro všechna $i \ne j \in \{1, \dots, n\}$.

Pokud data nepocházejí ze spojitého rozdělení nebo se v nich nacházejí shodná pozorování vzniklá vlivem zaokrouhlování, pořadí nelze stanovit jednoznačně. V takovém případě je možné všem shodným pozorováním přiřadit jejich průměrné pořadí nebo jim jejich pořadí stanovit náhodně. Většina výsledků odvozených pro pořadí pocházející ze spojitého rozdělení však pro takto upravená pořadí neplatí.

1.4 Transformovaný náhodný výběr

1.4.1 Transformace pozorování

Mějme náhodný výběr $X_1, \dots, X_n$ z rozdělení s distribuční funkcí F_X, hustotou f_X a nosičem S_X. Uvažujme ryze monotonní* diferencovatelnou funkci $g : S_X \to \mathbb{R}$ a definujme $Y_i = g(X_i)$.

* Nemonotonním transformacím se obvykle vyhýbáme, protože by mohly ztotožnit pozorování, která byla původně výrazně odlišná.

Potom $Y_1, \ldots, Y_n$ je náhodný výběr z rozdělení s hustotou f_Y. Kdyby rozdělení F_X bylo spojité a kdybychom znali f_X, spočítali bychom hustotu f_Y z tvrzení P.5.3.

Transformace pozorování se ve statistice používají dosti často. Běžný důvod pro provedení transformace bývá, že původní náhodný výběr $X_1, \ldots, X_n$ příliš porušuje předpoklady metod, které bychom chtěli použít (například normalitu, symetrii hustoty, existenci momentů apod.). Najdeme tedy vhodnou funkci g takovou, že $Y_i = g(X_i)$ splňuje předpoklady lépe než původní pozorování a pracujeme s náhodným výběrem $Y_1, \ldots, Y_n$ namísto původního náhodného výběru $X_1, \ldots, X_n$. Mezi nejčastěji používané transformace kladných náhodných veličin patří např. $g(x) = \log x$ nebo $g(x) = \sqrt{x}$.

1.4.2 Transformace stabilizující rozptyl

Jinou motivací pro použití transformace může být snaha stabilizovat rozptyl. Nechť platí $\mathsf{E}\, X_i = \mu$ a rozptyl $\mathsf{var}\, X_i$ je funkcí střední hodnoty, tj. $\mathsf{var}\, X_i = \sigma^2(\mu)$. Hledáme transformaci g takovou, aby rozptyl transformovaných pozorování $g(X_i)$ nezávisel na střední hodnotě. Použijeme lineární aproximaci Taylorovým rozvojem kolem μ

$$g(X_i) \approx g(\mu) + (X_i - \mu) g'(\mu).$$

Spočítáme-li střední hodnoty a rozptyl levé i pravé strany, dostaneme

$$\mathsf{E}\, g(X_i) \approx g(\mu), \qquad \mathsf{var}\, g(X_i) \approx \sigma^2(\mu)[g'(\mu)]^2.$$

Vliv μ na rozptyl $g(X_i)$ bude (přibližně) eliminován, pokud bude platit $g'(\mu)\sigma(\mu) = c$ pro každé μ a nějaké pevné $c \in \mathbb{R}$, neboli

$$g(x) = c \int \frac{1}{\sigma(x)}\, dx.$$

Pro tuto funkci g bude platit $\mathsf{var}\, g(X_i) \approx c^2$. Po provedení transformace stabilizující rozptyl se obvykle rozdělení transformovaného výběru výrazně přiblíží rozdělení normálnímu.

Příklady.

1. Nechť $\sigma^2(\mu) = q\mu$, $q > 0$. Tento vztah platí např. pro rozdělení $\mathsf{Po}(\mu)$ nebo $\Gamma(a, p)$ pro pevné a a proměnlivé p. Pak $\int x^{-1/2}\, dx = 2\sqrt{x}$ a transformace stabilizující rozptyl je $g(X_i) = \sqrt{X_i}$.

2. Nechť $\sigma^2(\mu) = q\mu(1-\mu)$. Tento vztah platí např. pro rozdělení $\mathsf{Alt}(\mu)$, $\mathsf{Bi}(n, \mu)$ nebo $\mathsf{B}(\alpha, \beta)$. Pak $\int [x(1-x)]^{-1/2}\, dx = 2 \arcsin \sqrt{x}$ a transformace stabilizující rozptyl je $g(X_i) = \arcsin \sqrt{X_i}$.

Zde končí předn. 5 (19.10.)

1.4.3 Vliv transformace na parametry

Pokud používáme transformace, musíme si uvědomovat, že řada parametrů rozdělení F_X původního náhodného výběru se po transformaci změní takovým způsobem, že je už nedokážeme identifikovat.

Například střední hodnota $\mu_X = \mathsf{E}\, X_i$ se změní na $\mu_Y = \mathsf{E}\, g(X_i)$. Pokud neznáme rozdělení X_i, nemůžeme pak z μ_Y spočítat původní střední hodnotu μ_X, ledaže by g byla lineární

funkce. Nechť je g rostoucí a ryze konkávní funkce, pak platí z Jensenovy nerovnosti (Věta P.2.5) $\mu_Y < g(\mu_X)$ a zpětná transformace $g^{-1}(\mu_Y)$ dává hodnotu ostře menší než μ_X. U ryze konvexní funkce je tomu naopak.

Spočítáme-li tedy výběrový průměr $\overline{Y}_n$ z transformovaného náhodného výběru, bude konvergovat podle Věty 1.2(ii) k μ_Y. Zpětná transformace $g^{-1}(\overline{Y}_n)$ bude konvergovat k $g^{-1}(\mu_Y) \neq \mu_X$. Obecně nelze nalézt funkci h takovou, aby $h(\overline{Y}_n)$ konvergovalo k μ_X. Zajímá-li nás konkrétní hodnota μ_X, nemůžeme tedy data transformovat. Podobné je to s rozptylem a vyššími momenty: po transformaci už obvykle nezjistíme, jaký byl rozptyl původních pozorování.

Příklad. Nechť $X_i \sim \mathsf{N}(\mu_X, \sigma_X^2)$. Definujme $Y_i = \exp\{X_i\}$. Potom Y_i má tzv. *logaritmicko-normální rozdělení* $\mathsf{LN}(\mu_X, \sigma_X^2)$ s momenty

$$\mu_Y = \mathsf{E}\, \exp\{X_i\} = \mathrm{e}^{\mu_X + \sigma_X^2/2},$$
$$\sigma_Y^2 = \mathsf{var}\, \exp\{X_i\} = (\mathrm{e}^{\sigma_X^2} - 1)\mathrm{e}^{2\mu_X + \sigma_X^2}.$$

Spočítáme-li $\log \mu_Y$, dostaneme $\mu_X + \sigma_X^2/2 > \mu_X$.

Některé jiné parametry však tento problém nemají. Například medián nebo kterýkoli jiný kvantil lze snadno získat zpětnou transformací: Nechť m_X je medián X_i a m_Y je medián Y_i, nechť g je ryze rostoucí funkce. Pak platí $m_Y = g(m_X)$, tj. m_X lze identifikovat zpětnou transformací $g^{-1}(m_Y)$.

Pořadí jsou invariantní vůči ryze rostoucím transformacím, takže statistiky závisející pouze na pořadích nabývají stejné hodnoty, ať už jsou počítány z původního nebo transformovaného náhodného výběru.

1.4.4 Standardizace

Speciálním druhem transformace je tzv. *standardizace*. Máme náhodný výběr $X_1, \ldots, X_n$ a spočítáme $\overline{X}_n$ a S_n^2. Potom definujeme náhodné veličiny $Z_1, \ldots, Z_n$ vztahem

$$Z_i = \frac{X_i - \overline{X}_n}{S_n}.$$

Tyto veličiny mají výběrový průměr 0 a výběrový rozptyl 1, ale nepředstavují náhodný výběr, neboť nejsou nezávislé. Jelikož však $\overline{X}_n \xrightarrow{\mathsf{P}} \mathsf{E}\, X_i$ a $S_n \xrightarrow{\mathsf{P}} \sqrt{\mathsf{var}\, X_i}$ pro $n \to \infty$, při dostatečně velkém počtu pozorování se $Z_1, \ldots, Z_n$ chovají jako nezávislé veličiny s nulovou střední hodnotou a jednotkovým rozptylem.

Standardizace se používá tehdy, pokud se chceme zbavit prvních dvou momentů a soustředit se na jiné aspekty rozdělení F_X.

2 Odhadování parametrů

2.1 Bodový odhad

2.1.1 Definice bodového odhadu

Máme náhodný výběr $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$, model $\mathcal{F}$ a parametr $\theta = t(F) \in \mathbb{R}$ pro $F \in \mathcal{F}$, který chceme v daném modelu odhadnout. Nechť $F_X \in \mathcal{F}$ je skutečné rozdělení náhodného vektoru $\boldsymbol{X}_i$ a $\theta_X \equiv t(F_X)$ je skutečná hodnota hledaného parametru.

Definice 2.1 *Odhadem parametru* $\theta_X \equiv t(F_X)$ rozumíme libovolnou měřitelnou funkci dat $\widehat{\theta}_n \equiv T_n(\boldsymbol{X}) \equiv T_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$.[*]

Poznámka. Odhad je statistika ve smyslu definice 1.3. Odhad nesmí záviset na neznámých parametrech.

2.1.2 Vlastnosti odhadů

Definice 2.2 (Nestrannost a konsistence) Mějme náhodný výběr $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ z rozdělení $F_X \in \mathcal{F}$ a odhad $\widehat{\theta}_n \equiv T_n(\boldsymbol{X})$ parametru $\theta_X \equiv t(F_X)$.

(i) Řekneme, že odhad $\widehat{\theta}_n$ je *nestranný odhad*[†] parametru θ_X v modelu $\mathcal{F}$, právě když $\mathsf{E}\,\widehat{\theta}_n = \theta_X$ pro každé n (pro něž je odhad definován) a pro každé rozdělení $F_X \in \mathcal{F}$.

(ii) Řekneme, že odhad $\widehat{\theta}_n$ je *konsistentní odhad*[‡] parametru θ_X v modelu $\mathcal{F}$, právě když $\widehat{\theta}_n \xrightarrow{\mathsf{P}} \theta_X$ při $n \to \infty$ pro každé rozdělení $F_X \in \mathcal{F}$.

Poznámka.

- Vlastnosti odhadů musíme zkoumat v kontextu daného modelu. Snadno se může stát, že odhad $\widehat{\theta}_n$ je nestranný a konsistentní v nějakém modelu $\mathcal{F}$, ale v jiném modelu $\mathcal{F}'$ tyto vlastnosti nemá.
- Nestrannost má platit pro každý počet pozorování n, pro nějž je odhad definován (např. u výběrového rozptylu pro $n \geq 2$). Nestrannost ale nezaručuje, že se odhad při zvětšujícím se rozsahu výběru přibližuje k hledanému parametru. Pro některé modely neexistují rozumné (nebo vůbec žádné) nestranné odhady.
- Konsistence je asymptotická vlastnost, která nic neříká o chování odhadu při konečném n. (Příklad: $\widehat{\theta}_n = 21{,}5$ pro $n \leq 10^{10}$, $\widehat{\theta}_n = \overline{X}_n$ pro $n > 10^{10}$ je konsistentní odhad $\theta_X = \mathsf{E}\,X_i$.)
- Odhady, které nejsou nestranné, ale jsou konsistentní, se ve statistice běžně používají. Odhady, které nejsou konsistentní, nepoužíváme, neboť vlastně odhadují „něco jiného“.

[*] Angl. *estimator, estimate* [†] Angl. *unbiased estimator* [‡] Angl. *consistent estimator*

Příklady.

1. *Odhad parametru* $\theta_X = \mathsf{E}\, X_i$ *v modelu* $\mathcal{F} = \mathcal{L}^1$*:*
   - Průměr $\overline{X}_n$ je nestranný a konsistentní odhad θ_X [plyne z věty 1.2, (i) a (ii)].
   - Odhad $\widehat{\theta}_n = X_1$ je nestranný odhad θ_X, ale není konsistentní.
2. *Odhad parametru* $\theta_X = \mathsf{var}\, X_i$ *v modelu* $\mathcal{F} = \mathcal{L}^2$*:*
   - Výběrový rozptyl S_n^2 je nestranný a konsistentní odhad θ_X [plyne z věty 1.6, (i) a (ii)].
   - Odhad $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ je konsistentní odhad θ_X, ale není nestranný.
3. *Odhad parametru* $\theta_X = \mathsf{P}[X_i = 0]$ *v modelu* $\mathcal{F} = \{\mathsf{Po}(\lambda),\, \lambda > 0\}$*:*
   - Odhad $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{0\}}(X_i)$ je nestranný a konsistentní odhad θ_X.
   - Odhad $\widetilde{\theta}_n = (\frac{n-1}{n})^{\sum_{i=1}^n X_i}$ je také nestranný a konsistentní odhad θ_X.
4. *Odhad parametru* $\theta_X = \mathrm{e}^{-2\lambda_X}$ *v modelu* $\mathcal{F} = \{\mathsf{Po}(\lambda),\, \lambda > 0\}$ *pro* $n = 1$*:*
   Jediný nestranný odhad jest $\widehat{\theta} = (-1)^{X_1}$, jeho možné hodnoty jsou -1 a 1. Hledaný parametr $\mathrm{e}^{-2\lambda_X}$ však nabývá pouze hodnot z intervalu $(0, 1)$.

Definice 2.3 (Vychýlení) Nechť odhad $\widehat{\theta}_n \equiv T_n(\boldsymbol{X})$ parametru θ_X má konečnou střední hodnotu. Rozdíl $\mathsf{E}\,(\widehat{\theta}_n - \theta_X)$ nazýváme *vychýlením* odhadu $\widehat{\theta}_n$.[*]

Věta 2.1 Nechť $\widehat{\theta}_n$ je odhad parametru θ_X, pro nějž platí $\mathsf{E}\,\widehat{\theta}_n \to \theta_X$ (vychýlení konverguje k nule) a $\mathsf{var}\,\widehat{\theta}_n \to 0$ pro $n \to \infty$. Pak je $\widehat{\theta}_n$ konsistentní odhad θ_X.

Zde končí předn. 6 (20.10.)

Poznámka. Opačná implikace neplatí. Existují běžně používané konsistentní odhady, pro něž platí $\mathsf{E}\,\widehat{\theta}_n = \infty$ pro každé konečné n. S příklady takových odhadů se setkáme později.

Definice 2.4 Nechť odhad $\widehat{\theta}_n \equiv T_n(\boldsymbol{X})$ parametru θ_X má konečný rozptyl.

(i) Výraz
$$\mathrm{MSE}(\widehat{\theta}_n) = \mathsf{E}\,(\widehat{\theta}_n - \theta_X)^2$$
nazýváme *střední čtvercovou chybou* odhadu $\widehat{\theta}_n$.[†]

(ii) Výraz
$$\mathrm{SE}(\widehat{\theta}_n) = \sqrt{\mathsf{var}\,\widehat{\theta}_n}$$
nazýváme *směrodatnou chybou*[‡] odhadu $\widehat{\theta}_n$.

Poznámka.
- Pozor na jemné rozdíly v terminologii. Pojem *směrodatná odchylka* (standard deviation, SD) obvykle znamená odmocninu z rozptylu jednoho pozorování náhodného výběru, tj. $\sqrt{\mathsf{var}\, X_i}$. Pojem *směrodatná chyba* (standard error, SE) obvykle znamená odmocninu z rozptylu nějakého odhadu spočítaného z celého náhodného výběru. Někteří autoři však tyto pojmy používají odlišně.

[*] Angl. *bias* [†] Angl. *mean square error, MSE* [‡] Angl. *standard error, SE*

- Střední čtvercová chyba i směrodatná chyba jsou míry *přesnosti* odhadu. Směrodatná chyba do přesnosti nezahrnuje vychýlení, zatímco střední čtvercová chyba ano.
- Platí: $\mathrm{MSE}(\widehat{\theta}_n) = \mathrm{SE}^2(\widehat{\theta}_n) + [\mathsf{E}\,(\widehat{\theta}_n - \theta_X)]^2$.
- Střední čtvercová chyba je jedno z nejvhodnějších kritérií pro porovnávání odhadů. Máme-li několik různých odhadů téhož parametru v tomtéž modelu, snažíme se mezi nimi najít ten, který má nejmenší MSE.
- MSE často nelze spočítat, proto se nahrazuje asymptotickou střední čtvercovou chybou AMSE. Její definice je však složitější a nebudeme ji zde uvádět.

Příklad. *Odhad parametru* $\sigma_X^2 = \mathsf{var}\, X_i$ *v modelu* $\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$. Platí: $\mathrm{MSE}(S_n^2) > \mathrm{MSE}(\widehat{\sigma}_n^2)$.

2.2 Volba parametru

Parametr $\theta = t(F)$, který se snažíme odhadovat, může být v principu cokoli. Ne všechny parametry však dávají smysl v kontextu daného praktického problému, který řešíme. Musíme tedy rozlišovat, které parametry pro daný problém má smysl odhadovat a které ne. To záleží na významu hodnot měřených veličin, na tom, jak byly získány, zpracovány atd. Statistické metody, kterými se budeme zabývat, budeme rozlišovat podle toho, pro jaký typ měření jsou určeny. Přitom budeme uvažovat následující typy dat, neboli *škály měření*[*].

2.2.1 Kvantitativní data

Náhodnou veličinu X nazveme *kvantitativní*[†], pokud její hodnoty mají konkrétní numerický význam (např. počet, procento, délka, objem, hmotnost, úroková míra, koncentrace látky, energie, teplota, doba trvání, velikost úhlu, zeměpisná šířka, kalendářní rok). U kvantitativních veličin existuje smysluplné uspořádání jejich hodnot (teplota 10 °C je vyšší než −11,4 °C) a rozdíly jejich hodnot mají reálnou interpretaci. Kvantitativní veličiny mohou být jak diskrétní tak spojité.

Kvantitativní veličiny můžeme dále dělit na dvě podskupiny: *intervalové* a *poměrové*. **Poměrové veličiny** jsou typicky nezáporné s jasně definovanou nulovou hodnotou a interpretovatelnými podíly. Například hmotnost 0 kg je jednoznačně daná a hmotnost 20 kg je čtyřikrát více než 5 kg. Příklady poměrových veličin jsou počet, délka, objem, hmotnost, úroková míra, koncentrace látky, energie, doba trvání, teplota měřená v Kelvinech. **Intervalové veličiny** jsou kvantitativní veličiny, které nejsou poměrové, to jest nemají pevně definovanou nulu nebo nemají interpretovatelné podíly. Například směr daný azimutem je intervalová veličina, neboť azimut 360° není šestkrát větší než 60°. Podobně teplota měřená v °C je intervalová veličina neboť 16 °C není čtyřikrát vyšší teplota než 4 °C. Kalendářní rok je také intervalová veličina, protože nemá smysl počítat podíl letošního roku a roku vašeho narození.

2.2.2 Kategoriální data

Náhodnou veličinu X nazveme *kategoriální*[‡], pokud její hodnoty kódují příslušnost (neboli *klasifikaci*) subjektu do určité kategorie, neboli jedné z několika disjunktních množin. Ka-

[*] Angl. *measurement scales* [†] Angl. *quantitative* [‡] Angl. *categorical*

tegoriální veličiny jsou vždy diskrétní a mají konečný počet K možných hodnot, obvykle $1, \ldots, K$ nebo $0, \ldots, K-1$. Hodnoty kategoriálních veličin nemají přímou numerickou interpretaci, slouží pouze k rozlišení konečného počtu možných stavů. Jednotlivým stavům říkáme *úrovně*[*] nebo *kategorie*.

Kategoriální veličiny dále dělíme na *nominální*[†] a *ordinální*[‡]. U **nominálních veličin** neexistuje ani žádné uspořádání jejich kategorií – nelze říci, že kategorie j předchází kategorii $j+1$. Příkladem nominální veličiny je třeba bydliště kategorizované jako kraj (1 = Praha, 2 = Středočeský kraj, …, 14 = Zlínský kraj) nebo sociální postavení (1 = nezletilý; 2 = student; 3 = zaměstnanec; 4 = živnostník; 5 = nezaměstnaný; 6 = důchodce). **Ordinální veličiny** mají v nějakém smyslu uspořádáné kategorie, takže lze tvrdit, že kategorie j předchází kategorii $j+1$, nebo že je menší, horší apod. Příkladem ordinální veličiny je třeba odpověď na otázku s možnostmi 1 = ostře nesouhlasím, 2 = spíše nesouhlasím, 3 = nevím, 4 = spíše souhlasím, 5 = naprosto souhlasím. Jiný příklad je veličina nejvyšší dosažené vzdělání kódovaná jako 1 = nižší než základní; 2 = základní; 3 = učební obor; 4 = středoškolské s maturitou; 5 = bakalářské; 6 = magisterské; 7 = doktorské.

2.2.3 Binární data

Binární[§] veličiny jsou speciálním případem kategoriálních veličin, kde $K = 2$. Klasifikují tedy pozorování do jednoho ze dvou možných stavů. Jejich hodnoty se obvykle volí jako 0 vs. 1, případně 1 vs. 2. Příkladem binární veličiny je pravdivostní hodnota výroku (0 = pravda, 1 = lež), realizace náhodného jevu (0 = nenastal/neúspěch, 1 = nastal/úspěch) nebo pohlaví (1 = samec, 2 = samice).

2.2.4 Volba parametru v závislosti na typu dat

Pro nominální veličiny obecně nemá smysl uvažovat parametry jako $\mathsf{E}\,X$, $\mathsf{var}\,X$, distribuční funkci, kvantily, kovariance a korelace, zkrátka žádné charakteristiky, které závisejí na kódování a uspořádaní jednotlivých kategorií. Tyto parametry jsou sice řádně definovány, ale nemají žádnou praktickou interpretaci. Jediné parametry, které u nominálních veličin interpretaci mají, jsou pravděpodobnosti jednotlivých kategorií, čili $p_j = \mathsf{P}[X = j]$ pro všechny možné hodnoty j.

Výjimkou jsou binární veličiny. Znamená-li např. hodnota 0 neúspěch a hodnota 1 úspěch, pak $\mathsf{E}\,X = \mathsf{P}[X = 1]$, tedy střední hodnota je zároveň pravděpodobnost úspěchu.

U ordinálních veličin má díky uspořádání jejich hodnot smysl distribuční funkce. Často je možné přikládat jim intervalovou interpretaci (doktorské vzdělání je o dva stupně vyšší než bakalářské), ale obvykle jim nelze dávat poměrovou interpretaci (nelze říci, že magisterské vzdělání je dvakrát vyšší než učební obor). Ordinálním veličinám se někdy přiřazují neceločíselné hodnoty, tzv. *skóry*. Např. ordinální veličinu můžeme vytvořit tak, že vezmeme kvantitativní veličinu Z a seskupíme ji podle zvolených dělicích bodů, např. $X = 1$ pokud $Z \in \langle 0,\ 5)$, $X = 2$ pokud $Z \in \langle 5,\ 20)$, $X = 3$ pokud $Z \in \langle 20,\ 100)$ a $X = 4$ pokud $Z \geq 100$. Takové veličiny běžně vznikají v dotaznících, kde respondent dostane na výběr jednu ze čtyř možností namísto toho, aby musel zapsat přesné číslo. Výsledná veličina X je zjevně ordinální. Namísto hodnot $1, \ldots, 4$ bychom ale mohli za hodnoty X vzít prostředky intervalů, z kterých hodnoty X vznikly, tedy $2{,}5$; $12{,}5$ a 60 pro první tři intervaly. S posledním je zjevně

[*] Angl. *levels* [†] Angl. *nominal* [‡] Angl. *ordinal* [§] Angl. *binary*

potíž, neboť nemá pravý okraj – jeho skóru bychom museli nějak doplnit, například vzít 150. Takto zakódovaná veličina X je nejen ordinální, ale má některé vlastnosti veličiny kvantitativní.

Ordinální veličiny můžeme vždy analyzovat jako by byly nominální, ale často je možné na ně používat metody určené pro kvantitativní veličiny, odhadovat jejich střední hodnotu nebo počítat jejich rozdíly. Existují také speciální metody určené právě pro ordinální veličiny, s těmi se ale zatím nesetkáme.

Náš výklad statistických metod počínaje kapitolou 4 bude rozlišovat metody pro kvantitativní data, kde budeme pracovat s charakteristikami jako je střední hodnota, rozptyl, medián, distribuční funkce, kovariance apod., a metody pro nominální data, kde budeme pracovat s pravděpodobnostmi jednotlivých kategorií.

2.3 Intervalový odhad

2.3.1 Definice

Definice 2.5 Interval $B = B_n(\boldsymbol{X}) \subset \mathbb{R}$ se nazývá *intervalový odhad* parametru $\theta_X \in \mathbb{R}$ *o spolehlivosti* $1-\alpha$, právě když $\mathsf{P}[B \ni \theta_X] = 1-\alpha$. Interval B se nazývá *asymptotický intervalový odhad* parametru $\theta_X \in \mathbb{R}$ *o (přibližné) spolehlivosti* $1-\alpha$, právě když $\mathsf{P}[B \ni \theta_X] \to 1-\alpha$ pro $n \to \infty$.

Poznámka.

- Interval B je náhodný (spočítaný z dat), zatímco parametr θ_X je pevný. Výraz $B \ni \theta_X$ čteme „interval B pokrývá (skutečnou hodnotu) θ_X“.
- Intervalovému odhadu se běžně říká i jinak, např. *interval spolehlivosti s pravděpodobností pokrytí* $1-\alpha$ nebo $(1-\alpha)100$*-procentní konfidenční interval* pro parametr θ.* Číslo $\alpha \in (0,1)$ je předem zvolené; obvykle se bere $\alpha = 0{,}05$ a počítají se 95procentní intervaly. Můžeme se však setkat i s intervaly, jež mají pokrytí 90 % či 99 %.
- Ne vždy je možné či vhodné počítat přesné intervaly spolehlivosti. Často se spokojujeme s intervaly asymptotickými, jejichž pokrytí se pro velké rozsahy výběru blíží k požadované hodnotě.
- Intervalové odhady zde definujeme pouze pro reálné parametry. Podobný koncept však lze zavést i pro vektorové parametry; hledáme náhodnou množinu B, která pokrývá skutečnou hodnotu se zadanou pravděpodobností. Této množině pak říkáme oblast spolehlivosti. Tvar množiny B lze ale potom volit mnoha různými způsoby.

Zde končí předn. 7 (26.10.)

Poznámka. Rozeznáváme intervalové odhady oboustranné a jednostranné (levo- a pravostranné).

- Interval tvaru (C_L, C_U), kde C_L a C_U jsou dvě náhodné veličiny splňující $\mathsf{P}[C_L < C_U] = 1$, $C_L > -\infty$ a $C_U < \infty$, nazýváme oboustranný interval spolehlivosti. Obvykle jej sestrojujeme tak, aby platilo (alespoň asymptoticky)

$$\mathsf{P}[\theta_X < C_L] = \frac{\alpha}{2}, \quad \mathsf{P}[\theta_X > C_L, \theta_X < C_U] = 1-\alpha, \quad \mathsf{P}[\theta_X > C_U] = \frac{\alpha}{2}.$$

* Angl. *confidence interval with coverage probability/confidence level* $1-\alpha$

- Interval tvaru (C_L, ∞) nazýváme levostranný interval spolehlivosti. Máme $\mathsf{P}[C_L < \theta_X] = 1 - \alpha$.
- Interval tvaru $(-\infty, C_U)$ nazýváme pravostranný interval spolehlivosti. Máme $\mathsf{P}[\theta_X < C_U] = 1 - \alpha$.

Příklad (střední hodnota normálního rozdělení se známým rozptylem)**.** Vezměme si problém intervalového odhadu střední hodnoty pro normálně rozdělená data se známým rozptylem.
Data: $X_1, \ldots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \{\mathsf{N}(\mu, \sigma_X^2),\ \mu \in \mathbb{R},\ \sigma_X^2 \text{ známo}\}$
Odhadovaný parametr: $\theta_X = \mathsf{E}\, X_i \equiv \mu_X$
Postup:

1. Máme bodový odhad $\overline{X}_n$, který je nestranný a konsistentní pro μ_X. Víme, že $\overline{X}_n \sim \mathsf{N}(\mu_X, \sigma_X^2/n)$. Tudíž
$$\sqrt{n}\,\frac{\overline{X}_n - \mu_X}{\sigma_X} \sim \mathsf{N}(0, 1).$$
2. Vyjdeme z rovnosti
$$\mathsf{P}\left[u_{\frac{\alpha}{2}} < \sqrt{n}(\overline{X}_n - \mu_X)/\sigma_X < u_{1-\frac{\alpha}{2}}\right] = 1 - \alpha,$$
kde $u_\alpha = \Phi^{-1}(\alpha)$ je α-kvantil normovaného normálního rozdělení, a postupnými úpravami (s využitím symetrie hustoty $\mathsf{N}(0, 1)$ kolem 0) dojdeme k
$$\mathsf{P}\left[\overline{X}_n - \sigma_X u_{1-\frac{\alpha}{2}}/\sqrt{n} < \mu_X < \overline{X}_n + \sigma_X u_{1-\frac{\alpha}{2}}/\sqrt{n}\right] = 1 - \alpha.$$
3. Získali jsme oboustranný interval spolehlivosti (C_L, C_U). Jeho krajní body jsou
$$C_L = \overline{X}_n - \frac{\sigma_X}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \qquad C_U = \overline{X}_n + \frac{\sigma_X}{\sqrt{n}} u_{1-\frac{\alpha}{2}}. \tag{2.1}$$
Kvantily normovaného normálního rozdělení,které potřebujeme pro konstrukci intervalů spolehlivosti, jsou uvedeny v Tabulce 2.1.
Pro $\alpha = 0{,}05$ vezmeme kvantil $u_{0{,}975} \doteq 1{,}96$ a dostaneme 95% oboustranný interval spolehlivosti. To znamená, že tento interval pokrývá skutečnou střední hodnotu μ_X s pravděpodobností 0,95.
4. Jednostranný interval bychom získali drobnou modifikací kroku 2. Levostranný interval vyjde (C_L, ∞), kde $C_L = \overline{X}_n - \frac{\sigma_X}{\sqrt{n}} u_{1-\alpha}$. Pravostranný interval vyjde $(-\infty, C_U)$, kde $C_U = \overline{X}_n + \frac{\sigma_X}{\sqrt{n}} u_{1-\alpha}$. Jednostranné intervaly se od oboustranného liší hodnotou kvantilu normálního rozdělení (používají $u_{1-\alpha}$ namísto $u_{1-\frac{\alpha}{2}}$). Pro 95% jednostranný interval spolehlivosti bychom vzali kvantil $u_{0{,}95} \doteq 1{,}645$.

Tabulka 2.1: Vybrané hodnoty kvantilů normovaného normálního rozdělení.

κ	0,9	0,95	0,975	0,99	0,995
$u_\kappa = \Phi^{-1}(\kappa)$	1,282	1,645	1,960	2,326	2,576

Poznámka. Délka intervalu spolehlivosti závisí na:
- počtu pozorování n,
- rozptylu dat σ_X^2,
- pravděpodobnosti pokrytí $1-\alpha$.

Příklad. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{N}(\mu_X, \sigma_X^2)$, rozptyl σ_X^2 známe. Kolik pozorování potřebujeme, aby délka oboustranného intervalu spolehlivosti pro střední hodnotu μ_X nepřekročila stanovenou mez $d > 0$?

Máme $2u_{1-\alpha/2}\sigma_X/\sqrt{n} \leq d$. Tudíž potřebujeme alespoň $4u_{1-\alpha/2}^2\sigma_X^2/d^2$ pozorování.

Lemma 2.2 (interval spolehlivosti po transformaci parametrů) Je-li (C_L, C_U) interval spolehlivosti pro parametr θ_X s pravděpodobností pokrytí $1-\alpha$ a je-li ψ ryze rostoucí reálná funkce, pak $(\psi(C_L), \psi(C_U))$ je interval spolehlivosti pro parametr $\psi(\theta_X)$ s pravděpodobností pokrytí $1-\alpha$.

2.3.2 Konstrukce intervalových odhadů

Nechť $\boldsymbol{X} = (\boldsymbol{X}_1^\mathsf{T}, \ldots, \boldsymbol{X}_n^\mathsf{T})^\mathsf{T}$, kde $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ je náhodný výběr z rozdělení $F_X \in \mathcal{F}$. Odhadujeme parametr $\theta_X = t(F_X) \in \mathbb{R}$. Popišme si stručně obecný postup při konstrukci oboustranných intervalových odhadů pro θ_X.

1. Nalezneme funkci $\varphi(\boldsymbol{x}, \theta_X)$ takovou, že φ je prostá a spojitá funkce v argumentu θ_X pro každé $\boldsymbol{x}$ a rozdělení náhodné veličiny $Z_n \equiv \varphi(\boldsymbol{X}, \theta_X)$ je známé alespoň asymptoticky (nezávisí ani na θ_X ani na jiných neznámých parametrech). Náhodná veličina Z_n se nazývá *pivotální statistika*. Označíme F_Z distribuční funkci Z_n, $c_\alpha = F_Z^{-1}(\alpha)$ budiž α-kvantil rozdělení F_Z. Při konstrukci funkce φ můžeme vyjít např. z bodového odhadu parametru θ_X, jehož rozdělení většinou známe alespoň asymptoticky.

2. Zinvertujeme $\varphi(\boldsymbol{x}, \theta)$ jakožto funkci argumentu θ při pevném $\boldsymbol{x}$ – nechť existuje $\bar{\varphi}(\boldsymbol{x}, z)$ taková, že $\varphi(\boldsymbol{x}, \bar{\varphi}(\boldsymbol{x}, z)) = z$ a $\bar{\varphi}(\boldsymbol{x}, \varphi(\boldsymbol{x}, \theta)) = \theta$ pro všechna $\boldsymbol{x}$, z a θ.

3. Máme $\mathsf{P}[c_{\alpha/2} < Z_n < c_{1-\alpha/2}] = 1-\alpha$. Aplikací funkce $\bar{\varphi}(\boldsymbol{x}, \cdot)$ na obě nerovnosti (předpokládajíce, že je klesající funkcí druhého argumentu z, což je v praxi častý případ) dostaneme
$$\mathsf{P}[\bar{\varphi}(\boldsymbol{X}, c_{1-\alpha/2}) < \theta_X < \bar{\varphi}(\boldsymbol{X}, c_{\alpha/2})] = 1-\alpha.$$

4. Získali jsme interval spolehlivosti (C_L, C_U) s pravděpodobností pokrytí $1-\alpha$, kde $C_L = \bar{\varphi}(\boldsymbol{X}, c_{1-\alpha/2})$ a $C_U = \bar{\varphi}(\boldsymbol{X}, c_{\alpha/2})$.

Příklad (střední hodnota normálního rozdělení s neznámým rozptylem)**.** Vezměme si problém intervalového odhadu střední hodnoty pro normálně rozdělená data s neznámým rozptylem.
Data: $X_1, \ldots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0\}$
Odhadovaný parametr: $\theta_X = \mathsf{E}\, X_i \equiv \mu_X$
Postup:

Tabulka 2.2: Vybrané hodnoty kvantilů $t_f(\kappa)$ rozdělení t s f stupni volnosti.

	κ				
f	0,9	0,95	0,975	0,99	0,995
5	1,476	2,015	2,571	3,365	4,032
10	1,372	1,812	2,228	2,764	3,169
15	1,341	1,753	2,131	2,602	2,947
25	1,316	1,708	2,060	2,485	2,787
100	1,290	1,660	1,984	2,364	2,626
∞	1,282	1,645	1,960	2,326	2,576

Odhad $\overline{X}_n$ je nestranný a konsistentní pro μ_X, odhad S_n^2 je nestranný a konsistentní pro $\sigma_X^2 \equiv \mathrm{var}\, X_i$. Z věty 1.10 víme, že

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_X}{S_n} \sim t_{n-1}.$$

Vezmeme tedy T_n jako pivotální statistiku, F_Z je distribuční funkce rozdělení t_{n-1} a $c_\alpha = t_{n-1}(\alpha)$ (α-kvantil rozdělení t_{n-1}). Vybrané kvantily t rozdělení jsou uvedeny v Tabulce 2.2. Jak je vidět, už pro $n - 1 = 25$ jsou jen o málo větší než kvantily normovaného normálního rozdělení, k nimž konvergují při počtu stupňů volnosti rostoucím nade všechny meze. Větší hodnoty t kvantilů proti kvantilům normovaného normálního rozdělení používaným v úvodním příkladě odrážejí zvýšenou variabilitu pivotální statistiky způsobenou neznalostí skutečného rozptylu.

Vyjdeme z rovnosti

$$\mathsf{P}\left[t_{n-1}(\alpha/2) < \sqrt{n}(\overline{X}_n - \mu_X)/S_n < t_{n-1}(1-\alpha/2)\right] = 1-\alpha$$

a stejným postupem jako u normálního rozdělení se známým rozptylem dojdeme k intervalu

$$\left(\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1}\Big(1-\frac{\alpha}{2}\Big),\ \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1}\Big(1-\frac{\alpha}{2}\Big)\right). \tag{2.2}$$

který má pravděpodobnost pokrytí přesně $1-\alpha$. Kvůli vyšším hodnotám t kvantilů je tento interval o něco širší než interval (2.1) při známém rozptylu.

Příklad (střední hodnota libovolného rozdělení s konečným rozptylem)**.** Vezměme si problém intervalového odhadu střední hodnoty bez předpokladu normality dat.
Data: $X_1, \ldots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \mathcal{L}^2$ (všechna rozdělení s konečným rozptylem)
Odhadovaný parametr: $\theta_X = \mathsf{E}\, X_i \equiv \mu_X$
Postup:

Odhad $\overline{X}_n$ je nestranný a konsistentní pro μ_X, odhad S_n^2 je nestranný a konsistentní pro $\sigma_X^2 \equiv \mathrm{var}\, X_i$. Z věty 1.9 víme, že

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_X}{S_n} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Vezmeme tedy T_n jako pivotální statistiku.

Vyjdeme z limitního vztahu (zdůvodněného konvergencí v distribuci pivotální statistiky)

$$\mathsf{P}\Big[u_{\frac{\alpha}{2}} < \sqrt{n}(\overline{X}_n - \mu_X)/S_n < u_{1-\frac{\alpha}{2}}\Big] \to 1-\alpha \quad \text{při } n \to \infty.$$

Jelikož pro $n \to \infty$ kvantil $t_{n-1}(\alpha)$ konverguje k u_α (pro libovolné $0 < \alpha < 1$), máme

$$\mathsf{P}\Big[t_{n-1}(\alpha/2) < \sqrt{n}(\overline{X}_n - \mu_X)/S_n < t_{n-1}(1-\alpha/2)\Big] \to 1-\alpha \quad \text{při } n \to \infty.$$

Proto interval (2.2), který byl přesným intervalem spolehlivosti pro μ_X u výběru z normálního rozdělení, je zároveň asymptotickým intervalem spolehlivosti pro μ_X pro data pocházející z jakéhokoli rozdělení s konečným rozptylem.

Zde končí předn. 8 (27.10.)

Příklad (alternativní rozdělení)**.** Ukažme si nyní jeden možný způsob odvození asymptotického intervalového odhadu pro pravděpodobnost úspěchu v alternativním rozdělení. (Několik dalších intervalových odhadů pro tento problém si ukážeme později.)

Data: $X_1, \ldots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \{\mathsf{Alt}(p),\ p \in (0,1)\}$
Odhadovaný parametr: $p_X = \mathsf{E}\, X_i = \mathsf{P}[X_i = 1]$
Postup:

Jelikož odhadujeme pravděpodobnost, vyjdeme z empirické relativní četnosti $\widehat{p}_n = \overline{X}_n$, která je nestranným a konsistentním odhadem p (věta 1.3). Z centrální limitní věty (tvrzení P.7.11) víme, že $\sqrt{n}(\widehat{p}_n - p_X) \xrightarrow{\mathsf{D}} \mathsf{N}(0,\ p_X(1-p_X))$. Tudíž

$$\sqrt{n}\,\frac{\widehat{p}_n - p_X}{\sqrt{p_X(1-p_X)}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Levá strana je nelineární funkcí p_X, ale můžeme si ji zjednodušit. Z konsistence $\widehat{p}_n$ a věty o spojité transformaci (tvrzení P.7.3) víme, že

$$\sqrt{\widehat{p}_n(1-\widehat{p}_n)} \xrightarrow{\mathsf{P}} \sqrt{p_X(1-p_X)}.$$

Ze Sluckého věty (tvrzení P.7.6) dostaneme

$$\sqrt{n}\,\frac{\widehat{p}_n - p_X}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} = \frac{\sqrt{p_X(1-p_X)}}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}\sqrt{n}\,\frac{\widehat{p}_n - p_X}{\sqrt{p_X(1-p_X)}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1). \tag{2.3}$$

Vezmeme tedy $Z_n = \sqrt{n}\,\frac{\widehat{p}_n - p_X}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}$, $F_Z = \Phi$ a $c_\alpha = u_\alpha$ (α-kvantil normovaného normálního rozdělení).

Vyjdeme z limitního vztahu

$$\mathsf{P}\left[-u_{1-\frac{\alpha}{2}} < \sqrt{n}\,\frac{\widehat{p}_n - p_X}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} < u_{1-\frac{\alpha}{2}}\right] \to 1-\alpha$$

(pro $n \to \infty$) a postupnými úpravami dojdeme k

$$\mathsf{P}\left[\widehat{p}_n - \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}} < p_X < \widehat{p}_n + \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right] \to 1-\alpha.$$

Získali jsme tedy interval

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

jehož pravděpodobnost pokrytí konverguje k $1-\alpha$ pro $n \to \infty$.

Příklad (rozptyl a směrodatná odchylka normálního rozdělení)**.** Vezměme si problém intervalového odhadu směrodatné odchylky v normálním rozdělení.

Data: $X_1, \ldots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0\}$
Odhadovaný parametr: $\sigma_X = \sqrt{\mathsf{var}\, X_i}$
Postup:

Zabývejme se nejprve rozptylem σ_X^2. Jeho nestranný a konsistentní odhad je S_n^2. Z věty 1.8, část (i), víme, že

$$\frac{(n-1)S_n^2}{\sigma_X^2} \sim \chi^2_{n-1}.$$

Vezmeme tedy $Z_n = (n-1)S_n^2/\sigma_X^2$, $F_Z = \chi^2_{n-1}$ a $c_\alpha = \chi^2_{n-1}(\alpha)$, tj. α-kvantil rozdělení χ^2_{n-1} (viz Tabulka 2.3).

Vyjdeme z rovnosti

$$\mathsf{P}\left[\chi^2_{n-1}(\alpha/2) < \frac{(n-1)S_n^2}{\sigma_X^2} < \chi^2_{n-1}(1-\alpha/2)\right] = 1-\alpha$$

a postupnými úpravami dojdeme k

$$\mathsf{P}\left[\frac{(n-1)S_n^2}{\chi^2_{n-1}(1-\alpha/2)} < \sigma_X^2 < \frac{(n-1)S_n^2}{\chi^2_{n-1}(\alpha/2)}\right] = 1-\alpha.$$

Získali jsme interval spolehlivosti

$$\left(\frac{(n-1)S_n^2}{\chi^2_{n-1}(1-\alpha/2)},\ \frac{(n-1)S_n^2}{\chi^2_{n-1}(\alpha/2)}\right) \tag{2.4}$$

pro rozptyl σ_X^2 s pravděpodobností pokrytí $1-\alpha$.

Tabulka 2.3: Vybrané hodnoty kvantilů $\chi^2_f(\kappa)$ rozdělení χ^2 s f stupni volnosti.

	κ							
f	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99
5	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086
10	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209
15	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578
25	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314
100	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807

Interval spolehlivosti pro směrodatnou odchylku σ_X získáme aplikováním odmocniny na krajní body intervalu pro rozptyl (odmocnina je rostoucí funkce na $(0, \infty)$):

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

2.4 Empirické odhady a výběrové momenty

Mějme dán náhodný výběr $X_1, X_2, \ldots, X_n$ z rozdělení F_X. Ukažme si, jak lze odhadnout některé charakteristiky rozdělení F_X.

2.4.1 Empirická distribuční funkce

Zabývejme se nejprve odhadováním celé distribuční funkce $F_X(u)$ pro $u \in \mathbb{R}$. Pracujeme s modelem, který zahrnuje veškerá rozdělení na $\mathbb{R}$, tj. na distribuční funkci F_X neklademe vůbec žádné podmínky.

Definice 2.6 Funkci $\widehat{F}_n(u) \overset{\mathsf{df}}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, u\rangle}(X_i)$ nazýváme *empirická distribuční funkce*[*] náhodného výběru $X_1, X_2, \ldots, X_n$.

Poznámka. Hodnota $\widehat{F}_n$ v bodě u je rovna počtu pozorování, která nepřekročí u, dělenému celkovým počtem pozorování. Funkce $\widehat{F}_n$ je neklesající, zprava spojitá, po částech konstantní, skáče v pozorovaných hodnotách veličin X_i, velikosti skoků jsou dány počtem pozorování rovných u děleným celkovým počtem pozorování. Empirická distribuční funkce má všechny vlastnosti distribuční funkce diskrétního rozdělení.

Pro pevné u je hodnota $\widehat{F}_n(u)$ vlastně relativní četnost jevu $[X_i \leq u]$ spočítaná z n pozorování, přičemž pravděpodobnost tohoto jevu je $F_X(u)$. Z věty 1.3 rovnou dostaneme nejdůležitější vlastnosti empirické distribuční funkce.

Věta 2.3 (vlastnosti empirické distribuční funkce) Pro libovolné $u \in \mathbb{R}$ platí

(i) $\mathsf{E}\, \widehat{F}_n(u) = F_X(u)$ (nestrannost), $\mathsf{var}\, \widehat{F}_n(u) = \frac{F_X(u)[1-F_X(u)]}{n}$

(ii) $\widehat{F}_n(u) \xrightarrow{\mathsf{P}} F_X(u)$ (bodová konsistence)

(iii) $\sqrt{n}[\widehat{F}_n(u) - F_X(u)] \xrightarrow{\mathsf{D}} \mathsf{N}(0, F_X(u)[1 - F_X(u)])$ (asymptotická normalita)

(iv) $n\widehat{F}_n(u) \sim \mathsf{Bi}(n, F_X(u))$

(v) $\sup_{u \in \mathbb{R}} \left| \widehat{F}_n(u) - F_X(u) \right| \xrightarrow{\mathsf{P}} 0$ (stejnoměrná konsistence)

Poznámka.

- Z bodu (iii) předchozí věty lze odvodit asymptotický interval spolehlivosti pro $F_X(u)$ stejně jako v případě parametru alternativního rozdělení (viz str. 27).
- Bod (v) se někdy nazývá Glivenkova-Cantelliho věta. Nelze jej odvodit z věty 1.3 ani jiných výsledků, které máme k dispozici. Bude dokázán na jedné z pokročilejších přednášek z teorie pravděpodobnosti.

[*] Angl. *empirical distribution function*

2.4.2 Empirické odhady

Z empirické distribuční funkce lze odvodit odhady mnoha základních charakteristik rozdělení F_X. Nechť $\theta_X = t(F_X)$ je hledaný parametr. Umíme-li jej spočítat ze skutečné distribuční funkce F_X, můžeme jej stejným způsobem spočítat i z empirické distribuční funkce $\widehat{F}_n$. Dostaneme tak odhad $\widehat{\theta}_n \stackrel{\text{df}}{=} t(\widehat{F}_n)$. Těmto odhadům říkáme *empirické odhady*. Uvidíme, že v řadě případů mají empirické odhady rozumné vlastnosti.

Ukažme si tento postup nejprve na příkladě empirického odhadu střední hodnoty. Máme

$$\mathsf{E}\, X_i = \int_{-\infty}^{\infty} x \, dF_X(x).$$

Empirický odhad střední hodnoty získáme dosazením $\widehat{F}_n$ na místo neznámé funkce F_X. Dostaneme

$$\int_{-\infty}^{\infty} x \, d\widehat{F}_n(x) = \int_{-\infty}^{\infty} x \, d\Big(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x)}(X_i)\Big) = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x \, d\, \mathbb{1}_{\langle X_i,\infty)}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

kde jsme využili toho, že $\mathbb{1}_{\langle X_i,\infty)}(x)$ je pro pevné X_i vlastně distribuční funkcí konstanty nabývající hodnoty X_i s pravděpodobností 1. Došli jsme tedy k tomu, že empirickým odhadem střední hodnoty je aritmetický průměr, o němž již víme, že je nestranný a konsistentní.

2.4.3 Empirické odhady momentů

Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z rozdělení F_X a h je měřitelná reálná funkce taková, že $\mathsf{E}\,|h(X_i)| < \infty$. Dá se snadno ověřit, že empirickým odhadem parametru $\mathsf{E}\, h(X_i)$ je průměr naměřených hodnot $h(X_i)$, tj. $n^{-1}\sum_{i=1}^{n} h(X_i)$. Tento odhad je nestranný a konsistentní.

Odvoďme si empirický odhad rozptylu $\sigma_X^2 = \mathsf{E}\, X_i^2 - (\mathsf{E}\, X_i)^2$. Víme, že empirickým odhadem $\mathsf{E}\, X_i$ je $\overline{X}_n$ a empirickým odhadem $\mathsf{E}\, X_i^2$ je $n^{-1}\sum_{i=1}^{n} X_i^2$. Empirický odhad rozptylu tedy je $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.

Poznámka. Platí $S_n^2 = \frac{n}{n-1}\widehat{\sigma}_n^2$. Pro velká n je rozdíl mezi $\widehat{\sigma}_n^2$ a S_n^2 malý, neboť $\widehat{\sigma}_n^2 - S_n^2 \xrightarrow{\mathsf{P}} 0$. Jak plyne z věty 1.6, výběrový rozptyl S_n^2 je nestranný a konsistentní odhad σ_X^2. Empirický odhad rozptylu $\widehat{\sigma}_n^2$ je konsistentní, ale není nestranný.

Podobně můžeme odvodit empirické odhady pro momenty vyšších řádů. Empirické odhady necentrálních momentů $\mu'_k = \mathsf{E}\, X_i^k$ jsou

$$\widehat{\mu}'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Empirické odhady centrálních momentů $\mu_k = \mathsf{E}\,(X_i - \mathsf{E}\, X_i)^k$ jsou

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^k.$$

Empirické necentrální momenty jsou evidentně nestranné a konsistentní. Empirické centrální momenty jsou konsistentní, nikoli však obecně nestranné.

Empirický odhad šikmosti je

$$\widehat{\gamma}_3 = \frac{\widehat{\mu}_3}{(\widehat{\sigma}_n^2)^{3/2}},$$

empirický odhad špičatosti je

$$\widehat{\gamma}_4 = \frac{\widehat{\mu}_4}{\widehat{\sigma}_n^4}.$$

Oba jsou konsistentní (z věty o spojité transformaci, tvrz. P.7.3).

Zde končí předn. 9 (2.11.)

2.4.4 Empirický odhad kvantilu

Nechť α je předem dané číslo z intervalu $(0, 1)$. Kvantilová funkce rozdělení F_X je definována jako $F_X^{-1}(\alpha) = \inf\{x : F_X(x) \geq \alpha\}$; α-kvantilem rozdělení F_X rozumíme číslo $u_X(\alpha) = F_X^{-1}(\alpha)$. Pro α-kvantil platí

$$\lim_{h \searrow 0} F_X(u_X(\alpha) - h) \leq \alpha \quad \text{a} \quad F_X(u_X(\alpha)) \geq \alpha.$$

Jako empirický odhad použijeme hodnotu α-kvantilu empirické distribuční funkce, tedy $\widehat{F}_n^{-1}(\alpha) = \inf\{x : \widehat{F}_n(x) \geq \alpha\}$.

Definice 2.7 (Výběrový kvantil) Označme $k_\alpha = \alpha n$, pokud αn je celé číslo, $k_\alpha = [\alpha n] + 1$ pokud αn není celé číslo. *Empirický (výběrový) α-kvantil*[*] $\widehat{u}_n(\alpha)$ definujeme jako k_α-tou pořádkovou statistiku náhodného výběru $X_1, \ldots, X_n$, tedy $\widehat{u}_n(\alpha) = X_{(k_\alpha)}$.

Poznámka.

- Pro $\alpha = 0.5$ dostaneme *výběrový medián*[†]: $\widehat{m}_n = X_{(\frac{n+1}{2})}$ pro n liché a $\widehat{m}_n = X_{(n/2)}$ pro n sudé.
- Výběrový α-kvantil splňuje nerovnosti

  $$\lim_{h \searrow 0} \widehat{F}_n(\widehat{u}_n(\alpha) - h) \leq \alpha \quad \text{a} \quad \widehat{F}_n(\widehat{u}_n(\alpha)) \geq \alpha$$

  tj. alespoň $n\alpha$ pozorování je menší nebo rovno $\widehat{u}_n(\alpha)$ a zároveň alespoň $n(1-\alpha)$ pozorování je větší nebo rovno $\widehat{u}_n(\alpha)$.
- Existuje alespoň 10 různých definic výběrového α-kvantilu.

Následující lemma charakterizuje výběrový kvantil jako řešení minimalizačního problému (srovnej s Lemmatem 1.1); speciálně výběrový medián minimalizuje součet absolutních odchylek jednotlivých pozorování od libovolného reálného čísla.

Lemma 2.4

(i) Pro výběrový medián $\widehat{m}_n$ platí

$$\widehat{m}_n = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n} |X_i - c|.$$

[*] Angl. *empirical quantile, sample quantile* [†] Angl. *sample median*

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

$$\widehat{u}_n(\alpha) = \arg\min_{c\in\mathbb{R}} \sum_{i=1}^{n} \varrho_\alpha(X_i - c),$$

kde $\varrho_\alpha(u) = |u(\alpha - \mathbb{1}_{(-\infty,0)}(u))|$.

Poznámka. Minimalizační problém z části (ii) lze přepsat ve tvaru

$$\arg\min_{c\in\mathbb{R}} \Big[-(1-\alpha) \sum_{i:X_i<c} (X_i - c) + \alpha \sum_{i:X_i\geq c} (X_i - c) \Big].$$

Zavedeme-li značení $U_i = (X_i - c)\mathbb{1}(X_i \geq c)$, $V_i = -(X_i - c)\mathbb{1}(X_i < c)$, $\boldsymbol{U} = (U_1, \ldots, U_n)^\mathsf{T}$, $\boldsymbol{V} = (V_1, \ldots, V_n)^\mathsf{T}$, $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$,můžeme problém přepsat jako úlohu lineárního programování ve $(2n+1)$-dimensionálním prostoru

$$\min_{\boldsymbol{U},\boldsymbol{V},c} \alpha \mathbf{1}_n^\mathsf{T}\boldsymbol{U} + (1-\alpha)\mathbf{1}_n^\mathsf{T}\boldsymbol{V}$$

při omezeních

$$c\mathbf{1}_n + \boldsymbol{U} - \boldsymbol{V} = \boldsymbol{X}, \quad \boldsymbol{U} \geq 0, \quad \boldsymbol{V} \geq 0.$$

Tento minimalizační problém samozřejmě nemusí mít právě jedno řešení. Minima může být dosaženo na celém intervalu hodnot.

Vlastnosti výběrového kvantilu budeme dokazovat pouze pro spojitá rozdělení s ostře rostoucí distribuční funkcí F_X a hustotou f_X.

Věta 2.5 Nechť $\alpha \in (0,1)$. Nechť $X_1, \ldots, X_n$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F_X, spojitou kvantilovou funkcí F_X^{-1} a hustotou f_X, která je spojitá a nenulová v okolí $u_X(\alpha)$. Potom platí:

(i) $\widehat{u}_n(\alpha)$ je konsistentní odhad $u_X(\alpha)$;

(ii) $\sqrt{n}[\widehat{u}_n(\alpha) - u_X(\alpha)] \xrightarrow{\mathsf{D}} \mathsf{N}(0, V(\alpha))$, kde $V(\alpha) = \dfrac{\alpha(1-\alpha)}{f_X^2(u_X(\alpha))}$.

Poznámka. Asymptotický rozptyl $V(\alpha)$ výběrového kvantilu se špatně odhaduje, protože nemáme k disposici univerzálně použitelný a spolehlivý odhad hustoty.

V důkazu věty 2.5 se používá následující lemma, které se odvodí snadnou aplikací věty o transformaci náhodného vektoru (tvrzení P.5.4).

Lemma 2.6 Nechť $Z_1, \ldots, Z_{n+1}$ je náhodný výběr z rozdělení $\mathsf{Exp}(1)$. Vezměme nějaké $k \in \{1, \ldots, n\}$ a označme $U = \sum_{i=1}^{k} Z_i$, $V = \sum_{i=k+1}^{n+1} Z_i$. Potom náhodná veličina $\frac{U}{U+V}$ má rozdělení $\mathsf{B}(k, n-k+1)$.

Zde končí předn. 10 (3.11.)

2.4.5 Empirické odhady pro náhodné vektory

Empirické odhady prvních dvou momentů můžeme snadno rozšířit na náhodné vektory. Nechť $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ je náhodný výběr nezávislých k-rozměrných náhodných vektorů s rozdělením F_X, které má střední hodnotu $\boldsymbol{\mu}$ a rozptylovou matici Σ. Jednotlivé složky vektoru $\boldsymbol{X}_i$ budeme značit X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$.

Empirickým odhadem $\boldsymbol{\mu}$ je zřejmě vektor empirických odhadů jeho jednotlivých složek, čili k-rozměrný výběrový průměr

$$\overline{\boldsymbol{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i.$$

Empirický odhad rozptylové matice Σ bychom dostali z vyjádření

$$\Sigma = \mathsf{E}\, \boldsymbol{X}_i^{\otimes 2} - (\mathsf{E}\, \boldsymbol{X}_i)^{\otimes 2}$$

nahrazením středních hodnot jejich empirickými odhady, tj. průměry.

Výběrovou rozptylovou matici[*] si však zadefinujeme malinko jinak, jako vícerozměrnou obdobu výběrového rozptylu S_n^2:

$$\widehat{\Sigma}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^{\otimes 2}.$$

Poznámka.

- $\widehat{\Sigma}_n$ má na diagonále výběrové rozptyly jednotlivých složek, tj.

  $$S_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2,$$

  pro $j = 1, \ldots, k$, kde $\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$.

- Prvek (j, m) matice $\widehat{\Sigma}_n$ je dán výrazem

  $$S_{jm} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{im} - \overline{X}_m)$$

  pro $j = 1, \ldots, k$ a $m = 1, \ldots, k$, $j \neq m$. Tato náhodná veličina odhaduje kovarianci $\operatorname{cov}(X_{ij}, X_{im})$ mezi j-tou a m-tou složkou $\boldsymbol{X}_i$. Říkáme jí *výběrová kovariance*.

- $\widehat{\Sigma}_n$ je positivně semidefinitní a platí

  $$\widehat{\Sigma}_n = \frac{n}{n-1} \Big(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} - \overline{\boldsymbol{X}}_n^{\otimes 2} \Big).$$

Následující tvrzení ukazuje, že jak $\overline{\boldsymbol{X}}_n$ tak $\widehat{\Sigma}_n$ jsou nestranné a konsistentní odhady.

[*] Angl. *sample covariation matrix*

Tvrzení 2.7

(i) Je-li $\mathsf{E}\left|X_{ij}\right| < \infty$, pak $\mathsf{E}\,\overline{\boldsymbol{X}}_n = \boldsymbol{\mu}$ a $\overline{\boldsymbol{X}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\mu}$.

(ii) Je-li $\mathsf{var}\, X_{ij} < \infty$, pak $\mathsf{E}\,\widehat{\Sigma}_n = \Sigma$ a $\widehat{\Sigma}_n \xrightarrow{\mathsf{P}} \Sigma$.

Konsistence $\widehat{\Sigma}_n$ se ukáže stejně jako u S_n^2 (viz Věta 1.6(i)). Nestrannost lze dokázat např. následujícím způsobem:

$$\begin{aligned}
\mathsf{E}\,\widehat{\Sigma}_n &= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2} - \frac{1}{n^2}\mathsf{E}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\right)^{\otimes 2}\right] = \\
&= \frac{n}{n-1}\left(\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2} - \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2} - \frac{1}{n^2}\sum_{i}\sum_{j\neq i}\mathsf{E}\,\boldsymbol{X}_i\boldsymbol{X}_j^{\mathsf{T}}\right) = \\
&= \frac{n}{n-1}\left[\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2}\left(1-\frac{1}{n}\right) - \frac{n-1}{n}(\mathsf{E}\,\boldsymbol{X}_i)^{\otimes 2}\right] = \Sigma.
\end{aligned}$$

Vzpomeňme si na definici korelačního koeficientu mezi veličinami X_{ij} a X_{im}:

$$\varrho(X_{ij}, X_{im}) = \frac{\mathsf{cov}\,(X_{ij}, X_{im})}{\sqrt{\mathsf{var}\, X_{ij}\ \mathsf{var}\, X_{im}}}.$$

Je logické zavést výběrový korelační koeficient jakožto empirický odhad tohoto parametru vzniklý z empirických odhadů jeho jednotlivých komponent.

Definice 2.8 *Výběrový korelační koeficient*[*] $\widehat{\varrho}_{jm}$ veličin X_{ij} a X_{im}, $j = 1, \ldots, k$ a $m = 1, \ldots, k$, $j \neq m$, definujeme jako

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

Poznámka.

- $-1 \leq \widehat{\varrho}_{jm} \leq 1$.
- $\widehat{\varrho}_{jm} = 1$ (resp. -1) právě když existují konstanty $a \in \mathbb{R}$ a $b > 0$ (resp. $b < 0$) takové, že $X_{ij} = a + bX_{im}\ \forall\, i = 1, \ldots, n$.
- $\widehat{\varrho}_{jm}$ je konsistentní odhad korelačního koeficientu $\varrho(X_{ij}, X_{im})$ [věta o spojité transformaci], ale není nestranný.

Zde končí předn. 11 (9.11.)

2.5 Momentová metoda

Uvažujme nyní parametrický model: máme náhodný výběr $X_1, \ldots, X_n$ z rozdělení s hustotou $f(x; \boldsymbol{\theta}_X)$, kde tvar funkce $f(\cdot;\cdot)$ je známý a $\boldsymbol{\theta}_X$ je neznámý (vektorový) parametr, jenž leží v parametrickém prostoru $\Theta \subseteq \mathbb{R}^d$, $d \geq 1$. Pracujeme tedy s modelem

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

[*] Angl. *sample correlation coefficient*

Cílem je odhadnout parametr $\boldsymbol{\theta}_X$. Využijeme toho, že máme k dispozici konsistentní odhady momentů a že momenty rozdělení X_i obvykle umíme vyjádřit jako funkce neznámých parametrů. Budeme předpokládat, že $\mathsf{E}\,|X|^d < \infty$.

Uvažujme nejprve $d = 1$. Máme $\mathsf{E}\,X_i = g(\theta_X)$. Pokud je funkce g ryze monotonní, můžeme ji zinvertovat a dostaneme $\theta_X = g^{-1}(\mathsf{E}\,X_i)$. Víme, že $\overline{X}_n$ je konsistentní odhad a, pokud $\mathsf{var}\,X_i < \infty$, pak $\sqrt{n}(\overline{X}_n - g(\theta_X)) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \mathsf{var}\,X_i)$. Hledaný parametr θ_X můžeme odhadnout pomocí $\widehat{\theta}_n = g^{-1}(\overline{X}_n)$.

- Je-li g^{-1} spojitá funkce, pak $\widehat{\theta}_n$ je konsistentním odhadem θ_X [věta o spojité transformaci].
- Má-li g^{-1} spojitou derivaci, pak $\sqrt{n}(\widehat{\theta}_n - \theta_X) \xrightarrow{\mathsf{D}} \mathsf{N}(0, V(\theta_X))$. Asymptotický rozptyl $V(\theta_X)$ spočítáme pomocí Δ-metody (Věta P.7.12) a odhadneme pomocí $V(\widehat{\theta}_n)$.

Příklady.

1. $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{Po}(\lambda_X)$, $\mathsf{E}\,X_i = \lambda_X$. Momentovým odhadem parametru λ_X je $\widehat{\theta}_n = \overline{X}_n$.

2. $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{Geo}(p_X)$, $\mathsf{E}\,X_i = \frac{1-p_X}{p_X}$. Momentovým odhadem parametru p_X je $\widehat{\theta}_n = \frac{1}{1+\overline{X}_n}$. Platí $\sqrt{n}(\widehat{\theta}_n - p_X) \xrightarrow{\mathsf{D}} \mathsf{N}(0,\, p_X^2(1-p_X))$.

3. $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{R}(0, \theta_X)$, $\mathsf{E}\,X_i = \theta_X/2$. Momentovým odhadem parametru θ_X je $\widehat{\theta}_n = 2\overline{X}_n$. Platí $\sqrt{n}(\widehat{\theta}_n - \theta_X) \xrightarrow{\mathsf{D}} \mathsf{N}(0,\, \theta_X^2/3)$.

Nyní rozšíříme momentovou metodu na $d = 2$ parametry.

Vyjádříme $(\mathsf{E}\,X_i, \mathsf{var}\,X_i)^\mathsf{T} = g(\boldsymbol{\theta}_X)$. Řešíme jako soustavu dvou rovnic o dvou neznámých, z nichž se snažíme jednoznačně vyjádřit $\boldsymbol{\theta}_X$ jakožto funkci $\mathsf{E}\,X_i$ a $\mathsf{var}\,X_i$ (lze, pokud je funkce g prostá). Dostaneme $\boldsymbol{\theta}_X = g^{-1}(\mathsf{E}\,X_i, \mathsf{var}\,X_i)$.

- Víme, že $\overline{X}_n$ a S_n^2 jsou konsistentní odhady $\mathsf{E}\,X_i$ a $\mathsf{var}\,X_i$. Je-li g^{-1} spojitá, $\widehat{\boldsymbol{\theta}}_n = g^{-1}(\overline{X}_n, S_n^2)$ je konsistentní odhad $\boldsymbol{\theta}_X$.
- Z věty 1.6, část (iv) víme, že pokud $\mathsf{E}\,X_i^4 < \infty$, pak $\overline{X}_n$ a S_n^2 jsou sdruženě asymptoticky normální. Má-li g^{-1} spojitou derivaci, pak podle Δ-metody má i $\widehat{\boldsymbol{\theta}}_n$ asymptoticky sdružené normální rozdělení s rozptylovou maticí, kterou lze spočítat pomocí věty 1.6 a Δ-metody.

Příklady.

4. $X_1, \ldots, X_n$ je náhodný výběr z gama rozdělení s parametry a a p. Momentovou metodou dostaneme konsistentní a asymptoticky normální odhady

$$\widehat{a} = \frac{\overline{X}_n}{S_n^2} \quad \text{a} \quad \widehat{p} = \frac{\overline{X}_n^2}{S_n^2}.$$

5. $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{R}(\theta_1, \theta_2)$. Momentovou metodou dostaneme konsistentní a asymptoticky normální odhady

$$\widehat{\theta}_1 = \overline{X}_n - \sqrt{3S_n^2} \quad \text{a} \quad \widehat{\theta}_2 = \overline{X}_n + \sqrt{3S_n^2}.$$

6. $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{B}(\alpha, \beta)$. Momentovou metodou dostaneme konsistentní a asymptoticky normální odhady

$$\widehat{\alpha} = \overline{X}_n\left(\frac{\overline{X}_n(1-\overline{X}_n)}{S_n^2} - 1\right) \quad \text{a} \quad \widehat{\beta} = (1-\overline{X}_n)\left(\frac{\overline{X}_n(1-\overline{X}_n)}{S_n^2} - 1\right)$$

(odhady jsou smysluplné pouze pokud $S_n^2 < \overline{X}_n(1-\overline{X}_n)$).

Poznámka. Odhady získané momentovou metodou mívají větší asymptotický rozptyl než odhady metodou maximální věrohodnosti, která bude probírána v Matematické statistice 2.

3 Principy testování hypotéz

3.1 Základní pojmy a definice

Nechť $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ je náhodný výběr nezávislých k-rozměrných náhodných vektorů s rozdělením $F_X \in \mathcal{F}$, kde $\mathcal{F}$ je model. Nechť $\boldsymbol{\theta} = t(F) \in \mathbb{R}^d$ je charakteristika rozdělení, která nás zajímá (parametr), nechť $\Theta = \{t(F), F \in \mathcal{F}\} \subseteq \mathbb{R}^d$ označuje všechny možné hodnoty parametru v modelu $\mathcal{F}$ (nazývá se *parametrický prostor*[*]). Označme skutečný parametr jako $\boldsymbol{\theta}_X = t(F_X)$. Označme celá napozorovaná data symbolem $\boldsymbol{X} = (\boldsymbol{X}_1^\mathsf{T}, \ldots, \boldsymbol{X}_n^\mathsf{T})^\mathsf{T}$.

Příklady. Nově zaváděné pojmy a tvrzení budeme v celé této kapitole objasňovat na následujících příkladech.

A. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{N}(\theta_X, \sigma_0^2)$, kde $\sigma_0^2 > 0$ je známo. Máme tedy model
$$\mathcal{F} = \{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\}.$$

B. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{N}(\theta_X, \sigma_X^2)$, kde σ_X^2 není známo. Pracujeme s modelem
$$\mathcal{F}' = \{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R},\ \sigma^2 > 0\} \supset \mathcal{F}.$$

C. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení F_X s konečnými druhými momenty. Pracujeme s neparametrickým modelem
$$\mathcal{F}'' = \mathcal{L}^2 \supset \mathcal{F}' \supset \mathcal{F}.$$

Testovaným parametrem bude střední hodnota $\theta = \int x\, dF(x)$, jeho skutečná hodnota je $\theta_X = \mathsf{E}\, X_i$, dimenze d parametru θ je 1. Parametrický prostor je $\Theta = \mathbb{R}$.

Zvolme si nyní dvě neprázdné disjunktní podmnožiny Θ, které označíme Θ_0 a Θ_1. Řekněme, že nás nyní nezajímá konkrétní hodnota parametru $\boldsymbol{\theta}_X$, ale chceme pouze odpovědět na otázku, zdali $\boldsymbol{\theta}_X \in \Theta_0$ nebo $\boldsymbol{\theta}_X \in \Theta_1$.

Definice 3.1 (Hypotéza a alternativa)

- Množinu Θ_0 nazýváme [nulová] *hypotéza*[†], množinu Θ_1 nazýváme *alternativa*[‡]. Hypotézu označujeme obvykle symbolem H_0, alternativu symbolem H_1. Mluvíme o *testování* hypotézy $H_0 : \boldsymbol{\theta}_X \in \Theta_0$ proti alternativě $H_1 : \boldsymbol{\theta}_X \in \Theta_1$.
- Označme $\mathcal{F}_0 \overset{\mathsf{df}}{=} \{F \in \mathcal{F} : t(F) \in \Theta_0\}$, tj. všechna rozdělení v modelu $\mathcal{F}$, jejichž parametry splňují hypotézu. Jestliže $\mathcal{F}_0 = \{F_0\}$ (tj. v modelu existuje právě jedno rozdělení, které hypotézu splňuje), hypotézu nazýváme *jednoduchou*[§], jinak *složenou*[¶]. Jednoduchou hypotézu tedy dostaneme, pokud $\Theta_0 = \{\boldsymbol{\theta}_0\}$ je jednobodová množina a zároveň existuje právě jedno rozdělení $F_0 \in \mathcal{F}$ takové, že $t(F_0) = \boldsymbol{\theta}_0$.

[*] Angl. *parameter space* [†] Angl. *null hypothesis* [‡] Angl. *alternative hypothesis* [§] Angl. *simple null hypothesis*
[¶] Angl. *composite null hypothesis*

- Označme $\mathcal{F}_1 \stackrel{\text{df}}{=} \{F \in \mathcal{F} : t(F) \in \Theta_1\}$, tj. všechna rozdělení v modelu $\mathcal{F}$, jejichž parametry splňují alternativu. Jestliže $\mathcal{F}_1 = \{F_1\}$ (tj. v modelu existuje právě jedno rozdělení, které alternativu splňuje), alternativu nazýváme *jednoduchou*[*], jinak *složenou*[†]. Jednoduchou alternativu tedy dostaneme, pokud $\Theta_1 = \{\boldsymbol{\theta}_1\}$ je jednobodová množina a zároveň existuje právě jedno rozdělení $F_1 \in \mathcal{F}$ takové, že $t(F_1) = \boldsymbol{\theta}_1$.

Většinou bereme $\Theta_1 = \Theta_0^{\mathsf{c}}$ a $\mathcal{F}_1 = \mathcal{F}_0^{\mathsf{c}}$. V případech, kdy tomu tak není, tj. $\Theta_0 \cup \Theta_1 \subsetneq \Theta$, můžeme zúžit si model na $\mathcal{F}^0 = \{F \in \mathcal{F} : t(F) \in \Theta_0 \cup \Theta_1\}$. Předpokládat, že $\Theta_1 = \Theta_0^{\mathsf{c}}$ a $\mathcal{F}_1 = \mathcal{F}_0^{\mathsf{c}}$ tedy není na újmu obecnosti.

Zde končí předn. 12 (10.11.)

Uvažujme nyní jednorozměrný parametr θ a parametrický prostor $\Theta = \mathbb{R}$.

- Nejobvyklejší volba hypotézy je $\Theta_0 = \{\theta_0\}$ pro nějaké předem zvolené $\theta_0 \in \mathbb{R}$, tj. testujeme $H_0 : \theta_X = \theta_0$. Za alternativu volíme $\Theta_1 = \Theta_0^{\mathsf{c}}$, tj. $H_1 : \theta_X \neq \theta_0$. Výslednou proceduru pak nazýváme *oboustranný test*[‡], respektive *test proti oboustranné alternativě*.

- Jiná možnost je volit $\Theta_0 = (-\infty, \theta_0\rangle$, tj. testovat $H_0 : \theta_X \leq \theta_0$ proti $H_1 : \theta_X > \theta_0$, případně $\Theta_0 = \langle\theta_0, \infty)$, tj. testovat $H_0 : \theta_X \geq \theta_0$ proti $H_1 : \theta_X < \theta_0$. Tyto testy nazýváme *jednostranné testy*[§], respektive *testy proti jednostranné alternativě*. Všimněte si, že krajní hodnota θ_0 je pokaždé zahrnuta v nulové hypotéze.

Volba hypotézy je dána podstatou praktického problému, který řešíme. V některých případech volíme hypotézu značně odlišně od tří zmíněných možností. V této přednášce se však budeme zabývat pouze výše zmíněnými oboustrannými a jednostrannými testy.

Příklady. Uvažujme oboustranný test parametru $\theta = t(F) = \int x\, dF(x) \in \mathbb{R}$. Testujeme hypotézu $H_0 : \theta_X = \theta_0$ proti alternativě $H_1 : \theta_X \neq \theta_0$.

A. Model $\mathcal{F} = \{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\}$. V tomto modelu je $\mathcal{F}_0 = \{\mathsf{N}(\theta_0, \sigma_0^2)\}$, jedná se tedy o test jednoduché hypotézy. Alternativa je složená, $\mathcal{F}_1 = \{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R} \setminus \{\theta_0\}\}$.

B. Model $\mathcal{F}' = \{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R},\ \sigma^2 > 0\}$. V tomto modelu je hypotéza složená, $\mathcal{F}_0 = \{\mathsf{N}(\theta_0, \sigma^2),\ \sigma^2 > 0\}$, alternativa je také složená, $\mathcal{F}_1 = \{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R} \setminus \{\theta_0\},\ \sigma^2 > 0\}$.

C. Model $\mathcal{F}'' = \mathcal{L}^2$. V tomto modelu je hypotéza složená, $\mathcal{F}_0 = \{F \in \mathcal{L}^2 : t(F) = \theta_0\}$, alternativa je také složená, $\mathcal{F}_1 = \{F \in \mathcal{L}^2 : t(F) \neq \theta_0\}$.

Na základě náhodného výběru $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ chceme rozhodnout, zda H_0 platí nebo nikoli. Použijeme k tomu nějakou vhodně zvolenou funkci dat $S(\boldsymbol{X})$, které říkáme *testová statistika*[¶], a množinu C, které říkáme *kritický obor*[‖]. Testová statistika je obvykle jednorozměrná; kritický obor je pak nějaká podmnožina $\mathbb{R}$. Rozhodujeme se podle toho, jestli testová statistika padne do kritického oboru, či nikoli.

- Pokud $S(\boldsymbol{X}) \in C$, učiníme závěr, že *zamítáme* hypotézu H_0 ve prospěch alternativy H_1.

- Pokud $S(\boldsymbol{X}) \notin C$, učiníme závěr, že hypotézu H_0 *nemůžeme zamítnout* ve prospěch alternativy H_1.

[*] Angl. *simple alternative* [†] Angl. *composite alternative* [‡] Angl. *two-sided test* [§] Angl. *one-sided tests*
[¶] Angl. *test statistic* [‖] Angl. *critical region*

Definice 3.2 (Test) *Statistický test* je definován pomocí testové statistiky $S(\boldsymbol{X})$, kritického oboru C a výše uvedeného pravidla pro zamítání hypotézy. Dva testy $(S(\boldsymbol{X}), C)$ a $(S^*(\boldsymbol{X}), C^*)$ nazveme *ekvivalentní* právě když $S(\boldsymbol{X}) \in C \Leftrightarrow S^*(\boldsymbol{X}) \in C^*$ skoro jistě, tj. oba testy vydávají s pravděpodobností 1 totéž rozhodnutí.

Poznámka. Budeme vyžadovat, aby testová statistika splňovala následující podmínku:
Pokud $F_1 \neq F_2$ a $t(F_1) = t(F_2) = \boldsymbol{\theta}$, pak pro každou borelovskou množinu B platí

$$\int \mathbb{1}_B(S(\boldsymbol{x}))\, dF_1(x_1) \cdots dF_1(x_n) - \int \mathbb{1}_B(S(\boldsymbol{x}))\, dF_2(x_1) \cdots dF_2(x_n) \to 0 \quad \text{pro } n \to \infty,$$

tj. rozdělení testové statistiky $S(\boldsymbol{X})$ je stejné (nebo aspoň přibližně stejné), ať mají data rozdělení F_1 nebo F_2.

Platí-li tato podmínka, pak rozdělení testové statistiky nezávisí na jiných charakteristikách rozdělení F_X než na testovaném parametru $\boldsymbol{\theta}$. Můžeme tedy označit

$$\mathsf{P}_{\boldsymbol{\theta}}[S(\boldsymbol{X}) \in B] \overset{\mathrm{df}}{=} \int \mathbb{1}_B(S(\boldsymbol{x}))\, dF(x_1) \cdots dF(x_n),$$

kde F je libovolné rozdělení splňující $t(F) = \boldsymbol{\theta}$.

3.2 Hladina testu a síla testu

Při testování hypotéz mohou nastat čtyři situace v závislosti na tom, zdali hypotéza ve skutečnosti platí a zdali ji test zamítne.

- **Hypotéza platí, test ji nezamítne**, tj. $\boldsymbol{\theta} \in \Theta_0$ a $S(\boldsymbol{X}) \notin C$. V tomto případě test rozhodl správně.
- **Hypotéza platí, test ji zamítne**, tj. $\boldsymbol{\theta} \in \Theta_0$ a $S(\boldsymbol{X}) \in C$. V tomto případě test rozhodl nesprávně.
- **Hypotéza neplatí, test ji nezamítne**, tj. $\boldsymbol{\theta} \notin \Theta_0$ a $S(\boldsymbol{X}) \notin C$. V tomto případě test rozhodl nesprávně.
- **Hypotéza neplatí, test ji zamítne**, tj. $\boldsymbol{\theta} \notin \Theta_0$ a $S(\boldsymbol{X}) \in C$. V tomto případě test rozhodl správně.

Definice 3.3 (Chyba I. a II. druhu)
(i) Jestliže test zamítl platnou hypotézu, říkáme, že nastala *chyba I. druhu*[*].
(ii) Jestliže test nezamítl neplatnou hypotézu, říkáme, že nastala *chyba II. druhu*[†].

Chybám I. a II. druhu se obecně nelze vyhnout. Klasický statistický přístup k testování hypotéz spočívá v tom, že se snažíme omezit pravděpodobnost chyby I. druhu, zatímco na chybu II. druhu neklademe žádné striktní požadavky.

Definice 3.4 (Hladina testu) Nechť $\alpha \in (0, 1)$ je předem stanovené číslo.
(i) Jestliže kritický obor C splňuje podmínku

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \mathsf{P}_{\boldsymbol{\theta}}[S(\boldsymbol{X}) \in C] = \alpha,$$

říkáme, že test $(S(\boldsymbol{X}), C)$ má *hladinu* významnosti [‡] přesně α.

[*] Angl. *type I error* [†] Angl. *type II error* [‡] Angl. *significance level*

(ii) Jestliže kritický obor C splňuje podmínku

$$\sup_{\boldsymbol{\theta}\in\Theta_0} \mathsf{P}_{\boldsymbol{\theta}}[S(\boldsymbol{X}) \in C] \to \alpha \quad \text{pro } n \to \infty,$$

říkáme, že test $(S(\boldsymbol{X}), C)$ má hladinu α asymptoticky.

Poznámka.

- Je-li množina $\Theta_0 = \{\boldsymbol{\theta}_0\}$ jednobodová, pak můžeme přesnou hladinu testu psát jednodušeji
$$\alpha = \mathsf{P}_{\boldsymbol{\theta}_0}[S(\boldsymbol{X}) \in C].$$
- Zhruba řečeno, hladina testu je pravděpodobnost chyby prvního druhu, to jest pravděpodobnost zamítnutí platné hypotézy. Pokud hypotéza zahrnuje více než jednu hodnotu parametru, pak jde o nejhorší možnou pravděpodobnost chyby prvního druhu.

Klasický přístup k testování hypotéz můžeme shrnout takto:

1. Předem stanovíme požadovanou hladinu testu α, kterou má test dosáhnout buď přesně nebo asymptoticky.
2. Najdeme vhodnou testovou statistiku $S(\boldsymbol{X})$.
3. Kritický obor $C = C(\alpha)$ zvolíme v závislosti na α tak, aby hladina testu (přesná nebo asymptotická) byla právě α a přitom pravděpodobnost chyby II. druhu byla co nejmenší.

Poznámka.

- Hladina testu se volí malá, v praxi se obvykle bere $\alpha = 0{,}05$.
- Má-li testová statistika $S(\boldsymbol{X})$ diskrétní rozdělení, pak není možné dosáhnout zcela libovolné hladiny α. V tom případě se spokojujeme s hladinou $\alpha' < \alpha$, která je nejblíže k původně požadovanému α.
- Tato procedura zaručuje, že pravděpodobnost zamítnutí platné hypotézy nemůže být větší než zvolená tolerance α.

Terminologie.

- Testu, jehož skutečná hladina je menší než požadované α, se říká test *konservativní*. Testu, jehož skutečná hladina je větší než požadované α, se říká *antikonservativní*.
- Test, který požadované hladiny α dosahuje přesně, budeme nazývat *přesný test*. Test, který požadované hladiny α dosahuje jen asymptoticky, budeme nazývat *asymptotický test*.

Definice 3.5 (Síla testu) Zvolme $\boldsymbol{\theta} \in \Theta_1$ a uvažujme data s rozdělením F splňujícím $t(F) = \boldsymbol{\theta}$ a tím porušujícím hypotézu. Pak pravděpodobnost $\beta(\boldsymbol{\theta})$ zamítnutí neplatné hypotézy při hodnotě parametru $\boldsymbol{\theta}$, tj.

$$\beta(\boldsymbol{\theta}) = \mathsf{P}_{\boldsymbol{\theta}}[S(\boldsymbol{X}) \in C],$$

se nazývá *síla*[*] testu proti alternativě $\boldsymbol{\theta}$.

Poznámka. Síla testu je pravděpodobnost zamítnutí neplatné hypotézy při dané konkrétní alternativě $\boldsymbol{\theta}$. Síla závisí na alternativě, pro níž ji vyhodnocujeme. Síla je rovna doplňku pravděpodobnosti chyby II. druhu do jedničky. Síla testu nemá netriviální dolní hranici; o pravděpodobnosti chyby II. druhu nemůžeme předpokládat, že je malá.

[*] Angl. *power*

Funkci $\beta(\boldsymbol{\theta})$ můžeme snadno rozšířit i na $\boldsymbol{\theta} \in \Theta_0$.

Definice 3.6 (Silofunkce) Funkce

$$\beta(\boldsymbol{\theta}) = \mathsf{P}_{\boldsymbol{\theta}}[S(\boldsymbol{X}) \in C]$$

zobrazující celý parametrický prostor Θ do $\langle 0, 1\rangle$ se nazývá *silofunkce* testu.

Má-li test hladinu α, pak musí platit $\sup_{\boldsymbol{\theta}\in\Theta_0} \beta(\boldsymbol{\theta}) = \alpha$ (nebo $\to \alpha$ pro $n \to \infty$).

Poznámka (Interpretace výsledku testu)**.**

- Skončí-li test *zamítnutím hypotézy* H_0, znamená to, že rozdělení dat neodpovídá rozdělení, jaké by data měla za platnosti hypotézy. Pravděpodobnost chybného zamítnutí v případě, že hypotéza platí, je omezena shora hladinou α, která je malá. Hypotézu H_0 vyvracíme, prokázali jsme platnost alternativy H_1.
- Skončí-li test tím, že *hypotézu* H_0 *nemůžeme zamítnout*, znamená to, že rozdělení dat není dostatečně odlišné od rozdělení, jaké by data měla za platnosti hypotézy. Proto nemůžeme usoudit, že hypotéza H_0 platí a alternativa neplatí. Pravděpodobnost chybného rozhodnutí v případě, že hypotéza neplatí, může být značně velká. Tento výsledek tedy neznamená potvrzení platnosti hypotézy.
- Hypotéza H_0 a alternativa H_1 při testování nevystupují symetricky. Hypotézu můžeme vyvrátit ve prospěch alternativy, ale nemůžeme ji potvrdit nebo prokázat.

Abychom mohli stanovit kritický obor $C(\alpha)$, který dodržuje požadovanou hladinu α, musíme být schopni spočítat přesné nebo asymptotické rozdělení testové statistiky za platnosti hypotézy, a to nesmí záviset na neznámých charakteristikách rozdělení F_X. *Testovou statistiku* $S(\boldsymbol{X})$ tedy volíme tak, aby

(i) její rozdělení bylo citlivé na hodnotu testovaného parametru $\boldsymbol{\theta}$;
(ii) za platnosti H_0 její rozdělení nezáviselo na neznámých parametrech a bylo známo aspoň asymptoticky.

Máme-li testovou statistiku, *kritický obor* $C(\alpha)$ volíme tak, aby

(i) byla dodržena požadovaná hladina testu α;
(ii) v kritickém oboru byly zahrnuty ty hodnoty testové statistiky, které jsou za platnosti hypotézy méně pravděpodobné než za platnosti alternativy.

Kritický obor $C(\alpha)$ má ve většině případů jeden z následujících tvarů:

- $(c_U(\alpha), \infty)$, tj. zamítáme pro příliš velké hodnoty testové statistiky $S(\boldsymbol{X})$;
- $(-\infty, c_L(\alpha))$, tj. zamítáme pro příliš malé hodnoty testové statistiky $S(\boldsymbol{X})$;
- $(-\infty, c_L(\alpha)) \cup (c_U(\alpha), \infty)$, tj. zamítáme jak pro příliš malé tak pro příliš velké hodnoty testové statistiky $S(\boldsymbol{X})$;
- $(-\infty, -c_U(\alpha)) \cup (c_U(\alpha), \infty)$, tj. zamítáme pro příliš velké hodnoty $|S(\boldsymbol{X})|$.

Konstanty $c_L(\alpha)$ a $c_U(\alpha)$, které určují hranice kritického oboru, nazýváme *kritické hodnoty*[*].

Zde končí předn. 13 (16.11.)

[*] Angl. *critical values*

Příklad (A1)**.** Oboustranný test střední hodnoty normálního rozdělení se známým rozptylem.

Máme náhodný výběr $X_1, \ldots, X_n$ z rozdělení $F_X = \mathsf{N}(\theta_X, \sigma_0^2) \in \mathcal{F} = \{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\}$. Testujeme $H_0 : \theta_X = \theta_0$ proti $H_1 : \theta_X \neq \theta_0$.

Testovou statistiku založíme na bodovém odhadu parametru θ_X, tj. průměru. Víme, že

$$U_n = \sqrt{n}\,\frac{\overline{X}_n - \theta_0}{\sigma_0}$$

má za platnosti hypotézy H_0 rozdělení $\mathsf{N}(0, 1)$. Jestliže hypotéza neplatí, tj. $\theta_X - \theta_0 = \delta \neq 0$, pak

$$U_n = \sqrt{n}\,\frac{\overline{X}_n - \theta_X + \theta_X - \theta_0}{\sigma_0} = \sqrt{n}\,\frac{\overline{X}_n - \theta_X}{\sigma_0} + \sqrt{n}\,\frac{\delta}{\sigma_0}$$

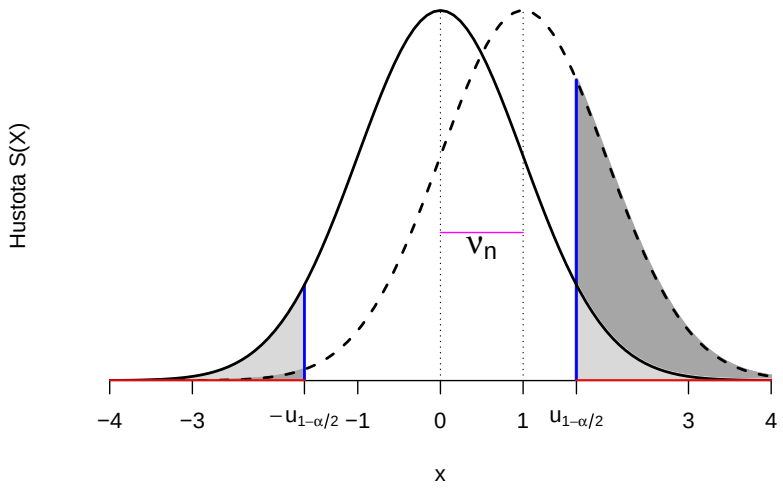
má rozdělení $\mathsf{N}(\nu_n, 1)$, kde $\nu_n = \sqrt{n}\delta/\sigma_0$. Je-li porušena hypotéza, pak se rozdělení testové statistiky posouvá pryč od nuly, a to tím dále, čím větší je n a $|\theta_X - \theta_0|$. Hodnoty testové statistiky, které jsou daleko od nuly, tedy povedou k zamítnutí hypotézy.

Kritický obor bude mít tvar $(-\infty, c_L(\alpha)) \cup (c_U(\alpha), \infty)$. Kritické hodnoty $c_L(\alpha)$ a $c_U(\alpha)$ určíme tak, aby $\mathsf{P}_{\theta_0}[U_n \in (-\infty, c_L(\alpha))] = \mathsf{P}_{\theta_0}[U_n \in (c_U(\alpha), \infty)] = \alpha/2$. To zaručí, že hladina testu je přesně rovna α. Odtud máme díky symetrii hustoty $c_U(\alpha) = -c_L(\alpha) = u_{1-\alpha/2}$. Test tedy funguje takto

$$\text{zamítni } H_0 : \theta_X = \theta_0 \quad \Longleftrightarrow \quad |U_n| = \sqrt{n}\,\frac{\left|\overline{X}_n - \theta_0\right|}{\sigma_0} > u_{1-\alpha/2},$$

tj. zamítáme hypotézu, pokud se $\overline{X}_n$ liší od hypotetické hodnoty θ_0 o více než $u_{1-\alpha/2}\sigma_0/\sqrt{n}$. Za kvantil $u_{1-\alpha/2}$ dosazujeme 1,96 pro $\alpha = 0{,}05$ a 1,645 pro $\alpha = 0{,}1$. Kritický obor a hustoty testové statistiky za hypotézy a za alternativy jsou zobrazeny na obrázku 3.1.

Obrázek 3.1: Hustota testové statistiky U_n za hypotézy a za alternativy pro $\nu_n = 1$ a $\alpha = 0{,}1$. Kritické hodnoty jsou vyznačeny modře, kritický obor červeně.



Spočítejme nyní silofunkci tohoto testu. Vezměme nějaké θ takové, že $\theta - \theta_0 = \delta \neq 0$. Pokud θ je skutečný parametr, pak rozdělení U_n je $\mathsf{N}(\nu_n, 1)$ a rozdělení $U_n - \nu_n$ je $\mathsf{N}(0, 1)$. Dostaneme tedy

$$\begin{aligned}\beta(\theta) &= \mathsf{P}_\theta[U_n \in C(\alpha)] = \mathsf{P}_\theta\left[U_n < -u_{1-\alpha/2}\right] + \mathsf{P}_\theta\left[U_n > u_{1-\alpha/2}\right] = \\ &= \mathsf{P}_\theta\left[U_n - \nu_n < -u_{1-\alpha/2} - \nu_n\right] + \mathsf{P}_\theta\left[U_n - \nu_n > u_{1-\alpha/2} - \nu_n\right] = \\ &= \Phi(-u_{1-\alpha/2} - \nu_n) + 1 - \Phi(u_{1-\alpha/2} - \nu_n).\end{aligned}$$

Protože $\Phi(-x) = 1 - \Phi(x)$, tento výsledek můžeme přepsat do tvaru

$$\beta(\theta) = \Phi(-u_{1-\alpha/2} - |\nu_n|) + 1 - \Phi(u_{1-\alpha/2} - |\nu_n|). \tag{3.1}$$

Pro $\theta = \theta_0$ dostaneme $\nu_n = 0$, a tedy $\beta(\theta_0) = \alpha$. Průběh silofunkce tohoto testu je zakreslen na obrázku 3.2.

Nechť δ je nenulové. Pak $|\nu_n|$ roste do nekonečna s rostoucím n a od určitého n počínaje bude $\Phi(-u_{1-\alpha/2} - |\nu_n|)$ zanedbatelné proti zbytku $\beta(\theta)$. Silofunkci tedy můžeme aproximovat výrazem

$$\beta(\theta) \approx 1 - \Phi\left(u_{1-\alpha/2} - \sqrt{n}\,\frac{|\delta|}{\sigma_0}\right) \tag{3.2}$$

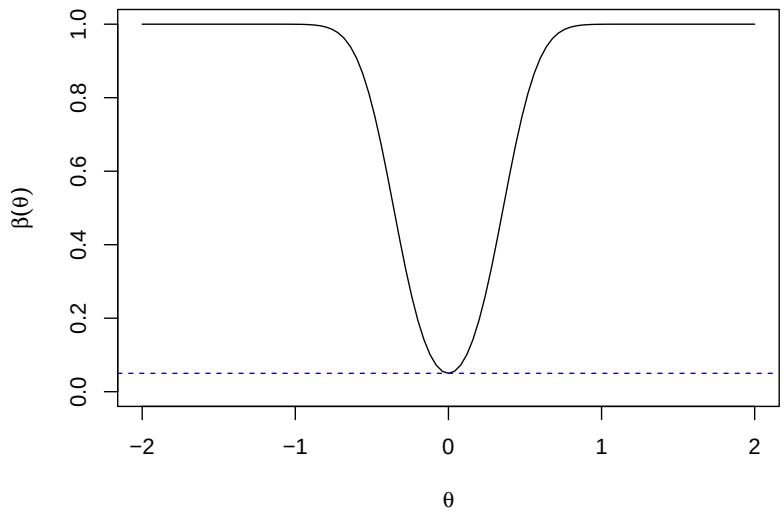
Odtud můžeme snadno spočítat, kolik pozorování je potřeba, aby test dosáhl síly alespoň β (například 0,95). Požadovaný rozsah výběru je

$$n \geq (u_{1-\alpha/2} + u_\beta)^2 \frac{\sigma_0^2}{\delta^2}. \tag{3.3}$$

Poznámka. Jak jsme viděli v předchozím příkladě, síla testu závisí na

- hladině testu α
- alternativě θ, respektive její vzdálenosti δ od hypotézy θ_0

Obrázek 3.2: Silofunkce oboustranného testu střední hodnoty normálního rozdělení se známým rozptylem pro $\theta_0 = 0$, $\sigma_0^2 = 1$, $n = 30$ a $\alpha = 0{,}05$.



- rozptylu pozorování σ_0^2
- počtu pozorování n

Z těchto faktorů je možné ovlivnit pouze počet pozorování. Chceme-li dosáhnout dostatečné síly, musíme získat alespoň takový počet pozorování, jaký je uveden v (3.3).

Poznámka. Všimněme si, že síla předchozího testu proti libovolné alternativě konverguje k 1 při $n \to \infty$ (viz (3.2)). Tuto vlastnost nazýváme *konsistence testu*. Konsistence je velmi žádoucí vlastnost, jinak totiž nemusíme být schopni dosáhnout požadované síly ani při velmi velkém počtu pozorování.

Definice 3.7 Test $(S(\boldsymbol{X}), C)$ na hladině α nazveme *konsistentním testem*[*], jestliže $\forall\theta \in \Theta_1$ $\lim_{n\to\infty} \beta(\theta) = 1$.

Zaveďme ještě jednu užitečnou vlastnost testů: *nestrannost*.

Definice 3.8 Test $(S(\boldsymbol{X}), C)$ na hladině α nazveme *nestranným testem*[†], jestliže $\forall\theta \in \Theta_1$ $\beta(\theta) \geq \alpha$.

Poznámka.

- Nenechte se zmást: pojmy nestrannost a konsistence testu mají jen velmi volný (pokud vůbec nějaký) vztah k pojmům nestrannost a konsistence odhadu.
- Nestrannost testu vyžaduje, aby síla proti každé alternativě byla alespoň α. Kdyby tomu tak nebylo, t.j. $\exists\theta \in \Theta_1$ taková, že $\beta(\theta) < \alpha$, test by tuto θ vlastně považoval za součást hypotézy.
- Test, který vždy zamítá H_0 s pravděpodobností α (bez ohledu na data) je nestranný. Nestranný test tedy existuje.
- Testy, které budeme uvádět v této přednášce, budou vždy nestranné i konsistentní. Tyto jejich vlastnosti nebudeme explicitně dokazovat. Kdybychom narazili na test, který některou z těchto vlastností nemá, upozorníme na to.

Příklad (A2)**.** JEDNOSTRANNÝ TEST STŘEDNÍ HODNOTY NORMÁLNÍHO ROZDĚLENÍ SE ZNÁMÝM ROZPTYLEM.

Máme náhodný výběr $X_1, \ldots, X_n$ z rozdělení $F_X = \mathsf{N}(\theta_X, \sigma_0^2) \in \mathcal{F} = \{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\}$. Testujeme $H_0 : \theta_X \leq \theta_0$ proti $H_1 : \theta_X > \theta_0$.

Testová statistika je stejná jako v příkladě A1

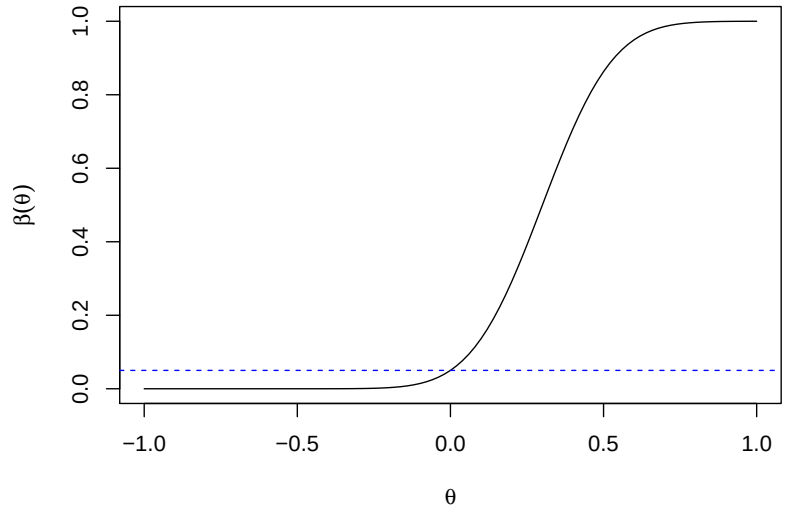
$$U_n = \sqrt{n}\,\frac{\overline{X}_n - \theta_0}{\sigma_0}.$$

Její rozdělení pro $\theta_X = \theta_0$ je $\mathsf{N}(0, 1)$. Pro hodnoty $\theta_X = \theta_0 + \delta$ máme $U_n \sim \mathsf{N}(\nu_n, 1)$, kde $\nu_n = \sqrt{n}\delta/\sigma_0$. Je-li porušena hypotéza, pak se rozdělení testové statistiky posouvá do kladných hodnot, a to tím dále, čím větší je n a δ. Příliš velké kladné hodnoty testové statistiky, tedy povedou k zamítnutí hypotézy.

Kritický obor bude mít tvar $C(\alpha) = (c_U(\alpha), \infty)$. Kritickou hodnotu $c_U(\alpha)$ určíme tak, aby $\mathsf{P}_{\theta_0}[U_n \in (c_U(\alpha), \infty)] = \alpha$. Jelikož $\mathsf{P}_\theta[U_n \in (c_U(\alpha), \infty)]$ je rostoucí funkce parametru θ, pro $\theta < \theta_0$ je pravděpodobnost zamítnutí ostře menší než α. Tento test proto splňuje podmínku $\sup_{\boldsymbol{\theta}\in\Theta_0} \mathsf{P}_\theta[U_n \in C] = \alpha$ a tudíž má hladinu α.

[*] Angl. *consistent test* [†] Angl. *unbiased test*

Obrázek 3.3: Silofunkce testu střední hodnoty normálního rozdělení se známým rozptylem proti pravostranné alternativě pro $\theta_0 = 0$, $\sigma_0^2 = 1$, $n = 30$ a $\alpha = 0{,}05$.



Dohromady dostáváme pravidlo

$$\text{zamítni } H_0 : \theta_X \le \theta_0 \quad \Longleftrightarrow \quad U_n = \sqrt{n}\,\frac{\overline{X}_n - \theta_0}{\sigma_0} > u_{1-\alpha},$$

tj. zamítáme hypotézu, pokud $\overline{X}_n$ je o více než $u_{1-\alpha}\sigma_0/\sqrt{n}$ větší než θ_0. Za kvantil $u_{1-\alpha/2}$ dosazujeme 1,645 pro $\alpha = 0{,}05$ a 1,282 pro $\alpha = 0{,}1$. Kritická hodnota pro jednostranný test na hladině α je stejná jako kritická hodnota pro oboustranný test na hladině $\alpha/2$. To je dáno tím, že nyní zamítáme hypotézu pouze v jednom chvostu rozdělení U_n.

Výpočet silofunkce je jednodušší než předtím. Vezměme nějaké θ takové, že $\theta - \theta_0 = \delta$ a dostaneme

$$\beta(\theta) = \mathsf{P}_\theta[U_n > u_{1-\alpha}] = \mathsf{P}_\theta[U_n - \nu_n > u_{1-\alpha} - \nu_n] = 1 - \Phi(u_{1-\alpha} - \nu_n).$$

Průběh silofunkce tohoto testu je zakreslen na obrázku 3.3. Počet pozorování, který je potřeba, aby test dosáhl síly alespoň β proti alternativě $\theta_0 + \delta$, $\delta > 0$, je

$$n \ge (u_{1-\alpha} + u_\beta)^2 \frac{\sigma_0^2}{\delta^2}.$$

Zde končí předn. 14 (23.11.)

Příklad (B)**.** Oboustranný test střední hodnoty normálního rozdělení s neznámým rozptylem.

Máme náhodný výběr $X_1, \ldots, X_n$ z rozdělení $F_X = \mathsf{N}(\theta_X, \sigma_X^2) \in \mathcal{F}' = \{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R},\ \sigma^2 > 0\}$. Testujeme $H_0 : \theta_X = \theta_0$ proti $H_1 : \theta_X \ne \theta_0$.

Nemůžeme použít testovou statistiku z příkladů (A1) a (A2), protože neznáme skutečný rozptyl σ_X^2. Pokud jej však nahradíme výběrovým rozptylem S_n^2 dostaneme statistiku

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \theta_0}{S_n},$$

která má v tomto modelu za platnosti hypotézy H_0 rozdělení t_{n-1} (viz věta 1.10 o T-statistice). Jestliže hypotéza neplatí, tj. $\theta_X - \theta_0 = \delta \neq 0$, pak lze hodnotu této statistiky vyjádřit jako

$$T_n = \frac{Z}{\sqrt{U/(n-1)}},$$

kde $Z \sim \mathsf{N}(\nu_n, 1)$, $\nu_n = \sqrt{n}\delta/\sigma_X$, $U \sim \chi^2_{n-1}$ a U, Z jsou nezávislé. Rozdělení této náhodné veličiny se nazývá *necentrální t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality ν_n*[*]. Jeho charakteristiky (hustota, distribuční funkce, momenty) mají komplikovaný tvar, ale stačí vědět, že pro velké n jej lze aproximovat rozdělením $\mathsf{N}(\nu_n, 1)$.

I zde tedy platí, že je-li porušena hypotéza, pak se rozdělení testové statistiky posouvá pryč od nuly, a to tím dále, čím větší je n a $|\theta_X - \theta_0|$. Hodnoty testové statistiky, které jsou daleko od nuly, tedy povedou k zamítnutí hypotézy.

Kritický obor bude mít tvar $(-\infty, c_L(\alpha)) \cup (c_U(\alpha), \infty)$. Zvolíme-li kritické hodnoty jako $c_U(\alpha) = -c_L(\alpha) = t_{n-1}(1-\alpha/2)$, test bude mít přesně hladinu α. Dostáváme pravidlo

$$\text{zamítni } H_0 : \theta_X = \theta_0 \quad \Longleftrightarrow \quad |T_n| = \sqrt{n}\,\frac{\left|\overline{X}_n - \theta_0\right|}{S_n} > t_{n-1}(1-\alpha/2).$$

To znamená, že hypotéza bude zamítnuta, pokud se bude průměr $\overline{X}_n$ lišit od hypotetické hodnoty θ_0 o více než $t_{n-1}(1-\alpha/2)S_n/\sqrt{n}$. Tento test se nazývá *jednovýběrový t-test*[†].

Silofunkci získáme podobným postupem jako v příkladě (1A). Vezměme nějaké θ takové, že $\theta - \theta_0 = \delta \neq 0$. Pokud θ je skutečný parametr, pak rozdělení T_n je necentrální t s $n-1$ stupni volnosti a parametrem necentrality $\nu_n = \sqrt{n}\delta/\sigma_X$. Označme distribuční funkci tohoto rozdělení G_n a počítejme

$$\begin{aligned}\beta(\theta) &= \mathsf{P}_\theta[T_n \in C(\alpha)] = \mathsf{P}_\theta\big[T_n < -t_{n-1}(1-\alpha/2)\big] + \mathsf{P}_\theta\big[T_n > t_{n-1}(1-\alpha/2)\big] = \\ &= G_n\big(-t_{n-1}(1-\alpha/2)\big) + 1 - G_n\big(t_{n-1}(1-\alpha/2)\big).\end{aligned}$$

Necentrální t-rozdělení nemá symetrickou hustotu, takže výsledek již nejde dále upravovat. Pokud je počet pozorování n dostatečně velký, můžeme aproximovat sílu pomocí vzorce (3.1) nebo (3.2).

Ze vzorce (3.2) lze získat aproximaci pro počet pozorování n potřebný k tomu, aby test dosáhl síly alespoň β. Požadovaný rozsah výběru je

$$n \geq (u_{1-\alpha/2} + u_\beta)^2 \frac{\sigma_X^2}{\delta^2} + 1,$$

Jednička se k výsledku přidává proto, aby trochu zkompenzovala nahrazení t rozdělení normálním. K výpočtu síly a rozsahu výběru je třeba znát skutečný rozptyl σ_X^2 nebo jej nahradit nějakým předběžným odhadem (tyto výpočty obvykle provádíme předtím, než získáme data).

Příklad (C)**.** Oboustranný test střední hodnoty libovolného rozdělení s konečným rozptylem.

Máme náhodný výběr $X_1, \ldots, X_n$ z rozdělení $F_X \in \mathcal{F}'' = \mathcal{L}^2$. Označme $\mathsf{E}\, X_i = \theta_X$, $\mathrm{var}\, X_i = \sigma_X^2$. Testujeme $H_0 : \theta_X = \theta_0$ proti $H_1 : \theta_X \neq \theta_0$.

[*] Angl. *non-central t distribution with $n-1$ degrees of freedom and noncentrality parameter ν_n* [†] Angl. *one-sample t-test*

Podle věty 1.9 (limitní věta o T statistice) má v tomto modelu náhodná veličina

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \theta_0}{S_n},$$

za platnosti hypotézy H_0 asymptoticky rozdělení $\mathsf{N}(0,1)$. Jestliže hypotéza neplatí, tj. $\theta_X - \theta_0 = \delta \neq 0$, pak

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \theta_X + \theta_X - \theta_0}{S_n} = \sqrt{n}\,\frac{\overline{X}_n - \theta_X}{S_n} + \sqrt{n}\,\frac{\delta}{S_n}$$

konverguje do $+\infty$ nebo $-\infty$ podle toho, jaké znaménko má δ. Hodnoty testové statistiky, které jsou daleko od nuly, tedy povedou k zamítnutí hypotézy.

Kritický obor bude mít tvar $(-\infty, c_L(\alpha)) \cup (c_U(\alpha), \infty)$. Kritické hodnoty $c_U(\alpha) = -c_L(\alpha) = u_{1-\alpha/2}$ zaručují, že hladina testu je asymptoticky rovna α. Místo kritické hodnoty $u_{1-\alpha/2}$ můžeme použít $t_{n-1}(1-\alpha/2)$, protože provádíme asymptotický test a $t_{n-1}(1-\alpha/2) \to u_{1-\alpha/2}$ pro $n \to \infty$.

Dostáváme pravidlo

$$\text{zamítni } H_0 : \theta_X = \theta_0 \quad \Longleftrightarrow \quad |T_n| = \sqrt{n}\,\frac{\left|\overline{X}_n - \theta_0\right|}{S_n} > t_{n-1}(1-\alpha/2),$$

Jedná se tedy opět o jednovýběrový t-test. Ukázali jsme, že jakožto asymptotický test jej můžeme použít pro libovolná data s konečným rozptylem.

3.3 P-HODNOTA

Posuzovat výsledek testu podle toho, zda $S(\boldsymbol{X})$ padne do C, není jediný ani nejběžnější způsob vyhodnocování testů. Výsledek testu se v praxi nejčastěji posuzuje pomocí tzv. p-hodnoty neboli dosažené hladiny testu.

Uvažujme hypotézu $H_0 : \theta_X = \theta_0$ proti alternativě $H_1 : \theta_X \neq \theta_0$ a test $(S(\boldsymbol{X}), C)$ s kritickým oborem tvaru $C = \mathbb{R} \setminus (c_L, c_U)$, kde $-\infty \leq c_L < c_U \leq \infty$. Označme $\boldsymbol{x} = (x_1, \ldots, x_n)$ pozorovanou realizaci náhodného výběru $\boldsymbol{X} = (X_1, \ldots, X_n)$ a $s_{\boldsymbol{x}} = S(\boldsymbol{x})$ realizovanou hodnotu testové statistiky $S(\boldsymbol{X})$, kterou jsme spočítali pro daný datový soubor. Označme dále symbolem F_0 distribuční funkci testové statistiky $S(\boldsymbol{X})$ za platnosti hypotézy. Značení $F_0(s-)$ znamená $\lim_{h \searrow 0} F_0(s-h) = \mathsf{P}_{\theta_0}[S(\boldsymbol{X}) < s]$.

Definice 3.9 (P-hodnota) *P-hodnotu*[*] neboli *dosaženou hladinu testu* definujeme jako

(i) $p(\boldsymbol{x}) = \mathsf{P}_{\theta_0}[S(\boldsymbol{X}) \geq s_{\boldsymbol{x}}] = 1 - F_0(s_{\boldsymbol{x}}-)$, pokud $c_L = -\infty$;

(ii) $p(\boldsymbol{x}) = \mathsf{P}_{\theta_0}[S(\boldsymbol{X}) \leq s_{\boldsymbol{x}}] = F_0(s_{\boldsymbol{x}})$, pokud $c_U = \infty$;

(iii) $p(\boldsymbol{x}) = 2\min(\mathsf{P}_{\theta_0}[S(\boldsymbol{X}) \geq s_{\boldsymbol{x}}], \mathsf{P}_{\theta_0}[S(\boldsymbol{X}) \leq s_{\boldsymbol{x}}]) = 2\min(1 - F_0(s_{\boldsymbol{x}}-), F_0(s_{\boldsymbol{x}}))$, pokud c_L a c_U jsou konečné a $F_0(c_L) = 1 - F_0(c_U-) = \alpha/2$.

Poznámka.

- P-hodnota je pravděpodobnost, že bychom za platnosti hypotézy napozorovali data, která by byla s hypotézou ve stejném nebo větším rozporu, než data, která analyzujeme.

[*] Angl. *p-value*

- Je-li hustota $S(\boldsymbol{X})$ je za platnosti hypotézy symetrická kolem 0 a $c_L = -c_U$ (častý případ v praxi), pak můžeme p-hodnotu počítat podle vzorce
$$p(\boldsymbol{x}) = \mathsf{P}_{\boldsymbol{\theta}_0}\left[|S(\boldsymbol{X})| \geq |s_{\boldsymbol{x}}|\right] = 2[1 - F_0(|s_{\boldsymbol{x}}|-)].$$
- Testujeme-li hypotézu $H_0 : \boldsymbol{\theta}_X \in \Theta_0$, kde $\Theta_0 \neq \emptyset$ není jednobodová množina, nahradíme $\mathsf{P}_{\boldsymbol{\theta}_0}$ v definici 3.9 výrazem $\sup_{\boldsymbol{\theta}\in\Theta_0} \mathsf{P}_{\boldsymbol{\theta}}$.
- Je-li distribuční funkce F_0 asymptotická, přidáme před $\mathsf{P}_{\boldsymbol{\theta}_0}$ nebo $\sup_{\boldsymbol{\theta}\in\Theta_0} \mathsf{P}_{\boldsymbol{\theta}}$ v definici 3.9 ještě $\lim_{n\to\infty}$. Pak tuto p-hodnotu nazveme *asymptotickou*.

Tvrzení 3.1 Nechť rozdělení testové statistiky $S(\boldsymbol{X})$ je spojité. Uvažujme test hypotézy H_0 proti alternativě H_1 daný pravidlem
$$H_0 \text{ zamítáme, jestliže } p(\boldsymbol{x}) \leq \alpha$$
$$H_0 \text{ nezamítáme, jestliže } p(\boldsymbol{x}) > \alpha,$$
Pak tento test má hladinu α (přesně nebo asymptoticky, podle toho, používáme-li přesnou nebo asymptotickou p-hodnotu).

Zde končí předn. 15 (24.11.)

Poznámka.
- Spočítáme-li p-hodnotu $p(\boldsymbol{x})$, můžeme hypotézu zamítnout na všech hladinách $\alpha' \geq p(\boldsymbol{x})$, ale nemůžeme ji zamítnout na hladinách $\alpha' < p(\boldsymbol{x})$. Proto se p-hodnotě říká *dosažená hladina testu*.
- Zamítáme-li pomocí p-hodnoty, nemusíme uvádět kritický obor a nemusíme jej přepočítávat, pokud se rozhodneme změnit hladinu testu (měnit hladinu testu poté, co je znám výsledek, však není legitimní).
- P-hodnotu můžeme chápat jako míru souladu dat s hypotézou. Pokud $p(\boldsymbol{x}) \ll \alpha$, data zamítají hypotézu s velkou „rezervou“.
- P-hodnotu není možné vykládat jako *„pravděpodobnost, že nulová hypotéza platí“*. Platnost nulové hypotézy totiž není náhodný, ale deterministický jev.

Příklad (C)**.** Máme náhodný výběr $X_1, \ldots, X_n$, $n = 26$, z rozdělení $F_X \in \mathcal{F}'' = \mathcal{L}^2$ se střední hodnotou $\mathsf{E}\, X_i = \theta_X$. Testujeme $H_0 : \theta_X = \theta_0$ proti $H_1 : \theta_X \neq \theta_0$. Testová statistika T_n má za platnosti hypotézy přibližně rozdělení t_{25}, které je symetrické kolem 0. Spočítali jsme testovou statistiku a její výsledek je $t = -1{,}37$. P-hodnota pro tento test se spočítá podle vzorce
$$p(\boldsymbol{x}) = \mathsf{P}_{\boldsymbol{\theta}_0}\left[|T_n| \geq |-1{,}37|\right] = 2[1 - F_{25}(1{,}37)] \doteq 0{,}183$$
kde F_{25} značí distribuční funkci rozdělení t_{25}. P-hodnota je 0,183. Testujeme-li na hladině $\alpha = 0{,}05$, nemůžeme zamítnout hypotézu, neboť $p(\boldsymbol{x}) > 0{,}05$. Kdybychom si však před provedením testu stanovili hladinu $\alpha' = 0{,}2$, hypotézu bychom zamítnout mohli.

Uvažujme nyní p-hodnotu $p(\boldsymbol{X})$ jakožto náhodnou veličinu, čili statistiku spočítanou z náhodného výběru $\boldsymbol{X}$. Lze ukázat, že za určitých předpokladů má p-hodnota spočítaná za platnosti hypotézy rovnoměrné rozdělení na intervalu (0, 1).

Tvrzení 3.2 Uvažujme test $(S(\boldsymbol{X}), C)$ hypotézy $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ proti $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$ s p-hodnotou $p(\boldsymbol{x})$. Nechť testová statistika $S(\boldsymbol{X})$ má spojité rozdělení a platí hypotéza. Pak $p(\boldsymbol{X}) \sim \mathsf{R}(0, 1)$.

Poznámka. Předchozí tvrzení neplatí, pokud je rozdělení testové statistiky diskrétní, ani tehdy, když hypotéza obsahuje více než jednu hodnotu parametru.

3.4 Dualita intervalových odhadů a testování hypotéz

Uvažujme náhodný výběr $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ z rozdělení $F_X \in \mathcal{F}$, kde $\mathcal{F}$ je model, nechť $\theta = t(F) \in \mathbb{R}$ je parametr a $\theta_X = t(F_X)$ je jeho skutečná hodnota. V kapitole 2.3 jsme řešili problém intervalového odhadu parametru θ_X, tj. hledali jsme náhodné veličiny C_L a C_U takové, že $\mathsf{P}[(C_L, C_U) \ni \theta_X] = 1 - \alpha$ (nebo $\to 1 - \alpha$).

V této kapitole se zabýváme testováním hypotéz, speciálně hypotézy $H_0 : \theta_X = \theta_0$ proti $H_1 : \theta_X \neq \theta_0$. Oba problémy se řeší postupy, které se v určitých rysech shodují, ale liší se v detailech.

Následující věta ukazuje, že mezi problémem testování hypotézy o parametru a problémem hledání intervalového odhadu pro ten samý parametr existuje jakási dualita. Intervalový odhad můžeme použít k testování hypotéz a test hypotézy můžeme převést na intervalový odhad.

Tvrzení 3.3 (Dualita intervalových odhadů a testování)

(i) Nechť je dán oboustranný interval spolehlivosti pro parametr θ_X s pravděpodobností pokrytí $1-\alpha$ (přesnou nebo asymptotickou), který má tvar $(C_L(\boldsymbol{X}), C_U(\boldsymbol{X}))$. Uvažujme test hypotézy $H_0 : \theta_X = \theta_0$ proti $H_1 : \theta_X \neq \theta_0$ založený na rozhodovacím pravidle

$$H_0 \text{ zamítáme, jestliže } \theta_0 \notin (C_L(\boldsymbol{X}), C_U(\boldsymbol{X}))$$
$$H_0 \text{ nezamítáme, jestliže } \theta_0 \in (C_L(\boldsymbol{X}), C_U(\boldsymbol{X})).$$

Pak tento test má hladinu α (přesně nebo asymptoticky).

(ii) Nechť je dán test hypotézy $H_0 : \theta_X = \theta$ proti $H_1 : \theta_X \neq \theta$ na hladině α (přesné nebo asymptotické). Sestavme množinu $B_{\boldsymbol{X}}$ obsahující všechny parametry $\theta \in \Theta$, pro něž se při pozorovaných datech $\boldsymbol{X}$ nezamítá hypotéza $H_0 : \theta_X = \theta$. Pak $\mathsf{P}[B_{\boldsymbol{X}} \ni \theta_X] = 1 - \alpha$ (nebo $\to 1 - \alpha$) a (je-li $B_{\boldsymbol{X}}$ interval) jedná se o interval spolehlivosti pro parametr $\boldsymbol{\theta}_X$ s pravděpodobností pokrytí $1 - \alpha$ (přesnou nebo asymptotickou).

Zde končí předn. 16 (30.11.)

Například interval spolehlivosti (2.2) pro střední hodnotu normálního rozdělení s neznámým rozptylem uvedený na straně 26 je duální k jednovýběrovému t-testu z příkladu (B) na str. 45. Jednovýběrový t-test zamítne na hladině α všechny hypotézy o θ, které neleží v $(1 - \alpha)100$ % intervalu spolehlivosti (2.2).

Tvrzení 3.3 říká, že umíme-li sestrojit interval spolehlivosti pro parametr, můžeme jej ihned využít k testování hypotéz o tomto parametru. Naopak máme-li test, můžeme s jeho pomocí sestrojit interval spolehlivosti. Tento krok je však pracnější, protože vyžaduje otestování všech možných hodnot parametru. Množina nezamítnutých hypotéz pak dává požadované pokrytí pro skutečný parametr, ale nemusí nutně tvořit interval.

4 Jednovýběrové a párové problémy pro kvantitativní data

V této kapitole uvažujeme náhodný výběr $X_1, \ldots, X_n$ kvantitativních veličin s distribuční funkcí F_X patřící do modelu $\mathcal{F}$. Zajímá nás parametr $\theta_X = t(F_X)$. Chceme testovat hypotézy o tomto parametru, případně pro něj sestrojit intervalový odhad.

4.1 Jednovýběrový Kolmogorovův-Smirnovův test

Jednovýběrový Kolmogorovův-Smirnovův test* testuje shodu distribuční funkce dat s určitou pevně danou distribuční funkcí. Je to neparametrický test, protože nepředpokládá žádný parametrický model.

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

Testovaný parametr: Celá distribuční funkce F_X

Hypotéza a alternativa:

$$H_0 : F_X(x) = F_0(x) \quad \forall x \in \mathbb{R}, \qquad H_1 : \exists x \in \mathbb{R} : F_X(x) \neq F_0(x),$$

kde F_0 je nějaká pevně specifikovaná spojitá distribuční funkce (bez neznámých parametrů).

Testová statistika je založena na empirické distribuční funkci $\widehat{F}_n$, s níž jsme se seznámili v kapitole 2.4.1 (viz str. 29). Její vlastnosti shrnuje věta 2.3. Empirická distribuční funkce je nestranným a konsistentním odhadem skutečné distribuční funkce v každém bodě. Navíc podle věty 2.3, bod (v), splňuje stejnoměrnou konsistenci, tj. $\sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F_X(x)\right| \xrightarrow{\mathsf{P}} 0$ při $n \to \infty$. Testová statistika přebírá tuto supremální normu a zachycuje s ní největší celkový rozdíl mezi $\widehat{F}_n(x)$ a $F_0(x)$.

Testová statistika:

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

Pokud hypotéza platí a F_0 je skutečná distribuční funkce dat, hodnota testové statistiky K_n bude blízko nuly. Hypotézu zamítneme, pokud se empirická distribuční funkce příliš liší od F_0, tj. pokud je testová statistika příliš velká.

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

Lemma 4.1 Platí

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

Poznámka. Předchozí lemma má několik důležitých důsledků.

* Angl. *one-sample Kolmogorov-Smirnov test*

- Testová statistika K_n se počítá pomocí Lemmatu 4.1, nikoli podle její definice. K jejímu výpočtu není třeba znát $\widehat{F}_n$.
- Platí-li hypotéza, $F_0(X_{(i)})$ má podle věty 1.13 beta rozdělení. Proto rozdělení K_n za platnosti hypotézy nezávisí na F_0.
- Z lemmatu 4.1 lze odvodit přesné rozdělení testové statistiky za platnosti hypotézy. Jedná se ovšem o netriviální výpočet, který je navíc i numericky obtížný. Proto se přesné rozdělení K_n zpravidla používá jen při velmi malém rozsahu výběru n.

Asymptotické rozdělení testové statistiky za platnosti hypotézy je určeno následujícím tvrzením, které rozšiřuje výsledek uvedený ve větě 2.3, bod (v).

Tvrzení 4.2 Nechť $X_1, \ldots, X_n$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F_X. Pak pro každé $y > 0$ platí

$$\mathrm{P}\left[\sqrt{n}\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F_X(x)\right| \le y\right] \to G(y) \equiv 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\mathrm{e}^{-2k^2y^2} \text{ pro } n \to \infty.$$

Funkce $G(y)$ je distribuční funkce. Určuje limitní rozdělení normalizované testové statistiky $\sqrt{n}K_n$ za platnosti hypotézy, tj. pro $F_X = F_0$. Toto rozdělení není normální, jak jsme byli doposud u limitních rozdělení zvyklí. Důkaz tvrzení 4.2 náleží do pokročilé teorie pravděpodobnosti, my jej neuvádíme.

Nyní již můžeme určit kritickou hodnotu pro zamítání H_0, aby měl test asymptotickou hladinu α. Označme α-kvantil rozdělení s distribuční funkcí G symbolem $k_\alpha = G^{-1}(\alpha)$. Hypotézu budeme zamítat, pokud $\sqrt{n}K_n$ překročí $k_{1-\alpha}$.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow \sqrt{n}K_n \ge k_{1-\alpha}.$$

Díky tvrzení 4.2 víme, že tento test má asymptoticky hladinu α.

Poznámka.

- Výhodou Kolmogorovova-Smirnovova testu je jeho universalita (reaguje na jakýkoli rozdíl v rozdělení dat proti hypotéze) a absence předpokladů o rozdělení F_X.
- F_0 musí být známa přesně (nesmí obsahovat neznámé parametry ani jejich odhady). Tímto testem tedy není možné testovat hypotézy jako *„Data mají (nějaké) normální rozdělení“*.
- Tento test má relativně malou sílu proti konkrétnímu typu porušení H_0 (např. změna střední hodnoty). Pokud tušíme, jaké porušení H_0 je pro danou aplikaci nejočekávanější nebo nejvíce relevantní, je lepší použít test, který je zaměřen na tento typ porušení H_0.
- Pro data z diskrétního rozdělení neplatí tvrzení 4.2. Je však možné spočítat pro ně přibližnou kritickou hodnotu Kolmogorovova-Smirnovova testu jiným způsobem.
- Tento test lze zformulovat i jako jednostranný proti alternativě $H_1' : F_X(x) \ge F_0(x)$, $\exists x \in \mathbb{R} : F_X(x) > F_0(x)$ nebo $H_1'' : F_X(x) \le F_0(x)$, $\exists x \in \mathbb{R} : F_X(x) < F_0(x)$. Jako testovou statistiku pak použijeme buď K_n^+ anebo K_n^- a zamítáme pro jejich velké hodnoty.

Obraťme nyní pozornost k problému sestrojení intervalu spolehlivosti pro distribuční funkci. Jestliže máme dané pevné $x \in \mathbb{R}$ a chceme intervalový odhad pouze pro hodnotu $F_X(x)$,

můžeme vyjít z věty 2.3, bod (iii), a použít postup uvedený v příkladě na str. 27 v kapitole 2.3.2. Dostaneme interval

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

jehož pravděpodobnost pokrytí konverguje k $1-\alpha$ pro $n \to \infty$.

Co když ale nemáme předem dané x, nýbrž chceme interval, který by pokryl hodnotu distribuční funkce kdekoli, třeba i v mnoha bodech zároveň? K tomu nemůžeme použít postup uvedený výše, ale znovu využijeme tvrzení 4.2. Máme totiž

$$\mathsf{P}\left[\sqrt{n}\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F_X(x)\right| \le k_{1-\alpha}\right] \to 1-\alpha$$

a také

$$\mathsf{P}\left[\sqrt{n}\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F_X(x)\right| \le k_{1-\alpha}\right] = \mathsf{P}\left[\sqrt{n}\left|\widehat{F}_n(x) - F_X(x)\right| \le k_{1-\alpha} \quad \forall x \in \mathbb{R}\right].$$

Sestavíme-li tedy intervaly

$$\left(\widehat{F}_n(x) - \frac{k_{1-\alpha}}{\sqrt{n}},\ \widehat{F}_n(x) + \frac{k_{1-\alpha}}{\sqrt{n}}\right),$$

bude pravděpodobnost, že skutečná distribuční funkce $F_X(x)$ leží uvnitř všech těchto intervalů pro všechna $x \in \mathbb{R}$, konvergovat k $1-\alpha$ při $n \to \infty$. Intervalům vytvářejícím oblast, v níž se se zadanou pravděpodobností nachází celý průběh nějaké neznámé funkce, se říká *pás spolehlivosti*[*]. Protože hranice pásu spolehlivosti pro distribuční funkci založené na Kolmogorovově-Smirnovově statistice mohou ležet mimo přirozený rozsah $\langle 0, 1\rangle$, předefinujeme dolní mez na $\max(0, \widehat{F}_n(x) - k_{1-\alpha}/\sqrt{n})$ a horní mez na $\min(1, \widehat{F}_n(x) + k_{1-\alpha}/\sqrt{n})$.[†]

Zde končí předn. 17 (1.12.)

4.2 Přesný jednovýběrový t-test

Jednovýběrový t-test[‡] porovnává střední hodnotu dat s nějakou zvolenou konstantou. V této kapitole předpokládáme normální rozdělení, test pak zachovává požadovanou hladinu přesně pro jakékoli $n \ge 3$. Tímto testem jsme se podrobně zabývali v Příkladě B na str. 45 (Oboustranný test střední hodnoty normálního rozdělení s neznámým rozptylem).

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

Testovaný parametr: Střední hodnota $\mu_X = \mathsf{E}\, X_i$

Hypotéza a alternativa:

$$H_0 : \mu_X = \mu_0, \qquad H_1 : \mu_X \neq \mu_0,$$

kde μ_0 je předem daná konstanta.

Testová statistika:

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{S_n},$$

[*] Angl. *confidence bounds* [†] Existuje samozřejmě řada jiných způsobů, jak sestavit pás spolehlivosti pro distribuční funkci. [‡] Angl. *one-sample t-test*

kde $\overline{X}_n$ je aritmetický průměr a S_n^2 je výběrový rozptyl.

Rozdělení testové statistiky za H_0:

$$T_n \sim t_{n-1}$$

(viz věta 1.10).

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow |T_n| \geq t_{n-1}(1-\alpha/2),$$

kde $t_{n-1}(1-\alpha/2)$ je $(1-\alpha/2)$-tý kvantil t-rozdělení s $n-1$ stupni volnosti.

P-hodnota: $p = 2(1 - F_n(|t|))$, kde t je pozorovaná hodnota testové statistiky T_n a F_n je distribuční funkce rozdělení t_{n-1}.

Interval spolehlivosti pro μ_X: Přesný interval spolehlivosti pro střední hodnotu normálního rozdělení je dán krajními body

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

Viz vzorec (2.2) na str. 26 a předcházející příklad.

Poznámka. Tento test lze převést na jednostranný test: zamítneme $H_0' : \mu_X \leq \mu_0$ proti $H_1' : \mu_X > \mu_0$, pokud testová statistika překročí kritickou hodnotu $t_{n-1}(1-\alpha)$. Zamítneme $H_0'' : \mu_X \geq \mu_0$ proti $H_1' : \mu_X < \mu_0$, pokud testová statistika nepřekročí kritickou hodnotu $-t_{n-1}(1-\alpha)$.

Viz též příklad A2. na str. 44.

4.3 Asymptotický jednovýběrový t-test

Jedná se o stejný test jako v předchozí kapitole, ale liší se jeho předpoklady. Nyní předpokládáme pouze existenci konečného druhého momentu. Test pak zachovává požadovanou hladinu přibližně pro $n \to \infty$. Tímto testem jsme se zabývali v Příkladě C na str. 46 (Oboustranný test střední hodnoty libovolného rozdělení s konečným rozptylem).

Model: $\mathcal{F} = \mathcal{L}^2$

Testovaný parametr: Střední hodnota $\mu_X = \mathsf{E}\, X_i$

Hypotéza a alternativa:

$$H_0 : \mu_X = \mu_0, \quad H_1 : \mu_X \neq \mu_0,$$

kde μ_0 je předem daná konstanta.

Testová statistika:

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{S_n},$$

kde $\overline{X}_n$ je aritmetický průměr a S_n^2 je výběrový rozptyl.

Rozdělení testové statistiky za H_0:

$$T_n \overset{\text{as.}}{\sim} \mathsf{N}(0,1)$$

(viz věta 1.9). Asymptotické rozdělení však lze aproximovat i rozdělením t_{n-1}.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow |T_n| \geq t_{n-1}(1-\alpha/2),$$

kde $t_{n-1}(1-\alpha/2)$ je $(1-\alpha/2)$-tý kvantil t-rozdělení s $n-1$ stupni volnosti. Hladina testu konverguje k α pro $n \to \infty$.

P-hodnota: $p = 2(1-F_n(|t|))$, kde t je pozorovaná hodnota testové statistiky T_n a F_n je distribuční funkce rozdělení t_{n-1}.

Interval spolehlivosti pro μ_X: Interval (2.2) má pravděpodobnost pokrytí konvergující k $1-\alpha$, jak je ukázáno v příkladě na str. 26.

Poznámka. Tento test lze převést na jednostranný test způsobem zmíněným v předchozí kapitole.

Poznámka. T-test nepotřebuje předpoklad normálního rozdělení, funguje jako asymptotický test pro libovolné rozdělení s konečným rozptylem. Pouze je potřeba mít k dispozici dostatek pozorování.

4.4 Jednovýběrový znaménkový test

Jednovýběrový znaménkový test* porovnává medián dat s pevně danou hodnotou. Je to neparametrický test, funguje pro jakékoli spojité rozdělení.

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

Testovaný parametr: Medián $m_X = F_X^{-1}(0.5)$

Hypotéza a alternativa:

$$H_0 : m_X = m_0, \quad H_1 : m_X \neq m_0,$$

kde m_0 je předem daná konstanta.

Testová statistika:

$$Y_n = \sum_{i=1}^{n} \mathbb{1}_{(0,\infty)}(X_i - m_0)$$

(počet pozorování větších než m_0).

Věta 4.3 Nechť $X_1, \ldots, X_n$ je náhodný výběr z libovolného spojitého rozdělení s mediánem m_X. Pak

(i)
$$\sum_{i=1}^{n} \mathbb{1}_{(0,\infty)}(X_i - m_X) \sim \mathrm{Bi}(n, 1/2)$$

(ii)
$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbb{1}_{(0,\infty)}(X_i - m_X) - \frac{1}{2} \right] \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1/4)$$

Poznámka. Věta 4.3 plyne z věty 1.3, části (iii) a (iv).

* Angl. *one-sample sign test*

Přesné rozdělení testové statistiky za H_0:

$$Y_n \sim \mathrm{Bi}(n, 1/2)$$

Kritický obor (přesný test): Hypotézu budeme zamítat pro příliš malé nebo příliš velké hodnoty Y_n.

$$H_0 \text{ zamítneme } \Leftrightarrow Y_n \leq c_{1n}(\alpha) \text{ nebo } Y_n \geq c_{2n}(\alpha)$$

kde $c_{1n}(\alpha)$ je největší celé číslo k_1, které splňuje $2^{-n}\sum_{j=0}^{k_1}\binom{n}{j} \leq \frac{\alpha}{2}$ a $c_{2n}(\alpha)$ je nejmenší celé číslo k_2, které splňuje $2^{-n}\sum_{j=k_2}^{n}\binom{n}{j} \leq \frac{\alpha}{2}$. (Ze symetrie binomického rozdělení plyne, že $c_{1n}(\alpha) + c_{2n}(\alpha) = n$.) Tento test má hladinu nejvýše α (přesné hladiny α nemusí být možné dosáhnout).

Asymptotické rozdělení testové statistiky za H_0:

$$\frac{2}{\sqrt{n}}\Big(Y_n - \frac{n}{2}\Big) \overset{\text{as.}}{\sim} \mathsf{N}(0, 1)$$

Kritický obor (asymptotický test): Hypotézu budeme zamítat pro příliš malé nebo příliš velké hodnoty Y_n.

$$H_0 \text{ zamítneme } \Leftrightarrow \left|\frac{2}{\sqrt{n}}Y_n - \sqrt{n}\right| \geq u_{1-\alpha/2}.$$

Poznámka.

- K výpočtu testové statistiky vlastně nepotřebujeme znát konkrétní hodnoty X_i. Stačí nám jen vědět, kolik z nich překročilo hodnotu m_0.
- Kdyby se stalo, že k hodnot X_i je přesně rovno m_0 (např. kvůli zaokrouhlování), můžeme tyto hodnoty vyloučit a provést znaménkový test na zbývajících $n - k$ pozorováních.
- Tento test lze převést na jednostranný test $H_0' : m_X \geq m_0$ (nebo $\leq m_0$).

4.5 Jednovýběrový Wilcoxonův test

Jednovýběrový Wilcoxonův test* porovnává medián nebo střední hodnotu dat s pevně danou konstantou. Je to neparametrický test, funguje pro za předpokladu symetrie hustoty.

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

Testovaný parametr: Střed symetrie δ_X

Poznámka. Model vyžaduje, aby hustota X_i byla symetrická kolem nějakého bodu δ_X. Pak musí platit $m_X = \delta_X$ a pokud $X_i \in \mathcal{L}^1$, pak i $\mathsf{E}\, X_i \equiv \mu_X = \delta_X$.

Hypotéza a alternativa:

$$H_0 : \delta_X = \delta_0, \quad H_1 : \delta_X \neq \delta_0,$$

kde δ_0 je předem daná konstanta.

* Angl. *one-sample Wilcoxon test, Wilcoxon signed rank test*

Poznámka. Za platnosti modelu $\mathcal{F}$ je hypotéza H_0 ekvivalentní hypotéze $H_0^* : m_X = \delta_0$ (test na medián). Pokud navíc $X_i \in \mathcal{L}^1$, pak je hypotéza H_0 též ekvivalentní hypotéze $H_0^{**} : \mu_X = \delta_0$ (test na střední hodnotu).

Testová statistika: Nechť $Z_i \overset{\text{df}}{=} X_i - \delta_0$. Definujme

$$W_S = \sum_{i \in \mathcal{I}} R_i,$$

kde $\mathcal{I} = \{i \in \{1, \ldots, n\} : Z_i > 0\}$ je množina všech indexů takových, že Z_i má kladné znaménko a $R_1, R_2, \ldots, R_n$ jsou pořadí absolutních hodnot $|Z_i|$ mezi všemi absolutními hodnotami $|Z_1|, \ldots, |Z_n|$.

Poznámka. Testová statistika W_S jednovýběrového Wilcoxonova testu může nabývat hodnot $0, 1, \ldots, n(n+1)/2$. Spočítá se následujícím způsobem:

1. Spočítáme odchylky $Z_i = X_i - \delta_0$ a určíme množinu indexů $\mathcal{I}$.
2. Seřadíme všechny Z_i podle jejich absolutní hodnoty od nejmenší do největší; získáme uspořádaný výběr
$$0 < |Z_{(1)}| < |Z_{(2)}| < \cdots < |Z_{(n)}|.$$
3. Určíme pořadí R_i náhodné veličiny $|Z_i|$ mezi všemi $|Z_{(1)}|, \ldots, |Z_{(n)}|$. Platí $|Z_i| = |Z_{(R_i)}|$.
4. Sečteme pořadí R_i pro $i \in \mathcal{I}$.

Velikost množiny $\mathcal{I}$ je rovna počtu pozorování, pro něž platí $X_i > \delta_0$ (srv. s testovou statistikou znaménkového testu).

Tvrzení 4.4 Nechť $X_1, \ldots, X_n$ je náhodný výběr z libovolného spojitého rozdělení splňujícího model $\mathcal{F}$ a nechť platí $H_0 : \delta_X = \delta_0$. Pak

(i)
$$\mathsf{E}\, W_S = \frac{n(n+1)}{4}, \quad \operatorname{var} W_S = \frac{n(n+1)(2n+1)}{24}.$$

(ii)
$$\frac{W_S - \mathsf{E}\, W_S}{\sqrt{\operatorname{var} W_S}} \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1).$$

Zde končí předn. 18 (7.12.)

Poznámka.

- Důkaz asymptotické normality vynecháváme.
- Hypotézu budeme zamítat pro příliš malé nebo příliš velké hodnoty W_S.
- Není-li n příliš velké, lze nalézt i přesné rozdělení testové statistiky W_S (numericky nebo v tabulkách).

Asymptotické rozdělení testové statistiky za H_0:

$$\frac{W_S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \overset{\text{as.}}{\sim} \mathsf{N}(0, 1)$$

Kritický obor (asymptotický test):

$$H_0 \text{ zamítneme } \Leftrightarrow \frac{\left|W_S - \frac{n(n+1)}{4}\right|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \geq u_{1-\alpha/2}.$$

Poznámka. Jednovýběrový Wilcoxonův test bere v úvahu i velikost odchylek od δ_0, nikoli jen jejich znaménko (jako znaménkový test). Jeho síla pro testování mediánu je obecně větší než síla znaménkového testu. Hladinu však dodržuje pouze tehdy, je-li rozdělení jednotlivých pozorování symetrické, zatímco znaménkový test takový předpoklad nevyžaduje.

4.6 Jednovýběrový χ^2 test na rozptyl

Jednovýběrový χ^2 test na rozptyl* je přesný test vyžadující normální rozdělení pozorovaných dat.

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

Testovaný parametr: Rozptyl $\sigma_X^2 = \mathrm{var}\, X_i$.

Hypotéza a alternativa:

$$H_0 : \sigma_X^2 = \sigma_0^2, \quad H_1 : \sigma_X^2 \neq \sigma_0^2,$$

kde σ_0^2 je předem daná konstanta.

Testová statistika:

$$\frac{(n-1)S_n^2}{\sigma_0^2},$$

kde S_n^2 je výběrový rozptyl (viz definice 1.4).

Asymptotické rozdělení testové statistiky za H_0:

$$\frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

podle věty 1.8 (i).

Kritický obor: Hypotézu zamítneme, pokud se výběrový rozptyl příliš liší od hypotetického rozptylu, tj. pokud je testová statistika buď moc velká nebo moc malá.

$$H_0 \text{ zamítneme } \Leftrightarrow \frac{(n-1)S_n^2}{\sigma_0^2} \leq \chi^2_{n-1}(\alpha/2) \text{ nebo } \frac{(n-1)S_n^2}{\sigma_0^2} \geq \chi^2_{n-1}(1-\alpha/2),$$

kde $\chi^2_{n-1}(\alpha/2)$ a $\chi^2_{n-1}(1-\alpha/2)$ jsou po řadě $(\alpha/2)$-tý a $(1-\alpha/2)$-tý kvantil χ^2 rozdělení s $n-1$ stupni volnosti.

P-hodnota: $p = 2\min(1 - F_n(s), F_n(s))$, kde s je pozorovaná hodnota testové statistiky a F_n je distribuční funkce rozdělení χ^2_{n-1}.

Interval spolehlivosti pro σ_X^2: (viz (2.4))

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

Poznámka.

- Při porušení předpokladu normality tento test nedodržuje hladinu ani asymptoticky.
- Tento test lze převést na jednostranný test: Hypotéza $H_0' : \sigma_X^2 \leq \sigma_0^2$ se zamítá pouze pro příliš velké hodnoty testové statistiky, kritická hodnota je $\chi^2_{n-1}(1-\alpha)$. Hypotéza $H_0'' : \sigma_X^2 \geq \sigma_0^2$, se zamítá pouze pro příliš malé hodnoty testové statistiky, kritická hodnota je $\chi^2_{n-1}(\alpha)$.

* Angl. *one-sample chi-square variance test*

4.7 Párové testy

Uvažujme náhodný výběr

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

dvousložkových náhodných vektorů s dvourozměrnou distribuční funkcí. Chceme porovnat nějakou charakteristiku marginálního rozdělení F_X náhodné veličiny X_i se stejnou charakteristikou marginálního rozdělení F_Y náhodné veličiny Y_i. Pozorování X_i a Y_i ovšem nejsou nezávislá.

Hlavní myšlenka párových testů je jednoduchá: Vezmeme rozdíly $Z_i = X_i - Y_i$ (jež tvoří náhodný výběr z nějakého jednorozměrného rozdělení) a na ně provedeme vhodný jednovýběrový test. Musíme se však zamyslet na tím, jestli hypotéza testovaná jednovýběrovým testem provedeným na Z_i má nějakou rozumnou interpretaci pro porovnání rozdělení X_i a Y_i. Někdy tomu tak je, ale v řadě případů taková interpretace neexistuje (např. párový Kolmogorovův-Smirnovův test rozumnou interpretaci nemá).

Nechť například jednovýběrový test provedený na rozdíly Z_i testuje střední hodnotu, třeba $H_0 : \mathsf{E}\, Z_i = 0$. Tato hypotéza je splněna právě tehdy, když $\mathsf{E}\, X_i = \mathsf{E}\, Y_i$ a výsledný test tedy testuje rovnost středních hodnot X_i a Y_i.

U jiných charakteristik toto neplatí: testujeme-li nulovost mediánu Z_i, neznamená to bez dalších předpokladů, že se za platnosti této hypotézy rovnají mediány X_i a Y_i. Testování rozptylu Z_i jednovýběrovým testem pak neříká vůbec nic o tom, jak a v čem se liší rozdělení X_i od rozdělení Y_i.

Párové testy lze použít pouze na intervalové a poměrové veličiny, jinak by rozdíly hodnot neměly smysluplnou interpretaci. Typicky je používáme na uspořádané dvojice měření téže veličiny na dvou přirozeně spárovaných jednotkách (např. levé oko – pravé oko, manžel – manželka) nebo dvě opakovaná měření téže veličiny na téže jednotce (např. před zásahem – po zásahu, loni – letos).

4.8 Přesný párový t-test

Párový t-test* se provádí jako jednovýběrový t-test na rozdíly Z_i. Předpokládá se normalita rozdílů Z_i, nikoli nutně normalita původních pozorování X_i a Y_i.

Model: $\mathcal{F} = \{Z_i = X_i - Y_i \sim \mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0\}$

Testované parametry: Střední hodnoty $\mu_X = \mathsf{E}\, X_i$ a $\mu_Y = \mathsf{E}\, Y_i$.

Hypotéza a alternativa:

$$H_0 : \mu_X - \mu_Y = d_0, \quad H_1 : \mu_X - \mu_Y \neq d_0,$$

kde d_0 je předem daná konstanta (obvykle $d_0 = 0$).

Testová statistika:

$$T_n = \sqrt{n}\, \frac{\overline{Z}_n - d_0}{S_n^{(Z)}},$$

kde $\overline{Z}_n$ je aritmetický průměr rozdílů Z_i (což je rovno $\overline{X}_n - \overline{Y}_n$) a $S_n^{(Z)}$ je výběrová směrodatná odchylka rozdílů Z_i.

* Angl. *paired t-test*

Rozdělení testové statistiky za H_0:

$$T_n \sim t_{n-1}$$

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow |T_n| \geq t_{n-1}(1-\alpha/2),$$

kde $t_{n-1}(1-\alpha/2)$ je $(1-\alpha/2)$-tý kvantil t-rozdělení s $n-1$ stupni volnosti.

P-hodnota: $p = 2(1 - F_n(|t|))$, kde t je pozorovaná hodnota testové statistiky a F_n je distribuční funkce rozdělení t_{n-1}.

Interval spolehlivosti pro $\mu_X - \mu_Y$: Vypracujte samostatně.

4.9 Asymptotický párový t-test

Jde o párový t-test provedený za slabších předpokladů konečného druhého momentu Z_i. Jeho vlastnosti jsou stejné, ale platí pouze asymptoticky.

Model: $\mathcal{F} = \{Z_i = X_i - Y_i \in \mathcal{L}^2\}$

Testované parametry: Střední hodnoty $\mu_X = \mathsf{E}\, X_i$ a $\mu_Y = \mathsf{E}\, Y_i$.

Hypotéza a alternativa:

$$H_0 : \mu_X - \mu_Y = d_0, \quad H_1 : \mu_X - \mu_Y \neq d_0,$$

kde d_0 je předem daná konstanta (obvykle $d_0 = 0$).

Testová statistika:

$$T_n = \sqrt{n}\,\frac{\overline{Z}_n - d_0}{S_n^{(Z)}},$$

kde $\overline{Z}_n$ je aritmetický průměr rozdílů Z_i (což je rovno $\overline{X}_n - \overline{Y}_n$) a $S_n^{(Z)}$ je výběrová směrodatná odchylka rozdílů Z_i.

Rozdělení testové statistiky za H_0:

$$T_n \overset{\text{as.}}{\sim} \mathsf{N}(0,1)$$

Asymptotické rozdělení však lze aproximovat i rozdělením t_{n-1}.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow |T_n| \geq t_{n-1}(1-\alpha/2),$$

kde $t_{n-1}(1-\alpha/2)$ je $(1-\alpha/2)$-tý kvantil t-rozdělení s $n-1$ stupni volnosti.

P-hodnota: $p = 2(1 - F_n(|t|))$, kde t je pozorovaná hodnota testové statistiky a F_n je distribuční funkce rozdělení t_{n-1}.

4.10 Párový znaménkový test

Párový znaménkový test* se provádí jako jednovýběrový znaménkový test na rozdíly Z_i. Předpokládá se spojitost rozdílů Z_i.

Model: $\mathcal{F} = \{Z_i$ má jakékoli spojité rozdělení$\}$

* Angl. *paired sign test*

Testovaný parametr: Medián m_Z rozdílu $Z_i = X_i - Y_i$.

Hypotéza a alternativa:

$$H_0 : m_Z = 0, \quad H_1 : m_Z \neq 0.$$

Poznámka.

1. Medián Z_i obecně nelze vyjádřit pomocí mediánů X_i a Y_i.
2. H_0 platí právě když $\mathsf{P}[X_i \leq Y_i] = \mathsf{P}[X_i \geq Y_i] = 1/2$, tj. X_i je s poloviční pravděpodobností větší než Y_i a s poloviční pravděpodobností menší než Y_i.
3. Má-li navíc Z_i konečnou střední hodnotu a hustotu symetrickou kolem 0, pak musí platit $\mathsf{E}\, Z_i = \mathsf{E}\, X_i - \mathsf{E}\, Y_i = 0$. Za těchto dodatečných předpokladů je H_0 ekvivalentní hypotéze o rovnosti středních hodnot X_i a Y_i.
4. Není to test shody mediánů X_i a Y_i.

Testová statistika:

$$Y_n = \sum_{i=1}^{n} \mathbb{1}_{(0,\infty)}(Z_i)$$

(počet párů, kde $X_i > Y_i$).

Přesné rozdělení testové statistiky za H_0:

$$Y_n \sim \mathsf{Bi}(n, 1/2)$$

Kritický obor (přesný test): Viz jednovýběrový znaménkový test.

Asymptotické rozdělení testové statistiky za H_0:

$$\frac{2}{\sqrt{n}}\Big(Y_n - \frac{n}{2}\Big) \overset{\text{as.}}{\sim} \mathsf{N}(0, 1)$$

Kritický obor (asymptotický test):

$$H_0 \text{ zamítneme } \Leftrightarrow \left|\frac{2}{\sqrt{n}} Y_n - \sqrt{n}\right| \geq u_{1-\alpha/2}.$$

Poznámka. Výhodou párového znaménkového testu je, že nevyžaduje vyčíslení rozdílu mezi X_i a Y_i. Stačí informace o tom, že X_i je „lepší“ než Y_i, resp. X_i je „horší“ než Y_i. Tento test je vhodný pro aplikace, v nichž může být určení konkrétních hodnot X_i a Y_i problematické.

Zde končí předn. 19 (8.12.)

4.11 Párový Wilcoxonův test

Párový Wilcoxonův test* porovnává střední hodnoty X_i a Y_i. Kvůli interpretaci hypotézy vyžaduje jak symetrii rozdělení Z_i tak konečnou střední hodnotu. Je to neparametrický test založený na pořadích.

Model: $\mathcal{F} = \{Z_i$ má spojité rozdělení s konečnou střední hodnotou a s hustotou f splňující $\exists \delta \in \mathbb{R} : f(\delta - x) = f(\delta + x) \quad \forall x \in \mathbb{R}\}$

* Angl. *paired Wilcoxon test, Wilcoxon signed rank test*

Poznámka. Předpoklad o symetrické hustotě se týká rozdílů Z_i, nikoli původních pozorování X_i a Y_i. Předpoklady symetrie a konečné střední hodnoty zajišťují, že $\delta_X \overset{\text{df}}{=} \mathsf{E}\, Z_i = \mathsf{E}\, X_i - \mathsf{E}\, Y_i$.

Testované parametry: Střední hodnoty $\mu_X = \mathsf{E}\, X_i$ a $\mu_Y = \mathsf{E}\, Y_i$.
Hypotéza a alternativa:

$$H_0 : \mu_X - \mu_Y = \delta_0, \quad H_1 : \mu_X - \mu_Y \neq \delta_0,$$

kde δ_0 je předem daná konstanta (obvykle $\delta_0 = 0$).
Testová statistika:

$$W_S = \sum_{i \in \mathcal{I}} R_i,$$

kde $\mathcal{I} \subset \{1, \ldots, n\}$ je množina všech indexů takových, že $Z_i^* \overset{\text{df}}{=} X_i - Y_i - \delta_0$ má kladné znaménko pro $i \in \mathcal{I}$, a $R_1 < R_2 < \cdots < R_n$ jsou pořadí náhodných veličin $\left|Z_i^*\right|$ mezi všemi $\left|Z_1^*\right|, \ldots, |Z_n^*|$.
Vlastnosti testové statistiky a kritický obor: viz jednovýběrový Wilcoxonův test.

Poznámka. K testování hypotézy H_0 je asymptotický párový t-test vhodnější než párový Wilcoxonův test, protože nevyžaduje symetrii hustoty.

5 Dvouvýběrové problémy pro kvantitativní data

Mějme dva *nezávislé* náhodné výběry: nechť $X_1, \ldots, X_n$ je náhodný výběr s distribuční funkcí F_X a $Y_1, \ldots, Y_m$ je náhodný výběr s distribuční funkcí F_Y. Model $\mathcal{F}$ specifikuje množinu uvažovaných distribučních funkcí F_X a F_Y. Máme daný parametr $\theta = t(F)$, jehož hodnotu chceme pro oba výběry porovnat. Označme si $\theta_X = t(F_X)$ a $\theta_Y = t(F_Y)$. Obvykle chceme testovat hypotézu $H_0 : \theta_X = \theta_Y$ proti alternativě $H_1 : \theta_X \neq \theta_Y$, případně sestrojit intervalový odhad pro rozdíl $\theta_X - \theta_Y$.

Dvouvýběrový problém lze zformulovat i jiným způsobem. Mějme náhodný výběr z dvourozměrného rozdělení

$$\begin{pmatrix} Z_1 \\ G_1 \end{pmatrix}, \ldots, \begin{pmatrix} Z_N \\ G_N \end{pmatrix},$$

kde Z_j jsou hodnoty nezávislých stejně rozdělených měření a G_j má alternativní rozdělení s parametrem $p_G \in (0, 1)$. Indikátor G_j určuje, do které z porovnávaných skupin j-té pozorování patří (jestliže $G_j = 0$, pak do první skupiny, jinak do druhé). Přeznačíme-li si měření Z_j na X_i anebo Y_i podle toho, do jaké skupiny dané pozorování patří

$$(X_1, \ldots, X_n) \overset{\mathrm{df}}{=} (Z_j : G_j = 0) \quad \text{a} \quad (Y_1, \ldots, Y_m) \overset{\mathrm{df}}{=} (Z_j : G_j = 1),$$

získáme dva nezávislé výběry podle první formulace problému. Chceme porovnat podmíněné rozdělení Z_j v obou skupinách, tj. zajímají nás podmíněné distribuční funkce $F_X(x) = \mathsf{P}\left[Z_j \leq x \,\middle|\, G_j = 0 \right]$ a $F_Y(x) = \mathsf{P}\left[Z_j \leq x \,\middle|\, G_j = 1 \right]$, případně jejich parametry $\theta_X = t(F_X)$ a $\theta_Y = t(F_Y)$. Tato druhá formulace dvouvýběrového problému je totožná s první, až na to, že rozsahy výběrů n a m nejsou konstanty, ale náhodné veličiny s binomickým rozdělením ($n = \sum_{j=1}^{N}(1 - G_j) \sim \mathsf{Bi}(N, 1 - p_G)$). Analýzu však provádíme stejně, jako by rozsahy výběrů byly pevné. To nám umožní předpoklad

$$\mathsf{P}\left[Z_j \leq x \,\middle|\, G_j, n, m \right] = \mathsf{P}\left[Z_j \leq x \,\middle|\, G_j \right]$$

to jest, že rozdělení měření v obou skupinách nemají nic společného s rozsahy výběrů. Pochopitelně, kdybychom rozsahy výběrů měnili podle toho, co už jsme napozorovali, analýza dat by to musela brát v úvahu a nemohla by postupovat takhle jednoduchým způsobem.

Data podle první formulace získáme tak, že si předem stanovíme, kolik měření z každé skupiny budeme mít, a pak napozorujeme požadovaný počet veličin pro každou skupinu zvlášť. Data podle druhé formulace vzniknou, pokud stanovíme celkový počet pozorování $N = n + m$, učiníme N pozorování a u každého pozorování teprve dodatečně určíme, do které skupiny patří.

Obě formulace se trochu liší v pojetí asymptotických výsledků. U druhé formulace stačí vzít $N \to \infty$. U první formulace potřebujeme $n \to \infty$ a $m \to \infty$, ale navíc ještě musíme

předpokládat, že rozsahy obou výběrů konvergují do nekonečna stejně rychle, tj. $n/m \to q$, kde $0 < q < \infty$.

Všechny metody uváděné v této kapitole se hodí pro obě formulace dvouvýběrového problému.

5.1 Dvouvýběrový Kolmogorovův-Smirnovův test

Dvouvýběrový Kolmogorovův-Smirnovův test* je rozšířením jednovýběrového testu stejného názvu. Je to neparametrický test, funguje pro jakákoli dvě spojitá rozdělení.

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

Testované parametry: celé distribuční funkce F_X a F_Y

Hypotéza a alternativa:

$$H_0 : F_X(x) = F_Y(x) \quad \forall x \in \mathbb{R}, \qquad H_1 : \exists x \in \mathbb{R} : F_X(x) \neq F_Y(x).$$

Testujeme, zdali oba výběry pocházejí z téhož rozdělení.

Testová statistika:

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

kde $\widehat{F}_X$ je empirická distribuční funkce náhodného výběru $X_1, \ldots, X_n$ a $\widehat{F}_Y$ je empirická distribuční funkce náhodného výběru $Y_1, \ldots, Y_m$.

Tvrzení 5.1 Nechť $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ jsou nezávislé náhodné výběry ze spojitého rozdělení s distribuční funkcí F_0. Pak platí

$$\mathsf{P}\Big[\sqrt{\frac{mn}{n+m}} K_{n,m} \leq y\Big] \to G(y) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} \mathrm{e}^{-2k^2y^2} \text{ pro } m, n \to \infty.$$

Poznámka.

- Hypotézu zamítneme, pokud se empirické distribuční funkce obou výběrů od sebe příliš liší, tj. pokud je testová statistika velká.
- Tvrzení 5.1 implikuje, že za platnosti hypotézy konverguje $\sqrt{\frac{mn}{n+m}} K_{n,m}$ v distribuci k náhodné veličině s distribuční funkcí $G(y)$, která je stejná jako u jednovýběrového Kolmogorovova-Smirnovova testu (viz tvrzení 4.2). To nám umožní určit kritickou hodnotu pro zamítání H_0.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow \sqrt{\frac{mn}{n+m}} K_{n,m} \geq k_{1-\alpha},$$

kde $k_{1-\alpha} = G^{-1}(1-\alpha)$ je $(1-\alpha)$-kvantil rozdělení s distribuční funkcí G.

Podle tvrzení 5.1 má tento test asymptotickou hladinu α.

Poznámka.

* Angl. *two-sample Kolmogorov-Smirnov test*

- Je možné spočítat i přesnou kritickou hodnotu dvouvýběrového Kolmogorovova-Smirnovova testu pro spojitá rozdělení s malými rozsahy výběru n, m.
- Výhodou tohoto testu je jeho universalita (reaguje na jakýkoli rozdíl v rozděleních obou skupin) a absence omezujících předpokladů. Nevýhodou tohoto testu je, že má malou sílu proti specifickým druhům porušení H_0. Zajímá-li nás pouze určitý typ porušení H_0 (třeba rozdíl ve střední hodnotě), je lepší použít test, který je zaměřen přímo na určitý parametr.

5.2 Přesný dvouvýběrový t-test

Dvouvýběrový t-test* porovnává střední hodnoty obou výběrů za předpokladu, že data mají normální rozdělení a rozptyly jsou v obou výběrech stejné. Test pak zachovává požadovanou hladinu přesně pro jakékoli n, $m \geq 3$.

Model:

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

Oba výběry mají normální rozdělení s totožným rozptylem, mohou se lišit pouze střední hodnotou.

Testované parametry: Střední hodnoty $\mu_X = \mathsf{E}\, X_i$ a $\mu_Y = \mathsf{E}\, Y_i$

Hypotéza a alternativa:

$$H_0 : \mu_X = \mu_Y + \delta_0, \qquad H_1 : \mu_X \neq \mu_Y + \delta_0.$$

Testujeme, zdali se střední hodnoty obou výběrů liší o δ_0 (obvykle se klade $\delta_0 = 0$).

Testová statistika:

$$T_{n,m} = \sqrt{\frac{nm}{n+m}}\, \frac{\overline{X}_n - \overline{Y}_m - \delta_0}{S_{n,m}},$$

kde $\overline{X}_n$ a $\overline{Y}_m$ jsou aritmetické průměry obou výběrů a

$$S^2_{n,m} \overset{\mathsf{df}}{=} \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2\right] = \frac{n-1}{n+m-2}S^2_X + \frac{m-1}{n+m-2}S^2_Y$$

je nestranný odhad společného rozptylu σ^2 spočítaný z obou výběrů (vážený průměr obou výběrových rozptylů).

Věta 5.2 Nechť $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ jsou nezávislé náhodné výběry z normálních rozdělení se středními hodnotami μ_X a μ_Y a se shodným rozptylem σ^2. Pak

$$\sqrt{\frac{nm}{n+m}}\, \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{S_{n,m}} \sim t_{n+m-2}.$$

Poznámka.

- Z věty 5.2 plyne, že za platnosti modelu $\mathcal{F}$ a hypotézy $H_0 : \mu_X - \mu_Y = \delta_0$ má $T_{n,m}$ rozdělení t_{n+m-2}.

* Angl. *two-sample t-test*

- Hypotézu budeme zamítat, pokud se výběrové průměry obou skupin od sebe příliš liší, tj. pokud je testová statistika buď moc velká nebo moc malá.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow |T_{n,m}| \geq t_{n+m-2}(1-\alpha/2),$$

kde $t_{n+m-2}(1-\alpha/2)$ je $(1-\alpha/2)$-tý kvantil t-rozdělení s $n+m-2$ stupni volnosti.

P-hodnota: $p = 2(1-F(|t|))$, kde t je pozorovaná hodnota testové statistiky $T_{n,m}$ a F je distribuční funkce rozdělení t_{n+m-2}.

Interval spolehlivosti pro $\mu_X - \mu_Y$: Z věty 5.2 lze odvodit přesný interval spolehlivosti pro rozdíl středních hodnot obou výběrů. Dostaneme

$$P\left[\overline{X}_n - \overline{Y}_m - S_{n,m}\sqrt{\frac{1}{n}+\frac{1}{m}}\, t_{n+m-2}(1-\alpha/2) < \mu_X - \mu_Y < \overline{X}_n - \overline{Y}_m + S_{n,m}\sqrt{\frac{1}{n}+\frac{1}{m}}\, t_{n+m-2}(1-\alpha/2)\right] = 1-\alpha.$$

Poznámka. Tento test lze snadno upravit na jednostranný.

Zde končí předn. 20 (14.12.)

5.3 Asymptotický dvouvýběrový z-test

Nyní upravíme dvouvýběrový t-test tak, aby se obešel bez předpokladu normality i bez shodných rozptylů. Půjde o asymptotický test.

Model:

$$\mathcal{F} = \{F_X \in \mathcal{L}^2, F_Y \in \mathcal{L}^2\}$$

Testované parametry: Střední hodnoty $\mu_X = \mathsf{E}\, X_i$ a $\mu_Y = \mathsf{E}\, Y_i$

Hypotéza a alternativa:

$$H_0 : \mu_X = \mu_Y + \delta_0, \quad H_1 : \mu_X \neq \mu_Y + \delta_0.$$

Testujeme, zdali se střední hodnoty obou výběrů liší o δ_0 (obvykle se klade $\delta_0 = 0$).

Testová statistika:

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

kde $\overline{X}_n$, $\overline{Y}_m$ jsou aritmetické průměry obou výběrů a S_X^2, S_Y^2 jsou výběrové rozptyly.

Věta 5.3 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 rozptyly. Pak

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$$

Poznámka.

- Hypotézu budeme zamítat, pokud se výběrové průměry obou skupin od sebe příliš liší, tj. pokud je testová statistika buď moc velká nebo moc malá.
- Věta 5.3 implikuje, že za platnosti modelu $\mathcal{F}$ a hypotézy H_0 má $Z_{n,m}$ asymptoticky rozdělení $N(0,1)$.

Poznámka. Nechť oba výběry mají stejný rozsah, tj. $m = n$. Potom

$$\sqrt{S_X^2/n + S_Y^2/m} = \sqrt{\frac{2}{n}}\sqrt{S_X^2/2 + S_Y^2/2} = \sqrt{\frac{n+n}{n^2}}S_{n,m}.$$

V tomto případě tedy vždy platí $Z_{n,m} = T_{n,m}$, tj. testové statistiky dvouvýběrového t-testu a z-testu jsou totožné. Jelikož Věta 5.3 platí i bez předpokladu shodných rozptylů, při $n = m$ dostatečně velkém lze rozdělení $T_{n,m}$ za platnosti H_0 aproximovat rozdělením t_{n+m-2} bez ohledu na to, jsou-li rozptyly stejné nebo ne. *Dvouvýběrový t-test tedy funguje alespoň asymptoticky i tehdy, pokud jsou rozptyly v obou výběrech různé, ale počty pozorování jsou shodné (nebo aspoň velmi podobné).*

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow |Z_{n,m}| \geq u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ je $(1-\alpha/2)$-tý kvantil normovaného normálního rozdělení.

P-hodnota: $p = 2(1-\Phi(|z|))$, kde z je pozorovaná hodnota testové statistiky $Z_{n,m}$ a Φ je distribuční funkce rozdělení $\mathsf{N}(0,1)$.

Interval spolehlivosti pro $\mu_X - \mu_Y$: Z věty 5.3 lze odvodit asymptotický interval spolehlivosti pro rozdíl středních hodnot obou výběrů. Dostaneme

$$P\left[\overline{X}_n - \overline{Y}_m - \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}\,u_{1-\alpha/2} < \mu_X - \mu_Y < \overline{X}_n - \overline{Y}_m + \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}\,u_{1-\alpha/2}\right] \to 1-\alpha.$$

Poznámka. Existují i lepší aproximace kritických hodnot pro tento test založené na t-rozdělení s počtem stupňů volnosti, který závisí na počtu pozorování v obou skupinách a výběrových rozptylech. Takových aproximací je několik[*]. Jedna z variant této aproximace, tzv. Welchův test[†], je implementována v R jako standardní metoda testování rovnosti středních hodnot dvou výběrů (provádí jej funkce `t.test`). Welchův test lze chápat jako variantu dvouvýběrového z-testu s vylepšenými kritickými hodnotami i jako zobecnění dvouvýběrového t-testu na výběry s nestejnými rozptyly.

Poznámka. Někdy se doporučuje před použitím dvouvýběrového t-testu otestovat shodnost rozptylů obou výběrů, např. testem uvedeným v kap. 5.5 níže, nebo tzv. Leveneovým testem (neuvádíme). Pokud test zamítne rovnost rozptylů, použijeme Welchův test, jinak použijeme dvouvýběrový t-test. Od používání takového postupu spíše odrazujeme. Jedná se o tzv. dvoufázový test, kdy celkový výsledek testu závisí na třech různých vzájemně závislých testových statistikách. Není ničím zaručeno, že celková hladina takové testovací procedury je rovna požadované hodnotě α. Pokud si nejsme jisti shodností rozptylů nebo normalitou dat, provedeme raději rovnou Welchův test. Ani jeden z předpokladů dvouvýběrového t-testu pak není třeba nijak ověřovat.

[*] lze je nalézt např. v knize Anděl: *Statistické metody*, Matfyzpress, Praha, 1998, kap. 8.1. [†] Angl. *Welch test*

5.4 Dvouvýběrový Wilcoxonův test

Dvouvýběrový Wilcoxonův test* je neparametrický test založený na pořadích.

Model: $\mathcal{F} = \{X \sim F_X$ spojitá d.f., $Y \sim F_Y$, $\exists \delta \in \mathbb{R} : F_X(x) = F_Y(x - \delta)\ \forall x \in \mathbb{R}\}$
(tzv. model posunutí v poloze).

Testovaný parametr: Posunutí δ_X.

Hypotéza a alternativa:

$$H_0 : \delta_X = 0, \quad H_1 : \delta_X \neq 0.$$

Poznámka.

- Na rozdíl od jednovýběrového a párového Wilcoxonova testu nevyžadujeme symetrii hustoty.
- Pokud platí model $\mathcal{F}$ a hypotéza H_0, rozdělení X a Y jsou totožná. Potom platí $m_X = m_Y$ a $\mathsf{E}\, X = \mathsf{E}\, Y$ (existují-li střední hodnoty). To jest, za platnosti modelu $\mathcal{F}$ lze dvouvýběrový Wilcoxonův test chápat jako test rovnosti středních hodnot i mediánů. Všimněte si, že nejsou-li rozptyly X a Y totožné, model $\mathcal{F}$ nemůže platit.

Testová statistika:

$$W_{n,m} = \sum_{i=1}^{n} R_i,$$

kde $R_1, R_2, \ldots, R_n$ jsou pořadí náhodných veličin X_i ve spojeném náhodném výběru $X_1, \ldots, X_n, Y_1, \ldots, Y_m$.

Poznámka. Testová statistika $W_{n,m}$ může nabývat hodnot $n(n+1)/2, \ldots, mn + n(n+1)/2$. Spočítá se následujícím způsobem:

1. Vezmeme spojený výběr $(Z_1, \ldots, Z_{n+m}) \overset{\mathsf{df}}{=} (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$.
2. Seřadíme všechny Z_j nejmenší do největší; získáme uspořádaný výběr

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

3. Určíme pořadí R_i náhodné veličiny X_i mezi všemi $Z_{(1)}, \ldots, Z_{(n+m)}$. Platí $X_i = Z_{(R_i)}$.
4. Sečteme pořadí R_i pro $i = 1, \ldots, n$.

Tvrzení 5.4 Platí-li model $\mathcal{F}$ a hypotéza H_0, pak

(i)
$$\mathsf{E}\, W_{n,m} = \frac{n(m+n+1)}{2}, \quad \mathsf{var}\, W_{n,m} = \frac{mn(m+n+1)}{12}.$$

(ii) Pokud $n, m \to \infty$,
$$\frac{W_{n,m} - \mathsf{E}\, W_{n,m}}{\sqrt{\mathsf{var}\, W_{n,m}}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Poznámka.

- Hypotézu budeme zamítat pro příliš malé nebo příliš velké hodnoty $W_{n,m}$.
- Předchozí tvrzení dává návod k nalezení kritických hodnot pro zamítání hypotézy, které zaručují asymptotickou hladinu α.

* Angl. *two-sample Wilcoxon test, Wilcoxon rank-sum test*

- Nejsou-li n a m příliš velká, lze nalézt i přesné rozdělení testové statistiky $W_{n,m}$ (numericky nebo v tabulkách).

Kritický obor (asymptotický test):

$$H_0 \text{ zamítneme } \Leftrightarrow \frac{\left|W_{n,m} - \frac{n(m+n+1)}{2}\right|}{\sqrt{\frac{mn(m+n+1)}{12}}} \geq u_{1-\alpha/2}.$$

Mannova-Whitneyho formulace Wilcoxonova testu

Test ekvivalentní s Wilcoxonovým lze získat i následující úvahou. Uvažujme všechny dvojice (X_i, Y_j) pro $i = 1, \ldots, n$ a $j = 1, \ldots, m$ a spočtěme, kolik z nich splňuje podmínku $X_i < Y_j$:

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

Náhodná veličina $W^*_{n,m}$, tzv. *Mannova-Whitneyho statistika*, může nabývat hodnot $0, \ldots, nm$.

Následující tvrzení ukazuje, že mezi dvouvýběrovou Wilcoxonovou statistikou $W_{n,m}$ a Mannovou-Whitneyho statistikou $W^*_{n,m}$ je deterministický lineární vztah. Můžeme tedy snadno spočítat momenty $W^*_{n,m}$.

Zde končí předn. 21 (15.12.)

Tvrzení 5.5

(i) $W_{n,m} + W^*_{n,m} = mn + n(n+1)/2$.
(ii) Platí-li H_0, pak $\mathsf{E}\, W^*_{n,m} = nm/2$.
(iii) Platí-li H_0 a model $\mathcal{F}$, pak $\mathsf{var}\, W^*_{n,m} = \mathsf{var}\, W_{n,m} = mn(m+n+1)/12$.
(iv) Pokud $\min(n, m) \to \infty$ a platí H_0, pak $(mn)^{-1} W^*_{n,m} \xrightarrow{\mathsf{P}} 1/2$.

Rozeberme si důsledky tvrzení 5.5. Část (i) říká, že testy založené na dvouvýběrové Wilcoxonově statistice a Mannově-Whitneyho statistice jsou ekvivalentní. Části (ii) a (iii) uvádějí momenty Mannovy-Whitneyho statistiky za hypotézy. Všimněte si, že rozptyl máme spočítaný pouze za předpokladu, že $F_X = F_Y$. Část (iv) ukazuje, že $W^*_{n,m}/(nm)$ je konsistentním odhadem parametru $\theta_{XY} = \mathsf{P}[X < Y]$, kde X a Y jsou nezávislé náhodné veličiny s distribučními funkcemi F_X a F_Y. Pokud $F_X = F_Y$, lze snadno ukázat, že $\theta_{XY} = 1/2$. Parametr θ_{XY} však může nabývat hodnoty $1/2$ i pro dvě rozdělení, která nejsou totožná.

Zkusme nyní shrnout, co je známo o chování dvouvýběrového Wilcoxonova testu (respektive Mannova-Whitneyho testu) v případě, kdy nečiníme žádné předpoklady o rozděleních F_X a F_Y.

Model: $\mathcal{F}^* = \{X \sim F_X \text{ spojitá d.f.}, Y \sim F_Y \text{ spojitá d.f.}\}$

Testovaný parametr: $\theta_{XY} = \mathsf{P}[X < Y]$

Hypotéza a alternativa:

$$H_0^* : \theta_{XY} = \frac{1}{2}, \quad H_1^* : \theta_{XY} \neq \frac{1}{2}.$$

V tomto obecném modelu lze tedy Wilcoxonův test interpretovat jako test hypotézy H_0^* o parametru $\mathsf{P}[X < Y]$. Bohužel hypotéza H_0^* je těžko interpretovatelná. Nelze ji vykládat jako

rovnost středních hodnot nebo mediánů nebo jakýchkoli jiných charakteristik obou rozdělení. Obecně tedy dvouvýběrový Wilcoxonův test není testem rovnosti mediánů ani testem rovnosti středních hodnot.

Stane-li se například, že $\mathsf{E}\,X_i = \mathsf{E}\,Y_j$ a přitom $\mathsf{P}[X_i < Y_j] \neq 1/2$, Wilcoxonův test zamítne hypotézu rovnosti středních hodnot s pravděpodobností konvergující k jedné při $m, n \to \infty$. Naopak, pokud $\mathsf{E}\,X_i \neq \mathsf{E}\,Y_j$ a přitom $\mathsf{P}[X_i < Y_j] = 1/2$, Wilcoxonův test při jakkoli velkém rozsahu výběrů zamítá hypotézu pouze s pravděpodobností rovnou jeho hladině.

Další potíž s Wilcoxonovou testovou statistikou v obecném modelu je, že její rozptyl za hypotézy nelze počítat podle tvrzení 5.4. Její rozptyl je ve skutečnosti jiný.[*] Kritické hodnoty spočítané pro Wilcoxonův test v modelu $\mathcal{F}$ tedy v obecném modelu nefungují.

Tyto úvahy vedou k jednoznačnému závěru: *Chceme-li testovat rovnost středních hodnot bez dalších předpokladů na tvar rozdělení obou výběrů, použijeme dvouvýběrový z-test nebo Welchův test, nikoli Wilcoxonův test.*

Poznámka. Někdy se doporučuje před použitím dvouvýběrového t-testu na porovnání středních hodnot otestovat normalitu obou výběrů (populární je tzv. Shapiro-Wilkův test normality, který neuvádíme). Pokud test zamítne normalitu, použijeme Wilcoxonův test, jinak použijeme dvouvýběrový t-test. Od používání takového postupu zásadně odrazujeme. Jak víme, jedná se o dva testy, které testují rozdílné hypotézy, nemůžeme je tedy použít na ten samý problém. Pokud si nejsme jisti normalitou dat, provedeme raději rovnou Welchův test, který normalitu nevyžaduje a testuje právě tu hypotézu, která byla zadána.

5.5 Dvouvýběrový F test shody rozptylů

Dvouvýběrový F test shody rozptylů[†] je přesný test porovnávající rozptyly dvou nezávislých výběrů za předpokladu normálního rozdělení.

Model: $\mathcal{F} = \{X_i \sim \mathsf{N}(\mu_X, \sigma_X^2), Y_i \sim \mathsf{N}(\mu_Y, \sigma_Y^2), \mu_X, \mu_Y \in \mathbb{R}, \sigma_X^2 > 0, \sigma_Y^2 > 0\}$

Testované parametry: Rozptyly $\sigma_X^2 = \mathrm{var}\, X_i$ a $\sigma_Y^2 = \mathrm{var}\, Y_j$.

Hypotéza a alternativa:

$$H_0 : \sigma_X^2 = \sigma_Y^2, \quad H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

Testová statistika:

$$F_{n,m} = \frac{S_X^2}{S_Y^2},$$

kde S_X^2 je výběrový rozptyl výběru $X_1, \ldots, X_n$ a S_Y^2 je výběrový rozptyl výběru $Y_1, \ldots, Y_n$.

Poznámka.

- Z věty 1.11 plyne, že testová statistika má za platnosti modelu a hypotézy přesně rozdělení $F_{n-1,m-1}$.
- Hypotézu zamítneme, pokud se výběrové rozptyly příliš liší, tj. pokud je testová statistika buď moc velká nebo moc malá.
- Při porušení předpokladu normality tento test nedodržuje hladinu ani asymptoticky.

[*] Lze jej spočítat a odhadnout, ale tím se tu zabývat nebudeme. [†] Angl. *two-sample F test of equality of variances*

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow F_{n,m} \leq F_{n-1,m-1}(\alpha/2) \text{ nebo } F_{n,m} \geq F_{n-1,m-1}(1-\alpha/2),$$

kde $F_{n-1,m-1}(\alpha/2)$ a $F_{n-1,m-1}(1-\alpha/2)$ jsou po řadě $(\alpha/2)$-tý a $(1-\alpha/2)$-tý kvantil F rozdělení s $n-1$ a $m-1$ stupni volnosti.

P-hodnota: $p = 2\min(1-F(s), F(s))$, kde s je pozorovaná hodnota testové statistiky a F je distribuční funkce rozdělení $F_{n-1,m-1}$.

Interval spolehlivosti pro σ_X^2/σ_Y^2: Z věty 1.11 lze odvodit interval spolehlivosti pro podíl rozptylů. Dostaneme

$$\mathsf{P}\left[\frac{S_X^2}{S_Y^2}\,\frac{1}{F_{n-1,m-1}(1-\frac{\alpha}{2})} < \sigma_X^2/\sigma_Y^2 < \frac{S_X^2}{S_Y^2}\,\frac{1}{F_{n-1,m-1}(\frac{\alpha}{2})}\right] = 1-\alpha.$$

Poznámka. Tento test lze převést na jednostranný test: Hypotéza $H_0' : \sigma_X^2 \leq \sigma_Y^2$ se zamítá pouze pro příliš velké hodnoty testové statistiky, kritická hodnota je $F_{n-1,m-1}(1-\alpha)$. Hypotéza $H_0'' : \sigma_X^2 \geq \sigma_Y^2$, se zamítá pouze pro příliš malé hodnoty testové statistiky, kritická hodnota je $F_{n-1,m-1}(\alpha)$.

6 Jednovýběrové problémy pro binární a kategoriální data

V této kapitole a v kapitole následující se budeme zabývat *kategoriálními veličinami*. Pojem kategoriální veličina byl vyložen v kapitole 2.2.2. Stručně řečeno, jde o diskrétní veličinu nabývající konečně mnoha hodnot, typicky $1, \ldots, K$, jejíž hodnoty nemusí mít numerickou interpretaci, ale označují členství v nějaké skupině (kategorii). Parametry používané v analýze kategoriálních dat jsou typicky pravděpodobnosti jednotlivých hodnot.

6.1 Alternativní a binomické rozdělení

Alternativní rozdělení je nejjednodušším modelem pro kategoriální veličinu, která nabývá pouze dvou hodnot zakódovaných jako 0 a 1. Nechť $p_X \in (0, 1)$ je pravděpodobnost, že daný jedinec je klasifikován do kategorie 1.

Nechť $Y_1, \ldots, Y_n$ je náhodný výběr z alternativního rozdělení $\mathsf{Alt}(p_X)$ zaznamenávající klasifikaci n jedinců do kategorií 0 a 1. Označme počet jedinců klasifikovaných do skupiny 1 jako $X_n = \sum_{i=1}^n Y_i$. Tato veličina má rozdělení $\mathsf{Bi}(n, p_X)$ (viz věta 1.3(iv)). Počet jedinců klasifikovaných do skupiny 0 je $n - X_n \sim \mathsf{Bi}(n, 1 - p_X)$.

Nestranným a konsistentním odhadem parametru p_X je relativní četnost $\widehat{p}_n = X_n/n = n^{-1} \sum_{i=1}^n Y_i = \overline{Y}_n$. Jeho vlastnosti vycházejí z vlastností průměru a jsou shrnuty ve větě 1.3.

6.1.1 Clopperova-Pearsonova metoda

Nejprve se budeme zabývat metodami pro sestrojení intervalu spolehlivosti pro pravděpodobnost p_X a pro testování hypotéz o p_X založenými na přesném rozdělení statistiky X_n, tj. $\mathsf{Bi}(n, p_X)$.

Uvažujme hypotézu $H_0 : p_x = p_0$ proti alternativě $H_0 : p_x \neq p_0$. Stanovme kritický obor

$$H_0 \text{ zamítneme } \Leftrightarrow X_n \leq c_L(\alpha) \text{ nebo } X_n \geq c_U(\alpha),$$

kde $c_L(\alpha)$ je největší celé číslo, které splňuje

$$\sum_{j=0}^{c_L(\alpha)} \binom{n}{j} p_0^j (1 - p_0)^{n-j} \leq \frac{\alpha}{2}$$

a $c_U(\alpha)$ je nejmenší celé číslo, které splňuje

$$\sum_{j=c_U(\alpha)}^{n} \binom{n}{j} p_0^j (1 - p_0)^{n-j} \leq \frac{\alpha}{2}.$$

Tento test (zvaný *Clopperův-Pearsonův*) má nejvyšší možnou dosažitelnou hladinu, jež nepřesahuje α (vzhledem k diskrétnímu rozdělení testové statistiky nelze vždy dosáhnout stanovené hladiny α).

Nyní řešme úlohu sestavení intervalu spolehlivosti pro p_x s pravděpodobností pokrytí nepřekračující $1-\alpha$. Podle tvrzení 3.3(ii) (dualita intervalů spolehlivosti a testování), můžeme sestavit požadovaný interval spolehlivosti jako množinu obsahující všechny parametry $p \in (0,1)$, pro něž při pozorovaných datech X_n Clopperův-Pearsonův test nezamítá hypotézu $H_0 : p_X = p$. Tento interval je možné vyjádřit v explicitním tvaru pomocí následujících dvou tvrzení.

Tvrzení 6.1 Nechť $X_n \sim \mathrm{Bi}(n, p_X)$. Pak pro každé $k \in \{0, \dots, n\}$ platí

$$\mathsf{P}[X_n \geq k] = \sum_{j=k}^{n} \binom{n}{j} p_X^j (1-p_X)^{n-j} = \mathsf{P}[U_1 \leq p_X], \quad \text{kde } U_1 \sim B(k, n-k+1)$$

a

$$\mathsf{P}[X_n \leq k] = \sum_{j=0}^{k} \binom{n}{j} p_X^j (1-p_X)^{n-j} = \mathsf{P}[U_2 \leq 1-p_X], \quad \text{kde } U_2 \sim B(n-k, k+1).$$

Platnost první části tohoto tvrzení plyne z poslední rovnosti uvedené ve znění věty 1.13 o chování pořádkových statistik. Tam byla tato rovnost dokázána, i když v úplně jiném kontextu. Aplikujeme-li první část tvrzení 6.1 na $n - X_n$, dostaneme druhou část.

Další tvrzení popisuje funkční vztah mezi beta a F rozdělením. Lze jej dokázat pomocí věty o transformaci.

Tvrzení 6.2 Nechť $X \sim B(\alpha, \beta)$, kde α a β jsou nezáporná celá čísla. Pak $\frac{\beta X}{\alpha(1-X)} \sim F_{2\alpha,2\beta}$.

Povšimněte si, že jde o ryze monotonní transformaci z $(0,1)$ do $\mathbb{R}^+$. Pomocí těchto dvou tvrzení dostaneme následující výsledek.

Věta 6.3 Množina všech parametrů $p \in (0,1)$ takových, že při pozorovaných datech X_n Clopperův-Pearsonův test nezamítá hypotézu $H_0 : p_X = p$ (a tedy interval spolehlivosti pro p_X s pravděpodobností pokrytí nepřekračující $1-\alpha$) je interval

$$\left(\frac{X_n q_L(\alpha)}{X_n q_L(\alpha) + n - X_n + 1}, \frac{(X_n+1) q_U(\alpha)}{(X_n+1) q_U(\alpha) + n - X_n} \right),$$

kde $q_L(\alpha)$ je $\alpha/2$-kvantil rozdělení $F_{2X_n, 2(n-X_n+1)}$ a $q_U(\alpha)$ je $(1-\alpha/2)$-kvantil $F_{2(X_n+1), 2(n-X_n)}$. Pokud $X_n = 0$, položíme dolní mez intervalu rovnou 0, pokud $X_n = n$, položíme horní mez intervalu rovnou 1.

Tento interval se nazývá *Clopperův-Pearsonův interval spolehlivosti* pro parametr binomického rozdělení. Výhodou tohoto intervalu je, že pravděpodobnost pokrytí zaručeně nepřevýší požadovanou hodnotu $1-\alpha$ ani při malém rozsahu výběru. Jeho nevýhodou je, že jeho pravděpodobnost pokrytí může být o hodně vyšší než $1-\alpha$ a že mívá příliš velkou délku.

Nyní se můžeme vrátit ke Clopperově-Pearsonově testu hypotézy $H_0 : p_x = p_0$ proti alternativě $H_0 : p_x \neq p_0$. Místo toho, abychom složitě počítali kritické hodnoty $c_L(\alpha)$ a $c_U(\alpha)$, spočítáme Clopperův-Pearsonův interval spolehlivosti podle věty 6.3 a H_0 zamítneme, pokud p_0 v tomto intervalu neleží.

6.1.2 Klasická asymptotická metoda

V příkladě uvedeném v kapitole 2.3.2 na str. 27 jsme odvodili asymptotický interval spolehlivosti pro p_X založený na bodě (iii) věty 1.3 a Sluckého větě. Podle (2.3) platí

$$Z_n = \sqrt{n}\,\frac{\widehat{p}_n - p_X}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Toto tvrzení lze použít k odvození asymptotického testu hypotézy $H_0 : p_X = p_0$ proti alternativě $H_1 : p_X \neq p_0$ s kritickým oborem

$$H_0 \text{ zamítneme } \Leftrightarrow \sqrt{n}\frac{|\widehat{p}_n - p_0|}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} \geq u_{1-\alpha/2}. \tag{6.1}$$

Interval spolehlivosti pro p_X z kapitoly 2.3.2 má tvar

$$\left(\widehat{p}_n - \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}},\ \widehat{p}_n + \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right). \tag{6.2}$$

Nevýhodou tohoto přístupu je, že vyžaduje relativně velký počet pozorování (doporučuje se alespoň 5 úspěchů a alespoň 5 neúspěchů), konvergence k normalitě je pomalá a krajní body intervalu spolehlivosti mohou být menší než 0 nebo větší než 1.

6.1.3 Wilsonova metoda

Wilsonova metoda je založena přímo na bodě (iii) věty 1.3

$$W_n = \sqrt{n}\,\frac{\widehat{p}_n - p_X}{\sqrt{p_X(1-p_X)}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$$

bez aplikace Sluckého věty. Za platnosti hypotézy $H_0 : p_X = p_0$ známe p_X a toho využijeme k sestavení kritického oboru

$$H_0 \text{ zamítneme } \Leftrightarrow \sqrt{n}\frac{|\widehat{p}_n - p_0|}{\sqrt{p_0(1-p_0)}} \geq u_{1-\alpha/2}.$$

Tento test se nazývá *Wilsonův*.

Zde končí předn. 22 (21.12)

Interval spolehlivosti pro p_X založíme na pivotální statistice W_n, tj. vyjdeme z

$$\mathsf{P}\left[u_{1-\alpha/2} \leq \sqrt{n}\,\frac{\widehat{p}_n - p_X}{\sqrt{p_X(1-p_X)}} \leq u_{1-\alpha/2}\right] \to 1-\alpha$$

a nerovnosti uvnitř upravíme tak, abychom uprostřed dostali p_X a na okrajích meze intervalu spolehlivosti. K tomu je nutné vyřešit kvadratickou rovnici pro p_X. Výsledkem je asymptotický interval s krajními body

$$\left[\widehat{p}_n + \frac{u^2}{2n} \mp u\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n} + \frac{u^2}{4n^2}}\right]\frac{1}{1+u^2/n}$$

kde u je zkrácené značení pro $u_{1-\alpha/2}$. Tento interval se též nazývá *Wilsonův*. V literatuře se uvádí, že Wilsonův test a interval dává přesnější výsledky než metody z kapitoly 6.1.2.

Je zajímavé si povšimnout, že střed Wilsonova intervalu lze vyjádřit jako vážený průměr $w_n\widehat{p}_n + (1-w_n)1/2$, kde $w_n = (1+u^2/n)^{-1} \to 1$ pro $n \to \infty$. Počítáme-li 95% interval spolehlivosti, pak střed Wilsonova intervalu je zhruba $(X_n+2)/(n+4)$.

6.1.4 Logitová metoda

Logitová metoda je založena na šanci místo na pravděpodobnosti.

Definice 6.1 Nechť úspěch nastává s pravděpodobností p. Podíl $\frac{p}{1-p}$ pravděpodobnosti úspěchu a neúspěchu se nazývá *šance*[*] na úspěch.

Pojem šance se běžně používá při kursových sázkách.

Zvolme jako odhadovaný parametr logaritmus šance $\theta_X = \log \frac{p_X}{1-p_X}$. Tomuto parametru se běžně říká *logit*, transformace $g(x) = \log \frac{x}{1-x}$ se nazývá *logitová*. Logitová transformace $g(x)$ je rostoucí a spojitě diferencovatelná pro $x \in (0, 1)$ a zobrazuje interval $(0, 1)$ na $\mathbb{R}$. Inversní transformace je $g^{-1}(y) = \frac{\exp\{y\}}{1+\exp\{y\}}$. Logaritmus šance θ_X tedy může nabývat libovolné hodnoty v $\mathbb{R}$ a můžeme z ní vyjádřit pravděpodobnost p_X jako $p_X = \exp\{\theta_X\}/(1 + \exp\{\theta_X\})$.

Logaritmus šance θ_X odhadneme transformací $g(\widehat{p}_n)$ odhadu $\widehat{p}_n$. Dostaneme odhad

$$\widehat{\theta}_n = \log \frac{\widehat{p}_n}{1 - \widehat{p}_n},$$

který je podle tvrzení P.7.3 konsistentním (ne však nestranným) odhadem θ_X.

Asymptotické rozdělení $\widehat{\theta}_n$ získáme aplikací bodu (iii) věty 1.3 a delta metody (věta P.7.11).

Věta 6.4 Nechť $p_X \in (0, 1)$. Pak platí

(i)
$$\sqrt{n}(\widehat{\theta}_n - \theta_X) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \frac{1}{p_X} + \frac{1}{1 - p_X}),$$

(ii)
$$\sqrt{\frac{X_n(n - X_n)}{n}}(\widehat{\theta}_n - \theta_X) \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1).$$

Označme $D_n = \sqrt{\frac{n}{X_n(n-X_n)}}$. Je to vlastně odhad směrodatné chyby $\widehat{\theta}_n$.

Na základě věty 6.4 můžeme sestrojit asymptotický test hypotézy $H_0 : p_X = p_0$. Označme $\theta_0 = \log \frac{p_0}{1-p_0}$. Hypotézu H_0 můžeme přepsat jako $H_0 : \theta_X = \theta_0$ a zamítáme ji ve prospěch alternativy $H_1 : \theta_X \neq \theta_0$ pokud

$$\frac{1}{D_n} \left|\widehat{\theta}_n - \theta_0\right| \geq u_{1-\alpha/2}.$$

Tento test nazveme *logitový*.

Interval spolehlivosti pro θ_X s pravděpodobností pokrytí konvergující k $1 - \alpha$ má tvar

$$\left(\widehat{\theta}_n - u_{1-\frac{\alpha}{2}} D_n,\ \widehat{\theta}_n + u_{1-\frac{\alpha}{2}} D_n\right).$$

Aplikujeme-li ryze rostoucí funkci g^{-1} na oba krajní body tohoto intervalu, dostaneme asymptotický $100(1 - \alpha)$-procentní interval spolehlivosti pro p_X ve tvaru

$$\left(\frac{\frac{\widehat{p}_n}{1-\widehat{p}_n} e^{-u_{1-\alpha/2} D_n}}{1 + \frac{\widehat{p}_n}{1-\widehat{p}_n} e^{-u_{1-\alpha/2} D_n}},\ \frac{\frac{\widehat{p}_n}{1-\widehat{p}_n} e^{u_{1-\alpha/2} D_n}}{1 + \frac{\widehat{p}_n}{1-\widehat{p}_n} e^{u_{1-\alpha/2} D_n}} \right). \tag{6.3}$$

Interval (6.3) nazýváme *logitový*. Oba jeho krajní body jistě leží uvnitř $(0, 1)$. Navíc konvergence $\widehat{\theta}_n$ k normálnímu rozdělení je rychlejší než konvergence $\widehat{p}_n$, takže limitní aproximace založená na $\widehat{\theta}_n$ je přesnější než aproximace založená na $\widehat{p}_n$. Logitová metoda patří spolu s Wilsonovou k metodám doporučovaným v literatuře.

[*] Angl. *odds*

6.2 Multinomické rozdělení

Multinomické rozdělení zobecňuje binomické rozdělení na situaci, kdy kategoriální veličina může nabývat více než dvou hodnot.

Multinomické rozdělení: definice a vlastnosti

Definice 6.2 (Multinomické rozdělení) Nechť $K \geq 2$ a $n \geq 1$ jsou přirozená čísla a $\boldsymbol{p} = (p_1, \ldots, p_K)^\mathsf{T}$ je vektor konstant splňující $p_k > 0\ \forall k$ a $\sum_{k=1}^K p_k = 1$. Náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_K)^\mathsf{T}$ má multinomické rozdělení $\mathsf{Mult}_K(n, \boldsymbol{p})$, právě když jeho hustota vzhledem k součinové čítací míře na $\mathbb{Z}^K$ je

$$\mathsf{P}[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K] = \begin{cases} \dfrac{n!}{x_1! \cdots x_K!} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K} & \sum_{k=1}^K x_k = n \\ & x_k \geq 0\ \forall k \\ 0 & \text{jinak.} \end{cases}$$

Multinomické rozdělení je rozdělení počtu pozorování přidělených do každé z K možných přihrádek v n nezávislých experimentech, přičemž v každém experimentu jsou pravděpodobnosti přiřazení do jednotlivých přihrádek dány složkami vektoru pravděpodobností $\boldsymbol{p}$.

Věta 6.5 (Rozklad multinomického rozdělení.) Nechť $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ jsou nezávislé náhodné vektory s rozdělením $\mathsf{Mult}_K(1, \boldsymbol{p})$. Pak $\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{Y}_i \sim \mathsf{Mult}_K(n, \boldsymbol{p})$.

Věta 6.6 (Vlastnosti multinomického rozdělení.) Nechť $\boldsymbol{X} \sim \mathsf{Mult}_K(n, \boldsymbol{p})$. Pak

(i) $X_k \sim \mathsf{Bi}(n, p_k)$,
(ii) $\mathsf{E}\, X_k = np_k$, $\mathsf{var}\, X_k = np_k(1 - p_k)$,
(iii) $\mathsf{cov}\,(X_j, X_k) = -np_jp_k$,
(iv) $\mathsf{var}\, \boldsymbol{X} = n\,[\mathsf{diag}\,(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}] = n\,\mathsf{diag}\,(\sqrt{\boldsymbol{p}})(I_K - \sqrt{\boldsymbol{p}}^{\otimes 2})\mathsf{diag}\,(\sqrt{\boldsymbol{p}})$, kde $\sqrt{\boldsymbol{p}} = (\sqrt{p_1}, \ldots, \sqrt{p_K})^\mathsf{T}$.

Zde končí předn. 23 (22.12)

Poznámka.

- Matice $I_K - \sqrt{\boldsymbol{p}}^{\otimes 2}$ je idempotentní. Její hodnost (a stopa) je rovna $K - 1$.
- Matice $\mathsf{var}\, \boldsymbol{X}$ je singulární, její hodnost je $K - 1$.

Věta 6.7 (Asymptotické vlastnosti multinomického rozdělení.)
Nechť $\boldsymbol{X} \sim \mathsf{Mult}_K(n, \boldsymbol{p})$. Pak

(i)
$$\boldsymbol{Z}_n \overset{\mathsf{df}}{=} \frac{1}{\sqrt{n}} \mathsf{diag}\,(\sqrt{\boldsymbol{p}})^{-1}(\boldsymbol{X} - n\boldsymbol{p}) \xrightarrow{\mathsf{D}} \mathsf{N}_K(\boldsymbol{0}, I_K - \sqrt{\boldsymbol{p}}^{\otimes 2}),$$

(ii)
$$\boldsymbol{Z}_n^\mathsf{T} \boldsymbol{Z}_n = \sum_{k=1}^K \frac{(X_k - np_k)^2}{np_k} \xrightarrow{\mathsf{D}} \chi^2_{K-1}$$

Odhady parametrů multinomického rozdělení

Pro odhadování jednotlivých parametrů p_k, testování hypotéz o p_k a konstrukci intervalových odhadů pro p_k můžeme použít metody popsané v kapitole 6.1, neboť podle věty 6.6(i) platí $X_k \sim \mathrm{Bi}(n, p_k)$,

Celý vektor $\boldsymbol{p}$ odhadneme pomocí $\widehat{\boldsymbol{p}}_n = \boldsymbol{X}/n$. Sdružené asymptotické rozdělení odhadu $\widehat{\boldsymbol{p}}_n$ získáme z věty 6.7(i):

$$\sqrt{n}(\widehat{\boldsymbol{p}}_n - \boldsymbol{p}) \xrightarrow{\mathsf{D}} \mathsf{N}_K(\boldsymbol{0}, \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}).$$

Pro libovolný vektor konstant $\boldsymbol{c}$ o délce K, platí

$$\sqrt{n}(\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - \boldsymbol{c}^\mathsf{T}\boldsymbol{p}) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \boldsymbol{c}^\mathsf{T}[\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}]\boldsymbol{c}).$$

Pokud $\boldsymbol{c}^\mathsf{T}[\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}]\boldsymbol{c} \neq 0$ a $V_c \overset{\mathsf{df}}{=} \boldsymbol{c}^\mathsf{T}[\operatorname{diag}(\widehat{\boldsymbol{p}}_n) - \widehat{\boldsymbol{p}}_n^{\otimes 2}]\boldsymbol{c} \neq 0$, dostaneme ze Sluckého věty

$$\sqrt{n}\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - \boldsymbol{c}^\mathsf{T}\boldsymbol{p}}{\sqrt{V_c}} \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1). \tag{6.4}$$

Odtud můžeme snadno odvodit asymptotické testy hypotéz $H_0 : \boldsymbol{c}^\mathsf{T}\boldsymbol{p} = \gamma_0$. Vezmeme testovou statistiku

$$T_c = \sqrt{n}\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - \gamma_0}{\sqrt{V_c}},$$

která má podle (6.4) za platnosti hypotézy asymptoticky normované normální rozdělení a H_0 zamítneme právě když $|T_c| \geq u_{1-\alpha/2}$.

Asymptotický interval spolehlivosti pro $\boldsymbol{c}^\mathsf{T}\boldsymbol{p}$ založený na konvergenci (6.4) jest

$$\left(\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - \sqrt{\frac{V_c}{n}}u_{1-\alpha/2},\ \boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n + \sqrt{\frac{V_c}{n}}u_{1-\alpha/2}\right).$$

Vektor $\boldsymbol{c}$ vybereme tak, aby součin $\boldsymbol{c}^\mathsf{T}\boldsymbol{p}$ vytvořil lineární kombinaci parametrů, která nás v dané aplikaci zajímá. Chceme-li například vědět, zdali pravděpodobnosti první a poslední kategorie jsou stejné, a sestrojit interval spolehlivosti pro rozdíl jejich hodnot, zvolíme $\boldsymbol{c} = (1, 0, \ldots, 0, -1)^\mathsf{T}$ a $\gamma_0 = 0$.

χ^2 test dobré shody pro multinomické rozdělení

Pojmem χ^2 *test dobré shody*[*] rozumíme test hypotézy $H_0 : \boldsymbol{p} = \boldsymbol{p}^0$ založený na větě 6.7(ii). Tato hypotéza říká, že pravděpodobnosti kategorií $\boldsymbol{p} = (p_1, \ldots, p_K)^\mathsf{T}$ jsou rovny předem stanoveným hypotetickým pravděpodobnostem $\boldsymbol{p}^0 = (p_1^0, \ldots, p_K^0)^\mathsf{T}$, tj. $p_k = p_k^0$ pro všechna $k = 1, \ldots, K$.

Platí-li hypotéza H_0, pak testová statistika

$$\chi^2 = \sum_{k=1}^{K}\frac{(X_k - np_k^0)^2}{np_k^0}$$

[*] Angl. χ^2 *test of goodness of fit*

má podle věty 6.7(ii) asymptotické rozdělení χ^2_{K-1}. Testová statistika porovnává pozorovanou četnost X_k v kategorii k s četností np_k^0 očekávanou za platnosti hypotézy. Velké hodnoty testové statistiky svědčí proti H_0. Hypotézu H_0 zamítneme, pokud

$$\chi^2 = \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{np_k^0} \geq \chi^2_{K-1}(1-\alpha), \tag{6.5}$$

kde $\chi^2_{K-1}(1-\alpha)$ značí $(1-\alpha)$-kvantil rozdělení χ^2_{K-1}.

Poznámka. Asymptotická aproximace χ^2 rozdělením vyžaduje, aby celkový počet pozorování n byl dostatečně velký. Jako jednoduché orientační pravidlo můžeme vzít např. požadavek, aby očekávané četnosti np_k^0 překročily 5 ve všech kategoriích $k = 1, \ldots, K$. Vyskytují-li se v hodnotách $\boldsymbol{X}$ velmi malé četnosti nebo nuly, χ^2 aproximace může být velmi nepřesná.

Poznámka. Vezmeme-li $K = 2$, $p_1^0 \equiv p_0$, $X_2 = n - X_1$, $p_2^0 = 1 - p_0$, dostaneme

$$\chi^2 = \frac{(X_1 - np_0)^2}{np_0} + \frac{[n - X_1 - n(1-p_0)]^2}{n(1-p_0)} = \left[\sqrt{n}\frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)}}\right]^2,$$

takže testová statistika χ^2 testu pro $K = 2$ kategorie je rovna čtverci Wilsonovy testové statistiky uvedené v kapitole 6.1.3.

Příklad (Je kostka pravidelná?)**.** Hodíme n-krát kostkou a zaznamenáme, kolikrát padly výsledky 1–6: dostaneme četnosti $X_1, \ldots, X_6$. Nastavíme $p_k^0 = 1/6$, $k = 1, \ldots, 6$. Zamítne-li χ^2 test hypotézu H_0, prokázali jsme, že na kostce nepadají všechna čísla stejně často.

Příklad (Rodí se děti během roku rovnoměrně?)**.** Máme dány počty dětí narozených v jednotlivých měsících během kalendářního roku: $X_1, \ldots, X_{12}$. Nastavíme $p_k^0 = m_k/365$, kde m_k je počet dní v měsíci k. Zamítne-li χ^2 test hypotézu H_0, prokázali jsme, že děti se nerodí během roku rovnoměrně.

Příklad (Pochází náhodný výběr z distribuční funkce F_0?)**.** Mějme náhodný výběr $Z_1, \ldots, Z_n$. Zajímá nás, zdali pochází z rozdělení s distribuční funkcí $F_0(x) = F(x; \boldsymbol{\theta}_0)$, kde $\boldsymbol{\theta}_0$ je známo.

Stanovíme si intervaly $(a_{k-1}, a_k\rangle$, $k = 1, \ldots, K$, $a_0 = -\infty$, $a_K = \infty$ tak, že jejich počet K je výrazně menší než n a do každého z intervalů padne dostatečný počet pozorování. Spočítáme, kolik pozorování padlo do k-tého intervalu: $X_k = \sum_{i=1}^n \mathbb{1}_{(a_{k-1}, a_k\rangle}(Z_i)$. Pochází-li náhodný výběr $Z_1, \ldots, Z_n$ z rozdělení s distribuční funkcí $F_0(x) = F(x; \boldsymbol{\theta}_0)$, potom vektor $\boldsymbol{X} = (X_1, \ldots, X_K)^\mathsf{T}$ má multinomické rozdělení $\mathrm{Mult}_K(n, \boldsymbol{p}^0)$, kde pravděpodobnosti jednotlivých kategorií jsou $p_k^0 = F(a_k; \boldsymbol{\theta}_0) - F(a_{k-1}; \boldsymbol{\theta}_0)$.

Provedeme test hypotézy $H_0 : \boldsymbol{p} = \boldsymbol{p}^0$ testem dobré shody podle vzorce (6.5). Zamítne-li test hypotézu H_0, prokázali jsme, že náhodný výběr $Z_1, \ldots, Z_n$ nepochází z rozdělení $F(x; \boldsymbol{\theta}_0)$.

χ^2 test dobré shody pro multinomické rozdělení s odhadnutými parametry

Jak jsme viděli v předchozím příkladě, pravděpodobnosti kategorií p_k^0 mohou záviset na vektoru parametrů $\boldsymbol{\theta}_0$. Test dobré shody můžeme provést podle vzorce (6.5) jen tehdy, pokud tyto parametry známe. V praxi je ovšem někdy neznáme, můžeme je nanejvýš odhadnout. Nyní si ukážeme, jak upravit test dobré shody pro takové případy.

Uvažujme model $\mathcal{F}_0$: Nechť náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_K)^\mathsf{T}$ má multinomické rozdělení $\mathsf{Mult}_K(n, \boldsymbol{p}(\boldsymbol{\theta}_X))$, kde $\boldsymbol{\theta}_X \in \Theta \subset \mathbb{R}^d$ je neznámý d-rozměrný parametr, $d < K$, a $\boldsymbol{p}$ je funkce zobrazující Θ do $(0,1)^K$ taková, že $\boldsymbol{p}(\boldsymbol{\theta})^\mathsf{T}\mathbf{1}_K = 1$ pro všechna $\boldsymbol{\theta} \in \Theta$ (součet všech složek $\boldsymbol{p}(\boldsymbol{\theta})$ je vždy 1). Zajímá nás, zdali rozdělení $\boldsymbol{X}$ lze popsat tímto modelem nebo ne.

Příklad. V nějaké populaci se určitý gen vyskytuje ve dvou variantách (alelách) A (např. tmavé oči) a a (např. světlé oči). Mezi všemi geny v celé populaci tvoří alela A podíl $\theta_X \in (0,1)$ a alela a $1-\theta_X$. Každý jedinec má dva exempláře příslušného genu (jeden po otci, jeden po matce). Pokud se geny míchají nezávisle (platí tzv. Hardyho-Weinbergovo ekvilibrium), pravděpodobnosti tří možných variant genotypu jedince jsou:

Genotyp	Pravděpodobnost
AA	θ_X^2
Aa	$2\theta_X(1-\theta_X)$
aa	$(1-\theta_X)^2$

Pozorujeme genotypy n nezávislých jedinců a označíme X_1, X_2, X_3 počty jedinců s genotypem (po řadě) AA, Aa, aa. Platí-li Hardyho-Weinbergovo ekvilibrium, pak vektor $\boldsymbol{X} = (X_1, X_2, X_3)^\mathsf{T}$ má rozdělení $\mathsf{Mult}_3(n, \boldsymbol{p}(\theta_X))$, kde $\boldsymbol{p}(\theta_X) = (\theta_X^2, 2\theta_X(1-\theta_X), (1-\theta_X)^2)^\mathsf{T}$. Na základě pozorování $\boldsymbol{X}$ chceme otestovat, zdali se populace nachází v Hardyho-Weinbergově ekvilibriu.

Zde končí předn. 24 (4.1.)

Parametry $\boldsymbol{\theta}_X$ potřebujeme odhadnout. K tomu lze použít např. metodu maximální věrohodnosti, která vede k soustavě d rovnic o d neznámých $\widehat{\boldsymbol{\theta}}_n$:

$$\sum_{k=1}^{K} \frac{X_k}{p_k(\widehat{\boldsymbol{\theta}}_n)} \frac{\partial p_k(\widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{6.6}$$

Uvažujme testování hypotézy

$$H_0 : \exists \boldsymbol{\theta}_X \in \Theta \quad \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta}_X) \qquad \text{(model } \mathcal{F}_0 \text{ platí)}$$

proti alternativě

$$H_1 : \forall \boldsymbol{\theta}_X \in \Theta \quad \boldsymbol{p} \neq \boldsymbol{p}(\boldsymbol{\theta}_X) \qquad \text{(model } \mathcal{F}_0 \text{ neplatí)}.$$

Nejprve získáme odhad $\widehat{\boldsymbol{\theta}}_n$ parametru $\boldsymbol{\theta}_X$ vyřešením soustavy (6.6). Poté můžeme otestovat hypotézu H_0 testem dobré shody s odhadnutými parametry namísto parametrů skutečných. Rozdělení testové statistiky je stále χ^2, ale ztrácí se jeden stupeň volnosti za každý odhadovaný parametr.

Tvrzení 6.8 Platí-li hypotéza H_0, pak testová statistika

$$\chi^2 = \sum_{k=1}^{K} \frac{[X_k - n p_k(\widehat{\boldsymbol{\theta}}_n)]^2}{n p_k(\widehat{\boldsymbol{\theta}}_n)}$$

má asymptoticky rozdělení χ^2_{K-d-1}, kde d je počet odhadovaných parametrů.

Platnost tohoto tvrzení plyne z teorie maximální věrohodnosti, která bude vysvětlena v navazující přednášce. Testová statistika porovnává pozorovanou četnost X_k v kategorii k s četností $np_k(\widehat{\boldsymbol{\theta}}_n)$ očekávanou za platnosti hypotézy; velké hodnoty testové statistiky svědčí proti H_0. Hypotézu H_0 zamítneme, pokud

$$\chi^2 = \sum_{k=1}^{K} \frac{[X_k - np_k(\widehat{\boldsymbol{\theta}}_n)]^2}{np_k(\widehat{\boldsymbol{\theta}}_n)} \geq \chi^2_{K-d-1}(1-\alpha), \tag{6.7}$$

kde $\chi^2_{K-d-1}(1-\alpha)$ značí $(1-\alpha)$-kvantil rozdělení χ^2_{K-d-1}.

Poznámka. I zde je nutné mít dostatečně velký počet pozorování v každé složce vektoru $\boldsymbol{X}$.

Příklad (Pochází náhodný výběr z dané parametrické rodiny rozdělení?)**.** Mějme náhodný výběr $Z_1, \ldots, Z_n$. Zajímá nás, zdali pochází z rozdělení $F_X(x) = F(x; \boldsymbol{\theta}_X)$, kde $\boldsymbol{\theta}_X \in \Theta$ není známo (např. nějaké normální, gama nebo Poissonovo rozdělení).

Stanovíme si intervaly $(a_{k-1}, a_k\rangle$, $k = 1, \ldots, K$, $a_0 = -\infty$, $a_K = \infty$ tak, že jejich počet K je výrazně menší než n a do každého z intervalů padne dostatečný počet pozorování. Spočítáme, kolik pozorování padlo do k-tého intervalu: $X_k = \sum_{i=1}^{n} \mathbb{1}_{(a_{k-1}, a_k\rangle}(Z_i)$.

Pochází-li náhodný výběr $Z_1, \ldots, Z_n$ z rozdělení s distribuční funkcí $F(x; \boldsymbol{\theta}_X)$, potom vektor $\boldsymbol{X} = (X_1, \ldots, X_K)^\top$ má multinomické rozdělení $\mathsf{Mult}_K(n, \boldsymbol{p}(\boldsymbol{\theta}_X))$, kde pravděpodobnosti jednotlivých kategorií jsou $p_k(\boldsymbol{\theta}_X) = F(a_k; \boldsymbol{\theta}_X) - F(a_{k-1}; \boldsymbol{\theta}_X)$.

Řešením soustavy (6.6) získáme odhad $\widehat{\boldsymbol{\theta}}_n$ parametru $\boldsymbol{\theta}_X$. Provedeme test hypotézy H_0 testem dobré shody podle vzorce (6.7). Zamítne-li test hypotézu, prokázali jsme, že náhodný výběr $Z_1, \ldots, Z_n$ nepochází z dané rodiny rozdělení.

7 Dvouvýběrové kategoriální problémy a kontingenční tabulky

7.1 Dvouvýběrové kategoriální problémy

Nyní se budeme zabývat porovnáním dvou nezávislých binomických veličin $X_1 \sim \mathsf{Bi}(n, p_1)$ a $X_2 \sim \mathsf{Bi}(m, p_2)$. Chceme zjistit, zdali a jakým způsobem se liší pravděpodobnosti p_1 a p_2. Jejich odlišnost můžeme vyjádřit různými způsoby, z toho nám vyplyne několik variant odhadů a testů.

Pokud veličiny X_1 a X_2 udávají počty nějakých negativních událostí (smrt, nemoc, ztráta zaměstnání, porucha, bankrot) parametry p_1 a p_2 nazýváme *riziky* události v obou populacích. Pravděpodobnosti (rizika) p_1 a p_2 můžeme odhadnout relativními četnostmi $\widehat{p}_1 = X_1/n$, $\widehat{p}_2 = X_2/m$. Jejich vlastnosti shrnuje věta 1.3.

U všech asymptotických výsledků budeme stejně jako v kapitole 5 předpokládat $n \to \infty$, $m \to \infty$ a $n/m \to q$, kde $0 < q < \infty$. Výsledky uváděné v této kapitole však platí i tehdy, je-li pevný pouze celkový počet pororování $n + m$, zatímco rozsahy výběrů n a m jsou náhodné (viz diskuse na str. 62).

7.1.1 Rozdíly pravděpodobností, nárůst rizika

Odlišnost obou rozdělení můžeme vyjádřit např. *rozdílem pravděpodobností (rizik)*[*] $d_X = p_1 - p_2$, jež říká, o kolik je větší riziko v populaci 1 než v populaci 2. Tento parametr může nabývat hodnot -1 až 1, nulová hodnota odpovídá totožným pravděpodobnostem v obou populacích.

Nestranným a konsistentním odhadem parametru d_X je $\widehat{d} = \widehat{p}_1 - \widehat{p}_2$. Z věty 1.3(iii) a z nezávislosti X_1 a X_2 dostaneme technikou velmi podobnou důkazu věty 5.3 tento výsledek:

Tvrzení 7.1

$$\frac{\widehat{d} - d_X}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}} \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1).$$

Pro asymptotický test hypotézy $H_0 : d_X = 0$ proti alternativě $H_1 : d_X \neq 0$ použijeme testovou statistiku

$$T_d = \frac{\widehat{d}}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}}$$

a hypotézu zamítneme pokud $|T_d| \geq u_{1-\alpha/2}$.

[*] Angl. *risk difference, excess risk*

Z tvrzení 7.1 dostaneme postupnými úpravami

$$P\Big[\widehat{d} - \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}\, u_{1-\alpha/2} < d_X < \widehat{d} + \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}\, u_{1-\alpha/2}\Big] \to 1-\alpha.$$

Odtud získáme asymptotický interval spolehlivosti pro rozdíl pravděpodobností d_X.

Tato metoda je obdobou asymptotického dvouvýběrového z-testu pro kvantitativní data.

7.1.2 Podíly pravděpodobností, relativní riziko

Jiný způsob, jak vyjádřit odlišnost pravděpodobností (rizik), je *relativní riziko*[*] $r_X = p_1/p_2$. Tento parametr říká, kolikrát je větší riziko v populaci 1 než v populaci 2 a může nabývat hodnot v intervalu $(0, \infty)$. Pravděpodobnosti (rizika) v obou populacích jsou totožné právě když $r_X = 1$.

Konsistentním (nikoli nestranným) odhadem parametru r_X je $\widehat{r} = \widehat{p}_1/\widehat{p}_2$. Zlogaritmováním dostaneme $\log \widehat{r} = \log \widehat{p}_1 - \log \widehat{p}_2$. Věta 1.3(iii) a delta metoda dává

$$\sqrt{n}(\log \widehat{p}_1 - \log p_1) \xrightarrow{\mathsf{D}} \mathsf{N}\Big(0, \frac{1-p_1}{p_1}\Big)$$

a

$$\sqrt{m}(\log \widehat{p}_2 - \log p_2) \xrightarrow{\mathsf{D}} \mathsf{N}\Big(0, \frac{1-p_2}{p_2}\Big).$$

Odtud a z nezávislosti X_1 a X_2 dostaneme toto tvrzení:

Tvrzení 7.2

$$\frac{\log \widehat{r} - \log r_X}{\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Chceme otestovat, jestli $\log r_X = 0$ neboli $r_X = 1$. Pro asymptotický test hypotézy $H_0 : r_X = 1$ proti alternativě $H_1 : r_X \neq 1$ použijeme testovou statistiku

$$T_r = \frac{\log \widehat{r}}{\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}}$$

a hypotézu zamítneme pokud $|T_r| \geq u_{1-\alpha/2}$.

Z tvrzení 7.2 dostaneme postupnými úpravami

$$P\Big[\widehat{r}\exp\Big\{-\sqrt{\tfrac{1-\widehat{p}_1}{n\widehat{p}_1} + \tfrac{1-\widehat{p}_2}{m\widehat{p}_2}}\, u_{1-\alpha/2}\Big\} < r_X < \widehat{r}\exp\Big\{\sqrt{\tfrac{1-\widehat{p}_1}{n\widehat{p}_1} + \tfrac{1-\widehat{p}_2}{m\widehat{p}_2}}\, u_{1-\alpha/2}\Big\}\Big] \to 1-\alpha,$$

což nám dává asymptotický interval spolehlivosti pro relativní riziko r_X.

[*] Angl. *relative risk*

7.1.3 Poměr šancí

Třetím možným způsobem vyjádření odlišnosti dvou pravděpodobností je *poměr šancí*[*]

$$o_X = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

Tento parametr říká, kolikrát je větší šance v populaci 1 než v populaci 2. Může nabývat hodnot v intervalu $(0, \infty)$. Pravděpodobnosti (rizika) v obou populacích jsou totožná právě když $o_X = 1$.

Konsistentním (nikoli nestranným) odhadem parametru o_X je

$$\widehat{o} = \frac{\widehat{p}_1(1-\widehat{p}_2)}{\widehat{p}_2(1-\widehat{p}_1)} = \frac{X_1(m-X_2)}{X_2(n-X_1)}.$$

Zlogaritmováním dostaneme $\log \widehat{o} = \log \widehat{p}_1 - \log(1-\widehat{p}_1) - \log \widehat{p}_2 + \log(1-\widehat{p}_2)$. Z věty 6.4(i) a z nezávislosti X_1 a X_2 plyne následující tvrzení:

Tvrzení 7.3 Nechť

$$\widehat{V}_o = \frac{1}{n\widehat{p}_1} + \frac{1}{n(1-\widehat{p}_1)} + \frac{1}{m\widehat{p}_2} + \frac{1}{m(1-\widehat{p}_2)} = \\ = \frac{1}{X_1} + \frac{1}{n-X_1} + \frac{1}{X_2} + \frac{1}{m-X_2}.$$

Pak

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

Pravděpodobnosti (šance) v obou populacích jsou totožné právě když $o_X = 1$ neboli $\log o_X = 0$. Pro asymptotický test hypotézy $H_0 : o_X = 1$ proti alternativě $H_1 : o_X \neq 1$ použijeme testovou statistiku

$$T_o = \frac{\log \widehat{o}}{\sqrt{\widehat{V}_o}}$$

a hypotézu zamítneme pokud $|T_o| \geq u_{1-\alpha/2}$.

Asymptotický interval spolehlivosti pro poměr šancí o_X je dán faktem

$$P\left[\widehat{o}\exp\left\{-\sqrt{\widehat{V}_o}\,u_{1-\alpha/2}\right\} < o_X < \widehat{o}\exp\left\{\sqrt{\widehat{V}_o}\,u_{1-\alpha/2}\right\}\right] \to 1-\alpha,$$

který plyne z tvrzení 7.3.

7.2 Kontingenční tabulky

Nechť $X \in \{1, \ldots, J\}$ a $Z \in \{1, \ldots, K\}$ jsou dvě kategoriální veličiny. Uvažujme náhodný výběr $(X_1, Z_1)^\mathsf{T}, \ldots, (X_N, Z_N)^\mathsf{T}$ o rozsahu N (pevném). Označme počet jedinců klasifikovaných do j-té kategorie veličiny X a k-té kategorie veličiny Z jako $n_{jk} = \sum_{i=1}^{N} \mathbb{1}\{X_i = j, Z_i = k\}$,

[*] Angl. *odds ratio*

$j = 1, \ldots, J$, $k = 1, \ldots, K$. Náhodnou veličinu n_{jk} nazýváme *pozorovanou četností*[*] pro kombinaci kategorií j a k. Označme $p_{jk} = \mathsf{P}[X = j, Z = k]$ a $\boldsymbol{p} = (p_{11}, \ldots, p_{JK})^\mathsf{T}$. Vzhledem k tomu, že pozorované četnosti byly vyvořeny klasifikací N nezávislých jedinců do JK kategorií, náhodný vektor $\boldsymbol{n} = (n_{11}, \ldots, n_{JK})^\mathsf{T}$ musí mít multinomické rozdělení $\mathsf{Mult}_{JK}(N, \boldsymbol{p})$. Protože pracujeme s multinomickým rozdělením, můžeme používat všechny výsledky z kapitoly 6.2. Odhadem pravděpodobnosti p_{ij} je n_{ij}/N. Odhadem vektoru $\boldsymbol{p}$ je $\widehat{\boldsymbol{p}}_n = \boldsymbol{n}/N$.

Označme dále

$$n_{j+} = \sum_{k=1}^{K} n_{jk}, \quad n_{+k} = \sum_{j=1}^{J} n_{jk}, \quad n_{++} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = N,$$
$$p_{j+} = \sum_{k=1}^{K} p_{jk}, \quad p_{+k} = \sum_{j=1}^{J} p_{jk}, \quad p_{++} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{jk} = 1.$$

Pravděpodobnosti p_{jk} určují sdružené rozdělení X a Z, pravděpodobnosti $p_{j+} = \mathsf{P}[X = j]$ určují marginální rozdělení X, pravděpodobnosti $p_{+k} = \mathsf{P}[Z = k]$ určují marginální rozdělení Z.

Pozorované četnosti můžeme sestavit do tabulky, kterou nazýváme *kontingenční tabulka*[†].

	$Z = 1$	…	$Z = K$	Σ
$X = 1$	n_{11}	…	n_{1K}	n_{1+}
$X = 2$	n_{21}	…	n_{2K}	n_{2+}
…	…	…	…	…
$X = J$	n_{J1}	…	n_{JK}	n_{J+}
Σ	n_{+1}	…	n_{+K}	N

Podobně můžeme sestavit tabulku pravděpodobností, která popisuje sdružené rozdělení vektoru $(X, Z)^\mathsf{T}$ i marginální rozdělení veličin X a Z.

	$Z = 1$	…	$Z = K$	Σ
$X = 1$	p_{11}	…	p_{1K}	p_{1+}
$X = 2$	p_{21}	…	p_{2K}	p_{2+}
…	…	…	…	…
$X = J$	p_{J1}	…	p_{JK}	p_{J+}
Σ	p_{+1}	…	p_{+K}	1

Označme ještě podmíněné pravděpodobnosti

$$\mathsf{P}\left[X = j \mid Z = k\right] = p_{j(k)} = \frac{p_{jk}}{p_{+k}},$$
$$\mathsf{P}\left[Z = k \mid X = j\right] = p_{(j)k} = \frac{p_{jk}}{p_{j+}}.$$

7.2.1 Kontingenční tabulky 2×2

Nejprve se budeme zabývat speciálním případem $J = 2$ a $K = 2$, kdy obě veličiny mohou nabývat pouze dvou hodnot. Výsledná kontingenční tabulka obsahuje 2×2 četnosti:

[*] Angl. *observed frequency* [†] Angl. *contingency table*

	$Z=1$	$Z=2$	Σ
$X=1$	n_{11}	n_{12}	n_{1+}
$X=2$	n_{21}	n_{22}	n_{2+}
Σ	n_{+1}	n_{+2}	N

	$Z=1$	$Z=2$	Σ
$X=1$	p_{11}	p_{12}	p_{1+}
$X=2$	p_{21}	p_{22}	p_{2+}
Σ	p_{+1}	p_{+2}	1

Tuto situaci jsme vlastně řešili v kapitole 7.1. Představme si, že veličina Z určuje číslo výběru: máme jeden výběr hodnot náhodné veličiny X z jedinců splňujících $Z=1$ a druhý výběr náhodné veličiny X z jedinců splňujících $Z=2$. V prvním výběru bylo n_{11} hodnot $X=1$ (úspěch) a n_{21} hodnot $X=2$ (neúspěch), celkem n_{+1} pozorování. Pravděpodobnost úspěchu v 1. výběru je $p_{1(1)}=p_{11}/p_{+1}$. V druhém výběru bylo n_{12} hodnot $X=1$ (úspěch) a n_{22} hodnot $X=2$ (neúspěch), celkem n_{+2} pozorování. Pravděpodobnost úspěchu v 2. výběru je $p_{1(2)}=p_{12}/p_{+2}$.

Značení zavedené v kapitole 7.1 můžeme snadno převést na značení používané nyní a naopak. Naše kontingenční tabulka přepsaná do značení z kapitoly 7.1 vypadá takto:

	$Z=1$	$Z=2$	Σ
$X=1$	X_1	X_2	X_1+X_2
$X=2$	$n-X_1$	$m-X_2$	$n+m-X_1-X_2$
Σ	n	m	$n+m$

Rozdíl proti situaci v kapitole 7.1 spočívá v tom, že tam byly oba výběry nezávislé, zatímco nyní uvažujeme jeden výběr z multinomického rozdělení se čtyřmi možnými hodnotami. Tehdy byly rozsahy obou výběrů n, m pevné, nyní jsou to binomické náhodné veličiny a pouze celkový počet pozorování $N=n+m$ je pevný. Znovu jsme narazili na dvě různé formulace dvouvýběrového problému, podobně jako v kapitole 5 o dvouvýběrových testech pro nominální data. Stejně jako tam, i tady je jedno, kterou formulaci používáme a jakým způsobem byla kontingenční tabulka vytvořena. Všechny studované metody platí pro obě dvě formulace.

Kapitola 7.1 vysvětluje, jak porovnat riziko události $[X=1]$ pro různé hodnoty Z. Můžeme použít tři způsoby porovnání:

- *rozdíl pravděpodobností* $d_X=p_{1(1)}-p_{1(2)}$ odhadneme pomocí $\widehat{d}=\frac{n_{11}}{n_{+1}}-\frac{n_{12}}{n_{+2}}$;
- *podíl pravděpodobností* $r_X=p_{1(1)}/p_{1(2)}$ odhadneme pomocí $\widehat{r}=\frac{n_{11}n_{+2}}{n_{12}n_{+1}}$;
- *poměr šancí* $o_X=\frac{p_{1(1)}(1-p_{1(2)})}{p_{1(2)}(1-p_{1(1)})}$ odhadneme pomocí $\widehat{o}=\frac{n_{11}n_{22}}{n_{12}n_{21}}$ (proto se poměru šancí někdy říká *křížový poměr*[*]).

Metody pro testování těchto parametrů a konstrukci intervalů spolehlivosti jsou uvedeny v kapitole 7.1.

Náhodné veličiny X a Z jsou nezávislé, právě když pro každé $j,k\in\{1,2\}$ platí

$$\mathsf{P}[X=j,Z=k]=\mathsf{P}[X=j]\,\mathsf{P}[Z=k] \quad \text{neboli} \quad p_{jk}=p_{j+}p_{+k}$$

anebo ekvivalentně

$$\mathsf{P}\left[X=j\mid Z=k\right]=\mathsf{P}[X=j] \quad \text{neboli} \quad p_{j(k)}=p_{j+}.$$

[*] Angl. *cross ratio*

Jelikož $p_{2(k)} = 1 - p_{1(k)}$, nezávislost platí právě když $p_{1(1)} = p_{1(2)}$, což je ekvivalentní kterémukoli ze vztahů

$$d_X = 0, \quad r_X = 1, \quad o_x = 1.$$

Test na nulovost rozdílu rizik nebo jednotkovost relativního rizika či poměru šancí je v této situaci zároveň testem nezávislosti X a Z.

Testování nezávislosti χ^2 testem

Jiný způsob, jak otestovat nezávislost X a Z poskytuje χ^2 test dobré shody pro multinomické rozdělení s odhadnutými parametry založený na tvrzení 6.8. Pokud platí hypotéza, že X a Z jsou nezávislé, pravděpodobnosti $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})^\mathsf{T}$ specifikující multinomické rozdělení vektoru $\boldsymbol{n}$ jsou vlastně funkcemi pouze dvou parametrů p_{1+} a p_{+1}. Máme tedy $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta}_X)$, kde $\boldsymbol{\theta}_X = (p_{1+}, p_{+1})^\mathsf{T}$. Maximálně věrohodný odhad parametru $\boldsymbol{\theta}_X$ za hypotézy nezávislosti je $\widehat{\boldsymbol{\theta}}_n = (\widehat{p}_{1+}, \widehat{p}_{+1})^\mathsf{T} = (n_{1+}/N, n_{+1}/N)^\mathsf{T}$, což jsou empirické relativní četnosti jevů $[X = 1]$ a $[Z = 1]$. Maximálně věrohodný odhad vektoru $\boldsymbol{p}$ za hypotézy nezávislosti jest

$$\begin{aligned}
p_{11}(\widehat{\boldsymbol{\theta}}_n) &= \widehat{p}_{1+}\widehat{p}_{+1} = \frac{n_{1+}n_{+1}}{N^2} \\
p_{12}(\widehat{\boldsymbol{\theta}}_n) &= \widehat{p}_{1+}(1 - \widehat{p}_{+1}) = \widehat{p}_{1+}\widehat{p}_{+2} = \frac{n_{1+}n_{+2}}{N^2} \\
p_{21}(\widehat{\boldsymbol{\theta}}_n) &= (1 - \widehat{p}_{1+})\widehat{p}_{+1} = \widehat{p}_{2+}\widehat{p}_{+1} = \frac{n_{2+}n_{+1}}{N^2} \\
p_{22}(\widehat{\boldsymbol{\theta}}_n) &= (1 - \widehat{p}_{1+})(1 - \widehat{p}_{+1}) = \widehat{p}_{2+}\widehat{p}_{+2} = \frac{n_{2+}n_{+2}}{N^2}
\end{aligned}$$

Očekávané četnosti v kontingenční tabulce za platnosti hypotézy jsou $N p_{jk}(\widehat{\boldsymbol{\theta}}_n) = N\widehat{p}_{j+}\widehat{p}_{+k} = n_{j+}n_{+k}/N$. Počet odhadovaných parametrů je $d = 2$.

Testová statistika je

$$\chi^2 = \sum_{j=1}^{2}\sum_{k=1}^{2} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}.$$

Za platnosti hypotézy nezávislosti má asymptoticky rozdělení χ^2_{4-d-1}, kde $d = 2$, tj. χ^2_1. Hypotézu nezávislosti zamítneme, pokud $\chi^2 \geq \chi^2_1(1-\alpha)$.

Zde končí předn. 25 (5.1.)

7.2.2 Kontingenční tabulky $2 \times K$

Nyní rozšíříme zkoumanou situaci na případ $J = 2$ a $K \geq 2$. Kontingenční tabulka obsahuje $2 \times K$ četností:

	$Z=1$	$Z=2$	$\ldots$	$Z=K$	Σ
$X=1$	n_{11}	n_{12}	$\ldots$	n_{1K}	n_{1+}
$X=2$	n_{21}	n_{22}	$\ldots$	n_{2K}	n_{2+}
Σ	n_{+1}	n_{+2}	$\ldots$	n_{+K}	N

	$Z=1$	$Z=2$	$\ldots$	$Z=K$	Σ
$X=1$	p_{11}	p_{12}	$\ldots$	p_{1K}	p_{1+}
$X=2$	p_{21}	p_{22}	$\ldots$	p_{2K}	p_{2+}
Σ	p_{+1}	p_{+2}	$\ldots$	p_{+K}	N

Toto je zobecnění situace řešené v kapitole 7.1. Můžeme si ji představit i tak, že máme (po sloupcích) K výběrů z binomického rozdělení s potenciálně různými pravděpodobnostmi úspěchu p_{1k}/p_{+k} nebo máme (po řádcích) dva výběry z multinomického rozdělení s potenciálně různými vektory pravděpodobností

$$(p_{11}/p_{1+}, p_{12}/p_{1+}, \ldots, p_{1K}/p_{1+})^\mathsf{T} \quad \text{a} \quad (p_{21}/p_{2+}, p_{22}/p_{2+}, \ldots, p_{2K}/p_{2+})^\mathsf{T}.$$

Testování nezávislosti χ^2 testem

X a Z jsou nezávislé, právě když $p_{1(1)} = p_{1(2)} = \ldots = p_{1(K)}$. To vyžaduje, aby pro kterékoli dvě skupiny $Z = k_1$ a $Z = k_2$ byl rozdíl rizik 0 nebo relativní riziko či poměr šancí 1. Zatímco zobecnit testování pomocí rozdílů rizik, jednotkovosti relativního rizika či poměrů šancí na tento případ by vyžadovalo další práci, χ^2 test nezávislosti lze zobecnit snadno.

Pokud platí hypotéza, že X a Z jsou nezávislé náhodné veličiny, pravděpodobnosti $\boldsymbol{p} = (p_{11}, p_{21}, \ldots, p_{1K}, p_{2K})^\mathsf{T}$ specifikující multinomické rozdělení vektoru $\boldsymbol{n}$ jsou funkcemi p_{1+} a $p_{+1}, \ldots, p_{+(K-1)}$, celkem K parametrů. Máme tedy $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta}_X)$, kde $\boldsymbol{\theta}_X = (p_{1+}, p_{+1}, \ldots, p_{+(K-1)})^\mathsf{T}$. Maximálně věrohodný odhad parametru $\boldsymbol{\theta}_X$ za hypotézy nezávislosti je roven marginálním empirickým četnostem

$$\widehat{\boldsymbol{\theta}}_n = (\widehat{p}_{1+}, \widehat{p}_{+1}, \ldots, \widehat{p}_{+(K-1)})^\mathsf{T} = (n_{1+}/N, n_{+1}/N, \ldots, n_{+(K-1)}/N)^\mathsf{T}.$$

Maximálně věrohodné odhady složek vektoru $\boldsymbol{p}$ za hypotézy nezávislosti jsou

$$p_{jk}(\widehat{\boldsymbol{\theta}}_n) = \widehat{p}_{j+}\widehat{p}_{+k} = \frac{n_{j+}n_{+k}}{N^2},$$

$j = 1, 2$, $k = 1, \ldots, K$. Očekávané četnosti v kontingenční tabulce za platnosti hypotézy jsou $N p_{jk}(\widehat{\boldsymbol{\theta}}_n) = N\widehat{p}_{j+}\widehat{p}_{+k} = n_{j+}n_{+k}/N$.

Testová statistika je

$$\chi^2 = \sum_{j=1}^{2}\sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}.$$

Za platnosti hypotézy nezávislosti má asymptoticky rozdělení χ^2_{2K-K-1}, tj. χ^2_{K-1}. Hypotézu nezávislosti zamítneme, pokud $\chi^2 \geq \chi^2_{K-1}(1-\alpha)$.

Test nezávislosti zároveň testuje i hypotézu, že K výběrů z binomického rozdělení má stejné pravděpodobnosti úspěchu (jde tedy o K-výběrový test na binomické rozdělení) a hypotézu, že dva výběry z multinomického rozdělení mají stejné vektory pravděpodobností (jde tedy o dvouvýběrový test na multinomické rozdělení).

7.2.3 Kontingenční tabulky $J \times K$

Zobecnění na situaci $J \geq 2$ a $K \geq 2$ je nyní snadné. Kontingenční tabulka obsahuje $J \times K$ četností:

	$Z = 1$	$\ldots$	$Z = K$	Σ
$X = 1$	n_{11}	$\ldots$	n_{1K}	n_{1+}
$X = 2$	n_{21}	$\ldots$	n_{2K}	n_{2+}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$X = J$	n_{J1}	$\ldots$	n_{JK}	n_{J+}
Σ	n_{+1}	$\ldots$	n_{+K}	N

	$Z=1$	…	$Z=K$	Σ
$X=1$	p_{11}	…	p_{1K}	p_{1+}
$X=2$	p_{21}	…	p_{2K}	p_{2+}
…	…	…	…	…
$X=J$	p_{J1}	…	p_{JK}	p_{J+}
Σ	p_{+1}	…	p_{+K}	1

Můžeme si ji představit i tak, že máme (po sloupcích) K výběrů z multinomického rozdělení Mult_J s potenciálně různými vektory pravděpodobností nebo (po řádcích) J výběrů z multinomického rozdělení Mult_K s potenciálně různými vektory pravděpodobností.

Testování nezávislosti χ^2 testem

Nezávislost X a Z platí, právě když $p_{j(1)} = p_{j(2)} = \ldots = p_{j(K)}$ pro všechna $j = 1, \ldots, J$. To vyžaduje, aby v kterékoli podtabulce 2×2 obsahující hodnoty $X = j_1, j_2$ a $Z = k_1, k_2$ byl rozdíl rizik 0 nebo relativní riziko či poměr šancí 1.

Pokud platí hypotéza, že X a Z jsou nezávislé náhodné veličiny, pravděpodobnosti $\boldsymbol{p} = (p_{11}, \ldots, p_{JK})^\mathsf{T}$ specifikující multinomické rozdělení vektoru $\boldsymbol{n}$ jsou funkcemi $d = J + K - 2$ parametrů $\boldsymbol{\theta}_X = (p_{1+}, \ldots, p_{(J-1)+}, p_{+1}, \ldots, p_{+(K-1)})^\mathsf{T}$. Maximálně věrohodný odhad parametru $\boldsymbol{\theta}_X$ za hypotézy nezávislosti je

$$\widehat{\boldsymbol{\theta}}_n = (\widehat{p}_{1+}, \ldots, \widehat{p}_{(J-1)+}, \widehat{p}_{+1}, \ldots, \widehat{p}_{+(K-1)})^\mathsf{T} = (n_{1+}/N, \ldots, n_{(J-1)+}/N, n_{+1}/N, \ldots, n_{+(K-1)}/N)^\mathsf{T}.$$

Maximálně věrohodné odhady složek vektoru $\boldsymbol{p}$ za hypotézy nezávislosti vyjdou

$$p_{jk}(\widehat{\boldsymbol{\theta}}_n) = \widehat{p}_{j+}\widehat{p}_{+k} = \frac{n_{j+}n_{+k}}{N^2},$$

$j = 1, \ldots, J$, $k = 1, \ldots, K$. Očekávané četnosti v kontingenční tabulce za platnosti hypotézy jsou opět $N p_{jk}(\widehat{\boldsymbol{\theta}}_n) = N\widehat{p}_{j+}\widehat{p}_{+k} = n_{j+}n_{+k}/N$.

Testová statistika χ^2 testu nezávislosti má tvar

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}.$$

Podle tvrzení 6.8 má tato statistika za platnosti hypotézy nezávislosti asymptoticky χ^2 rozdělení s počtem stupňů volnosti $JK - (J + K - 2) - 1$, tj. $(J-1)(K-1)$. Hypotézu nezávislosti zamítneme, pokud $\chi^2 \geq \chi^2_{(J-1)(K-1)}(1-\alpha)$.

Test nezávislosti zároveň testuje i hypotézu že K výběrů z multinomického rozdělení má stejné vektory pravděpodobností (jde tedy o K-výběrový test na shodnost parametrů K multinomických rozdělení).

8 Analýza rozptylu

Dvouvýběrové testy ověřují, jestli se dvě skupiny nezávislých pozorování liší v nějaké charakteristice, nejčastěji ve střední hodnotě. Jak ale porovnat střední hodnoty, je-li skupin více? Pro kategoriální data (binomické či multinomické rozdělení) jsme problém porovnání několika skupin řešili v minulé kapitole. Nyní budeme studovat tento problém u kvantitativních náhodných veličin.

Máme $p \geq 2$ nezávislých náhodných výběrů (skupin)

$$\begin{aligned} &Y_{11}, \ldots, Y_{1n_1} \text{ z rozdělení } F_1, \\ &Y_{21}, \ldots, Y_{2n_2} \text{ z rozdělení } F_2, \\ &\vdots \\ \text{a} \quad &Y_{p1}, \ldots, Y_{pn_p} \text{ z rozdělení } F_p. \end{aligned}$$

Pozorování označujeme Y_{ij}, kde i je číslo výběru jdoucí od 1 do p a j je index pozorování v rámci daného výběru běžící od 1 do n_i, kde n_i je rozsah i-tého výběru. Označme $N = \sum_{i=1}^{p} n_i$ a $\boldsymbol{n} = (n_1, \ldots, n_p)^\mathsf{T}$. Platí $\mathbf{1}_p^\mathsf{T} \boldsymbol{n} = N$.

8.1 Analýza rozptylu – jednoduché třídění

Budeme předpokládat platnost modelu, který požaduje, aby všechny výběry měly normální rozdělení s totožným rozptylem. Jednotlivé skupiny se tedy mohou navzájem lišit pouze střední hodnotou.

Model:

$$\mathcal{F} = \{F_i = \mathsf{N}(\mu_i, \sigma^2),\ \mu_i \in \mathbb{R},\ i = 1, \ldots, p,\ \sigma^2 > 0\}$$

Parametr μ_i označuje střední hodnotu i-té skupiny, tj. $\mu_i = \mathsf{E}\, Y_{ij}$. Budeme se zabývat otázkou, zdali všechny skupiny mají stejnou střední hodnotu.

Testované parametry: Střední hodnoty $\mu_i = \mathsf{E}\, Y_{ij}$

Hypotéza a alternativa:

$$H_0 : \mu_1 = \cdots = \mu_p, \qquad H_1 : \exists i \neq j : \mu_i \neq \mu_j.$$

Značení. Nechť $Y_{i+} \overset{\text{df}}{=} \sum_{j=1}^{n_i} Y_{ij}$ a $\overline{Y}_{i+} \overset{\text{df}}{=} n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$ jsou součty a průměry jednotlivých skupin, nechť $Y_{++} \overset{\text{df}}{=} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}$ je celkový součet a $\overline{Y}_{++} \overset{\text{df}}{=} N^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}$ je celkový průměr. Všimněte si, že $\overline{Y}_{++}$ je vážený průměr skupinových průměrů $\overline{Y}_{i+}$ s vahami n_i, tj.

$$\overline{Y}_{++} = \frac{\sum_{i=1}^{p} n_i \overline{Y}_{i+}}{\sum_{i=1}^{p} n_i}.$$

Označme dále pozorování ve skupinách $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\mathsf{T}$, $i = 1, \ldots, p$, všechna pozorování $\boldsymbol{Y} = (\boldsymbol{Y}_1^\mathsf{T}, \ldots, \boldsymbol{Y}_p^\mathsf{T})^\mathsf{T}$ a průměry skupin $\overline{\boldsymbol{Y}} = (\overline{Y}_{1+}, \ldots, \overline{Y}_{p+})^\mathsf{T}$.

Náš přístup bude založen na několika druzích součtů čtverců, které zavádí následující definice.

Definice 8.1 Součty čtverců v analýze rozptylu

- $SS_C \overset{\text{df}}{=} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{++})^2$ nazýváme *celkový součet čtverců*[*].
- $SS_A \overset{\text{df}}{=} \sum_{i=1}^{p} n_i (\overline{Y}_{i+} - \overline{Y}_{++})^2$ nazýváme *součet čtverců skupin*[†].
- $SS_e \overset{\text{df}}{=} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2$ nazýváme *residuální součet čtverců*[‡].

Věta 8.1 Platí

$$SS_C = SS_A + SS_e.$$

Poznámka. SS_C měří celkovou variabilitu dat. Tu můžeme rozložit na variabilitu mezi jednotlivými skupinami vyjadřující jejich vzájemnou odlišnost (SS_A) a variabilitu uvnitř jednotlivých skupin SS_e.

Jelikož $\overline{Y}_{i+}$ je odhadem μ_i a $\overline{Y}_{++}$ je odhadem celkové střední hodnoty (za H_0), bude za platnosti hypotézy SS_A malé vzhledem k SS_e. Pokud je SS_A velké vzhledem k SS_e, znamená to, že se průměry jednotlivých skupin od sebe příliš liší a hypotézu o rovnosti středních hodnot bychom měli zamítat.

Označme $\mathbb{A}_i = \mathbb{I}_{n_i} - \frac{1}{n_i}\mathbf{1}_{n_i}^{\otimes 2}$, $\mathbb{C} = \operatorname{diag}(\boldsymbol{n}) - \frac{1}{N}\boldsymbol{n}^{\otimes 2}$,

$$\mathbb{H} = \begin{pmatrix} \mathbf{1}_{n_1}^\mathsf{T} & 0 & \dots & 0 \\ 0 & \mathbf{1}_{n_2}^\mathsf{T} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{1}_{n_p}^\mathsf{T} \end{pmatrix}, \qquad \mathbb{A} = \begin{pmatrix} \mathbb{A}_1 & 0 & \dots & 0 \\ 0 & \mathbb{A}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbb{A}_p \end{pmatrix}.$$

Následující lemma ukazuje, že SS_A a SS_e lze přepsat jako kvadratické formy.

Lemma 8.2 Platí

(i) $\overline{\boldsymbol{Y}} = \operatorname{diag}(\boldsymbol{n})^{-1}\mathbb{H}\boldsymbol{Y}$;
(ii) $\mathbf{1}_N^\mathsf{T}\mathbb{A} = \mathbf{0}^\mathsf{T}$, $\mathbf{1}_p^\mathsf{T}\mathbb{C} = \mathbf{0}^\mathsf{T}$;
(iii) $SS_e = \boldsymbol{Y}^\mathsf{T}\mathbb{A}\boldsymbol{Y} = (\boldsymbol{Y} - c\mathbf{1}_N)^\mathsf{T}\mathbb{A}(\boldsymbol{Y} - c\mathbf{1}_N)$ pro libovolné $c \in \mathbb{R}$;
(iv) $SS_A = \overline{\boldsymbol{Y}}^\mathsf{T}\mathbb{C}\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}} - c\mathbf{1}_p)^\mathsf{T}\mathbb{C}(\overline{\boldsymbol{Y}} - c\mathbf{1}_p)$ pro libovolné $c \in \mathbb{R}$ a
(v) $SS_A = \boldsymbol{Y}^\mathsf{T}\mathbb{B}\boldsymbol{Y} = (\boldsymbol{Y} - c\mathbf{1}_N)^\mathsf{T}\mathbb{B}(\boldsymbol{Y} - c\mathbf{1}_N)$ pro libovolné $c \in \mathbb{R}$, kde $\mathbb{B} = \mathbb{H}^\mathsf{T}\operatorname{diag}(\boldsymbol{n})^{-1}\mathbb{C}\operatorname{diag}(\boldsymbol{n})^{-1}\mathbb{H}$.

Zde končí předn. 26 (11.1.)

Věta 8.3 (rozdělení součtů čtverců) Za platnosti modelu $\mathcal{F}$ máme

(i)
$$\frac{SS_e}{\sigma^2} \sim \chi^2_{N-p}, \qquad \mathsf{E}\,\frac{SS_e}{N-p} = \sigma^2.$$

(ii) Platí-li navíc hypotéza H_0, pak
$$\frac{SS_C}{\sigma^2} \sim \chi^2_{N-1}, \qquad \mathsf{E}\,\frac{SS_C}{N-1} = \sigma^2.$$

[*] Angl. *total sum of squares* [†] Angl. *between group sum of squares* [‡] Angl. *residual sum of squares, error sum of squares*

(iii) Platí-li navíc hypotéza H_0, pak

$$\frac{SS_A}{\sigma^2} \sim \chi^2_{p-1}, \quad \mathsf{E}\,\frac{SS_A}{p-1} = \sigma^2.$$

(iv) SS_A a SS_e jsou nezávislé.

Poznámka.

- $SS_e/(N-p)$ je vždy nestranným odhadem rozptylu σ^2 (bez ohledu na platnost hypotézy nebo předpoklad normality).
- $SS_A/(p-1)$ je nestranným odhadem rozptylu pouze za hypotézy (ať už je rozdělení Y_{ij} normální nebo ne). Pokud hypotéza neplatí, lze ukázat pomocí lemmatu 1.5, že

$$\mathsf{E}\,\frac{SS_A}{p-1} = \sigma^2 + \frac{1}{p-1}\sum_{i=1}^{p} n_i(\mu_i - \overline{\mu})^2,$$

  kde $\overline{\mu} = N^{-1}\sum_{i=1}^{p} n_i\mu_i$. Porušení hypotézy se tedy projeví na SS_A zvýšením jeho střední hodnoty.
- Tato metoda se nazývá analýza rozptylu* kvůli tomu, jakým způsobem je sestavena testová statistika. Účelem analýzy rozptylu není analyzovat rozptyl.

Testová statistika:

$$F_A = \frac{SS_A}{p-1}\Big/\frac{SS_e}{N-p}$$

Hypotézu budeme zamítat pro příliš velké hodnoty F_A.

Věta 8.4 Za platnosti modelu $\mathcal{F}$ a hypotézy H_0 platí $F_A \sim F_{p-1,N-p}$.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow F_A \geq F_{p-1,N-p}(1-\alpha)$$

kde $F_{p-1,N-p}(1-\alpha)$ je $(1-\alpha)$-tý kvantil F rozdělení s $p-1$ a $N-p$ stupni volnosti.

Poznámka.

- Tento test se nazývá F test analýzy rozptylu. Je to přesný test rovnosti středních hodnot v $p \geq 2$ nezávislých výběrech. Vyžaduje normální rozdělení a stejný rozptyl ve všech výběrech.
- Pokud rozdělení dat není normální, ale rozptyly ve všech skupinách jsou stejné, F test analýzy rozptylu dodržuje hladinu alespoň asymptoticky.
- Pro případ nestejných rozptylů navrhl zobecnění testové statistiky a aproximaci jejího rozdělení Welch. Jde vlastně o zobecnění dvouvýběrového Welchova testu na více výběrů. Publikované simulační studie ukazují, že porušení předpokladu shodných rozptylů nemá zásadní vliv na chování F testu analýzy rozptylu, pokud je počet pozorování ve všech skupinách přibližně stejný.

P-hodnota: $1-F^*(s)$, kde s je pozorovaná hodnota testové statistiky a F^* je distribuční funkce rozdělení $F_{p-1,N-p}$.

* Angl. *analysis of variance, ANOVA*

Poznámka. Výsledky analýzy rozptylu se tradičně uvádějí formou tabulky.

Zdroj měnlivosti	Součet čtverců	Stupňů volnosti	Podíl	F
Skupina	SS_A	$p-1$	$\frac{SS_A}{p-1}$	$\frac{SS_A}{p-1} / \frac{SS_e}{N-p}$
Residuální	SS_e	$N-p$	$\frac{SS_e}{N-p}$	
Celkový	SS_C	$N-1$		

Tvrzení 8.5 Pokud $p = 2$, pak platí

$$F_A = T^2_{n_1,n_2},$$

kde F_A je testová statistika analýzy rozptylu a $T^2_{n_1,n_2}$ je čtverec testové statistiky dvouvýběrového t-testu.

Pro porovnání dvou skupin je tedy analýza rozptylu ekvivalentní dvouvýběrovému t-testu.

Analýza rozptylu se dále zobecňuje na vícenásobné třídění. Tato zobecnění se probírají v předmětu Lineární regrese. Např. dvojné třídění spočívá v tom, že se pozorování klasifikují do pq skupin podle dvou kategoriálních veličin s p a q hodnotami. Zajímá nás, zdali některá z obou kategoriálních veličin ovlivňuje střední hodnotu pozorování.

8.2 Mnohonásobná porovnávání

V analýze rozptylu porovnáváme mezi sebou střední hodnoty p skupin. Pokud F test analýzy rozptylu zamítne hypotézu, že všechny skupiny mají stejnou střední hodnotu, pak usoudíme, že alespoň některé skupiny se od sebe liší ve středních hodnotách. Nevíme ovšem, kolik takových odlišných skupin je, ani které to jsou.

Kdybychom chtěli porovnat střední hodnoty pouze dvou skupin, třeba skupin i a j, použili bychom dvouvýběrový t-test. Mohli bychom pak provést dvouvýběrové testy pro všech $p(p-1)/2$ možných dvojic skupin a otestovat všechny hypotézy $H_0^{ij} : \mu_i = \mu_j$ na hladině α. Potom ale pravděpodobnost, že alespoň jednu hypotézu zamítneme za podmínky, že všechny hypotézy platí, není rovna α, ale je větší.

Problém současného testování více hypotéz se ve statistice často nazývá *problém mnohonásobných porovnávání*[*] nebo *mnohonásobného testování*[†]. Tento problém lze převést na problém konstrukce několika intervalů spolehlivosti pro různé parametry tak, aby pravděpodobnost, že *všechny* intervaly pokrývají hledané parametry byla $1-\alpha$. Pak hovoříme o *simultánních intervalech spolehlivosti*[‡].

V této kapitole si uvedeme nejprve jeden obecný přístup k tomuto problému a pak speciální metodu pro porovnávání středních hodnot několika nezávislých výběrů.

8.2.1 Bonferroniho metoda

Představme si obecný problém mnohonásobného testování: máme m hypotéz $H_0^1, \ldots, H_0^m$, které chceme otestovat. Hypotéza H_0^i bude testována testem s testovou statistikou T_i a kritickým oborem C_i zvoleným tak, aby každý test měl hladinu α_0. Pro každé $i \in \{1, \ldots, m\}$ tedy

[*] Angl. *multiple comparisons* [†] Angl. *multiple testing* [‡] Angl. *simultaneous confidence intervals*

platí

$$\mathsf{P}_{H_0^i}[T_i \in C_i] = \alpha_0.$$

Celková pravděpodobnost zamítnutí alespoň jedné hypotézy za předpokladu, že všechny platí, je

$$\mathsf{P}_{\bigcap H_0^i}\left(\bigcup_{i=1}^m [T_i \in C_i]\right) = \alpha_C.$$

Pochopitelně α_C je větší než α_0, často výrazně.

Máme danou celkovou hladinu α a chceme zaručit, že $\alpha_C \leq \alpha$. K tomu použijeme následující lemma.

Lemma 8.6 (Booleova nerovnost) Pro jakékoli náhodné jevy $A_1, \ldots, A_n$ platí

$$\mathsf{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathsf{P}(A_i).$$

Booleova nerovnost je triviální pro $n = 2$, pro vyšší n se snadno dokáže matematickou indukcí.

Máme tedy

$$\alpha_C = \mathsf{P}_{\bigcap H_0^i}\left(\bigcup_{i=1}^m [T_i \in C_i]\right) \leq m\alpha_0.$$

Zvolíme-li $\alpha_0 = \alpha/m$, pak musí platit $\alpha_C \leq \alpha$. Chceme-li tedy provést m testů tak, aby celková hladina všech testů (pravděpodobnost zamítnutí alespoň jedné hypotézy za podmínky, že všechny platí) byla nejvýše α, provedeme jednotlivé dílčí testy na hladině α/m. Podobně, chceme-li sestrojit m intervalů spolehlivosti tak, aby pravděpodobnost, že všechny intervaly pokryjí hledané parametry, byla alespoň $1 - \alpha$, stačí stanovit pravděpodobnost pokrytí jednotlivých dílčích intervalů na $1 - \alpha/m$. Tento přístup k mnohonásobnému testování a konstrukci simultánních intervalů spolehlivosti se nazývá *Bonferroniho metoda*[*].

Výhodou Bonferroniho metody je její jednoduchopst a universalita. Její nevýhodou je, že úprava hladiny α na α/m je téměř vždy příliš přísná. Bonferroniho metoda tedy dává testy s malou silou a zbytečně široké intervaly spolehlivosti.

Aplikace Bonferroniho metody na mnohonásobná porovnávání v analýze rozptylu vypadá takto: provedeme $p(p-1)/2$ dvouvýběrových t-testů pro všechny možné dvojice skupin a otestujeme všechny hypotézy $H_0^{ij} : \mu_i = \mu_j$ na hladině $2\alpha/[p(p-1)]$. Pokud je některá z těchto hypotéz zamítnuta, prohlásíme střední hodnoty daných dvou skupin za významně odlišné na celkové hladině α.

Máme-li například $\alpha = 0.05$ a $p = 6$ skupin, provádíme 15 testů rovnosti středních hodnot pro 15 dvojic různých skupin na hladině $0.05/15 \doteq 0,0033$. To je natolik nízká hladina, že může být obtížné najít kterékoli dvě odlišné skupiny, přestože F test analýzy rozptylu zamítá hypotézu, že všechny střední hodnoty jsou stejné.

Zde končí předn. 27 (12.1.)

8.2.2 Tukeyova metoda

Pozn.: Tato část nebyla v roce 2016/17 přednášena.

Mějme nezávislé náhodné veličiny $Z_i \sim \mathsf{N}(\mu, \sigma^2)$ pro $i = 1, \ldots, m$. Nechť S^2 je odhad rozptylu σ^2 takový, že S^2 je nezávislé na $Z_1, \ldots, Z_m$ a pro nějaké přirozené k platí $kS^2/\sigma^2 \sim \chi^2_k$.

[*] Angl. *Bonferroni correction*

Definujme tak řečené *studentisované rozpětí*[*] jako

$$Q = \frac{\max_{i=1,\dots,m} Z_i - \min_{i=1,\dots,m} Z_i}{S}.$$

Lze ukázat, že náhodná veličina Q má rozdělení závisející pouze na hodnotách m a k. Označme kvantilovou funkci tohoto rozdělení $q_{m,k}(\alpha)$. (Vzorce pro hustotu a kvantilovou funkci studentisovaného rozpětí nebudeme uvádět.)[†]

Studentizovaného rozpětí lze použít k sestrojení simultánních intervalů spolehlivosti pro rozdíly středních hodnot. Tento postup se nazývá *Tukeyova metoda.*[‡]

Věta 8.7 (Tukeyova) Nechť $Z_1, \dots, Z_m$ jsou nezávislé náhodné veličiny s rozdělením $Z_i \sim \mathsf{N}(\mu_i, \sigma^2)$. Nechť S^2 je odhad rozptylu σ^2 takový, že S^2 je nezávislé na $Z_1, \dots, Z_m$ a pro nějaké přirozené k platí $kS^2/\sigma^2 \sim \chi^2_k$. Pak

$$\mathsf{P}\Big[Z_i - Z_j - Sq_{m,k}(1-\alpha) \le \mu_i - \mu_j \le Z_i - Z_j + Sq_{m,k}(1-\alpha) \quad \forall i \ne j \in \{1,\dots,m\}\Big] = 1-\alpha.$$

Tukeyovu větu lze snadno použít i na testování hypotéz. Hypotézu $H_0^{ij} : \mu_i = \mu_j$ zamítneme, pokud $\left|Z_i - Z_j\right| > Sq_{m,k}(1-\alpha)$. Hypotézu $H_0 : \mu_1 = \ldots = \mu_m$ zamítneme na celkové hladině α, pokud pro alespoň jednu dvojici $i \ne j$ platí $\left|Z_i - Z_j\right| > Sq_{m,k}(1-\alpha)$.

Tukeyovu větu můžeme přímo aplikovat na mnohonásobná porovnávání v analýze rozptylu, pokud rozsah výběru všech skupin je totožný, tj. $n_1 = \cdots = n_p \equiv n$. Pak totiž $\overline{Y}_{1+}, \dots, \overline{Y}_{p+}$ jsou nezávislé náhodné veličiny s rozdělením $\overline{Y}_{i+} \sim \mathsf{N}(\mu_i, \sigma^2/n)$. Za S^2, odhad σ^2/n vezmeme $SS_e/[n(N-p)]$. Máme $k = N - p$. Hypotézu $H_0^{ij} : \mu_i = \mu_j$ zamítneme, pokud

$$\left|\overline{Y}_{i+} - \overline{Y}_{j+}\right| > \sqrt{\frac{SS_e}{N-p}}\sqrt{\frac{1}{n}}q_{p,N-p}(1-\alpha). \tag{8.1}$$

Pokud rozsahy všech výběrů nejsou stejné, nemůžeme Tukeyovu větu přímo použít, protože nejsou splněny její předpoklady. Lze ale dokázat, že pokud výraz $\sqrt{\frac{1}{n}}$ v (8.1) nahradíme výrazem $\sqrt{\frac{1}{2n_i} + \frac{1}{2n_j}}$, celková pravděpodobnost zamítnutí některé z platných hypotéz H_0^{ij} nepřekročí α. Tukeyova metoda tedy po této úpravě stále funguje, pouze se stává poněkud konservativní.

8.3 Kruskalův-Wallisův test

Pozn.: Tato část nebyla v roce 2016/17 přednášena.

Kruskalův-Wallisův test je zobecněním dvouvýběrového Wilcoxonova testu na porovnání $p \ge 2$ výběrů. I nadále používáme značení zavedené na začátku kapitoly Analýza rozptylu.
Model: $\mathcal{F} = \{F_i \text{ je spojitá d.f. taková, že } F_i(x) = F(x - \delta_i)\ \forall x \in \mathbb{R}\}$

[*] Angl. *studentized range* [†] Studentisované rozpětí se někdy definuje jako $Q/\sqrt{2}$. Na to je třeba dávat pozor při používání tabelovaných nebo softwarem vypočtených hodnot $q_{m,k}(\alpha)$. Pro kontrolu můžeme porovnat rozdělení Q při $m = 2$ s rozdělením $|T|$, kde $T \sim t_k$. Pro naši definici jsou tato dvě rozdělení totožná. [‡] Angl. *Tukey method, Tukey's range test, Tukey's HSD (honest significant difference) test.*

Jde o p spojitých rozdělení navzájem posunutých v poloze. Bez újmy na obecnosti můžeme položit $\delta_1 = 0$.

Hypotéza a alternativa:

$$H_0 : \delta_1 = \cdots = \delta_p = 0, \quad H_1 : \exists i : \delta_i \neq 0.$$

Poznámka. Pokud platí model $\mathcal{F}$ a hypotéza H_0, rozdělení ve všech skupinách jsou totožná. Potom platí mezi p skupinami rovnost veškerých charakteristik. Nejsou-li rozptyly ve všech skupinách totožné, model $\mathcal{F}$ nemůže platit.

Testová statistika:

Lze ukázat, že testová statistika dvouvýběrového Wilcoxonova testu je ekvivalentní čitateli testové statistiky dvouvýběrového t testu (tj. rozdílu průměrů), pokud do ní místo původních pozorování dosadíme jejich pořadí. Se stejnou logikou můžeme použít čitatel testové statistiky F testu analýzy rozptylu (tj. SS_A), do nějž dosadíme pořadí namísto původních pozorování.

Nechť R_{ij} je pořadí pozorování Y_{ij} ve spojeném výběru $Y_{11}, \ldots, Y_{pn_p}$. Položme $R_{i+} = \sum_{j=1}^{n_i} R_{ij}$ a $\overline{R}_{i+} = n_i^{-1} R_{i+}$. Celkový průměr všech pořadí je $\overline{R}_{++} = N^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n_i} R_{ij} = (N+1)/2$. Dosazením do vzorce pro SS_A dostaneme

$$\sum_{i=1}^{p} n_i \Big(\overline{R}_{i+} - \frac{N+1}{2}\Big)^2 = \sum_{i=1}^{p} \frac{1}{n_i} \Big(R_{i+} - n_i \frac{N+1}{2}\Big)^2 =$$

$$= \sum_{i=1}^{p} \frac{1}{n_i} \Big(R_{i+}^2 - R_{i+} n_i (N+1) + n_i^2 \frac{(N+1)^2}{4}\Big) = \sum_{i=1}^{p} \frac{R_{i+}^2}{n_i} - \frac{N(N+1)^2}{4}.$$

Tento výraz podělíme $N(N+1)/12$ a tím dostaneme testovou statistiku Q Kruskalova-Wallisova testu:

$$Q = \frac{12}{N(N+1)} \sum_{i=1}^{p} \frac{R_{i+}^2}{n_i} - 3(N+1).$$

Tvrzení 8.8 (Hájek & Šidák, 1967) Platí-li model $\mathcal{F}$ a hypotéza H_0 a všechna n_i konvergují do ∞ stejně rychle, pak

$$Q \xrightarrow{\mathsf{D}} \chi^2_{p-1}.$$

Statistika $Q/(p-1)$ má tedy za platnosti hypotézy stejné asymptotické rozdělení jako F_A. Hypotézu budeme zamítat pro příliš velké hodnoty Q.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow Q \geq \chi^2_{p-1}(1-\alpha).$$

Poznámka. Neplatí-li model posunutí v poloze, Kruskalův-Wallisův test ověřuje hypotézy $H_0^{*(ij)} : \mathsf{P}[Y_{ik} < Y_{jl}] = 1/2$ pro všechna $i \neq j$ (viz diskuse dvouvýběrového Wilcoxonova testu na str. 69). Tuto hypotézu nelze interpretovat jako rovnost středních hodnot nebo mediánů. Limitní rozdělení testové statistiky Q navíc platí pouze za předpokladu posunutí v poloze, při porušení tohoto předpokladu nejsou kritické hodnoty správné. Nejsme-li si jistí platností modelu posunutí v poloze, Kruskalův-Wallisův test raději nepoužíváme.

9 Korelační analýza

Pozn.: Tato kapitola nebyla v roce 2016/17 přednášena.

Uvažujme náhodný výběr

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

dvousložkových náhodných vektorů, kde obě veličiny jsou spojité a $n \geq 3$.

9.1 Výběrový korelační koeficient

Chceme otestovat korelaci mezi oběma složkami, případně sestrojit interval spolehlivosti pro korelační koeficient definovaný jako

$$\varrho = \varrho(X_i, Y_i) = \frac{\operatorname{cov}(X_i, Y_i)}{\sqrt{\operatorname{var} X_i \operatorname{var} Y_i}}.$$

Jeho konsistentním odhadem je výběrový korelační koeficient

$$\widehat{\varrho} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2 \sum_{i=1}^n (Y_i - \overline{Y}_n)^2}} = \frac{\sum_{i=1}^n X_i Y_i - n\overline{X}_n \overline{Y}_n}{\sqrt{\left(\sum_{i=1}^n X_i^2 - n\overline{X}_n^2\right)\left(\sum_{i=1}^n Y_i^2 - n\overline{Y}_n^2\right)}} \tag{9.1}$$

zavedený v definici 2.8.

Ukážeme si nejprve bez důkazu rozdělení korelačního koeficientu za předpokladu normality a nezávislosti.

Tvrzení 9.1 Nechť $\binom{X_i}{Y_i}$, $i = 1, \ldots, n$ je náhodný výběr z dvourozměrného normálního rozdělení s kladnými rozptyly a nulovou korelací mezi složkami. Pak platí

$$T = \sqrt{n-2}\frac{\widehat{\varrho}}{\sqrt{1-\widehat{\varrho}^2}} \sim t_{n-2}.$$

Tohoto tvrzení můžeme použít pro otestování hypotézy $H_0 : \varrho = 0$ proti alternativě $H_1 : \varrho \neq 0$ za předpokladu normality. Hypotézu H_0 zamítneme, pokud $|T| \geq t_{n-2}(1-\alpha/2)$. Tento test má přesně hladinu α.

Tvrzení 9.1 však nelze rozšířit na testování hypotéz $H_0 : \varrho = \varrho_0$, kde $\varrho_0 \neq 0$. Nelze z něj také sestrojit interval spolehlivosti. Navíc předpoklad normality je zásadní, nemůžeme jej ignorovat ani pro velmi velká n.

Jinou metodu založenou na transformaci funkcí

$$\operatorname{arctgh} x = \frac{1}{2} \log \frac{1+x}{1-x}$$

navrhl R. A. Fisher. Říká se jí Fisherova Z-transformace. Fisher ukázal platnost následujícího tvrzení.

Tvrzení 9.2 Nechť $\binom{X_i}{Y_i}$, $i = 1, \ldots, n$ je náhodný výběr z dvourozměrného normálního rozdělení s korelačním koeficientem ϱ. Pak pro výběrový korelační koeficient $\widehat{\varrho}$ platí

$$\sqrt{n-3}(\operatorname{arctgh}\widehat{\varrho} - \operatorname{arctgh}\varrho) \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Chceme-li otestovat hypotézu $H_0 : \varrho = \varrho_0$ proti alternativě $H_1 : \varrho \neq \varrho_0$, spočítáme testovou statistiku

$$Z = \sqrt{n-3}(\operatorname{arctgh}\widehat{\varrho} - \operatorname{arctgh}\varrho_0)$$

a H_0 zamítneme na hladině α, pokud $|Z_n| \geq u_{1-\alpha/2}$. Přibližný interval spolehlivosti pro ϱ získáme z intervalu spolehlivosti pro $\operatorname{arctgh}\varrho$ zpětnou transformací pomocí funkce $\operatorname{tgh} x = \frac{\exp(2x)-1}{\exp(2x)+1}$. Dostaneme interval

$$(\operatorname{tgh}(\operatorname{arctgh}\widehat{\varrho} - u_{1-\alpha/2}/\sqrt{n-3}),\ \operatorname{tgh}(\operatorname{arctgh}\widehat{\varrho} + u_{1-\alpha/2}/\sqrt{n-3})).$$

I Fisherova Z-transformace spoléhá na normalitu, pro jiná rozdělení tvrzení 9.2 neplatí. Pokud data nejsou normální, lze vytvořit asymptotické testy a intervaly spolehlivosti např. metodou *bootstrap*, s níž se budete moci seznámit v předmětu „Moderní statistické metody".

9.2 Spearmanův korelační koeficient

Spearmanův korelační koeficient vychází z výrazu (9.1), ale dosazuje do něj pořadí namísto původních pozorování. Označme R_i pořadí pozorování X_i v náhodném výběru $X_1, \ldots, X_n$ a označme S_i pořadí pozorování Y_i v náhodném výběru $Y_1, \ldots, Y_n$. Pokud X_i je nezávislé na Y_i pro každé i, pak by neměly být závislosti ani mezi pořadími R_i a S_i.

Spearmanův korelační koeficient dostaneme dosazením R_i místo X_i a S_i místo Y_i v (9.1):

$$\widehat{\varrho}_S = \frac{\sum_{i=1}^n R_i S_i - n\overline{R}_n\overline{S}_n}{\sqrt{\left(\sum_{i=1}^n R_i^2 - n\overline{R}_n^2\right)\left(\sum_{i=1}^n S_i^2 - n\overline{R}_n^2\right)}}.$$

S použitím vztahů $\overline{R}_n = \overline{S}_n = (n+1)/2$, $\sum_{i=1}^n R_i^2 = \sum_{i=1}^n S_i^2 = n(n+1)(2n+1)/6$ a $\sum(R_i - S_i)^2 = \sum R_i^2 + \sum S_i^2 + 2\sum R_i S_i$ přepíšeme $\widehat{\varrho}_S$ v jednodušším tvaru.

Definice 9.1 Nechť $\binom{X_i}{Y_i}$, $i = 1, \ldots, n$, je náhodný výběr ze spojitého dvourozměrného rozdělení, R_i je pořadí pozorování X_i v náhodném výběru $X_1, \ldots, X_n$ a S_i je pořadí pozorování Y_i v náhodném výběru $Y_1, \ldots, Y_n$. Náhodnou veličinu

$$\widehat{\varrho}_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^n (R_i - S_i)^2$$

nazýváme *Spearmanův korelační koeficient*[*] veličin X a Y.

Spearmanův korelační koeficient rozhodně není odhadem teoretického korelačního koeficientu ϱ. Z definice 9.1 vidíme, že Spearmanův korelační koeficient nabývá hodnoty 1 tehdy a jen tehdy, pokud $R_i = S_i$ pro každé i. Pořadí X_i a Y_i jsou si rovna, právě když existuje ostře

[*] Angl. *Spearman correlation coefficient*

rostoucí funkce h taková, že $X_i = h(Y_i)$ pro každé i. Spearmanův korelační koeficient má tedy hodnotu jedna, právě když X_i je ostře rostoucí transformací Y_i, zatímco výběrový korelační koeficient $\widehat{\varrho}$ nabývá hodnoty jedna, právě když X_i je rostoucí lineární transformací Y_i. Naopak, Spearmanův korelační koeficient nabývá své minimální hodnoty -1, právě když $R_i = n + 1 - S_i$, tj. když existuje ostře klesající funkce h taková, že $X_i = h(Y_i)$ pro každé i.

Lze ukázat, že jsou-li X_i a Y_i nezávislé pro každé i, pak

$$\mathsf{E}\,\widehat{\varrho}_S = 0 \quad \text{a} \quad \mathsf{var}\,\widehat{\varrho}_S = \frac{1}{n-1}.$$

Za nezávislosti tedy $\widehat{\varrho}_S$ konverguje v pravděpodobnosti k nule. Dokonce lze ukázat, že za nezávislosti platí

$$\sqrt{n-1}\,\widehat{\varrho}_S \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Toho lze využít ke konstrukci testu hypotézy nezávislosti mezi X_i a Y_i. Hypotézu zamítneme, pokud $\sqrt{n-1}\,|\widehat{\varrho}_S| \geq u_{1-\alpha/2}$. Test je asymptotický a nepředpokládá normalitu. Není však citlivý vůči některým alternativám, například pro $X_i \in (-1, 1)$ nebude konsistentní vůči alternativám $Y_i = X_i^2$.

Pozn.: Termíny uvedené *kurzívou* nebyly v roce 2016/17 přednášeny.

Rejstřík