NMFM301 Statistika pro finanční matematiky

Přehledový větník

Michal Kulich

Naposledy upraveno dne 27. září 2014.

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fysikální fakulta University Karlovy

Tento dokument poskytuje přehled všech vět, definic, tvrzení a poznámek probíraných v přednášce „NMFM301 Statistika pro finanční matematiky“ v rámci bakalářského studia oboru „Finanční matematika“ na MFF UK. Nejedná se o učební text, protože zde nejsou uvedeny všechny příklady a důkazy probírané na přednášce ani látka ze cvičení. Při přípravě na zkoušku je nutno tento materiál doplnit poznámkami z přednášek a cvičení.

Odkazy na potřebné definice, věty a tvrzení z teorie pravděpodobnosti (začínající písmenem P) se týkají příručky „Souhrn teorie pravděpodobnosti pro obor Finanční matematika“, která je k dispozici na webových stránkách předmětu NMFM301. Např. tvrzení P.2.2 nebo definici P.6.1 lze najít ve 2., resp. 6. kapitole zmíněného Souhrnu.

Autor bude povděčen za upozornění na případné překlepy a nejasnosti, které laskavý čtenář nalezne kdekoli v tomto dokumentu.

Michal Kulich
kulich@karlin.mff.cuni.cz

Dáno v Karlíně dne 7. ledna 2014

1 Náhodný výběr

1.1 Definice náhodného výběru

Definice 1.1. Posloupnost $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ nezávislých stejně rozdělených náhodných veličin, z nichž každá má distribuční funkci F_0, nazýváme *náhodný výběr z rozdělení* F_0.* Konstantu n nazýváme *rozsah výběru*.†

Prvky náhodného výběru mohou být reálné náhodné veličiny i náhodné vektory. 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* for 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. O rozdělení F_0 se potřebujeme dozvědět jeho charakteristiky, které nazýváme *parametry*. Jedná se o nějakou konstantu (nebo vektor konstant) $\boldsymbol{\theta}_0 \in \mathbb{R}^k$, kterou bychom uměli zjistit, kdybychom znali F_0. 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. Model $\mathcal{F}$ může být množina všech [diskrétních, spojitých] rozdělení na $\mathbb{R}$ s konečnou střední hodnotou [s konečným rozptylem]. Hledané parametry mohou být např. $\mathsf{E}\, X_i$, $\mathsf{var}\, X_i$, $\mathrm{P}\left[X \leq x\right] \equiv F_0(x)$, $F_0^{-1}(\alpha)$. 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 $\boldsymbol{\theta} \equiv t(F) : F \in \mathcal{F}$.
2. Model $\mathcal{F}$ může být množina 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 složek $\boldsymbol{\theta}$.

* Angl. *random sample from distribution* F_0 † Angl. *sample size*

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

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 ty se dále zpracovávají. 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. Statistika je náhodná veličina (náhodný vektor, je-li vícerozměrná).

Mezi nejčastěji používané statistiky patří výběrový průměr a výběrový rozptyl. Uvažujme nyní výběr reálných náhodných veličin, nikoli vektorů.

Definice 1.4.

(i) Veličina $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ se nazývá *výběrový průměr*† náhodného výběru $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$.

(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 $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$.

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

Vlastnosti výběrového průměru

Pracujme v širokém modelu $\mathcal{F} = \mathcal{L}^2$ (všechna rozdělení s konečnými druhými momenty). Označme $\mu \equiv \mathsf{E}\, X_i$ a $\sigma^2 = \mathsf{var}\, X_i$.

Věta 1.1 (Vlastnosti průměru)**.** Nechť $\mathsf{var}\, X_i < \infty$. Pak platí

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

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

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

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

Vlastnosti výběrového rozptylu

Pracujme v širokém modelu $\mathcal{F} = \mathcal{L}^2$ (všechna rozdělení s konečnými druhými momenty). Označme $\mu \equiv \mathsf{E}\,X_i$ a $\sigma^2 = \mathsf{var}\,X_i$.

Poznámka. Výběrový rozptyl lze přepsat jako

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

Pro výpočet je tento vzorec většinou výhodnější než definice.

Věta 1.2 (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

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \sigma^4(\gamma_2 - 1)),$$

kde γ_2 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 $\Sigma = \begin{pmatrix}\sigma^2 & \sigma^3\gamma_1 \\ \sigma^3\gamma_1 & \sigma^4(\gamma_2 - 1)\end{pmatrix}$ a γ_1 je šikmost rozdělení X_i.

Poznámka. Věta 1.2(iii) říká, že variabilita výběrového rozptylu asymptoticky závisí na špičatosti pozorování. Věta 1.2(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 $\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\}$.

Věta 1.3 (Vlastnosti výběrového rozptylu za normality)**.** Nechť $X_i \sim \mathsf{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$. Pak platí

(i)
$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1} \tag{1.3}$$

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

Poznámka. Pro velká n máme z definice χ^2 rozdělení a z centrální limitní věty $\chi^2_{n-1} \dot{\sim} \mathsf{N}(n-1, 2(n-1))$. Odtud a z (1.3) dostaneme pro velké n

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

a nakonec

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

Uvědomíme-li si, že špičatost normálního rozdělení je 3, vidíme, že tvrzení (i) z věty 1.3 dává v podstatě stejný výsledek, jako tvrzení (iii) z věty 1.2. Věta 1.3(i) udává přesné rozdělení S_n^2 pro normální data, zatímco věta 1.2(iii) udává asymptotické rozdělení S_n^2 pro normální i nenormální data,

Poznámka. Věta 1.3(ii) říká, že jsou-li data normální, $\overline{X}_n$ a S_n^2 jsou nezávislé pro každé konečné $n > 1$.

Věta 1.4. 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 1.5 (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 1.5 udává přesné rozdělení statistiky T pro normální data, zatímco věta 1.4 udává asymptotické rozdělení téže statistiky pro normální i nenormální data. 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 1.6 (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 $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ a $S_Y^2 = \frac{1}{m-1}\sum_{j=1}^m (Y_j - \overline{Y}_m)^2$. 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, $\mathrm{P}[X_i = X_j] = 0$ 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.

(i) Hodnoty $X_1, \dots, X_n$ lze jednoznačně určit z n-tice pořádkových statistik a n-tice pořadí.
(ii) První pořádková statistika je minimum, n-tá pořádková statistika je maximum všech veličin náhodného výběru.
(iii) Platí $R_i = \sum_{j=1}^n \mathbb{I}_{(0,\infty)}(X_i - X_j)$.
(iv) 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.

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

Věta 1.7. Sdružená hustota náhodného vektoru $(X_{(1)}, \dots, X_{(n)})^{\mathsf{T}}$ vzhledem k Lebesgueově míře jest

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

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

Věta 1.8. Distribuční funkce k-té pořádkové statistiky jest

$$F_{(k)}(x) = \mathrm{P}\left[X_{(k)} \le x\right] = \sum_{j=k}^{n} \binom{n}{i} F^i(x)[1 - F(x)]^{n-i} = \\ = \frac{1}{B(k, n-k+1)} \int_0^{F(x)} t^{k-1}(1-t)^{n-k} dt.$$

Důsledek.

- 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

$$\mathsf{E}\, X_{(k)} = \frac{k}{n+1}, \quad \mathsf{var}\, X_{(k)} = \frac{k(n-k+1)}{(n+2)(n+1)^2}.$$

- Nechť mají X_i jakékoli spojité rozdělení s ryze rostoucí distribuční funkcí F. Nechť $Z \sim \mathsf{B}(k, n-k+1)$. Pak $\mathrm{P}\left[X_{(k)} \le x\right] = \mathrm{P}\left[Z \le F(x)\right] = \mathrm{P}\left[F^{-1}(Z) \le x\right]$, tj. $X_{(k)}$ má stejné rozdělení jako $F^{-1}(Z)$.

Věta 1.9. 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 1.10. 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 1.11. Platí

(i) $\mathrm{P}\left[R_i = k\right] = \frac{1}{n}$ pro všechna $i, k \in \{1, \dots, n\}$.
(ii) $\mathrm{P}\left[R_i = k, R_j = m\right] = \frac{1}{n(n-1)}$ pro všechna $i \neq j, k \neq m \in \{1, \dots, n\}$.
(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 $i \neq j \in \{1, \dots, n\}$.

2 Základy teorie odhadu

2.1 Bodový odhad

Definice bodového odhadu

Máme náhodný výběr $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, model $\mathcal{F}$ a parametr $\boldsymbol{\theta} = t(F) \in \mathbb{R}^k$ 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 $\boldsymbol{\theta}_X \equiv t(F_X)$ je skutečná hodnota hledaného parametru.

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

Poznámka. Odhad je statistika ve smyslu definice 1.3.

Vlastnosti odhadů

Definice 2.2 (Nestrannost a konsistence)**.**

1. Odhad $\widehat{\boldsymbol{\theta}}_n \equiv \boldsymbol{T}_n(\boldsymbol{X})$ nazveme *nestranným odhadem* parametru $\boldsymbol{\theta}_X$ právě když $\mathsf{E}\,\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_X$ pro každé n.†
2. Odhad $\widehat{\boldsymbol{\theta}}_n \equiv \boldsymbol{T}_n(\boldsymbol{X})$ nazveme *konsistentním odhadem* parametru $\boldsymbol{\theta}_X$ právě když $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_X$ pro $n \to \infty$.‡

Poznámka.

- Nestrannost má platit pro každé n. Nestrannost 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 kvalitě 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, lze v praxi používat a často se s nimi setkáme. Odhady, které nejsou konsistentní, lze považovat za nevhodné.

---

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

Příklad.

1. *Odhad* $\theta_X = \mathsf{E}\, X_i$ *v modelu* $\mathcal{F} = \{$*všechna rozdělení s konečnou střední hodnotou*$\}$*:*
   - Průměr $\overline{X}_n$ je nestranný a konsistentní odhad θ_X [plyne z věty 1.1, (i) a (ii)].
   - Odhad $\widehat{\theta}_n = X_1$ je nestranný odhad θ_X, ale není konsistentní.
2. *Odhad* $\theta_X = \mathsf{var}\, X_i$ *v modelu* $\mathcal{F} = \{$*všechna rozdělení s konečným rozptylem*$\}$*:*
   - Výběrový rozptyl S_n^2 je nestranný a konsistentní odhad θ_X [plyne z věty 1.2, (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* $\theta_X = \mathrm{P}\,[X_i = 0]$ *v modelu* $\mathcal{F} = \{\mathsf{Po}(\lambda), \lambda > 0\}$*:*
   - Odhad $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{0\}}(X_i)$ je nestranný a konsistentní odhad θ_X.
   - Odhad $\widetilde{\theta}_n = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^n X_i}$ je také nestranný a konsistentní odhad θ_X.
4. *Odhad* $\theta_X = \mathrm{e}^{-2\lambda_X}$ *v modelu* $\mathcal{F} = \{\mathsf{Po}(\lambda), \lambda > 0\}$ *pro* $n = 1$*:*
   Jediný nestranný odhad jest $\widehat{\theta} = (-1)^{X_1}$, který ale nikdy nenabývá hodnoty přípustné pro $\mathrm{e}^{-2\lambda_X}$.

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

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.

Definice 2.4 (Střední čtvercová odchylka)**.** Nechť odhad $\widehat{\boldsymbol{\theta}}_n \equiv \boldsymbol{T}_n(\boldsymbol{X})$ parametru $\boldsymbol{\theta}_X$ má konečný rozptyl. Výraz

$$\mathrm{MSE}_{\widehat{\boldsymbol{\theta}}_n} = \mathsf{E}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X)^{\otimes 2}$$

nazýváme *střední čtvercovou odchylkou* odhadu $\widehat{\boldsymbol{\theta}}_n$.†

Poznámka.

- Platí: $\mathsf{E}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X)^{\otimes 2} = \mathsf{var}\,\widehat{\boldsymbol{\theta}}_n + [\mathsf{E}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X)]^{\otimes 2}$.
- Střední čtvercová odchylka MSE je jedno z nejobvyklejších kritérií pro porovnávání odhadů. Máme-li několik různých odhadů téhož parametru v tomtéž modelu, obvykle si vybereme si ten, který má nejmenší MSE.
- MSE většinou nelze spočítat. Častěji se používá asymptotická střední čtvercová odchylka AMSE, ale její definice je složitější a nebudeme ji zde uvádět.

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

* Angl. *bias* † Angl. *mean square error, MSE*

2.2 Intervalový odhad

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 $\mathrm{P}[B \ni \theta_X] = 1-\alpha$. Interval B se nazývá *přibližný (asymptotický) intervalový odhad* parametru $\theta_X \in \mathbb{R}$ *o spolehlivosti* $1-\alpha$, právě když $\mathrm{P}[B \ni \theta_X] \to 1-\alpha$ 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 95-tiprocentní 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 přibližnými, 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 region spolehlivosti. Tvar množiny B lze ale potom volit mnoha různými způsoby.

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

- Interval tvaru (C_L, C_U), kde C_L a C_U jsou dvě náhodné veličiny splňující $\mathrm{P}[C_L < C_U] = 1$, $C_L > -\infty$ a $C_U < \infty$, nazýváme oboustranný interval spolehlivosti. Obvykle jej sestrojujeme tak, aby platilo (alespoň asymptoticky)
$$\mathrm{P}[\theta_X < C_L] = \frac{\alpha}{2}, \quad \mathrm{P}[\theta_X > C_L, \theta_X < C_U] = 1-\alpha, \quad \mathrm{P}[\theta_X > C_U] = \frac{\alpha}{2}.$$
- Interval tvaru (C_L, ∞) nazýváme levostranný interval spolehlivosti. Máme $\mathrm{P}[C_L < \theta_X] = 1-\alpha$.
- Interval tvaru $(-\infty, C_U)$ nazýváme pravostranný interval spolehlivosti. Máme $\mathrm{P}[\theta_X < C_U] = 1-\alpha$.

Příklad (střední hodnota normálního rozdělení se známým rozptylem)**.** Vezměme si problém intervalového odhadu střední hodnoty pro normálně rozdělená data se známým rozptylem.

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

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
$$\mathrm{P}\left[u_{\frac{\alpha}{2}} < \sqrt{n}(\overline{X}_n - \mu_X)/\sigma_X < u_{1-\frac{\alpha}{2}}\right] = 1-\alpha,$$
kde u_α je α-kvantil normovaného normálního rozdělení, a postupnými úpravami (s využitím symetrie hustoty $\mathsf{N}(0,1)$ kolem 0) dojdeme k
$$\mathrm{P}\left[\overline{X}_n - \sigma_X u_{1-\frac{\alpha}{2}}/\sqrt{n} < \mu_X < \overline{X}_n + \sigma_X u_{1-\frac{\alpha}{2}}/\sqrt{n}\right] = 1-\alpha.$$
3. Získali jsme oboustranný interval spolehlivosti (C_L, C_U). Jeho hranice jsou
$$C_L = \overline{X}_n - \frac{\sigma_X}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \qquad C_U = \overline{X}_n + \frac{\sigma_X}{\sqrt{n}} u_{1-\frac{\alpha}{2}}.$$
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 $u_{1-\frac{\alpha}{2}}$).

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 intervalu spolehlivosti pro střední hodnotu μ_X nepřekročila stanovenou mez $d > 0$?

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

Poznámka (transformace parametrů)**.** Je-li (C_L, C_U) interval spolehlivosti pro parametr θ_X s pravděpodobností pokrytí $1-\alpha$ a je-li ψ rostoucí reálná funkce, pak $(\psi(C_L), \psi(C_U))$ je interval spolehlivosti pro parametr $\psi(\theta_X)$ s pravděpodobností pokrytí $1-\alpha$.

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 $\theta_X = t(F_X) \in \mathbb{R}$.

Intervalový odhad parametru θ_X můžeme sestrojit postupem, který si zde popíšeme (pro případ konstrukce oboustranných intervalových odhadů):

1. Nalezneme funkci $\varphi(\boldsymbol{x}, \theta_X)$ takovou, že φ je prostá funkce 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 $\mathrm{P}\left[c_{\alpha/2} < Z_n < c_{1-\alpha/2}\right] = 1-\alpha$. Aplikací funkce $\bar{\varphi}(\boldsymbol{x}, \cdot)$ na obě nerovnosti (předpokládaje, že je rostoucí funkcí argumentu z) dostaneme

$$\mathrm{P}\left[\bar{\varphi}(\boldsymbol{X}, c_{\alpha/2}) < \theta_X < \bar{\varphi}(\boldsymbol{X}, c_{1-\alpha/2})\right] = 1 - \alpha.$$

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

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:

Odhad $\overline{X}_n$ je nestranný a konsistentní pro μ_X, odhad S_n^2 je nestranný a konsistentní pro $\sigma_X^2 \equiv \mathsf{var}\, X_i$. Z věty 1.5 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}).

Vyjdeme z rovnosti

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

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

$$\left(\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1}\Big(1-\frac{\alpha}{2}\Big),\ \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1}\Big(1-\frac{\alpha}{2}\Big)\right). \tag{2.1}$$

který má pravděpodobnost pokrytí přesně $1-\alpha$.

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 $\sigma_X^2 \equiv \mathsf{var}\, X_i$. Z věty 1.4 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 rovnosti

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

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

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

Proto interval (2.1), který byl přesným intervalem spolehlivosti pro μ_X u výběru z normálního rozdělení, je zároveň přibližným intervalem spolehlivosti pro μ_X pro data pocházející z jakéhokoli rozdělení s konečným rozptylem.

Příklad (alternativní rozdělení)**.** Vezměme si problém intervalového odhadu pravděpodobnosti úspěchu v alternativním rozdělení.

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

Jelikož odhadujeme střední hodnotu, vyjdeme z bodového odhadu $\widehat{p}_n = \overline{X}_n$, který je nestranný a konsistentní (věta 1.1). Z centrální limitní věty (tvrzení P.7.10) 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 a nemonotonní funkcí p_X, proto by se odtud p_X špatně vyjadřovalo. Z konsistence $\widehat{p}_n$ a věty o spojité transformaci (tvrzení P.7.3) však 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

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

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 rovnosti

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

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

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

Získali jsme tedy požadovaný interval. Jeho krajní body jsou

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

a jeho 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 1.3, část (i), víme, že

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

Vezmeme tedy $Z_n = (n-1)S_n^2/\sigma_X^2$, $F_Z = \chi_{n-1}^2$ a $c_\alpha = \chi_{n-1}^2(\alpha)$ (α-kvantil rozdělení χ_{n-1}^2).

Vyjdeme z rovnosti

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

a postupnými úpravami dojdeme k

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

Získali jsme tedy interval spolehlivosti pro rozptyl σ_X^2 s pravděpodobností pokrytí $1-\alpha$. Jeho krajní body jsou

$$\left(\frac{(n-1)S_n^2}{\chi_{n-1}^2(1-\alpha/2)}, \frac{(n-1)S_n^2}{\chi_{n-1}^2(\alpha/2)}\right). \tag{2.2}$$

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

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

3 Metody pro odhadování parametrů

3.1 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.

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 3.1. Funkci $\widehat{F}_n(u) \overset{\mathsf{df}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,u\rangle}(X_i)$ 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. $\widehat{F}_n$ 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.

Věta 3.1 (vlastnosti empirické distribuční funkce)**.** Pro libovolné $u \in \mathbb{R}$ platí

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

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

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

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

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

Poznámka. Z bodu (iv) předchozí věty lze odvodit přibližný interval spolehlivosti pro $F_X(u)$. Interval se odvodí stejně jako v příkladě na interval spolehlivosti pro parametr alternativního rozdělení (viz str. 14).

[*] Angl. *empirical distribution function*

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 $\widehat{\theta}_n \stackrel{\mathsf{df}}{=} t(\widehat{F}_n)$. 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

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

kde jsme využili toho, že $\mathbb{I}_{\langle X_i,\infty)}(x)$ 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.

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 1.2, 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

$$\widehat{\gamma}_1 = \frac{\widehat{\mu}_3}{(\widehat{\sigma}_n^2)^{3/2}},$$

empirický odhad špičatosti je

$$\widehat{\gamma}_2 = \frac{\widehat{\mu}_4}{\widehat{\sigma}_n^4}.$$

Oba jsou konsistentní (z věty o spojité transformaci).

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 3.2 (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.

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

- 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.

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

Věta 3.2. 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) $\sqrt{n}\big[\widehat{u}_n(\alpha) - u_X(\alpha)\big] \xrightarrow{\mathsf{D}} \mathsf{N}(0, V(\alpha))$, 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 3.2 se používá následující lemma, které se odvodí snadnou aplikací věty o transformaci náhodného vektoru (tvrzení P.5.4).

Lemma 3.3. 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 $\mathsf{B}(k, n-k+1)$.

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 výběrový průměr

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

Empirickým odhadem Σ je *výběrová rozptylová matice*[*]

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

[*] Angl. *sample covariation matrix*

Tvrzení 3.4.

- Je-li $\mathsf{E}\,|X_{ij}| < \infty$, pak $\mathsf{E}\,\overline{\boldsymbol{X}}_n = \boldsymbol{\mu}$ a $\overline{\boldsymbol{X}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\mu}$.
- Je-li $\mathsf{var}\,X_{ij} < \infty$, pak $\mathsf{E}\,\widehat{\Sigma}_n = \Sigma$ a $\widehat{\Sigma}_n \xrightarrow{\mathsf{P}} \Sigma$.

Poznámka.

- $\widehat{\Sigma}_n$ má na diagonále odhady rozptylu jednotlivých složek $\boldsymbol{X}_i$, tj.

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

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

- $\widehat{\Sigma}_n$ má mimo diagonálu odhady kovariancí dvojic složek $\boldsymbol{X}_i$, tj.

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

  pro $j = 1, \ldots, k$ a $m = 1, \ldots, k$, $j \neq m$. Těmto odhadům $\mathsf{cov}\,(X_{ij}, X_{im})$ říkáme *výběrové kovariance.*

- $\widehat{\Sigma}_n$ má všechny vlastnosti rozptylové matice, např. je positivně semidefinitní.
- Platí

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

Definice 3.3. *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.

- $-1 \leq \widehat{\varrho}_{jm} \leq 1$.
- $\widehat{\varrho}_{jm}$ je konsistentní odhad korelačního koeficientu $\varrho(X_{ij}, X_{im})$.
- $\widehat{\varrho}_{jm}$ není nestranný.

[*] Angl. *sample correlation coefficient*

3.2 Momentová metoda

Uvažujme nyní parametrický model: máme náhodný výběr $X_1, \dots, X_n$ 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

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

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 a odhadneme pomocí $V(\widehat{\theta}_n)$.

Příklady.

1. $X_1, \dots, X_n$ 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. $X_1, \dots, X_n$ 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. $X_1, \dots, X_n$ 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 1.2, 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 1.2 a Δ-metody.

Příklady.

1. $X_1, \ldots, X_n$ je náhodný výběr z gama rozdělení s parametry a a p. Momentovou metodou dostaneme konsistentní a asymptoticky normální odhady
$$\widehat{a} = \frac{\overline{X}_n}{S_n^2} \quad \text{a} \quad \widehat{p} = \frac{\overline{X}_n^2}{S_n^2}.$$
2. $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{R}(\theta_1, \theta_2)$. Momentovou metodou dostaneme konsistentní a asymptoticky normální odhady
$$\widehat{\theta}_1 = \overline{X}_n - \sqrt{3S_n^2} \quad \text{a} \quad \widehat{\theta}_2 = \overline{X}_n + \sqrt{3S_n^2}.$$
3. $X_1, \ldots, X_n$ 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 $S_n^2 < \overline{X}_n(1-\overline{X}_n)$).

3.3 Metoda maximální věrohodnosti

Stále pracujeme s náhodným výběrem $\boldsymbol{X} = (X_1, \ldots, X_n)$ z rozdělení s hustotou $f(x|\boldsymbol{\theta}_X)$ vzhledem k míře μ a parametrickým modelem
$$\mathcal{F} = \{\text{rozdělení s hustotou } f(x|\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}.$$
Předpokládáme, že pro libovolné $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ platí $f(x|\boldsymbol{\theta}_1) \neq f(x|\boldsymbol{\theta}_2)$. Tomuto požadavku se říká *identifikovatelnost modelu*[*].

Princip maximální věrohodnosti

Náhodný výběr $X_1, \ldots, X_n$ má sdruženou hustotu $\prod_{i=1}^n f(x_i|\boldsymbol{\theta}_X)$. Maximálně věrohodný odhad $\widehat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}_X$ je takový bod z Θ, který maximalisuje (přes všechny $\boldsymbol{\theta} \in \Theta$) sdruženou hustotu spočítanou v pozorovaných hodnotách $X_1, \ldots, X_n$.

Definice 3.4 (věrohodnost, maximálně věrohodný odhad)**.**

- Náhodnou funkci
$$L_n(\boldsymbol{\theta}) \overset{\mathsf{df}}{=} \prod_{i=1}^n f(X_i|\boldsymbol{\theta})$$
nazýváme *věrohodnostní funkcí* (zkráceně věrohodností)[†] pro parametr $\boldsymbol{\theta}$ v modelu $\mathcal{F}$.

[*] Angl. *model identifiability* [†] Angl. *likelihood function*

- *Maximálně věrohodný odhad*[*] parametru $\boldsymbol{\theta}_X$ v modelu $\mathcal{F}$ je definován jako

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta}).$$

- Náhodnou funkci

$$\ell_n(\boldsymbol{\theta}) \overset{\mathsf{df}}{=} \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(X_i|\boldsymbol{\theta})$$

  nazýváme *logaritmickou věrohodností*[†].

Poznámka. Jelikož logaritmus je ryze rostoucí funkce, $L_n(\boldsymbol{\theta})$ a $\ell_n(\boldsymbol{\theta})$ nabývají maxima v tomtéž bodě.

Značení. Označme $S_{\boldsymbol{\theta}} = \{x \in \mathbb{R} : f(x|\boldsymbol{\theta}) > 0\}$ nosič rozdělení z modelu $\mathcal{F}$ při hodnotě parametru $\boldsymbol{\theta}$.

Tvrzení 3.5. Pro každé $\boldsymbol{\theta} \neq \boldsymbol{\theta}_X$ platí

$$\mathrm{P}\left[\ell_n(\boldsymbol{\theta}_X) > \ell_n(\boldsymbol{\theta})\right] \to 1 \quad \text{při } n \to \infty.$$

Při vzrůstajícím počtu pozorování bude s velmi velkou pravděpodobností hodnota (logaritmické) věrohodnosti ve skutečném parametru větší než v jakémkoli jiném parametru.

Výpočet maximálně věrohodných odhadů

Maximálně věrohodný odhad obvykle hledáme diferenciací logaritmické věrohodnosti. Abychom našli maximum, položíme první derivaci rovnou nule a ověříme, že druhá derivace je záporná (negativně definitní, pokud model obsahuje více než jeden parametr).

Definice 3.5 (skóre, informace)**.**

- Náhodný vektor

$$\boldsymbol{U}(\boldsymbol{\theta}|X_i) \overset{\mathsf{df}}{=} \frac{\partial}{\partial\boldsymbol{\theta}} \log f(X_i|\boldsymbol{\theta})$$

  nazýváme *skórovou funkcí*[‡] pro parametr $\boldsymbol{\theta}$ v modelu $\mathcal{F}$.
- Náhodný vektor

$$\boldsymbol{U}_n(\boldsymbol{\theta}|\boldsymbol{X}) \overset{\mathsf{df}}{=} \sum_{i=1}^{n} \boldsymbol{U}(\boldsymbol{\theta}|X_i) = \sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\theta}} \log f(X_i|\boldsymbol{\theta})$$

  nazýváme *skórovou statistikou*[§].

[*] Angl. *maximum likelihood estimator* [†] Angl. *log-likelihood* [‡] Angl. *score function* [§] Angl. *score statistic*

- Náhodnou matici

$$I(\boldsymbol{\theta}|X_i) \stackrel{\mathsf{df}}{=} -\frac{\partial}{\partial \boldsymbol{\theta}^\mathsf{T}} \boldsymbol{U}(\boldsymbol{\theta}|X_i) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}} \log f(X_i|\boldsymbol{\theta})$$

  nazýváme příspěvkem i-tého pozorování do informační matice.
- Náhodnou matici

$$I_n(\boldsymbol{\theta}|\boldsymbol{X}) \stackrel{\mathsf{df}}{=} -\frac{1}{n} \frac{\partial}{\partial \boldsymbol{\theta}^\mathsf{T}} \boldsymbol{U}_n(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{\theta}|X_i)$$

  nazýváme *pozorovanou informační maticí*[*].
- Matici

$$I(\boldsymbol{\theta}) \stackrel{\mathsf{df}}{=} \mathsf{E}\, I(\boldsymbol{\theta}|X_i) = -\mathsf{E}\, \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}} \log f(X_i|\boldsymbol{\theta})$$

  nazýváme *očekávanou (Fisherovou) informační maticí*[†].

Je-li množina Θ otevřená, maximálně věrohodný odhad $\widehat{\boldsymbol{\theta}}_n$ řeší soustavu rovnic $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) = \mathbf{0}$ neboli

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i|\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}.$$

Této soustavě se říká *věrohodnostní rovnice*[‡].

Řešení věrohodnostní rovnice se často musí hledat numericky. Řešení však nemusí existovat nebo může existovat více řešení a ne každé z nich je nutně maximálně věrohodný odhad. Pokud platí $I_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) > 0$ (pozorovaná informace je positivně definitní v bodě $\widehat{\boldsymbol{\theta}}_n$), víme, že $\widehat{\boldsymbol{\theta}}_n$ je alespoň lokální maximum. Je-li $I_n(\boldsymbol{\theta}|\boldsymbol{X}) > 0$ pro každé $\boldsymbol{\theta} \in \Theta$, věrohodnostní funkce je konkávní a řešení věrohodnostní rovnice musí být hledaným globálním maximem.

Příklady.

1. Data: $X_1, \ldots, X_n \sim \mathsf{Exp}(\lambda_X)$

   Model: $\mathcal{F} = \{\mathsf{Exp}(\lambda), \lambda > 0\}$

   Odhadovaný parametr: λ_X

   Výsledek: Maximálně věrohodný odhad $\widehat{\lambda}_n$ parametru λ_X jest $\frac{1}{\overline{X}_n}$.

2. Data: $X_1, \ldots, X_n \sim \mathsf{Alt}(p_X)$

   Model: $\mathcal{F} = \{\mathsf{Alt}(p), p > 0\}$

   Odhadovaný parametr: p_X

   Výsledek: Maximálně věrohodný odhad $\widehat{p}_n$ parametru p_X jest $\overline{X}_n$.

[*] Angl. *observed information matrix* [†] Angl. *expected (Fisher) information matrix* [‡] Angl. *likelihood equation*

3. Data: $X_1, \ldots, X_n \sim \mathsf{N}(\mu_X, \sigma_X^2)$

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

   Odhadované parametry: $\mu_X = \mathsf{E}\, X_i$, $\sigma_X^2 = \mathsf{var}\, X_i$

   Výsledek: Maximálně věrohodný odhad parametru μ_X jest $\overline{X}_n$. Maximálně věrohodný odhad parametru σ_X^2 jest $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

4. Data: $X_1, \ldots, X_n \sim \Gamma(a_X, p_X)$

   Model: $F_X \in \mathcal{F} = \{\Gamma(a, p), a,\ p > 0\}$

   Odhadované parametry: a_X, p_X

   Výsledek: Maximálně věrohodný odhad $\widehat{p}_n$ parametru p_X řeší rovnici

$$\log \widehat{p}_n - \frac{\Gamma'(\widehat{p}_n)}{\Gamma(\widehat{p}_n)} = \log \frac{\overline{X}_n}{\sqrt[n]{\prod X_i}}$$

   Maximálně věrohodný odhad parametru a_X jest $\widehat{a}_n = \frac{\widehat{p}_n}{\overline{X}_n}$.

Vlastnosti maximálně věrohodných odhadů

Maximálně věrohodné odhady jsou konsistentní a asymptoticky normální, pokud platí podmínky, kterým se v tomto kontextu říká *podmínky regularity*.

Podmínky (podmínky regularity pro maximálně věrohodné odhady)**.**

R1. Počet parametrů d v modelu $\mathcal{F}$ je konstantní.

R2. Nosič rozdělení $S_{\boldsymbol{\theta}} = \{x \in \mathbb{R} : f(x|\boldsymbol{\theta}) > 0\}$ nezávisí na parametru $\boldsymbol{\theta}$; všechna rozdělení v modelu $\mathcal{F}$ mají stejný nosič.

R3. Parametrický prostor Θ je otevřená množina.

R4. Informační matice je konečná a positivně definitní v okolí $\boldsymbol{\theta}_X$.

R5. Hustota $f(x|\boldsymbol{\theta})$ je dostatečně hladká funkce $\boldsymbol{\theta}$ (aspoň dvakrát spojitě diferencovatelná).

R6. Lze prohodit pořadí derivace a integrálu ve výrazech

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int h(x, \boldsymbol{\theta})\, d\mu(x) = \int \frac{\partial}{\partial \boldsymbol{\theta}} h(x, \boldsymbol{\theta})\, d\mu(x),$$

   kde $h(x, \boldsymbol{\theta})$ je buď $f(x|\boldsymbol{\theta})$ nebo $\partial f(x|\boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

U všech následujících tvrzení předpokládáme, že podmínky R1 – R6 jsou splněny.

Poznámka. Vezměme rovnost

$$\int_{-\infty}^{\infty} f(x|\boldsymbol{\theta})\, d\mu(x) = 1$$

a derivujme postupně dvakrát levou i pravou stranu podle $\boldsymbol{\theta}$. Z podmínky R6 plyne, že

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} f(x|\boldsymbol{\theta})\, d\mu(x) = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} f(x|\boldsymbol{\theta})\, d\mu(x) = \mathbf{0} \tag{3.1}$$

Věta 3.6 (konsistence maximálně věrohodného odhadu)**.** Existuje posloupnost řešení $\widehat{\boldsymbol{\theta}}_n$ věrohodnostní rovnice $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) = \mathbf{0}$ taková, že $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_X$.

Poznámka. Je-li logaritmická věrohodnost ryze konkávní, věrohodnostní rovnice má právě jedno řešení a to představuje konsistentní odhad. Není-li logaritmická věrohodnost ryze konkávní, věrohodnostní rovnice může mít více řešení; mezi nimi je jedno (to nejbližší k $\boldsymbol{\theta}_X$), které představuje konsistentní odhad. Ostatní řešení nemusí být blízko $\boldsymbol{\theta}_X$ a nemusí k němu konvergovat.

Věta 3.7 (vlastnosti skórové funkce a skórové statistiky)**.**

(i) $\mathsf{E}\, \boldsymbol{U}(\boldsymbol{\theta}_X|X_i) = 0$, $\mathsf{var}\, \boldsymbol{U}(\boldsymbol{\theta}_X|X_i) = I(\boldsymbol{\theta}_X)$.

(ii) $\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\theta}_X|\boldsymbol{X}) \xrightarrow{\mathsf{D}} \mathsf{N}_d(\mathbf{0}, I(\boldsymbol{\theta}_X))$.

Poznámka. Existují dva způsobu výpočtu Fisherovy informační matice: z definice 3.5 (minus střední hodnota druhé derivace logaritmu hustoty) anebo z věty 3.7 (rozptyl skórové funkce).

Věta 3.8 (asymptotická normalita maximálně věrohodného odhadu)**.** Nechť $\widehat{\boldsymbol{\theta}}_n$ je maximálně věrohodný odhad. Pak

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_d(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_X)).$$

Poznámka.

- Asymptotický rozptyl maximálně věrohodného odhadu je dán převrácenou hodnotou Fisherovy informace; čím větší informace tím větší přesnost odhadování.
- Asymptotický rozptyl maximálně věrohodného odhadu je v jistém smyslu optimální; odhady odvozené jinak, např. momentovou metodou, nemohou mít menší asymptotický rozptyl.

Věta 3.9 (asymptotické rozdělení věrohodnostního poměru)**.** Nechť $\widehat{\boldsymbol{\theta}}_n$ je maximálně věrohodný odhad. Pak

$$2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_X)) \xrightarrow{\mathsf{D}} \chi^2_d.$$

Věta 3.10 (transformace maximálně věrohodného odhadu)**.** Nechť $q : \Theta \to \mathbb{R}^k$ je spojitě diferencovatelná funkce. Označme $\boldsymbol{\nu}_X = q(\boldsymbol{\theta}_X)$ a $D(\boldsymbol{\theta}) = \partial q(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$. Pak $\widehat{\boldsymbol{\nu}}_n = q(\widehat{\boldsymbol{\theta}}_n)$ je maximálně věrohodný odhad parametru $\boldsymbol{\nu}_X$ a platí

$$\sqrt{n}(\widehat{\boldsymbol{\nu}}_n - \boldsymbol{\nu}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_k(\mathbf{0}, D(\boldsymbol{\theta}_X) I^{-1}(\boldsymbol{\theta}_X) D(\boldsymbol{\theta}_X)^{\mathsf{T}}).$$

Příklady (Najděte maximálně věrohodné odhady a určete jejich asymptotické rozdělení)**.**

Data: $X_1, \ldots, X_n$ jsou náhodný výběr z rozdělení s hustotou $f(x|\theta_X)$

1. $f(x|\theta) = \theta(1-x)^{\theta-1}\mathbb{I}_{(0,1)}(x), \quad \theta > 1.$

   Postup řešení:

$$\begin{aligned}
L_n(\theta) &= \theta^n \prod_{i=1}^{n} (1 - X_i)^{\theta-1} \\
\ell_n(\theta) &= n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log(1 - X_i) \\
U(\theta|X_i) &= \frac{1}{\theta} + \log(1 - X_i) \\
U_n(\theta|\boldsymbol{X}) &= \frac{n}{\theta} + \sum_{i=1}^{n} \log(1 - X_i) \\
\widehat{\theta}_n &= -\left[\frac{1}{n} \sum_{i=1}^{n} \log(1 - X_i)\right]^{-1} \\
I(\theta|X_i) &= \frac{1}{\theta^2} \\
I_n(\theta|\boldsymbol{X}) &= \frac{n}{\theta^2} > 0 \\
I(\theta_X) &= \frac{1}{\theta_X^2}
\end{aligned}$$

   Asymptotické rozdělení $\widehat{\theta}_n$ je

$$\sqrt{n}(\widehat{\theta}_n - \theta_X) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \theta_X^2)$$

2. $f(x|\theta) = \frac{1}{\sqrt{2\pi\theta}} \mathrm{e}^{-\frac{(x-\theta)^2}{2\theta}}, \quad \theta > 0.$

   Postup řešení:

$$
\begin{aligned}
L_n(\theta) &= (2\pi\theta)^{-n/2}\mathrm{e}^{-\sum_{i=1}^{n}\frac{(X_i-\theta)^2}{2\theta}} \\
\ell_n(\theta) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\theta - \sum_{i=1}^{n}\frac{(X_i-\theta)^2}{2\theta} \\
U(\theta|X_i) &= \frac{X_i^2}{2\theta^2} - \frac{1}{2\theta} - \frac{1}{2} \\
U_n(\theta|\boldsymbol{X}) &= \frac{1}{2\theta^2}\sum_{i=1}^{n}X_i^2 - \frac{n}{2\theta} - \frac{n}{2} \\
\widehat{\theta}_n &= \sqrt{\widehat{\mu}_2' + \frac{1}{4}} - \frac{1}{2} \\
I(\theta|X_i) &= \frac{X_i^2}{\theta^3} - \frac{1}{2\theta^2} \\
I_n(\theta|\boldsymbol{X}) &= \frac{\widehat{\mu}_2'}{\theta^3} - \frac{1}{2\theta^2} \\
I_n(\widehat{\theta}_n|\boldsymbol{X}) &= \frac{1}{\widehat{\theta}_n} + \frac{1}{2\widehat{\theta}_n^2} > 0 \\
I(\theta_X) &= \frac{2\theta_X + 1}{2\theta_X^2}
\end{aligned}
$$

Asymptotické rozdělení $\widehat{\theta}_n$ je

$$\sqrt{n}(\widehat{\theta}_n - \theta_X) \xrightarrow{\mathsf{D}} \mathsf{N}\big(0, \frac{2\theta_X^2}{2\theta_X + 1}\big)$$

4 Principy testování hypotéz

4.1 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 $\mathcal{F}$. Označme skutečný parametr jako $\boldsymbol{\theta}_X = t(F_X)$. Označme celá napozorovaná data symbolem $\boldsymbol{X} = (\boldsymbol{X}_1^\mathsf{T}, \ldots, \boldsymbol{X}_n^\mathsf{T})^\mathsf{T}$.

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 $\boldsymbol{\theta}_X \in \Theta_1$. (Většinou bereme $\Theta_1 = \Theta_0^\mathsf{c}$, ale to není naprosto nutné.)

Definice 4.1 (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 $t(F_0) = \boldsymbol{\theta}_0$. Jednoduchou hypotézu značíme $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$.
- Označme $\mathcal{F}_1 \overset{\mathsf{df}}{=} \{F \in \mathcal{F} : t(F) \in \Theta_1\}$, 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 $t(F_1) = \boldsymbol{\theta}_1$. Jednoduchou alternativu značíme $H_1 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_1$.

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 $\mathcal{C}$, které říkáme *kritický obor*. Testová statistika je obvykle jednorozměrná; kritický obor je pak nějaká podmnožina $\mathbb{R}$. Rozhodujeme se podle to, jestli testová statistika padne do kritického oboru, či nikoli.

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

Definice 4.2 (Test)**.** *Statistický test* je definován pomocí testové statistiky $S(\boldsymbol{X})$ a kritického oboru $\mathcal{C}$. Dva testy $(S(\boldsymbol{X}), \mathcal{C})$ a $(S^*(\boldsymbol{X}), \mathcal{C}^*)$ nazveme *ekvivalentní* právě když $S(\boldsymbol{X}) \in \mathcal{C} \Leftrightarrow S^*(\boldsymbol{X}) \in \mathcal{C}^*$ skoro jistě, tj. oba testy vydávají s pravděpodobností 1 totéž rozhodnutí.

Poznámka. Testovou statistiku volíme tak, aby její rozdělení bylo citlivé na hodnotu testovaného parametru, ale aby co nejméně záviselo na těch charakteristikách rozdělení $F \in \mathcal{F}$, které testovat nechceme. Proto 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

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

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

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

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

4.2 Hladina testu a síla testu

Definice 4.3 (Hladina testu)**.** Nechť $\alpha \in (0, 1)$ je předem stanovené číslo. Jestliže kritický obor $\mathcal{C}$ splňuje podmínku

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

(pravděpodobnost, že testová statistika padne do kritického oboru, mají-li data rozdělení F splňující nulovou hypotézu), říkáme, že test $(S(\boldsymbol{X}), \mathcal{C})$ má *hladinu*[*] α.

Jestliže kritický obor $\mathcal{C}$ splňuje podmínku

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

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

[*] Angl. *level*

Poznámka.

- Je-li množina $\Theta_0 = \{\boldsymbol{\theta}_0\}$ jednobodová, pak můžeme (asymptotickou) hladinu testu zapsat jednodušeji:

$$\alpha = \lim_{n\to\infty} \mathrm{P}_{\theta_0}[S(\boldsymbol{X}) \in \mathcal{C}]\,.$$

- Hladina testu je pravděpodobnost zamítnutí platné hypotézy (pokud je hypotéza jednoduchá) nebo maximalizovaná platnost zamítnutí platné hypotézy (pokud je hypotéza složená).
- Připouštíme pouze ty testy, které mají požadovanou hladinu, nebo jí dosahují alespoň přibližně při velkém rozsahu výběru n.
- Hladina se obvykle volí malá, v praxi je standartem $\alpha = 0.05$.
- Abychom mohli dodržet stanovenou hladinu, musíme být schopni spočítat přesné nebo asymptotické rozdělení testové statistiky za platnosti nulové hypotézy, a to nesmí záviset na neznámých charakteristikách rozdělení F_X.
- U některých testů (přesné testy s diskrétní testovou statistikou) není možné dosáhnout zcela libovolné hladiny — pak se většinou spokojujeme s nižší hladinou nejbližší k té, kterou bychom normálně požadovali.

Definice 4.4 (Síla testu)**.** Nechť $\boldsymbol{\theta} \in \Theta_1$. Pak

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

(pravděpodobnost, že testová statistika padne do kritického oboru, mají-li data rozdělení F porušující nulovou hypotézu) 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. Funkci $\beta(\boldsymbol{\theta})$ můžeme snadno rozšířit i na $\boldsymbol{\theta} \in \Theta_0$. Má-li test hladinu α, pak musí platit $\sup_{\boldsymbol{\theta}\in\Theta_0} \beta(\boldsymbol{\theta}) = \alpha$ (nebo $\to \alpha$ pro $n \to \infty$).

Definice 4.5 (Nestranný test)**.** Nechť test $(S(\boldsymbol{X}), \mathcal{C})$ má hladinu α a sílu $\beta(\boldsymbol{\theta})$. Test nazveme [asymptoticky] *nestranný*†, pokud pro každé $\boldsymbol{\theta} \in \Theta_1$ platí $\beta(\boldsymbol{\theta}) \geq \alpha$ $[\lim_{n\to\infty} \beta(\boldsymbol{\theta}) \geq \alpha]$.

Poznámka.

- Testy, které nejsou nestranné, nebudeme připouštět. Např. test, který vždy zamítne H_0 s pravděpodobností α zcela nezávisle na datech, je nestranný ($\beta(\boldsymbol{\theta}) = \alpha$ splňuje požadavek nestrannosti).
- Rádi bychom maximalizovali sílu mezi všemi testy dosahujícími požadovanou hladinu. Většinou však není možné maximalizovat sílu pro všechny alternativy zároveň, zvlášť je-li model $\mathcal{F}$ bohatý.

* Angl. *power* † Angl. *unbiased*

Testovou statistiku volíme tak, aby
(i) její rozdělení bylo co nejcitlivější 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* volíme tak, aby
(i) zahrnoval hodnoty testové statistiky, které jsou za platnosti hypotézy méně pravděpodobné než za platnosti alternativy;
(ii) byla dodržena požadovaná hladina testu.

Kritický obor $\mathcal{C}$ 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 $|S(\boldsymbol{X})|$.

Příklad (Test střední hodnoty normálního rozdělení se známým rozptylem)**.**

Data: $X_1, \ldots, X_n \sim \mathsf{N}(\mu_X, \sigma_0^2)$
Model: $\mathcal{F} = \{\mathsf{N}(\mu, \sigma_0^2), \mu \in \mathbb{R}, \sigma_0^2 \text{ známé}\}$
Problém: $H_0 : \mu_X = \mu_0$ proti $H_1 : \mu_X \neq \mu_0$

Testová statistika:

$$S(\boldsymbol{X}) = \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{\sigma_0}$$

Kritický obor: $(-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$

Síla testu: proti alternativě $\mu_1 = \mu_0 + \delta$, $\delta > 0$

$$\beta(\mu_1) \approx 1 - \Phi\left(u_{1-\alpha/2} - \sqrt{n}\,\frac{\delta}{\sigma}\right)$$

Požadovaný rozsah výběru: pro dosažení síly alespoň β proti alternativě $\mu_1 = \mu_0 + \delta$

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

Poznámka. Síla testu $(S(\boldsymbol{X}), \mathcal{C})$ závisí na
- hladině testu α
- alternativě $\boldsymbol{\theta}$, respektive její vzdálenosti od hypotézy Θ_0
- počtu pozorování n
- rozptylu pozorování $\mathsf{var}\, X_i$

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. Hypotézu H_0 vyvracíme, prokázali jsme platnost alternativy H_1. Pravděpodobnost chybného rozhodnutí v případě, že hypotéza platí, je omezena shora hladinou α, která je malá.
- Skončí-li test tím, že *hypotézu* H_0 *nemůžeme zamítnout*, znamená to pouze, 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, není omezena shora a může tedy být značně velká.
- Hypotéza H_0 a alternativa H_1 při testování vystupují asymetricky. Hypotézu můžeme někdy vyvrátit ve prospěch alternativy, ale nemůžeme ji nikdy potvrdit.

4.3 P-hodnota

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

Uvažujme hypotézu $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ proti alternativě $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$ a test $(S(\boldsymbol{X}), \mathcal{C})$ s kritickým oborem tvaru $\mathcal{C} = \mathbb{R} \setminus (c_L, c_U)$, kde $-\infty \leq c_L < c_U \leq \infty$. Hodnoty $S(\boldsymbol{X})$ v intervalu (c_L, c_U) tedy považujeme za hodnoty v souladu s hypotézou, ty ostatní hypotéze protiřečí. Označme $s_{\boldsymbol{x}}$ realizovanou hodnotu testové statistiky $S(\boldsymbol{X})$, kterou jsme napozorovali pro náš datový soubor. Označme dále symbolem F_0 distribuční funkci testové statistiky $S(\boldsymbol{X})$ za platnosti nulové hypotézy (přesnou nebo asymptotickou); pro jednoduchost předpokládejme, že $S(\boldsymbol{X})$ má spojité rozdělení. Chceme rozhodnout, jestli pozorovaná hodnota $s_{\boldsymbol{x}}$ testové statistiky stačí k zamítnutí nulové hypotézy na hladině α.

Definice 4.6 (P-hodnota)**.** *P-hodnotu*[*] neboli *dosaženou hladinu testu* definujeme jako

- $p(\boldsymbol{x}) = \mathrm{P}_{\boldsymbol{\theta}_0}[S(\boldsymbol{X}) \geq s_{\boldsymbol{x}}] = 1 - F_0(s_{\boldsymbol{x}})$ pokud $c_L = -\infty$;
- $p(\boldsymbol{x}) = \mathrm{P}_{\boldsymbol{\theta}_0}[S(\boldsymbol{X}) \leq s_{\boldsymbol{x}}] = F_0(s_{\boldsymbol{x}})$ pokud $c_U = \infty$;
- $p(\boldsymbol{x}) = 2\min(\mathrm{P}_{\boldsymbol{\theta}_0}[S(\boldsymbol{X}) \geq s_{\boldsymbol{x}}], \mathrm{P}_{\boldsymbol{\theta}_0}[S(\boldsymbol{X}) \leq s_{\boldsymbol{x}}]) = 2\min(1 - F_0(s_{\boldsymbol{x}}), F_0(s_{\boldsymbol{x}}))$ pokud c_L a c_U jsou konečné a $F_0(c_L) = 1 - F_0(c_U) = \alpha/2$.

Poznámka.

- P-hodnota je funkcí pozorovaných hodnot $\boldsymbol{x} = (x_1, \ldots, x_n)$ náhodného výběru $\boldsymbol{X} = (X_1, \ldots, X_n)$.

[*] Angl. *p-value*

- P-hodnotu můžeme slovně popsat jako pravděpodobnost, že bychom za platnosti hypotézy napozorovali data, která by byla s hypotézou ve stejném nebo větším rozporu, než analyzovaný náhodný výběr.
- 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 spočítat jako $p(\boldsymbol{x}) = \mathrm{P}_{\boldsymbol{\theta}_0}[|S(\boldsymbol{X})| \geq |s_{\boldsymbol{x}}|] = 2[1 - F_0(|s_{\boldsymbol{x}}|)]$.
- Je-li distribuční funkce F_0 asymptotická, přidáme před výraz definující p-hodnotu ještě $\lim_{n\to\infty}$.
- Testujeme-li hypotézu $H_0 : \boldsymbol{\theta}_X \in \Theta_0$, kde $\Theta_0 \neq \emptyset$ není jednobodová množina, přidáme před výraz definující p-hodnotu ještě $\sup_{\boldsymbol{\theta}\in\Theta_0}$.

Tvrzení 4.1. Zamítáme-li hypotézu podle pravidla

$$H_0 \text{ zamítáme, jestliže } p(\boldsymbol{x}) \leq \alpha$$
$$H_0 \text{ nezamítáme, jestliže } p(\boldsymbol{x}) > \alpha,$$

výsledný test má hladinu α (přesně nebo asymptoticky).

Poznámka.
- 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-hodnota do jisté míry vyjadřuje, s jakou rezervou k zamítnutí hypotézy došlo.
- Mezi laiky rozšířená představa o p-hodnotě jakožto *„pravděpodobnosti, že nulová hypotéza platí“* je zcela mylná a nesmyslná.

4.4 Intervalové odhady 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, který nás zajímá a $\theta_X = t(F_X)$ je jeho skutečná hodnota. V kapitole 2.2 jsme se zabývali problémem intervalového odhadu θ_X, tj. nalezení náhodných veličin C_L a C_U takových, že $\mathrm{P}[(C_L, C_U) \ni \theta_X] = 1 - \alpha$ (nebo $\to 1 - \alpha$). Nyní se zabýváme testováním; snažíme se rozhodnout, zdali θ_X nabývá nějaké zadané hodnoty θ_0 či nikoli. Oba problémy se řeší postupem, který vypadá na pohled dosti podobně, ale liší se v detailech – obě úlohy jsou principiálně odlišné. Nicméně mezi testováním hypotézy o parametru a intervalovým odhadem pro parametr existuje jakási dualita, kterou je dobré si uvědomovat a rozumět jí.

Tvrzení 4.2 (Ekvivalence intervalových odhadů a testování)**.**
1. 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 má výsledný test hladinu α (přesně nebo asymptoticky).

2. 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 $\mathrm{P}\left[B_{\boldsymbol{X}} \ni \theta_X\right] = 1 - \alpha$ (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).

4.5 Asymptotické testy založené na metodě maximální věrohodnosti

Vlastnosti maximálně věrohodných odhadů uvedené v kapitole 3.3 lze použít k odvozování asymptotických testů v parametrických modelech. Ukážeme si, jak takové testy vypadají.

Pracujeme s náhodným výběrem $X_1, \ldots, X_n$ z rozdělení* s hustotou $f(x|\boldsymbol{\theta}_X)$ vzhledem k míře μ a parametrickým modelem

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

Předpokládáme, že platí všechny podmínky zaručující platnost vět z kapitoly 3.3, tj. identifikovatelnost modelu a podmínky regularity R1 – R6 uvedené na str. 26.

Testování jednoduchých hypotéz

Chceme testovat hypotézu $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ proti $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$, kde $\boldsymbol{\theta}_0 \in \Theta$. Jde o jednoduchou hypotézu, neboť v modelu $\mathcal{F}$ je právě jedno rozdělení s hustotou $f(x|\boldsymbol{\theta}_0)$. Hypotéza H_0 znamená, že všechny parametry modelu nabývají nějakých předurčených hodnot obsažených ve vektoru $\boldsymbol{\theta}_0$.

Příklad.

$$\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$$
$$\boldsymbol{\theta}_X = (\mu_X, \sigma_X^2)^{\mathsf{T}}, \boldsymbol{\theta}_0 = (0, 1)^{\mathsf{T}}$$
$$H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0, \quad \text{tj. } \mu_X = 0 \text{ a } \sigma_X^2 = 1$$

* může jít i o výběr náhodných vektorů, ale ve značení to nezohledňujeme.

Pozorujeme data z nějakého normálního rozdělení a zajímá nás, jestli pocházejí z normovaného normálního rozdělení $\mathsf{N}(0,1)$.

Definujme si tři testové statistiky. Používáme značení zavedené v kapitole 3.3.

Definice 4.7.

(i) Statistika
$$\lambda_n = \frac{L_n(\widehat{\boldsymbol{\theta}}_n)}{L_n(\boldsymbol{\theta}_0)}$$
se nazývá *věrohodnostní poměr*[*].

(ii) Statistika
$$W_n = n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\mathsf{T} \widehat{I}_n(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$
se nazývá *Waldova statistika*[†].

(iii) Statistika
$$R_n = \frac{1}{n}\boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})^\mathsf{T} \widehat{I}_n^{-1}(\boldsymbol{\theta}_0)\boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})$$
se nazývá *Raova (skórová) statistika*[‡].

Poznámka. Symbolem $\widehat{I}_n$ rozumíme nějaký konsistentní odhad Fisherovy informační matice. Můžeme si vybrat jednu ze tří možností:

1. $\widehat{I}_n(\boldsymbol{\theta}) = I_n(\boldsymbol{\theta}|\boldsymbol{X}) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} \log f(\boldsymbol{\theta}|X_i)$ (výběrová informační matice)
2. $\widehat{I}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{U}(\boldsymbol{\theta}|X_i)^{\otimes 2}$ (výběrový rozptyl skórové funkce)
3. $\widehat{I}_n(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$ (Fisherova informační matice)

Do Waldovy statistiky se obvykle dosazuje $\widehat{I}_n(\widehat{\boldsymbol{\theta}}_n) = I_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X})$. Do Raovy statistiky se obvykle dosazuje $\widehat{I}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{U}(\boldsymbol{\theta}_0|X_i)^{\otimes 2}$.

Poznámka.

- Věrohodnostní poměr vyžaduje výpočet $\widehat{\boldsymbol{\theta}}_n$ a L_n nebo ℓ_n. Nevyžaduje výpočet $\boldsymbol{U}_n$ a $\widehat{I}_n$.
- Waldova statistika vyžaduje výpočet $\widehat{\boldsymbol{\theta}}_n$ a $\widehat{I}_n$. Nevyžaduje výpočet L_n a $\boldsymbol{U}_n$.
- Raova statistika vyžaduje výpočet $\boldsymbol{U}_n$ a $\widehat{I}_n$. Nevyžaduje výpočet $\widehat{\boldsymbol{\theta}}_n$ a L_n.

Poznámka. Je-li $d = 1$ (jeden parametr) a $\theta_0 = 0$, pak lze Waldovu statistiku přepsat jako
$$W_n = \left[\frac{\widehat{\theta}_n}{\sqrt{n^{-1}\widehat{I}_n^{-1}(\widehat{\theta}_n)}}\right]^2,$$
kde $n^{-1}\widehat{I}_n^{-1}(\widehat{\theta}_n)$ je odhad asymptotického rozptylu $\widehat{\theta}_n$.

[*] Angl. *likelihood ratio* [†] Angl. *Wald statistic* [‡] Angl. *Rao (score) statistic*

Věta 4.3. Nechť je splněna hypotéza $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$. Pak platí:

(i)
$$2 \log \lambda_n = 2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_0)) \xrightarrow{\mathsf{D}} \chi^2_d;$$

(ii)
$$W_n \xrightarrow{\mathsf{D}} \chi^2_d;$$

(iii)
$$R_n \xrightarrow{\mathsf{D}} \chi^2_d.$$

Poznámka. Platí-li H_0, očekáváme, že $\widehat{\boldsymbol{\theta}}_n$ je blízko $\boldsymbol{\theta}_0$, $L_n(\widehat{\boldsymbol{\theta}}_n)$ je blízko $L_n(\boldsymbol{\theta}_0)$ a $\boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})$ je blízko $\mathbf{0}$. Za platnosti H_0 mají všechny tři testové statistiky hodnoty blízko 0. Velké hodnoty statistik svědčí proti platnosti H_0.

Důsledek. Nechť $\chi^2_d(1-\alpha)$ je $(1-\alpha)$-kvantil rozdělení χ^2_d. Uvažujme testy hypotézy $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ proti proti $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$ dané pravidlem: H_0 zamítneme ve prospěch H_1, pokud

(i) $2 \log \lambda_n \geq \chi^2_d(1-\alpha)$ *(test poměrem věrohodností)*
(ii) $W_n \geq \chi^2_d(1-\alpha)$ *(Waldův test)*
(iii) $R_n \geq \chi^2_d(1-\alpha)$ *(skórový test)*

Potom každý z těchto tří testů má asymptoticky (pro $n \to \infty$) hladinu α.

Poznámka. Všechny tři testy jsou asymptoticky ekvivalentní; pro velké rozsahy výběru dávají velmi podobné výsledky. V menších rozsazích výběru se jejich výsledky mohou lišit. V takových situacích je nevhodnější test poměrem věrohodností, méně vhodný je skórový test, nejméně vhodný je Waldův test.

Je třeba si vybrat jeden z těchto testů *předtím* než spočítáme jejich testové statistiky. Není možně spočítat všechny tři a vybrat si tu největší. Taková procedura by porušovala hladinu testu mnohem výrazněji, než kterýkoli ze tří testů provedený samostatně.

Testování složených hypotéz

Velmi často se setkáváme s úlohou otestovat jeden parametr v modelu, který obsahuje více parametrů, ale ty ostatní nás nezajímají.

Příklad.

$$\begin{aligned}
&\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\} \\
&\boldsymbol{\theta}_X = (\mu_X, \sigma^2_X)^{\mathsf{T}}, \\
&H_0^* : \mu_X = 0, \quad H_1^* : \mu_X \neq 0
\end{aligned}$$

Pozorujeme data z nějakého normálního rozdělení a zajímá nás, jestli mají nulovou střední hodnotu.

Nejedná se o testování jednoduché hypotézy, protože v modelu je mnoho rozdělení, která H_0^* splňují.

Rozdělme si parametr $\boldsymbol{\theta}$ na část $\boldsymbol{\theta}_A$ obsahující prvních m složek $\boldsymbol{\theta}$ a část $\boldsymbol{\theta}_B$ obsahující zbylých $d-m$ složek $\boldsymbol{\theta}$. Máme tedy

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)^\mathsf{T} = (\theta_1, \ldots, \theta_m, \theta_{m+1}, \ldots, \theta_d)^\mathsf{T}$$

Chceme testovat hypotézu $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ proti $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$, kde $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\} \subset \Theta$. Zajímá nás tedy, zdali prvních m složek $\boldsymbol{\theta}_X$ nabylo hodnot stanovených vektorem $\boldsymbol{\theta}_{A0}$ bez ohledu na zbylých $d-m$ složek $\boldsymbol{\theta}_X$.

Podobně jako parametr $\boldsymbol{\theta}$ rozdělíme na prvních m složek (část A) a zbylých $d-m$ složek (část B) všechny vektory a matice vyskytující se ve značení pro teorii maximální věrohodnosti. Například

$$\widehat{\boldsymbol{\theta}}_n = \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{An} \\ \widehat{\boldsymbol{\theta}}_{Bn} \end{pmatrix}, \qquad \boldsymbol{U}_n(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{U}_{An}(\boldsymbol{\theta}) \\ \boldsymbol{U}_{Bn}(\boldsymbol{\theta}) \end{pmatrix}, \qquad I(\boldsymbol{\theta}) = \begin{pmatrix} I_{AA}(\boldsymbol{\theta}) & I_{AB}(\boldsymbol{\theta}) \\ I_{BA}(\boldsymbol{\theta}) & I_{BB}(\boldsymbol{\theta}) \end{pmatrix}, \quad \text{apod.}$$

Lemma 4.4 (Inverse blokové matice)**.** Nechť

$$I = \begin{pmatrix} I_{AA} & I_{AB} \\ I_{BA} & I_{BB} \end{pmatrix}$$

je matice o plné hodnosti. Pak existuje inversní matice k I a její tvar je

$$I^{-1} = \begin{pmatrix} I^{AA} & I^{AB} \\ I^{BA} & I^{BB} \end{pmatrix},$$

kde

$$\begin{aligned} I^{AA} &= I_{AA.B}^{-1}, \\ I^{AB} &= -I_{AA.B}^{-1} I_{AB} I_{BB}^{-1}, \\ I^{BA} &= -I_{BB.A}^{-1} I_{BA} I_{AA}^{-1}, \\ I^{BB} &= I_{BB.A}^{-1}, \\ I_{AA.B} &= I_{AA} - I_{AB} I_{BB}^{-1} I_{BA} \\ I_{BB.A} &= I_{BB} - I_{BA} I_{AA}^{-1} I_{AB}. \end{aligned}$$

Jestliže platí nulová hypotéza $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ víme, že $\boldsymbol{\theta}_{AX} = \boldsymbol{\theta}_{A0}$, neznáme však hodnotu $\boldsymbol{\theta}_{BX}$. Můžeme odhadnout $\boldsymbol{\theta}_{BX}$ pomocí metody maximální věrohodnosti aplikované na vnořený model (submodel)

$$\mathcal{F}_0 = \{\text{rozdělení s hustotou } f(x|(\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)),\ \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0},\ \boldsymbol{\theta}_B \in \Theta_B \subseteq \mathbb{R}^{d-m}\},$$

který má $d - m$ neznámých parametrů.

Označme maximálně věrohodný odhad parametru $\boldsymbol{\theta}_X$ v submodelu $\mathcal{F}_0$ jako $\widetilde{\boldsymbol{\theta}}_n = \begin{pmatrix} \widetilde{\boldsymbol{\theta}}_{An} \\ \widetilde{\boldsymbol{\theta}}_{Bn} \end{pmatrix}$, kde $\widetilde{\boldsymbol{\theta}}_{An} = \boldsymbol{\theta}_{A0}$ a $\widetilde{\boldsymbol{\theta}}_{Bn}$ řeší soustavu věrohodnostních rovnic

$$\boldsymbol{U}_{Bn}(\boldsymbol{\theta}_{A0}, \widetilde{\boldsymbol{\theta}}_{Bn}) = \mathbf{0}.$$

Fisherova informační matice pro pro $\boldsymbol{\theta}_B$ v tomto modelu je $I_{BB}(\boldsymbol{\theta}_X)$.

Podle vět 3.7 a 3.8 v modelu $\mathcal{F}_0$ platí

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_{Bn}(\boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}(\boldsymbol{\theta}_X))$$

a

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{Bn} - \boldsymbol{\theta}_{BX}) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}^{-1}(\boldsymbol{\theta}_X)),$$

zatímco podle těchtýž vět a lemmatu 4.4 v širším modelu $\mathcal{F}$ platí

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_{Bn}(\boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}(\boldsymbol{\theta}_X))$$

a

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{Bn} - \boldsymbol{\theta}_{BX}) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB.A}^{-1}(\boldsymbol{\theta}_X)),$$

kde

$$I_{BB.A}^{-1} = (I_{BB} - I_{BA} I_{AA}^{-1} I_{AB})^{-1} \geq I_{BB}^{-1}.$$

Asymptotický rozptyl maximálně věrohodného odhadu parametru $\boldsymbol{\theta}_{BX}$ tedy závisí na tom, jestli známe $\boldsymbol{\theta}_{AX}$ nebo ne. Pokud známe $\boldsymbol{\theta}_{AX}$ (za platnosti H_0^*), je asymptotický rozptyl odhadu $\widetilde{\boldsymbol{\theta}}_{Bn}$ obecně menší než asymptotický rozptyl odhadu $\widehat{\boldsymbol{\theta}}_{Bn}$ při neznalosti $\boldsymbol{\theta}_{AX}$.

Pokud však $I_{BA} = 0$ (odhady $\boldsymbol{\theta}_{AX}$ a $\boldsymbol{\theta}_{BX}$ jsou asymptoticky nezávislé), potom jsou asymptotické rozptyly $\widetilde{\boldsymbol{\theta}}_{Bn}$ a $\widehat{\boldsymbol{\theta}}_{Bn}$ stejné a nezáleží na tom, jestli $\boldsymbol{\theta}_{AX}$ známe, nebo ne.

Zobecněme si tři testové statistiky zavedené v předchozí kapitole na testování složené hypotézy $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ proti $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$, kde $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\} \subset \Theta$.

Definice 4.8.

(i) Statistika

$$\lambda_n^* = \frac{L_n(\widehat{\boldsymbol{\theta}}_n)}{L_n(\widetilde{\boldsymbol{\theta}}_n)} > 1$$

se nazývá *věrohodnostní poměr*.

(ii) Statistika
$$W_n^* = n(\widehat{\boldsymbol{\theta}}_{An} - \boldsymbol{\theta}_{A0})^\mathsf{T} \widehat{I}_{AA.B}(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_{An} - \boldsymbol{\theta}_{A0})$$
se nazývá *Waldova statistika.*

(iii) Statistika
$$R_n^* = \frac{1}{n}\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)^\mathsf{T} \widehat{I}_n^{-1}(\widetilde{\boldsymbol{\theta}}_n)\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)$$
se nazývá *Raova (skórová) statistika.*

Poznámka.
- Výrazem $\widehat{I}_{AA.B}$ ve Waldově statistice rozumíme inverzi levého horního bloku matice $\widehat{I}_n^{-1}$.
- Jelikož $\boldsymbol{U}_{Bn}(\widetilde{\boldsymbol{\theta}}_n) = \mathbf{0}$, Raova statistika jde přepsat do tvaru
$$R_n^* = \frac{1}{n}\boldsymbol{U}_{An}(\widetilde{\boldsymbol{\theta}}_n)^\mathsf{T} \widehat{I}_{AA.B}^{-1}(\widetilde{\boldsymbol{\theta}}_n)\boldsymbol{U}_{An}(\widetilde{\boldsymbol{\theta}}_n)$$
- Raova statistika nevyžaduje výpočet $\widehat{\boldsymbol{\theta}}_n$, stačí jí jen odhad $\widetilde{\boldsymbol{\theta}}_n$ pro zmenšený model (za hypotézy).

Věta 4.5. Nechť je splněna hypotéza $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$, kde $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\}$. Pak platí:

(i)
$$2\log\lambda_n^* = 2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\widetilde{\boldsymbol{\theta}}_n)) \xrightarrow{\mathsf{D}} \chi^2_m;$$

(ii)
$$W_n^* \xrightarrow{\mathsf{D}} \chi^2_m;$$

(iii)
$$R_n^* \xrightarrow{\mathsf{D}} \chi^2_m.$$

Poznámka. Platí-li H_0^*, očekáváme, že $\widehat{\boldsymbol{\theta}}_n$ je blízko $\widetilde{\boldsymbol{\theta}}_n$, $L_n(\widehat{\boldsymbol{\theta}}_n)$ je blízko $L_n(\widetilde{\boldsymbol{\theta}}_n)$ a $\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)$ je blízko $\mathbf{0}$. Za platnosti H_0^* mají všechny tři testové statistiky hodnoty blízko 0. Velké hodnoty statistik svědčí proti platnosti H_0^*.

Důsledek. Nechť $\chi^2_m(1-\alpha)$ je $(1-\alpha)$-kvantil rozdělení χ^2_m. Uvažujme testy hypotézy $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$, kde $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\}$, proti $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$ dané pravidlem: H_0^* zamítneme ve prospěch H_1^*, pokud
(i) $2\log\lambda_n^* \geq \chi^2_m(1-\alpha)$ *(test poměrem věrohodností)*
(ii) $W_n^* \geq \chi^2_m(1-\alpha)$ *(Waldův test)*
(iii) $R_n^* \geq \chi^2_m(1-\alpha)$ *(skórový test)*
Potom každý z těchto tří testů má asymptoticky (pro $n \to \infty$) hladinu α.

Poznámka. Počet stupňů volnosti v referenčním χ^2_m rozdělení je roven počtu testovaných parametrů.

5 Jednovýběrové a párové problémy pro nominální data

V této kapitole uvažujeme náhodný výběr $X_1, \ldots, X_n$ reálných veličin se spojitou distribuční funkcí F_X patřící do modelu $\mathcal{F}$. Zajímá nás parametr $\theta_X = t(F_X)$. Chceme testovat hypotézu $H_0 : \theta_X = \theta_0$ proti alternativě $H_1 : \theta_X \neq \theta_0$, případně sestrojit intervalový odhad θ_X.

5.1 Kolmogorovovův-Smirnovův test

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

Testovaný parametr: celá distribuční funkce F_X

Hypotéza a alternativa:

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

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

Testová statistika:

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

kde $\widehat{F}_n$ je empirická distribuční funkce náhodného výběru $X_1, \ldots, X_n$.

Nulovou hypotézu budeme zamítat, pokud se empirická distribuční funkce příliš liší od distribuční funkce za nulové hypotézy, 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 $K_n = \max(K_n^+, K_n^-)$.

Tvrzení 5.1. Platí

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

Poznámka. Předchozí tvrzení ukazuje několik zajímavých věcí.

- K výpočtu K_n není třeba počítat $\widehat{F}_n$.

- Rozdělení K_n za platnosti nulové hypotézy nezávisí na F_0 (platí-li H_0, $F_0(X_{(i)})$ má podle věty 1.8 beta rozdělení).

Tvrzení 5.2. Nechť $X_1, \ldots, X_n$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F_X. Pak pro každé $y \in \mathbb{R}$ platí

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

Z tvrzení 5.2 plyne, že $\sqrt{n}K_n$ za platnosti nulové hypotézy konverguje v distribuci k náhodné veličině s distribuční funkcí $F_K(y) = 1-2\sum_{k=1}^{\infty}(-1)^{k+1}\mathrm{e}^{-2k^2y^2}$. To nám umožní určit kritickou hodnotu pro zamítání H_0, aby měl test asymptotickou hladinu α.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow \sqrt{n}K_n \geq c_\alpha,$$

kde c_α je konstanta splňující rovnost

$$2\sum_{k=1}^{\infty}(-1)^{k+1}\mathrm{e}^{-2k^2c_\alpha^2} = \alpha.$$

Řešení této rovnice je třeba najít numericky.

Poznámka.

- Je možné spočítat i přibližnou kritickou hodnotu Kolmogorovova-Smirnovova testu pro diskrétní rozdělení anebo přesnou kritickou hodnotu pro spojité rozdělení a malé n.
- Výhodou tohoto testu je jeho universalita (reaguje na jakýkoli rozdíl v rozdělení dat proti nulové hypotéze) a absence předpokladů o rozdělení F_X.
- Nevýhodou Kolmogorovova-Smirnovova testu je to, že F_0 musí být známa přesně (nesmí obsahovat neznámé parametry ani jejich odhady) a to, že test má malou sílu v situacích, kdy některé druhy porušení H_0 jsou častější nebo důležitější než jiné. Pak je lepší použít test, který je specificky zaměřen na konkrétní typ porušení H_0.

5.2 Jednovýběrový t-test

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:

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

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 (viz definice 1.4).

Poznámka.

- Nulovou hypotézu budeme zamítat, pokud se výběrový průměr příliš liší od hypotetické střední hodnoty, tj. pokud je testová statistika buď moc velká nebo moc malá.
- Věta 1.5 implikuje, že za platnosti nulové hypotézy má T_n rozdělení 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.

Poznámka. Jednovýběrový t-test* je přesný test zaměřený na střední hodnotu. Vyžaduje normální rozdělení pozorovaných dat.

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

Interval spolehlivosti pro μ_X: Interval spolehlivosti pro střední hodnotu normálního rozdělení při neznámém rozptylu je

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

(viz (2.1) na str. 14 a předcházející příklad).

5.3 Jednovýběrový z-test

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.

* Angl. *one-sample t-test*

Poznámka.

- Testová statistika je naprosto stejná jako u t-testu. Věta 1.4 implikuje, že za platnosti nulové hypotézy má T_n asymptoticky normované normální rozdělení.
- Budeme-li zamítat H_0, pokud $|T_n| \geq u_{1-\alpha/2}$, test bude mít asymptotickou hladinu α. Nahradíme-li kvantil normálního rozdělení kvantilem rozdělení t_{n-1}, test bude mít stále asymptotickou hladinu α a bude mít lepší vlastnosti pro konečné n .

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.

Poznámka. Tento test je ekvivalentní jednovýběrovému t-testu, ale bez předpokladu normality dat. Zatímco jednovýběrový t-test je přesný, tento test je pouze asymptotický a vyžaduje tedy dostatečně velký počet pozorování (v praxi většinou stačí $n \geq 30$).

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

Interval spolehlivosti pro μ_X:

Interval (2.1) má pravděpodobnost poktytí konvergující k $1-\alpha$. Viz příklad na str. 14.

5.4 Jednovýběrový znaménkový test

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

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

Hypotéza a alternativa:

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

kde m_0 je předem daná konstanta.

Testová statistika:

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

(počet pozorování větších než m_0).

Věta 5.3. Nechť $X_1, \ldots, X_n$ je náhodný výběr z libovolného spojitého rozdělení s mediánem m_X. Pak

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

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

Poznámka.
- Hypotézu budeme zamítat pro příliš malé nebo příliš velké hodnoty Y_n.
- První část věty udává přesné rozdělení Y_n za platnosti hypotézy $H_0 : m_X = m_0$.
- Druhá část věty udává asymptotické rozdělení Y_n za platnosti hypotézy $H_0 : m_X = m_0$ při $n \to \infty$.

Kritický obor (přesný test):

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

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).

Kritický obor (asymptotický test):

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

5.5 Jednovýběrový Wilcoxonův test

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.

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:

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

kde $\mathcal{I} \subset \{1, \ldots, n\}$ je množina všech indexů takových, že $Z_i \overset{\mathsf{df}}{=} X_i - \delta_0$ má kladné znaménko pro $i \in \mathcal{I}$, a $R_1, R_2, \ldots, R_n$ jsou pořadí náhodných veličin $|Z_i|$ mezi všemi $|Z_1|, \ldots, |Z_n|$.

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
$$0 < \left|Z_{(1)}\right| < \left|Z_{(2)}\right| < \cdots < \left|Z_{(n)}\right|.$$
3. Určíme pořadí R_i náhodné veličiny $|Z_i|$ mezi všemi $\left|Z_{(1)}\right|, \ldots, \left|Z_{(n)}\right|$. Platí $|Z_i| = \left|Z_{(R_i)}\right|$.
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 5.4. 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)
$$\mathsf{E}\, W_S = \frac{n(n+1)}{4}, \quad \mathsf{var}\, W_S = \frac{n(n+1)(2n+1)}{24}.$$

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

Poznámka.

- Předchozí tvrzení dává návod k nalezení kritických hodnot pro zamítání nulové hypotézy, které zaručují asymptotickou hladinu α.
- 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).

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

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 žádný takový předpoklad nevyžaduje.

5.6 Jednovýběrový χ^2 test na rozptyl

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

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

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).

Poznámka.

- Z věty 1.2 (bod 3) víme, že testová statistika má za platnosti modelu a nulové hypotézy přesně rozdělení χ^2_{n-1}.
- Nulovou hypotézu budeme zamítat, pokud se výběrový rozptyl příliš liší od hypotetického rozptylu, tj. pokud je testová statistika buď moc velká nebo moc malá.

Kritický obor:

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

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.

Poznámka. Jednovýběrový χ^2 test rozptylu je přesný test. Vyžaduje normální rozdělení pozorovaných dat.

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^2_X: (viz (2.2))

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

5.7 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.

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.

5.8 Párový t-test

Párový t-test* je ekvivalentní jednovýběrovému t-testu provedenému na rozdíly Z_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:

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

* Angl. *paired t-test*

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.

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

Interval spolehlivosti pro $\mu_X - \mu_Y$: Samostatné cvičení.

5.9 Párový z-test

Párový z-test je ekvivalentní jednovýběrovému z-testu provedenému na rozdíly Z_i. Je to asymptotická verze párového t-testu na nenormální data.

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.

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

5.10 Párový znaménkový test

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

Testovaný parametr: Medián m_Z rozdílu Z_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 $\mathrm{P}[X_i \leq Y_i] = \mathrm{P}[X_i \geq Y_i] = 1/2$, 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.

Testová statistika:

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

(počet rozdílů větších než 0).

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

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 to, 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.

5.11 Párový Wilcoxonův test

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

Poznámka. Na rozdíl od jednovýběrového Wilcoxonova testu u párového testu† vyžadujeme, aby rozdíly $Z_i = X_i - Y_i$ měly konečnou střední hodnotu. Předpoklad o symetrické hustotě se týká rozdílů Z_i, nikoli původních pozorování X_i a Y_i. V modelu $\mathcal{F}$ musí platit $\mathsf{E}\, Z_i = \mathsf{E}\, X_i - \mathsf{E}\, Y_i = \delta_X$.

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

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 $Z_i^* \overset{\mathsf{df}}{=} X_i - Y_i - \delta_0$ má kladné znaménko pro $i \in \mathcal{I}$, a $R_1 < R_2 < \cdots < R_n$ jsou pořadí náhodných veličin $|Z_i^*|$ mezi všemi $|Z_1^*|, \ldots, |Z_n^*|$.

Vlastnosti testové statistiky a kritický obor: viz jednovýběrový Wilcoxonův test.

K testování hypotézy H_0 je asymptotický párový t-test (=párový z-test) vhodnější než párový Wilcoxonův test, protože nevyžaduje symetrii hustoty.

6 Dvouvýběrové problémy pro nominální data

Nyní budeme řešit situace, kdy máme k dispozici dva *nezávislé* náhodné výběry: $X_1, \ldots, X_n$ 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 $\theta_X - \theta_Y$.

Existuje ještě druhý způsob, jak zformulovat dvouvýběrový problém. Představme si, že pozorujeme náhodný výběr z dvourozměrného rozdělení

$$\binom{Z_1}{G_1}, \ldots, \binom{Z_N}{G_N},$$

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

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

získáme první formulaci problému (dva nezávislé výběry). Chceme porovnat podmíněné rozdělení Z_j v obou skupinách, tj. zajímají nás podmíněné distribuční funkce $F_X(x) = \mathrm{P}\,[\,Z_j \leq x \mid G_j = 0\,]$ a $F_Y(x) = \mathrm{P}\,[\,Z_j \leq x \mid G_j = 1\,]$, případně jejich parametry $\theta_X = t(F_X)$ a $\theta_Y = t(F_Y)$.

Data podle první formulace získáme obvykle tak, že si předem stanovíme, kolik měření z každé skupiny chceme mít, a pak napozorujeme příslušný 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 jsou ekvivalentní, až na to, že u první formulace jsou m a n pevná čísla, zatímco u druhé formulace jsou m a n náhodné veličiny s binomickým rozdělením ($n = \sum_{j=1}^{N}(1 - G_j) \sim \mathsf{Bi}(N, 1 - p_G)$). U druhé formulace se snáze

používají asymptotické výsledky pro $N \to \infty$. Chceme-li používat asymptotické metody u první formulace, musíme mít $n \to \infty$ i $m \to \infty$, ale navíc ještě musíme předpokládat, že $n/m \to q$, kde $0 < q < \infty$ (tj. rozsahy obou výběrů konvergují do nekonečna stejně rychle).

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

6.1 Dvouvýběrový Kolmogorovovův-Smirnovův test

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

Testované parametry: celé distribuční funkce F_X a F_Y

Hypotéza a alternativa:

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

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 6.1. 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

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

Poznámka.

- Nulovou hypotézu budeme zamítat, pokud se empirické distribuční funkce obou výběrů od sebe příliš liší, tj. pokud je testová statistika velká.
- Tvrzení 6.1 implikuje, že za platnosti nulové hypotézy konverguje $\sqrt{\frac{mn}{n+m}} K_{n,m}$ v distribuci k náhodné veličině s distribuční funkcí $1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} \mathrm{e}^{-2k^2x^2}$ (stejná, jako u jednovýběrového Kolmogorovova-Smirnovova testu). To nám umožní určit kritickou hodnotu pro zamítání H_0, aby měl test asymptotickou hladinu α (musí se spočítat numericky).

Kritický obor:

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

kde c_α je konstanta splňující rovnost

$$2\sum_{k=1}^{\infty}(-1)^{k+1}\mathrm{e}^{-2k^2c_\alpha^2} = \alpha.$$

Poznámka.

- 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 na tento konkrétní parametr.

6.2 Dvouvýběrový t-test

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:

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

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

Testová statistika:

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

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

$$\begin{aligned} S_{n,m}^2 \overset{\mathsf{df}}{=} & \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2\right] \\ = & \frac{n-1}{n+m-2}S_X^2 + \frac{m-1}{n+m-2}S_Y^2 \end{aligned}$$

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

Věta 6.2. 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. Pak

$$T \overset{\mathsf{df}}{=} \sqrt{\frac{mn}{n+m}} \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{S_{n,m}} \sim t_{n+m-2}$$

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 6.2 implikuje, že za platnosti modelu $\mathcal{F}$ a hypotézy H_0 má $T_{n,m}$ rozdělení t_{n+m-2}.

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.

Poznámka.

- Dvouvýběrový t-test* je přesný test zaměřený na střední hodnotu. Vyžaduje normální rozdělení pozorovaných dat a shodný rozptyl v obou výběrech.
- Nemají-li data normální rozdělení, věta 6.2 platí přibližně pro $m, n \to \infty$. Nemají-li data shodný rozptyl, věta 6.2 neplatí a test nemá správnou hladinu ani asymptoticky.

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

Interval spolehlivosti pro $\mu_X - \mu_Y$: Z věty 6.2 lze odvodit přesný interval spolehlivosti pro rozdíl středních hodnot obou výběrů. Dostaneme

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

6.3 Dvouvýběrový z-test

Model:

$$\mathcal{F} = \{F_X, F_Y \text{ mají konečné rozptyly}\}$$

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

* Angl. *two-sample t-test*

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 6.3. 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

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

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 6.3 implikuje, že za platnosti modelu $\mathcal{F}$ a hypotézy H_0 má $Z_{n,m}$ asymptoticky rozdělení $N(0,1)$.

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.

Poznámka. Dvouvýběrový z-test je asymptotický test zaměřený na střední hodnotu. Na rozdíl od dvouvýběrového t-testu nevyžaduje ani normální rozdělení pozorovaných dat ani shodný rozptyl v obou výběrech.

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

Interval spolehlivosti pro $\mu_X - \mu_Y$: Z věty 6.3 lze odvodit přibližný 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 (funkce `t.test`). Welchův test je vlastně náš dvouvýběrový z-test s vylepšenými kritickými hodnotami.

6.4 Dvouvýběrový Wilcoxonův test

Model: $\mathcal{F} = \{X \sim F_X, Y \sim F_Y$, kde $F_X(x) = F_Y(x - \delta)$ pro nějaké $\delta \in \mathbb{R}$ a F_X je libovolná spojitá d.f.$\}$
(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

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

* lze je nalézt např. v knize Anděl: *Statistické metody*, Matfyzpress, Praha, 1998, kap. 8.1.
† Angl. *Welch test* ‡ Angl. *two-sample Wilcoxon test*

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 6.4. 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 nulové hypotézy, které zaručují asymptotickou hladinu α.
- 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):

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

Mann-Whitneyho formulace Wilcoxonova testu

Uvažujme všechny dvojice (X_i, Y_j) pro $i = 1, \ldots, n$ a $j = 1, \ldots, m$. Spočtěme, kolik z nich splňuje podmínku $X_i < Y_j$:

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

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

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

Tvrzení 6.5.

(i)

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

(ii) Platí-li H_0,

$$\mathsf{E}\, W^*_{n,m} = \frac{nm}{2}, \quad \mathsf{var}\, W^*_{n,m} = \frac{mn(m+n+1)}{12}.$$

(iii) Pokud $\min(n,m) \to \infty$ a platí H_0, pak $(mn)^{-1} W^*_{n,m} \xrightarrow{\mathsf{P}} 1/2$

Testy založené na dvouvýběrové Wilcoxonově statistice a Mann-Whitneyho statistice jsou ekvivalentní, jeden z nich zamítá hypotézu tehdy a jen tehdy, zamítá-li druhý.

Mann-Whitneyho statistika lépe ukazuje, jakou hypotézu vlastně dvouvýběrový Wilcoxonův test testuje: i pokud neplatí model $\mathcal{F}$ a data mají zcela obecná rozdělení F_X a F_Y, lze ukázat, že $W^*_{n,m}/nm$ je nestranným a konsistentním odhadem parametru $\mathrm{P}[X_i < Y_j]$. Protože zamítáme hypotézu pro příliš velké i příliš malé hodnoty $W^*_{n,m}$ dvouvýběrový Wilcoxonův test v obecném modelu vlastně testuje

$$H_0^* : \mathrm{P}[X_i < Y_j] = \frac{1}{2} \quad \text{proti alternativě} \quad H_1^* : \mathrm{P}[X_i < Y_j] \neq \frac{1}{2}.$$

Hypotéza H_0^* je však těžko interpretovatelná. Nelze ji obecně vyjádřit pomocí rovnosti charakteristik obou rozdělení. Dvouvýběrový Wilcoxonův test proto obecně není ani test rovnosti mediánů ani test rovnosti středních hodnot. Může se totiž stát, že $\mathsf{E}\, X_i = \mathsf{E}\, Y_j$, ale $\mathrm{P}[X_i < Y_j] \neq 1/2$, tudíž Wilcoxonův test při dostatečně velkém rozsahu výběru zamítne hypotézu. Nebo naopak, v situaci, kdy $\mathsf{E}\, X_i \neq \mathsf{E}\, Y_j$, ale $\mathrm{P}[X_i < Y_j] = 1/2$, Wilcoxonův test při jakémkoli rozsahu výběru zamítá hypotézu pouze s pravděpodobností α.

Dvouvýběrový Wilcoxonův test tedy není vhodný pro testování rovnosti středních hodnot. Takový problém je lépe řešit asymptotickým z-testem (nebo jeho Welchovou aproximací).

6.5 Dvouvýběrový F test na rozptyl

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 $\sigma_X^2 = \mathsf{var}\, X_i$ a $\sigma_Y^2 = \mathsf{var}\, Y_j$.

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 1.6 plyne, že testová statistika má za platnosti modelu a nulové hypotézy přesně rozdělení $F_{n-1,m-1}$.
- Nulovou hypotézu budeme zamítat, pokud se výběrové rozptyly příliš liší, tj. pokud je testová statistika buď moc velká nebo moc malá.

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.

Poznámka. Dvouvýběrový F test na rozptyl je přesný test. Vyžaduje normální rozdělení v obou výběrech.

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 1.6 lze odvodit interval spolehlivosti pro podíl rozptylů. Dostaneme

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

7 Jednovýběrové problémy pro kategoriální data

V této kapitole a v kapitole následující se budeme zabývat *kategoriálními veličinami*. Pod pojmem kategoriální veličina rozumíme diskrétní náhodnou veličinu nabývající konečně mnoha hodnot, typicky $1, \dots, K$, jejíž hodnoty nemusí mít numerickou interpretaci, ale označují členství v nějaké skupině (kategorii). Například

$$Y = \begin{cases} 1 \ \dots \ \text{muž} \\ 2 \ \dots \ \text{žena} \end{cases} \qquad Y = \begin{cases} 1 \ \dots \ \text{základní vzdělání} \\ 2 \ \dots \ \text{střední vzdělání} \\ 3 \ \dots \ \text{VŠ vzdělání} \end{cases} \qquad Y = \begin{cases} 1 \ \dots \ \text{Praha} \\ 2 \ \dots \ \text{Středočeský kraj} \\ \quad \dots \\ 14 \ \dots \ \text{Zlínský kraj} \end{cases}$$

Veličiny, které nejsou kategoriální, nazýváme *nominální*. Nominální veličiny mohou být spojité i diskrétní; diskrétní nominální veličiny mohou být např. počty nějakých sledovaných událostí.

Charakteristiky rozdělení, které obvykle počítáme u nominálních veličin ($\mathsf{E}\, X$, $\mathsf{var}\, X$ atp.), nemají u kategoriálních veličin žádnou přirozenou interpretaci. Analýza kategoriálních dat se proto soustředí výhradně na pravděpodobnosti jednotlivých hodnot.

7.1 Alternativní a binomické rozdělení

Alternativní rozdělení je nejjednodušším modelem pro kategoriální veličinu, která nabývá pouze dvou hodnot. Chceme-li použít alternativní rozdělení, zakódujeme tyto dvě hodnoty jako 0 a 1 (v libovolném pořadí).

Nechť $Y_1, \dots, Y_n$ je náhodný výběr z alternativního rozdělení $\mathsf{Alt}(p_X)$, $p_X \in (0, 1)$, označující klasifikaci n jedinců do kategorie 0 a 1. Parametr p_X označuje pravděpodobnost, že libovolný jedinec je klasifikován do skupiny 1. Označíme-li $X_n = \sum_{i=1}^{n} Y_i$, dostaneme počet jedinců klasifikovaných do skupiny 1 a víme, že tato veličina má rozdělení $\mathsf{Bi}(n, p_X)$. Počet jedinců klasifikovaných do skupiny 0 je $n - X_n \sim \mathsf{Bi}(n, 1 - p_X)$.

Odhad parametru p_X a testování hypotézy $H_0 : p_X = p_0$

Nestranným a konsistentním odhadem parametru p_X je relativní četnost $\widehat{p}_n = X_n/n = n^{-1}\sum_{i=1}^n Y_i = \overline{Y}_n$. Tento odhad lze odvodit jako odhad empirický, momentovou metodou nebo metodou maximální věrohodnosti.

Věta 7.1. Nechť $p_X \in (0,1)$. Pak platí

(i) $n\widehat{p}_n = X_n \sim \mathsf{Bi}(n, p_X)$

(ii) $\mathsf{E}\,\widehat{p}_n = p_X$ (nestrannost), $\mathsf{var}\,\widehat{p}_n = \frac{p_X(1-p_X)}{n}$

(iii) $\widehat{p}_n \xrightarrow{\mathsf{P}} p_X$ (konsistence)

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

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

Poznámka. Se všemi částmi věty 7.1 jsme se seznámili již dříve.

Poznámka. Povšimněte si, že pro libovolné pevné n jest $\mathrm{P}[\widehat{p}_n = 1] > 0$ a $\mathrm{P}[\widehat{p}_n = 0] > 0$, z čehož plyne $\mathrm{P}[Z_n = \infty] > 0$ a $\mathsf{E}\,|Z_n| = \infty$ pro všechna n. Přesto však $Z_n \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$.

V kapitole 2.2 na str. 14 jsme odvodili přibližný interval spolehlivosti pro p_X založený na bodě (v) věty 7.1:

$$\left(\widehat{p}_n - \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}},\ \widehat{p}_n + \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right). \tag{7.1}$$

Totéž tvrzení lze použít k odvození přibližného testu hypotézy $H_0 : p_X = p_0$ proti alternativě $H_1 : p_X \neq p_0$. Tento test má kritický obor

$$H_0 \text{ zamítneme } \Leftrightarrow \sqrt{n}\frac{|\widehat{p}_n - p_0|}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} \geq u_{1-\alpha/2}. \tag{7.2}$$

Přesný test a přesný interval spolehlivosti lze založit na bodě (i) věty 7.1. Přesný test má 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 má hladinu nejvýše α (přesné hladiny α nemusí být možné dosáhnout).

Odhad šance a testování hypotéz o šanci

Alternativní přístupy k testování hypotéz a konstrukci intervalu spolehlivosti pro p_X jsou založeny na transformaci parametru p_X.

Podíl $\frac{p_X}{1-p_X}$ pravděpodobnosti úspěchu a neúspěchu se nazývá *šance*[*] na úspěch. Pojem šance se běžně používá při kursových sázkách.

Zvolme jako odhadovaný parametr logaritmus šance $\theta_X = \log \frac{p_X}{1-p_X}$. Funkce $g(x) = \log \frac{x}{1-x}$ je rostoucí a spojitě diferencovatelná pro $x \in (0,1)$ a zobrazuje interval $(0,1)$ na $\mathbb{R}$. Inversní transformace je $g^{-1}(y) = \frac{\exp\{y\}}{1+\exp\{y\}}$. Šance θ_X tedy může nabývat libovolné hodnoty v $\mathbb{R}$ a můžeme z ní vyjádřit pravděpodobnost p_X jako $p_X = \frac{\exp\{\theta_X\}}{1+\exp\{\theta_X\}}$.

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 (iv) věty 7.1 a delta metody (věta P.7.11).

Věta 7.2. 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 7.2 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 vyjádřit jako $H_0 : \theta_X = \theta_0$ a zamítáme ji ve prospěch alternativy $H_1 : \theta_X \neq \theta_0$ pokud

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

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

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

[*] Angl. *odds*

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

$$\left(\frac{\frac{\widehat{p}_n}{1-\widehat{p}_n} \mathrm{e}^{-u_{1-\alpha/2} D_n}}{1 + \frac{\widehat{p}_n}{1-\widehat{p}_n} \mathrm{e}^{-u_{1-\alpha/2} D_n}}, \frac{\frac{\widehat{p}_n}{1-\widehat{p}_n} \mathrm{e}^{u_{1-\alpha/2} D_n}}{1 + \frac{\widehat{p}_n}{1-\widehat{p}_n} \mathrm{e}^{u_{1-\alpha/2} D_n}} \right). \tag{7.3}$$

Interval (7.3) zaručuje, že oba jeho krajní body leží uvnitř $(0, 1)$, což neplatí v případě intervalu (7.1). 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 $\widehat{p}_n$.

7.2 Multinomické rozdělení

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

Multinomické rozdělení: definice a vlastnosti

Definice 7.1 (Multinomické rozdělení)**.** Nechť $K \geq 2$ a $n \geq 1$ jsou přirozená čísla a $\boldsymbol{p}_X = (p_1, \ldots, p_K)^\mathsf{T}$ 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 $\mathsf{Mult}_K(n, \boldsymbol{p}_X)$, právě když jeho hustota vzhledem k součinové čítací míře na $\mathbb{Z}^K$ je

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

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 $\boldsymbol{p}_X$.

Věta 7.3 (Rozklad multinomického rozdělení.)**.** Nechť $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ jsou nezávislé náhodné vektory s rozdělením $\mathsf{Mult}_K(1, \boldsymbol{p}_X)$. Pak $\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{Y}_i \sim \mathsf{Mult}_K(n, \boldsymbol{p}_X)$.

Věta 7.4 (Vlastnosti multinomického rozdělení.)**.** Nechť $\boldsymbol{X} \sim \mathsf{Mult}_K(n, \boldsymbol{p}_X)$. 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) $\mathsf{cov}\,(X_j, X_k) = -np_j p_k$,

(iv)

$$\begin{aligned}\mathsf{var}\,\boldsymbol{X} &= n\left[\mathrm{diag}\,(\boldsymbol{p}_X) - \boldsymbol{p}_X^{\otimes 2}\right] = \\ &= n\,\mathrm{diag}\,(\sqrt{\boldsymbol{p}_X})(I_K - \sqrt{\boldsymbol{p}_X}^{\otimes 2})\mathrm{diag}\,(\sqrt{\boldsymbol{p}_X}),\end{aligned}$$

kde $\sqrt{\boldsymbol{p}_X} = (\sqrt{p_1}, \ldots, \sqrt{p_K})^\mathsf{T}$.

Poznámka.

- Matice $I_K - \sqrt{\boldsymbol{p}_X}^{\otimes 2}$ je idempotentní. Její hodnost (a stopa) je rovna $K-1$.
- Matice $\mathsf{var}\,\boldsymbol{X}$ je singulární, její hodnost je $K-1$. Složky náhodného vektoru $\boldsymbol{X}$ splňují podmínku $\sum_{k=1}^K X_i = n$.

Věta 7.5 (Asymptotické vlastnosti multinomického rozdělení.)**.**
Nechť $\boldsymbol{X} \sim \mathsf{Mult}_K(n, \boldsymbol{p}_X)$. Pak

(i)

$$\boldsymbol{Z}_n \overset{\mathsf{df}}{=} \frac{1}{\sqrt{n}}\mathrm{diag}\,(\sqrt{\boldsymbol{p}_X})^{-1}(\boldsymbol{X} - n\boldsymbol{p}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_K(\boldsymbol{0}, I_K - \sqrt{\boldsymbol{p}_X}^{\otimes 2}),$$

(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 p_k, testování hypotéz o p_k a konstrukci intervalových odhadů pro p_k můžeme použít metody popsané v kapitole 7.1, neboť podle věty 7.4(i) platí $X_k \sim \mathsf{Bi}(n, p_k)$,

Celý vektor $\boldsymbol{p}_X$ odhadneme pomocí $\widehat{\boldsymbol{p}}_n = \boldsymbol{X}/n$. Sdružené asymptotické rozdělení odhadu $\widehat{\boldsymbol{p}}_n$ získáme z věty 7.5(i):

$$\sqrt{n}(\widehat{\boldsymbol{p}}_n - \boldsymbol{p}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_K(\boldsymbol{0}, \mathrm{diag}\,(\boldsymbol{p}_X) - \boldsymbol{p}_X^{\otimes 2}).$$

Pro libovolný vektor konstant $\boldsymbol{c}$ o délce K, platí

$$\sqrt{n}(\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - \boldsymbol{c}^\mathsf{T}\boldsymbol{p}_X) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \boldsymbol{c}^\mathsf{T}[\mathrm{diag}\,(\boldsymbol{p}_X) - \boldsymbol{p}_X^{\otimes 2}]\boldsymbol{c}).$$

Pokud $\boldsymbol{c}^\mathsf{T}[\mathrm{diag}\,(\boldsymbol{p}_X) - \boldsymbol{p}_X^{\otimes 2}]\boldsymbol{c} \neq 0$ a $V_c \overset{\mathsf{df}}{=} \boldsymbol{c}^\mathsf{T}[\mathrm{diag}\,(\widehat{\boldsymbol{p}}_n) - \widehat{\boldsymbol{p}}_n^{\otimes 2}]\boldsymbol{c} \neq 0$, dostaneme ze Sluckého věty

$$\sqrt{n}\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - \boldsymbol{c}^\mathsf{T}\boldsymbol{p}_X}{\sqrt{V_c}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1). \tag{7.4}$$

Odtud můžeme snadno odvodit přibližné testy hypotéz $H_0 : \boldsymbol{c}^\mathsf{T}\boldsymbol{p}_X = \gamma_0$. Vezmeme testovou statistiku

$$T_c = \sqrt{n}\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - \gamma_0}{\sqrt{V_c}},$$

která má podle (7.4) za platnosti hypotézy asymptoticky normované normální rozdělení a H_0 zamítneme právě když $|T_c| \geq u_{1-\alpha/2}$.

Přibližný interval spolehlivosti pro $\boldsymbol{c}^\mathsf{T}\boldsymbol{p}_X$ založený na konvergenci (7.4) 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 $\boldsymbol{c}^\mathsf{T}\boldsymbol{p}_X$ 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 $H_0 : \boldsymbol{p}_X = \boldsymbol{p}^0$ založený na větě 7.5(ii). Tato hypotéza říká, že pravděpodobnosti kategorií $\boldsymbol{p}_X = (p_1, \ldots, p_K)^\mathsf{T}$ 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

$$\chi^2 = \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{np_k^0}$$

má asymptoticky 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

$$\chi^2 = \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{np_k^0} \geq \chi^2_{K-1}(1-\alpha),$$

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

$$\chi^2 = \frac{(X_1 - np_0)^2}{np_0} + \frac{[n - X_1 - n(1-p_0)]^2}{n(1-p_0)} = \left[\sqrt{n}\frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)}}\right]^2,$$

* Angl. χ^2 *test of goodness of fit*

takže testová statistika χ^2 testu pro $K = 2$ kategorie je v podstatě čtvercem testové statistiky (7.2) pro asymptotický test hypotézy $H_0 : p_X = p_0$ (až na to, že ve jmenovateli je rozptyl za hypotézy, nikoli rozptyl odhadnutý).

Příklady.

1. *Je kostka pravidelná?* Hodíme n-krát kostkou a zaznamenáme, kolikrát padly výsledky 1 – 6: dostaneme četnosti $X_1, \dots, X_6$. Nastavíme $p_k^0 = 1/6$, $k = 1, \dots, 6$. Zamítne-li χ^2 test hypotézu H_0, prokázali jsme, že na kostce nepadají všechna čísla stejně často.

2. *Rodí se děti během roku rovnoměrně?* Máme dány počty dětí narozených v jednotlivých měsících během kalendářního roku: $X_1, \dots, X_{12}$. 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.

3. *Pochází náhodný výběr z distribuční funkce F_0?* Uvažujme náhodný výběr $Z_1, \dots, Z_n$. Stanovíme si intervaly $(a_{k-1}, a_k\rangle$, $k = 1, \dots, K$, $a_0 = -\infty$, $a_K = \infty$ tak, že jejich počet K je výrazně menší než počet pozorování n. Spočítáme, kolik pozorování padlo do k-tého intervalu: $X_k = \sum_{i=1}^n \mathbb{I}_{(a_{k-1},a_k\rangle}(Z_i)$. Pochází-li náhodný výběr $Z_1, \dots, Z_n$ z rozdělení s distribuční funkcí $F_0(x) = F(x; \boldsymbol{\theta}_0)$, kde $\boldsymbol{\theta}_0$ je známo, pravděpodobnosti jednotlivých intervalů jsou $p_k^0 = F(a_k; \boldsymbol{\theta}_0) - F(a_{k-1}; \boldsymbol{\theta}_0)$. Zamítne-li χ^2 test hypotézu H_0, prokázali jsme, že náhodný výběr $Z_1, \dots, Z_n$ nepochází z rozdělení $F(x; \boldsymbol{\theta}_0)$.

7.3 Modelování pravděpodobností v multinomickém rozdělení

V posledním z příkladů uvedených na konci minulé kapitoly pravděpodobnosti jednotlivých kategorií p_k^0 závisely na vektoru parametrů $\boldsymbol{\theta}_0$, který jsme museli znát, abychom mohli provést test dobré shody. Situace, kdy pravděpodobnosti kategorií závisejí na parametrech, není zdaleka ojedinělá, avšak tyto parametry jsou v praxi většinou neznámé.

Uvažujme následující model $\mathcal{F}_0$: Nechť $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ je náhodný výběr z rozdělení $\mathsf{Mult}_K(1, \boldsymbol{p}(\boldsymbol{\theta}_X))$, 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). Náhodný vektor $\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{Y}_i$ (pozorované četnosti kategorií $1, \dots, K$) pak má rozdělení $\mathsf{Mult}_K(n, \boldsymbol{p}(\boldsymbol{\theta}_X))$.

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	$(1-\theta_X)^2$

Budeme-li pozorovat genotypy n nezávislých jedinců a označíme-li X_1, X_2, X_3 počty jedinců s genotypem (po řadě) AA, Aa, aa, bude mít vektor $\boldsymbol{X} = (X_1, X_2, X_3)^\mathsf{T}$ 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 $\boldsymbol{X}$ můžeme chtít

- odhadnout parametr θ_X;
- otestovat, zdali se populace nachází v Hardyho-Weinbergově ekvilibriu.

Parametry $\boldsymbol{\theta}_X$ můžeme odhadnout metodou maximální věrohodnosti takto:

$$L_n(\boldsymbol{\theta}) = C\prod_{i=1}^{n}\prod_{k=1}^{K} p_k(\boldsymbol{\theta})^{Y_{ik}} = C\prod_{k=1}^{K} p_k(\boldsymbol{\theta})^{X_k}, \text{ kde } C \text{ je konstanta neobsahující } \boldsymbol{\theta},$$

$$\ell_n(\boldsymbol{\theta}) = \sum_{k=1}^{K} X_k \log p_k(\boldsymbol{\theta}) + \log C,$$

$$\boldsymbol{U}_i(\boldsymbol{\theta}) = \sum_{k=1}^{K} Y_{ik}\frac{\dot{p}_k(\boldsymbol{\theta})}{p_k(\boldsymbol{\theta})} = \dot{\boldsymbol{p}}(\boldsymbol{\theta})^\mathsf{T} D_{\boldsymbol{\theta}}^{-1}\boldsymbol{Y}_i.$$

Zavedli jsme značení $D_{\boldsymbol{\theta}} = \operatorname{diag}(\boldsymbol{p}(\boldsymbol{\theta}))$, $\dot{p}_k(\boldsymbol{\theta}) \overset{\mathsf{df}}{=} \frac{\partial p_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ a $\dot{\boldsymbol{p}}(\boldsymbol{\theta}) \overset{\mathsf{df}}{=} \frac{\partial \boldsymbol{p}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. Předpokládáme, že matice $\dot{\boldsymbol{p}}(\boldsymbol{\theta})$ má plnou hodnost d. Jelikož $\boldsymbol{p}(\boldsymbol{\theta})^\mathsf{T}\mathbf{1}_K = 1$, musí být $\dot{\boldsymbol{p}}(\boldsymbol{\theta})^\mathsf{T}\mathbf{1}_K = \mathbf{0}$.

Máme

$$\mathsf{E}\,\boldsymbol{U}_i(\boldsymbol{\theta}_X) = \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)^\mathsf{T} D_{\boldsymbol{\theta}_X}^{-1}\mathsf{E}\,\boldsymbol{Y}_i = \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)^\mathsf{T} D_{\boldsymbol{\theta}_X}^{-1}\boldsymbol{p}(\boldsymbol{\theta}_X) = \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)^\mathsf{T}\mathbf{1}_K = \mathbf{0}$$

a

$$\begin{aligned}\mathsf{var}\,\boldsymbol{U}_i(\boldsymbol{\theta}_X) &= \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)^\mathsf{T} D_{\boldsymbol{\theta}_X}^{-1}\mathsf{var}\,\boldsymbol{Y}_i D_{\boldsymbol{\theta}_X}^{-1}\dot{\boldsymbol{p}}(\boldsymbol{\theta}_X) = \\ &= \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)^\mathsf{T} D_{\boldsymbol{\theta}_X}^{-1}\dot{\boldsymbol{p}}(\boldsymbol{\theta}_X) - \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)^\mathsf{T} D_{\boldsymbol{\theta}_X}^{-1}\boldsymbol{p}(\boldsymbol{\theta}_X)\boldsymbol{p}(\boldsymbol{\theta}_X)^\mathsf{T} D_{\boldsymbol{\theta}_X}^{-1}\dot{\boldsymbol{p}}(\boldsymbol{\theta}_X) = \\ &= \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)^\mathsf{T} D_{\boldsymbol{\theta}_X}^{-1}\dot{\boldsymbol{p}}(\boldsymbol{\theta}_X) = A^\mathsf{T}A,\end{aligned}$$

kde $A \overset{\mathsf{df}}{=} D_{\boldsymbol{\theta}_X}^{-1/2}\dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)$. Dále

$$\boldsymbol{U}_n(\boldsymbol{\theta}) = \sum_{k=1}^{K} X_k\frac{\dot{\boldsymbol{p}}(\boldsymbol{\theta})}{p_k(\boldsymbol{\theta})} = \dot{\boldsymbol{p}}(\boldsymbol{\theta})^\mathsf{T}\operatorname{diag}^{-1}(\boldsymbol{p}(\boldsymbol{\theta}))\boldsymbol{X} = \dot{\boldsymbol{p}}(\boldsymbol{\theta})^\mathsf{T}\operatorname{diag}^{-1}(\boldsymbol{p}(\boldsymbol{\theta}))[\boldsymbol{X} - n\boldsymbol{p}(\boldsymbol{\theta})],$$

$$I(\boldsymbol{\theta}_X) = \mathsf{var}\,\boldsymbol{U}_i(\boldsymbol{\theta}_X) = \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)^\mathsf{T} D_{\boldsymbol{\theta}_X}^{-1}\dot{\boldsymbol{p}}(\boldsymbol{\theta}_X) = A^\mathsf{T}A.$$

Informační matice $I(\boldsymbol{\theta}_X)$ je positivně semidefinitní a má plnou hodnost d, tudíž existuje její inverse.

Maximálně věrohodný odhad $\widehat{\boldsymbol{\theta}}_n$ dostaneme řešením soustavy rovnic

$$\dot{\boldsymbol{p}}(\widehat{\boldsymbol{\theta}}_n)^{\mathsf{T}} D_{\widehat{\boldsymbol{\theta}}_n}^{-1} \boldsymbol{X} = \boldsymbol{0}. \tag{7.5}$$

Asymptotické rozdělení $\widehat{\boldsymbol{\theta}}_n$ je

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_d(0, (A^{\mathsf{T}}A)^{-1}).$$

Asymptotické rozdělení $p(\widehat{\boldsymbol{\theta}}_n)$ můžeme dostat delta metodou. Jeho asymptotický rozptyl je menší než asymptotický rozptyl odhadu $\widehat{\boldsymbol{p}}_n$ uvažovaného v kapitole 7.2.

Nyní zobecníme větu 7.5 na současný případ.

Věta 7.6 (Asymptotické vlastnosti multinomického rozdělení s odhadnutými parametry.)**.**

Nechť platí model $\mathcal{F}_0$, tj. $\boldsymbol{X} \sim \mathsf{Mult}_K(n, \boldsymbol{p}(\boldsymbol{\theta}_X))$, kde $\boldsymbol{\theta}_X \in \Theta \subset \mathbb{R}^d$, $d < K$. Nechť funkce $\boldsymbol{p}$ splňuje $\boldsymbol{p}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{1}_K = 1$, nechť $\dot{\boldsymbol{p}}(\boldsymbol{\theta})$ existuje a má plnou hodnost d. Nechť $\widehat{\boldsymbol{\theta}}_n$ je maximálně věrohodný odhad parametru $\boldsymbol{\theta}_X$ v modelu $\mathcal{F}_0$. Označme $D_{\boldsymbol{\theta}} = \mathrm{diag}\,(\boldsymbol{p}(\boldsymbol{\theta}))$ a $A = D_{\boldsymbol{\theta}_X}^{-1/2} \dot{\boldsymbol{p}}(\boldsymbol{\theta}_X)$. Pak

(i)

$$\boldsymbol{Z}_n^* \overset{\mathsf{df}}{=} \frac{1}{\sqrt{n}} D_{\widehat{\boldsymbol{\theta}}_n}^{-1/2} [\boldsymbol{X} - n\boldsymbol{p}(\widehat{\boldsymbol{\theta}}_n)] \xrightarrow{\mathsf{D}} \mathsf{N}_K(\boldsymbol{0}, I_K - \sqrt{\boldsymbol{p}_X}^{\otimes 2} - A(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}),$$

(ii)

$$\boldsymbol{Z}_n^{*\mathsf{T}} \boldsymbol{Z}_n^* = \sum_{k=1}^{K} \frac{[X_k - np_k(\widehat{\boldsymbol{\theta}}_n)]^2}{np_k(\widehat{\boldsymbol{\theta}}_n)} \xrightarrow{\mathsf{D}} \chi^2_{K-d-1}.$$

χ^2 test dobré shody pro multinomické rozdělení s odhadnutými parametry

Věta 7.6(ii) poskytuje nástroj pro testování hypotézy

$$H_0 : \exists \boldsymbol{\theta}_X \in \Theta \quad \boldsymbol{p}_X = \boldsymbol{p}(\boldsymbol{\theta}_X)$$

proti alternativě

$$H_1 : \forall \boldsymbol{\theta}_X \in \Theta \quad \boldsymbol{p}_X \neq \boldsymbol{p}(\boldsymbol{\theta}_X),$$

to jest testování platnosti menšího modelu $\mathcal{F}_0$ proti širšímu modelu, podle něhož má $\boldsymbol{X}$ zcela libovolné multinomické rozdělení.

Nejprve odhadneme parametr $\boldsymbol{\theta}_X$ vyřešením soustavy (7.5). Platí-li hypotéza H_0, pak testová statistika

$$\chi^2 = \sum_{k=1}^{K} \frac{[X_k - np_k(\widehat{\boldsymbol{\theta}}_n)]^2}{np_k(\widehat{\boldsymbol{\theta}}_n)}$$

má asymptoticky rozdělení χ^2_{K-d-1}, tj. proti situaci se známým parametrem ztrácíme jeden stupeň volnosti za každý odhadovaný parametr. 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

$$\chi^2 = \sum_{k=1}^{K} \frac{[X_k - np_k(\widehat{\boldsymbol{\theta}}_n)]^2}{np_k(\widehat{\boldsymbol{\theta}}_n)} \geq \chi^2_{K-d-1}(1-\alpha),$$

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. Uvažujme náhodný výběr $Z_1, \ldots, Z_n$. Chceme vědět, zdali tento náhodný výběr pochází z nějaké parametrické rodiny 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 počet pozorování n. Spočítáme, kolik pozorování padlo do k-tého intervalu: $X_k = \sum_{i=1}^{n} \mathbb{I}_{(a_{k-1}, a_k\rangle}(Z_i)$. Odhadneme parametr $\boldsymbol{\theta}_X$ řešením soustavy (7.5). Pochází-li náhodný výběr $Z_1, \ldots, Z_n$ z dané rodiny rozdělení, pravděpodobnosti jednotlivých intervalů jsou přibližně $p_k(\widehat{\boldsymbol{\theta}}_n) = F(a_k; \widehat{\boldsymbol{\theta}}_n) - F(a_{k-1}; \widehat{\boldsymbol{\theta}}_n)$. Zamítne-li χ^2 test hypotézu H_0, prokázali jsme, že náhodný výběr $Z_1, \ldots, Z_n$ nepochází z dané rodiny rozdělení.

8 Dvouvýběrové kategoriální problémy a kontingenční tabulky

8.1 Dvouvýběrové kategoriální problémy

Nyní se budeme zabývat porovnáním dvou nezávislých binomických veličin $X_1 \sim \mathsf{Bi}(n, p_1)$ a $X_2 \sim \mathsf{Bi}(m, p_2)$.

Cílem je zjistit, zdali a jakým způsobem se liší pravděpodobnosti p_1 a p_2. Jejich odlišnost můžeme vyjádřit různými způsoby, podle toho dostaneme několik alternativních variant odhadů a testů.

Pokud chápeme události, jejichž počty udávají X_1 a X_2, negativně (smrt, nemoc, nezaměstnanost, bankrot) můžeme interpretovat parametry p_1 a p_2 jako *rizika* události v obou populacích.

Pravděpodobnosti (rizika) p_1 a p_2 můžeme odhadnout metodami popsanými v kapitole 7.1, tj. $\widehat{p}_1 = X_1/n$, $\widehat{p}_2 = X_2/m$. Platí pro ně věta 7.1.

U všech asymptotických výsledků budeme podobně jako v kapitole 6 předpokládat $n \to \infty$, $m \to \infty$ a $n/m \to q$, kde $0 < q < \infty$.

8.1.1 Rozdíly pravděpodobností, nárůst rizika

První způsob, jak můžeme vyjádřit odlišnost obou rozdělení, je *rozdíl pravděpodobností (rizik)** $d_X = p_1 - p_2$. Tento parametr říká, o kolik je větší riziko v populaci 1 než v populaci 2. Může nabývat hodnot -1 až 1. Pravděpodobnosti (rizika) v obou populacích jsou totožná právě když $d_X = 0$.

Nestranným a konsistentním odhadem parametru d_X je $\widehat{d} = \widehat{p}_1 - \widehat{p}_2$. Z věty 7.1(iv) a z nezávislosti X_1 a X_2 dostaneme technikou velmi podobnou důkazu věty 6.3

Tvrzení 8.1.

$$\frac{\widehat{d} - d_X}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Pro asymptotický test hypotézy $H_0 : d_X = 0$ proti alternativě $H_1 : d_X \neq 0$

* Angl. *risk difference, excess risk*

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 $|T_d| \geq u_{1-\alpha/2}$.

Přibližný interval spolehlivosti pro d_X je

$$P\Bigg[\widehat{d} - \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}\, u_{1-\alpha/2} < d_X < \widehat{d} + \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}\, u_{1-\alpha/2}\Bigg] \to 1-\alpha.$$

Tato metoda je obdobou dvouvýběrového z-testu pro nominální data.

8.1.2 Podíly pravděpodobností, relativní riziko

Jiný způsob, jak vyjádřit odlišnost obou rozdělení, je *relativní riziko** $r_X = p_1/p_2$. Tento parametr říká, kolikrát je větší riziko v populaci 1 než v populaci 2. Může nabývat hodnot v intervalu $(0, \infty)$. Pravděpodobnosti (rizika) v obou populacích jsou totožná 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 7.1(iv) 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 8.2.

$$\frac{\log \widehat{r} - \log r_X}{\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Pravděpodobnosti (rizika) v obou populacích jsou totožné právě když $r_X = 1$ neboli $\log r_X = 0$. Pro asymptotický test hypotézy $H_0 : r_X = 1$ proti alternativě $H_1 : r_X \neq 1$ použijeme testovou statistiku

$$T_r = \frac{\log \widehat{r}}{\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}}$$

* Angl. *relative risk*

a hypotézu zamítneme pokud $|T_r| \geq u_{1-\alpha/2}$.

Přibližný interval spolehlivosti pro relativní riziko r_X je

$$P\left[\widehat{r}\exp\left\{-\sqrt{\tfrac{1-\widehat{p}_1}{n\widehat{p}_1}+\tfrac{1-\widehat{p}_2}{m\widehat{p}_2}}\,u_{1-\alpha/2}\right\} < r_X < \widehat{r}\exp\left\{\sqrt{\tfrac{1-\widehat{p}_1}{n\widehat{p}_1}+\tfrac{1-\widehat{p}_2}{m\widehat{p}_2}}\,u_{1-\alpha/2}\right\}\right] \to 1-\alpha.$$

8.1.3 Poměr šancí

Třetím možným způsobem vyjádření odlišnosti obou rozdělení 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 7.2(i) a z nezávislosti X_1 a X_2 plyne následující tvrzení:

Tvrzení 8.3. 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

$$T_o = \frac{\log\widehat{o}}{\sqrt{\widehat{V}_o}}$$

[*] Angl. *odds ratio*

a hypotézu zamítneme pokud $|T_o| \geq u_{1-\alpha/2}$.

Přibližný interval spolehlivosti pro poměr šancí o_X je

$$P\left[\widehat{o}\exp\left\{-\sqrt{\widehat{V}_o}\,u_{1-\alpha/2}\right\} < o_X < \widehat{o}\exp\left\{\sqrt{\widehat{V}_o}\,u_{1-\alpha/2}\right\}\right] \to 1-\alpha.$$

8.2 Kontingenční tabulky

Nechť $X \in \{1,\dots,J\}$ a $Z \in \{1,\dots,K\}$ jsou dvě kategoriální veličiny. Uvažujme náhodný výběr $(X_1,Z_1)^\mathsf{T},\dots,(X_N,Z_N)^\mathsf{T}$ o rozsahu N (pevném). Označme počet jedinců klasifikovaných do j-té kategorie veličiny X a k-té kategorie veličiny Z jako $n_{jk} = \sum_{i=1}^N \mathbb{I}\{X_i = j, Z_i = k\}$, $j = 1,\dots,J$, $k = 1,\dots,K$. Náhodnou veličinu n_{jk} nazýváme pozorovanou četností kategorie j,k. Nechť $p_{jk} = \mathrm{P}\left[X = j, Z = k\right]$.

Zavedeme-li vektory $\boldsymbol{n} = (n_{11},\dots,n_{JK})^\mathsf{T}$ a $\boldsymbol{p} = (p_{11},\dots,p_{JK})^\mathsf{T}$, pak $\boldsymbol{n} \sim \mathsf{Mult}_{JK}(N,\boldsymbol{p})$. Označme $m_{jk} = \mathsf{E}\, n_{jk} = Np_{jk}$.

Označme dále

$$n_{j+} = \sum_{k=1}^K n_{jk}, \quad n_{+k} = \sum_{j=1}^J n_{jk}, \quad n_{++} = \sum_{j=1}^J\sum_{k=1}^K n_{jk} = N,$$
$$m_{j+} = \sum_{k=1}^K m_{jk}, \quad m_{+k} = \sum_{j=1}^J m_{jk}, \quad m_{++} = \sum_{j=1}^J\sum_{k=1}^K m_{jk} = N,$$
$$p_{j+} = \sum_{k=1}^K p_{jk}, \quad p_{+k} = \sum_{j=1}^J p_{jk}, \quad p_{++} = \sum_{j=1}^J\sum_{k=1}^K p_{jk} = 1.$$

Pravděpodobnosti p_{jk} určují sdružené rozdělení X a Z. Pravděpodobnosti $p_{j+} = \mathrm{P}\left[X = j\right]$ určují marginální rozdělení X. Pravděpodobnosti $p_{+k} = \mathrm{P}\left[Z = k\right]$ určují marginální rozdělení Z.

Pozorované četnosti můžeme sestavit do tabulky, kterou nazýváme *kontingenční tabulka*[*].

	$Z=1$	$\dots$	$Z=K$	$\sum$
$X=1$	n_{11}	$\dots$	n_{1K}	n_{1+}
$X=2$	n_{21}	$\dots$	n_{2K}	n_{2+}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$X=J$	n_{J1}	$\dots$	n_{JK}	n_{J+}
$\sum$	n_{+1}	$\dots$	n_{+K}	N

Podobně můžeme sestavit tabulku pravděpodobností, která popisuje sdružené rozdělení vektoru $(X,Z)^\mathsf{T}$ i marginální rozdělení veličin X a Z.

[*] Angl. *contingency table*

	$Z=1$	$\dots$	$Z=K$	$\sum$
$X=1$	p_{11}	$\dots$	p_{1K}	p_{1+}
$X=2$	p_{21}	$\dots$	p_{2K}	p_{2+}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$X=J$	p_{J1}	$\dots$	p_{JK}	p_{J+}
$\sum$	p_{+1}	$\dots$	p_{+K}	1

Označme ještě podmíněné pravděpodobnosti

$$\mathrm{P}\left[X=j \mid Z=k\right] = p_{j(k)} = \frac{p_{jk}}{p_{+k}},$$
$$\mathrm{P}\left[Z=k \mid X=j\right] = p_{(j)k} = \frac{p_{jk}}{p_{j+}}.$$

8.2.1 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:

	$Z=1$	$Z=2$	$\sum$
$X=1$	n_{11}	n_{12}	n_{1+}
$X=2$	n_{21}	n_{22}	n_{2+}
$\sum$	n_{+1}	n_{+2}	N

	$Z=1$	$Z=2$	$\sum$
$X=1$	p_{11}	p_{12}	p_{1+}
$X=2$	p_{21}	p_{22}	p_{2+}
$\sum$	p_{+1}	p_{+2}	1

Tuto situaci jsme vlastně řešili v kapitole 8.1. 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 8.1 můžeme snadno převést na značení používané nyní a naopak. Například studovaná kontingenční tabulka přepsaná ve značení kapitoly 8.1 je

	$Z=1$	$Z=2$	$\sum$
$X=1$	X_1	X_2	X_1+X_2
$X=2$	$n-X_1$	$m-X_2$	$n+m-X_1-X_2$
$\sum$	n	m	$n+m$

Rozdíl proti situaci v kapitole 8.1 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 6 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.

Protože v naší současné formulaci pracujeme s multinomickým rozdělením, můžeme používat všechny výsledky z kapitoly 7.2 a 7.3. Odhadem pravděpodobnosti p_{ij} je n_{ij}/N. Odhadem vektoru $\boldsymbol{p}$ je $\widehat{\boldsymbol{p}}_n = \boldsymbol{n}/N$.

Kapitola 8.1 nám dává návod, jak porovnat riziko události $[X = 1]$ pro různé hodnoty Z. Můžeme použít tři způsoby porovnání:

- *rozdíl rizik* $d_X = p_{1(1)} - p_{1(2)}$ odhadneme pomocí $\widehat{d} = \frac{n_{11}}{n_{+1}} - \frac{n_{12}}{n_{+2}}$;
- *relativní riziko* $r_X = p_{1(1)}/p_{1(2)}$ 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 8.1.

Jsou-li náhodné veličiny X a Z nezávislé, musí pro každé $j, k \in \{1, 2\}$ platit

$$\mathrm{P}\left[X = j, Z = k\right] = \mathrm{P}\left[X = j\right]\mathrm{P}\left[Z = k\right] \quad \text{neboli} \quad p_{jk} = p_{j+}p_{+k}$$

anebo ekvivalentně

$$\mathrm{P}\left[X = j \mid Z = k\right] = \mathrm{P}\left[X = j\right] \quad \text{neboli} \quad p_{j(k)} = p_{j+}.$$

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

Jinou metodu jak otestovat nezávislost X a Z poskytuje χ^2 test dobré shody pro multinomické rozdělení s odhadnutými parametry založený na větě 7.6. Pokud

* Angl. *cross ratio*

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 $\widehat{\boldsymbol{\theta}}_n = (\widehat{p}_{1+}, \widehat{p}_{+1})^\mathsf{T} = (n_{1+}/N, n_{+1}/N)^\mathsf{T}$. 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 $Np_{jk}(\widehat{\boldsymbol{\theta}}_n) = N\widehat{p}_{j+}\widehat{p}_{+k} = n_{j+}n_{+k}/N$. Počet parametrů za nulové hypotézy 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 $\chi^2 \geq \chi^2_1(1 - \alpha)$.

8.2.2 Kontingenční tabulky 2 × K

Nyní rozšíříme zkoumanou situaci na případ $J = 2$ a $K \geq 2$. Kontingenční tabulka obsahuje $2 \times K$ četností:

	$Z = 1$	$Z = 2$	$\ldots$	$Z = K$	$\sum$
$X = 1$	n_{11}	n_{12}	$\ldots$	n_{1K}	n_{1+}
$X = 2$	n_{21}	n_{22}	$\ldots$	n_{2K}	n_{2+}
$\sum$	n_{+1}	n_{+2}	$\ldots$	n_{+K}	N

	$Z = 1$	$Z = 2$	$\ldots$	$Z = K$	$\sum$
$X = 1$	p_{11}	p_{12}	$\ldots$	p_{1K}	p_{1+}
$X = 2$	p_{21}	p_{22}	$\ldots$	p_{2K}	p_{2+}
$\sum$	p_{+1}	p_{+2}	$\ldots$	p_{+K}	N

Toto je zobecnění situace řešené v kapitole 8.1. 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

Jsou-li náhodné veličiny X a Z nezávislé, musí pro každé $j = 1, 2$ a $k = 1, \ldots, K$ platit

$$\mathrm{P}\left[X = j, Z = k\right] = \mathrm{P}\left[X = j\right] \mathrm{P}\left[Z = k\right] \quad \text{neboli} \quad p_{jk} = p_{j+}p_{+k}$$

anebo ekvivalentně

$$\mathrm{P}\left[X = j \mid Z = k\right] = \mathrm{P}\left[X = j\right] \quad \text{neboli} \quad p_{j(k)} = p_{j+}.$$

Jelikož $p_{2(k)} = 1 - p_{1(k)}$, nezávislost platí právě když $p_{1(1)} = p_{1(2)} = \ldots = p_{1(K)}$.

To vyžaduje, aby mezi kterýmikoli dvěma skupinami 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 vlastně funkcemi pouze K parametrů p_{1+} a $p_{+1}, \ldots, p_{+(K-1)}$. 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 $\widehat{\boldsymbol{\theta}}_n = (\widehat{p}_{1+}, \widehat{p}_{+1}, \ldots, \widehat{p}_{+(K-1)})^{\mathsf{T}} = (n_{1+}/N, n_{+1}/N, \ldots, n_{+(K-1)}/N)$ 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 $Np_{jk}(\widehat{\boldsymbol{\theta}}_n) = N\widehat{p}_{j+}\widehat{p}_{+k} = n_{j+}n_{+k}/N$. Počet parametrů za nulové hypotézy je $d = K$.

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).

8.2.3 Kontingenční tabulky J × K

Zobecnění na situaci $J \geq 2$ a $K \geq 2$ je nyní snadné. Kontingenční tabulka obsahuje $J \times K$ četností:

	$Z=1$	$\ldots$	$Z=K$	$\sum$
$X=1$	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+}
$\sum$	n_{+1}	$\ldots$	n_{+K}	N

	$Z=1$	$\ldots$	$Z=K$	$\sum$
$X=1$	p_{11}	$\ldots$	p_{1K}	p_{1+}
$X=2$	p_{21}	$\ldots$	p_{2K}	p_{2+}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$X=J$	p_{J1}	$\ldots$	p_{JK}	p_{J+}
$\sum$	p_{+1}	$\ldots$	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

Jsou-li náhodné veličiny X a Z nezávislé, musí pro každé $j = 1, \ldots, J$ a $k = 1, \ldots, K$ platit

$$\mathrm{P}\left[X = j, Z = k\right] = \mathrm{P}\left[X = j\right]\mathrm{P}\left[Z = k\right] \quad \text{neboli} \quad p_{jk} = p_{j+}p_{+k}$$

anebo ekvivalentně

$$\mathrm{P}\left[X = j \mid Z = k\right] = \mathrm{P}\left[X = j\right] \quad \text{neboli} \quad p_{j(k)} = p_{j+}.$$

Nezávislost 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 $X = j, j'$ a $Z = k, k'$ 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 vlastně funkcemi pouze $J - 1 + K - 1$ parametrů $p_{1+}, \ldots, p_{(J-1)+}$ a $p_{+1}, \ldots, p_{+(K-1)}$. Máme tedy $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta}_X)$, kde $\boldsymbol{\theta}_X = (p_{1+}, \ldots, p_{(J-1)+}, p_{+1}, \ldots, p_{+(K-1)})^\mathsf{T}$. Maximálně věrohodný odhad parametru $\boldsymbol{\theta}_X$ za hypotézy nezávislosti je

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_n &= (\widehat{p}_{1+}, \ldots, \widehat{p}_{(J-1)+}, \widehat{p}_{+1}, \ldots, \widehat{p}_{+(K-1)})^\mathsf{T} \\ &= (n_{1+}/N, \ldots, n_{(J-1)+}/N, n_{+1}/N, \ldots, n_{+(K-1)}/N)^\mathsf{T}.\end{aligned}$$

Maximálně věrohodné odhady složek vektoru $\boldsymbol{p}$ za hypotézy nezávislosti jsou

$$p_{jk}(\widehat{\boldsymbol{\theta}}_n) = \widehat{p}_{j+}\widehat{p}_{+k} = \frac{n_{j+}n_{+k}}{N^2},$$

$j = 1, \ldots, J$, $k = 1, \ldots, K$. Očekávané četnosti v kontingenční tabulce za platnosti hypotézy jsou $Np_{jk}(\widehat{\boldsymbol{\theta}}_n) = N\widehat{p}_{j+}\widehat{p}_{+k} = n_{j+}n_{+k}/N$. Počet parametrů za nulové hypotézy je $d = J + K - 2$.

Testová statistika je

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}.$$

Za platnosti hypotézy nezávislosti má asymptoticky rozdělení $\chi^2_{JK-(J+K-2)-1}$, tj. $\chi^2_{(J-1)(K-1)}$. Hypotézu nezávislosti zamítneme, pokud $\chi^2 \geq \chi^2_{(J-1)(K-1)}(1-\alpha)$.

Test nezávislosti zároveň testuje i hypotézu že K (J) výběrů z multinomického rozdělení má stejné vektory pravděpodobností (jde tedy o K-výběrový test na multinomické rozdělení).

9 Analýza rozptylu

Dvouvýběrové testy se hodí, chceme-li zjistit, jestli se dvě disjunktní 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 tento problém řešili v minulé kapitole. Nyní budeme studovat tento problém pro nominální (numerické) náhodné veličiny.

Máme $k \geq 2$ nezávislých náhodných výběrů

$$\begin{aligned} Y_{11}, \ldots, Y_{1n_1} &\text{ z rozdělení } F_1, \\ Y_{21}, \ldots, Y_{2n_2} &\text{ z rozdělení } F_2, \\ &\vdots \\ \text{a } Y_{k1}, \ldots, Y_{kn_k} &\text{ z rozdělení } F_k. \end{aligned}$$

Pozorování označujeme Y_{ij}, kde i je číslo výběru jdoucí od 1 do k a j je index pozorování v rámci daného výběru běžící od 1 do n_i, kde n_i je rozsah i-tého výběru.

Model $\mathcal{F}$ specifikuje množinu uvažovaných distribučních funkcí $F_1, \ldots, F_k$. Parametrem, který chceme porovnat, budiž střední hodnota. Označme si $\mu_i = \mathsf{E}\, Y_{ij}$ střední hodnotu i-tého výběru. Chceme testovat hypotézu

$$H_0 : \mu_1 = \cdots = \mu_k$$

proti alternativě

$$H_1 : \exists i \neq j : \mu_i \neq \mu_j.$$

Kdybychom chtěli porovnat střední hodnoty pouze dvou vybraných skupin i a j, mohli bychom použít třeba dvouvýběrový t-test nebo z-test. Šlo by takto provést dvouvýběrové testy pro všechny možné dvojice skupin a otestovat všechny hypotézy $H_0^{ij} : \mu_i = \mu_j$. Pokud by některý test zamítl H_0^{ij} na hladině α, pak střední hodnoty nemohou být u všech výběrů stejné. Bohužel, jak lze snadno nahlédnout, celková pravděpodobnost zamítnutí platné H_0 by byla mnohem větší než ono α, na němž provádíme jednotlivé dvouvýběrové testy hypotéz H_0^{ij}. Proto potřebujeme jinou metodu, která zaručí dodržení požadované hladiny. Metoda, kterou si nyní ukážeme, se nazývá *analýza rozptylu*[*]. (Vlastně půjde jen o nejjednodušší speciální případ analýzy rozptylu, tzv. jednoduché třídění – *one way ANOVA*.)

[*] Angl. *analysis of variance, ANOVA*

Model:

$$\mathcal{F} = \{F_i = \mathsf{N}(\mu_i, \sigma^2), \mu_i \in \mathbb{R}, i = 1, \dots, k, \sigma^2 > 0\}$$

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

Testované parametry: Střední hodnoty $\mu_i = \mathsf{E}\, Y_{ij}$

Hypotéza a alternativa:

$$H_0 : \mu_1 = \cdots = \mu_k, \quad H_1 : \exists i \neq j : \mu_i \neq \mu_j.$$

Testujeme, zdali mají všechny výběry stejnou střední hodnotu.

Značení. Označme $n = \sum_{i=1}^{k} n_i$. Nechť $Y_{i+} \overset{\mathsf{df}}{=} \sum_{j=1}^{n_i} Y_{ij}$ a $\overline{Y}_{i+} \overset{\mathsf{df}}{=} n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$ jsou součty a průměry jednotlivých výběrů, nechť $Y_{++} \overset{\mathsf{df}}{=} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$ je celkový součet a $\overline{Y}_{++} \overset{\mathsf{df}}{=} n^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$ je celkový průměr.

Definice 9.1. Součty čtverců v analýze rozptylu

- $SS_C \overset{\mathsf{df}}{=} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{++})^2$ nazýváme *celkový součet čtverců*[*].
- $SS_A \overset{\mathsf{df}}{=} \sum_{i=1}^{k} n_i (\overline{Y}_{i+} - \overline{Y}_{++})^2$ nazýváme *součet čtverců skupin*[†].
- $SS_e \overset{\mathsf{df}}{=} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2$ nazýváme *residuální součet čtverců*[‡].

Věta 9.1. Platí

$$SS_C = SS_A + SS_e.$$

Poznámka. 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.

Věta 9.2 (rozdělení součtů čtverců)**.** Za platnosti modelu $\mathcal{F}$ máme

1. $$\frac{SS_e}{\sigma^2} \sim \chi^2_{n-k}, \quad \mathsf{E}\, \frac{SS_e}{n-k} = \sigma^2.$$

2. Platí-li navíc hypotéza H_0, pak
$$\frac{SS_C}{\sigma^2} \sim \chi^2_{n-1}, \quad \mathsf{E}\, \frac{SS_C}{n-1} = \sigma^2.$$

[*] Angl. *total sum of squares* [†] Angl. *between group sum of squares* [‡] Angl. *residual sum of squares, error sum of squares*

3. Platí-li navíc hypotéza H_0, pak

$$\frac{SS_A}{\sigma^2} \sim \chi^2_{k-1}, \quad \mathsf{E}\,\frac{SS_A}{k-1} = \sigma^2.$$

4. Platí-li navíc hypotéza H_0, pak SS_A a SS_e jsou nezávislé.

Poznámka.

- $SS_e/(n-k)$ je vždy nestranným odhadem rozptylu σ^2 (bez ohledu na platnost hypotézy).
- $SS_A/(k-1)$ je nestranným odhadem rozptylu pouze za hypotézy. Pokud hypotéza neplatí, lze ukázat, že

$$\mathsf{E}\,\frac{SS_A}{k-1} = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2,$$

  kde $\overline{\mu} = n^{-1}\sum_{i=1}^{k} n_i\mu_i$.
- Tato metoda se nazývá analýza rozptylu kvůli tomu, jakým způsobem je sestavena testová statistika. Účelem analýzy rozptylu není analyzovat rozptyl.

Testová statistika:

$$F_A = \frac{SS_A}{k-1} \Big/ \frac{SS_e}{n-k}$$

Hypotézu budeme zamítat pro příliš velké hodnoty F_A.

Věta 9.3. Za platnosti modelu $\mathcal{F}$ a hypotézy H_0 platí $F_A \sim F_{k-1,n-k}$.

Kritický obor:

$$H_0 \text{ zamítneme } \Leftrightarrow F_A \geq F_{k-1,n-k}(1-\alpha)$$

kde $F_{k-1,n-k}(1-\alpha)$ je $(1-\alpha)$-tý kvantil F rozdělení s $k-1$ a $n-k$ stupni volnosti.

Poznámka. Tetno test se nazývá F test analýzy rozptylu. Je to přesný test rovnosti středních hodnot v $k \geq 2$ nezávislých výběrech. Vyžaduje normální rozdělení a stejný rozptyl ve všech výběrech.

P-hodnota: $p = 1 - F^*(s)$, kde s je pozorovaná hodnota testové statistiky a F^* je distribuční funkce rozdělení $F_{k-1,n-k}$.

Poznámka. Výsledky analýzy rozptylu se tradičně uvádějí formou tabulky.

Zdroj měnlivosti	Součet čtverců	Stupňů volnosti	Podíl	F
Skupina	SS_A	$k-1$	$\frac{SS_A}{k-1}$	$\frac{SS_A}{k-1} \Big/ \frac{SS_e}{n-k}$
Residuální	SS_e	$n-k$	$\frac{SS_e}{n-k}$	
Celkový	SS_C	$n-1$		

Věta 9.4. Pokud $k = 2$, 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 dvouvýběrovéto t-testu. Pro porovnání dvou skupin je tedy analýza rozptylu ekvivalentní dvouvýběrovému t-testu.

10 Základy regrese

10.1 Korelační analýza

Nechť $\binom{X}{Y}$ je spojitý náhodný vektor s dvourozměrnou distribuční funkcí F. Nechť

$$\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$$

je náhodný výběr z téže distribuční funkce o rozsahu $n \geq 4$. Korelační koeficient veličin X a Y

$$\varrho \equiv \varrho(X, Y) = \frac{\mathsf{cov}\,(X, Y)}{\sqrt{\mathsf{var}\, X\, \mathsf{var}\, Y}}$$

měří sílu lineární závislosti mezi nimi.

V definici 3.3 jsme zavedli výběrový korelační koeficient

$$\widehat{\varrho}_n = \frac{S_{XY}}{S_X S_Y},$$

kde

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

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

$$\text{a } S_Y^2 = \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2.$$

Víme, že se jedná o konsistentní odhad korelačního koeficientu ϱ.

Později v této kapitole dokážeme větu, s jejíž pomocí budeme moci odvodit následující tvrzení.

Věta 10.1. Nechť

$$\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$$

je náhodný výběr z dvourozměrného normálního rozdělení s nulovým korelačním koeficientem $\varrho(X, Y) = 0$. Pak pro výběrový korelační koeficient $\widehat{\varrho}_n$ platí

$$T_\varrho = \sqrt{n-2}\frac{\widehat{\varrho}_n}{\sqrt{1-\widehat{\varrho}_n^2}} \sim t_{n-2}.$$

Tato věta nám dává nástroj k testování hypotézy $H_0 : \varrho = 0$ proti alternativě $H_1 : \varrho \neq 0$ za předpokladu dvourozměrného normálního rozdělení. Hypotézu H_0 zamítneme na hladině α, pokud

$$|T_\varrho| \geq t_{n-2}(1-\alpha/2),$$

kde $t_{n-2}(1-\alpha/2)$ je $(1-\alpha/2)$-kvantil rozdělení t_{n-2}. Pomocí věty 10.1 ovšem nelze sestrojit interval spolehlivosti pro ϱ.

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 10.2. Nechť

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

je náhodný výběr z dvourozměrného normálního rozdělení s korelačním koeficientem ϱ. Pak pro výběrový korelační koeficient $\widehat{\varrho}_n$ platí

$$\sqrt{n-3}(\operatorname{arctgh}\widehat{\varrho}_n - \operatorname{arctgh}\varrho) \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Chceme-li otestovat hypotézu $H_0 : \varrho = \varrho_0$ proti alternativě $H_1 : \varrho \neq \varrho_0$, spočítáme testovou statistiku

$$Z_n = \sqrt{n-3}(\operatorname{arctgh}\widehat{\varrho}_n - \operatorname{arctgh}\varrho_0)$$

a H_0 zamítneme na hladině α, pokud

$$|Z_n| \geq u_{1-\alpha/2}.$$

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

$$\big(\operatorname{tgh}(\operatorname{arctgh}\widehat{\varrho}_n - u_{1-\alpha/2}/\sqrt{n-3}),\ \operatorname{tgh}(\operatorname{arctgh}\widehat{\varrho}_n + u_{1-\alpha/2}/\sqrt{n-3})\big).$$

10.2 Model lineární regrese

Lineární regrese zkoumá vztah mezi spojitou veličinou Y a vektorem $\boldsymbol{X}$, který může obsahovat jednu nebo více spojitých či diskrétních veličin. Předpokládáme, že hodnota vektoru $\boldsymbol{X}$ může ovlivňovat střední hodnotu Y, ale nikoli rozptyl Y. Zajímá nás, které komponenty $\boldsymbol{X}$ ovlivňují $\mathsf{E}\,Y$ a jakým způsobem. Můžeme také chtít předpovídat Y pro danou hodnotu $\boldsymbol{X}$.

Data se sestávají z n nezávislých pozorování vektorů $(Y_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$, kde každé X_i má $p < n$ složek $(X_{i1}, \dots, X_{ip})$.

Definice 10.1.
- Náhodnou veličinu Y_i nazýváme *odezva*. Alternativní název: *závisle proměnná*[*].
- Komponenty náhodného vektoru $\boldsymbol{X}_i$ nazýváme *regresory*. Alternativní názvy: *nezávisle proměnné, vysvětlující veličiny, prediktory, kovariáty*[†].

Poznámka. Původní data nemusí obsahovat přímo pozorování náhodných vektorů $\boldsymbol{X}_i$, ale nějaké jiné veličiny $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{iq})^\mathsf{T}$, z nichž $\boldsymbol{X}_i$ spočítáme nějakou transformací $\boldsymbol{X}_i = h(\boldsymbol{Z}_i)$. Jedním z problémů, které regresní analýza řeší, je určení vhodné transformace h původních dat $\boldsymbol{Z}_i$. My se tu ale tímto problémem zabývat nebudeme. Budeme předpokládat, že máme dány konkrétní již ztransformované regresory $\boldsymbol{X}_i$.

Definice 10.2. Řekneme, že data $(Y_i, \boldsymbol{X_i})$, $i = 1, \dots, n$, splňují *lineární regresní model*[‡], pokud platí

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \tag{10.1}$$

kde $\boldsymbol{\beta} = (\beta_1, \dots \beta_p)^\mathsf{T}$ je vektor neznámých parametrů a $\varepsilon_1, \dots, \varepsilon_n$ jsou nezávislé náhodné veličiny splňující $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{var}\,\varepsilon_i = \sigma^2$. Složky vektoru $\boldsymbol{\beta}$ nazýváme *regresní koeficienty*[§], náhodné veličiny ε_i nazýváme *chybové členy*[¶].

Model 10.1 můžeme přepsat několika dalšími způsoby. Například pomocí podmíněných momentů:

$$\begin{aligned} \mathsf{E}\,(Y_i \mid \boldsymbol{X}_i) &= \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \\ \mathsf{var}\,(Y_i \mid \boldsymbol{X}_i) &= \sigma^2 \end{aligned}$$

Tento zápis zdůrazňuje, že lineární regresní model vyjadřuje podmíněnou střední hodnotu Y_i, je-li dáno $\boldsymbol{X}_i$, pomocí lineárního vztahu a předpokládá, že rozptyl Y_i je konstatní a nezávisí na $\boldsymbol{X}_i$.

---

[*] Angl. *response, dependent variable, outcome* [†] Angl. *regressors, independent variable, explanatory variable, predictors, covariates* [‡] Angl. *linear regression model* [§] Angl. *regression coefficients* [¶] Angl. *error terms*

Značení. Nechť $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\mathsf{T}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\mathsf{T}$ a

$$X = \begin{pmatrix} \boldsymbol{X}_1^\mathsf{T} \\ \boldsymbol{X}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{X}_n^\mathsf{T} \end{pmatrix}.$$

Matice X se nazývá *regresní matice*; má n řádků a p sloupců.

Regresní matice obsahuje v řádcích regresory jednotlivých pozorování. Budeme předpokládat, že X má plnou hodnost, tj. $r(X) = p$ čili sloupce matice X jsou lineárně nezávislé. Model 10.1 nyní můžeme přepsat vektorově:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ a $\mathsf{var}\,\boldsymbol{\varepsilon} = \sigma^2 I_n$.

Poznámka. V regresní analýze většinou volíme $X_{i1} = 1$ pro všechna i. Parametr β_1 pak nazýváme *absolutní člen**.

Příklady.

Absolutní člen Nejjednodušší regresní model obsahuje pouze absolutní člen: $Y_i = \beta_1 + \varepsilon_i$. Platí-li tento model, $Y_1, \dots, Y_n$ jsou nezávislé veličiny se stejnou střední hodnotou a stejným rozptylem. Jsou-li $Y_1, \dots, Y_n$ náhodným výběrem z rozdělení s konečným druhým momentem, pak splňují tento regresní model. Regresní matice je sloupec délky n obsahující jedničky.

Dvě skupiny Nechť Z_i je náhodná veličina nabývající hodnot 0 nebo 1 pozorovaná spolu s Y_i. Definujme $X_{i2} = Z_i$. Regresní model jest $Y_i = \beta_1 + \beta_2 X_{i2} + \varepsilon_i$.

Platí-li tento model, $Y_1, \dots, Y_n$ se dělí na dvě skupiny podle hodnoty Z_i. Je-li $Z_i = 0$, pak $\mathsf{E}\,Y_i = \beta_1$. Je-li $Z_i = 1$, pak $\mathsf{E}\,Y_i = \beta_1 + \beta_2$. Parametr β_2 představuje rozdíl středních hodnot obou skupin. Jsou-li $Y_1, \dots, Y_n$ dva (spojené) nezávislé náhodné výběry z rozdělení s konečnými druhými momenty, pak splňují tento regresní model. Přidáme-li předpoklad normality, tento model odpovídá modelu pro dvouvýběrový t-test.

Regresní matice X obsahuje v j-tém řádku vektor $(1, Z_i)$.

k skupin Nechť Z_i je náhodná veličina nabývající hodnot $1, \dots, k$ pozorovaná spolu s Y_i. Definujme $X_{ij} = \mathbb{I}_{\{Z_i = j\}}$ pro $j = 2, \dots, k$. Regresní model jest $Y_i = \beta_1 + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$.

* Angl. *intercept*

Platí-li tento model, $Y_1, \dots, Y_n$ se dělí na k skupin podle hodnoty Z_i. Je-li $Z_i = 1$, pak $\mathsf{E}\, Y_i = \beta_1$. Je-li $Z_i = j > 1$, pak $\mathsf{E}\, Y_i = \beta_1 + \beta_j$. Parametr β_j představuje rozdíl středních hodnot j-té a první skupiny. Tvoří-li $Y_1, \dots, Y_n$ k (spojených) nezávislých náhodných výběrů z rozdělení s konečnými druhými momenty, pak splňují tento regresní model. Přidáme-li předpoklad normality, tento model odpovídá modelu pro analýzu rozptylu.

Regresní matice X obsahuje v j-tém řádku vektor $(1, \mathbb{I}_{\{Z_i=2\}}, \dots, \mathbb{I}_{\{Z_i=k\}})$.

Jednoduchá lineární regrese Nechť Z_i je nominální (nikoli kategoriální) náhodná veličina pozorovaná spolu s Y_i. Definujme $X_{i2} = Z_i$. Regresní model jest $Y_i = \beta_1 + \beta_2 X_{i2} + \varepsilon_i$. Tento model nazýváme *jednoduchá lineární regrese*. V tomto modelu je $\mathsf{E}\, Y_i$ lineární funkcí regresoru Z_i. Parametr β_2 se nazývá *směrnice regresní přímky*[*].

Dosadíme-li $Z_i = 0$, dostaneme $\beta_1 = \mathsf{E}\,(Y_i \mid Z_i = 0)$, tj. absolutní člen vyjadřuje střední hodnotu Y_i pro pozorování s nulovým regresorem. Dále,

$$\beta_2 = \mathsf{E}\,(Y_i \mid Z_i = x + 1) - \mathsf{E}\,(Y_i \mid Z_i = x),$$

čili β_2 vyjadřuje rozdíl ve střední hodnotě $\mathsf{E}\, Y_i$ po zvýšení regresoru X_i o jednu jednotku. Je-li $\beta_2 = 0$, znamená to, že regresor neovlivňuje střední hodnotu (ani rozptyl) Y_i.

Regresní matice X obsahuje v j-tém řádku vektor $(1, Z_i)$.

Kvadratická regrese Nechť Z_i je nominální (nikoli kategoriální) náhodná veličina pozorovaná spolu s Y_i. Definujme $X_{i2} = Z_i$ a $X_{i3} = Z_i^2$. Regresní model jest $Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$. Tento model nazýváme *kvadratická regrese*. V tomto modelu je $\mathsf{E}\, Y_i$ kvadratickou funkcí regresoru Z_i.

Je-li $\beta_3 = 0$, znamená to, že závislost střední hodnoty na Z_i není kvadratická, ale lineární. Je-li $\beta_2 = 0$ a $\beta_3 = 0$, znamená to, že regresor neovlivňuje střední hodnotu (ani rozptyl) Y_i.

Regresní matice X obsahuje v j-tém řádku vektor $(1, Z_i, Z_i^2)$.

Kvadratickou regresi lze snadno zobecnit na polynomiální regresi stupně p.

Po částech lineární regrese Nechť Z_i je nominální (nikoli kategoriální) náhodná veličina pozorovaná spolu s Y_i. Zvolme konstantu $c_0 \in \mathbb{R}$ a definujme $X_{i2} = Z_i$ a $X_{i3} = (Z_i - c_0)\mathbb{I}_{(c_0,\infty)}(Z_i)$. Regresní model jest $Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$. V tomto modelu je $\mathsf{E}\, Y_i$ spojitou po částech lineární funkcí regresoru Z_i se zlomem ve směrnici v bodě c_0.

[*] Angl. *slope*

Parametr β_2 vyjadřuje směrnici regresní přímky v intervalu $(-\infty, c_0)$. Parametr β_3 vyjadřuje změnu ve směrnici regresní přímky pro $Z_i > c_0$. Na intervalu (c_0, ∞) je směrnice $\beta_2 + \beta_3$.

Je-li $\beta_3 = 0$, znamená to, že směrnice se v bodě c_0 nemění a závislost střední hodnoty na Z_i je lineární. Je-li $\beta_2 = 0$, znamená to, že na intervalu $(-\infty, c_0)$ nezávisí střední hodnota Y_i na Z_i. Je-li $\beta_2 = 0$ a $\beta_3 = 0$, znamená to, že regresor neovlivňuje střední hodnotu (ani rozptyl) Y_i.

Regresní matice X obsahuje v j-tém řádku vektor $(1, Z_i, (Z_i - c_0)\mathbb{I}_{(c_0,\infty)}(Z_i))$.

Po částech konstantní regrese Nechť Z_i je nominální (nikoli kategoriální) náhodná veličina pozorovaná spolu s Y_i. Zvolme konstanty $c_0 = -\infty < c_1 < c_2 < \cdots < c_k = \infty$, $\mathcal{J}_k = (c_{k-1}, c_k\rangle$ a definujme $X_{ij} = \mathbb{I}_{\mathcal{J}_j}(Z_i)$, $j = 2, \ldots, k$. Regresní model jest $Y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$. V tomto modelu je $\mathsf{E}\, Y_i$ schodovitou po částech konstantní funkcí regresoru Z_i se skoky v bodech $c_1, \ldots, c_{k-1}$.

Je-li $Z_i < c_1$, pak $\mathsf{E}\, Y_i = \beta_1$. Je-li $Z_i \in \mathcal{J}_j$, $j > 1$, pak $\mathsf{E}\, Y_i = \beta_1 + \beta_j$. Parametr β_j představuje rozdíl středních hodnot na j-tém a prvním intervalu. Přidáme-li předpoklad normality, tento model odpovídá modelu pro analýzu rozptylu, kde skupiny jsou definovány pomocí hodnoty Z_i a dělicích bodů $c_1, \ldots, c_{k-1}$.

Regresní matice X obsahuje v j-tém řádku vektor $(1, \mathbb{I}_{\{Z_i \in \mathcal{J}_2\}}, \ldots, \mathbb{I}_{\{Z_i \in \mathcal{J}_k\}})$.

Interpretace parametrů obecného regresního modelu

Uvažujme regresní model

$$\mathsf{E}\,(Y_i \mid \boldsymbol{X}_i) = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip}.$$

Máme $\beta_1 = \mathsf{E}\,(Y_i \mid \boldsymbol{X}_i = 0)$, tj. absolutní člen vyjadřuje střední hodnotu Y_i pro pozorování s nulovou hodnotou všech regresorů.

Podívejme se nyní na parametr β_2. Dostaneme

$$\beta_2 = \mathsf{E}\,(Y_i \mid X_{i2} = x + 1, X_{i3} = x_3, \ldots, X_{ip} = x_p) - \mathsf{E}\,(Y_i \mid X_{i2} = x, X_{i3} = x_3, \ldots, X_{ip} = x_p)$$

čili parametr β_2 vyjadřuje rozdíl ve střední hodnotě $\mathsf{E}\, Y_i$ po zvýšení regresoru X_{i2} o jednu jednotku, přičemž všechny ostatní regresory zůstávají konstantní. Jde tedy o efekt regresoru X_{i2} očištěný od vlivu všech ostatních v modelu přítomných regresorů. Je-li $\beta_2 = 0$, znamená to, že kdyby ostatní regresory byly v celé populaci konstantní, neměl by regresor X_{i2} žádný vliv na střední hodnotu (ani rozptyl) Y_i.

To však neznamená, že X_{i2} sám o sobě nemá vliv na $\mathsf{E}\, Y_i$ – tento vliv je ale zcela zprostředkován (vysvětlen) závislostí mezi X_{i2} a ostatními regresory.

Tato interpretace neplatí v situaci, kdy regresory $X_{i3}, \ldots, X_{ip}$ jsou funkcemi regresoru X_{i2}. Pak by nebylo možné, aby se regresor X_{i2} změnil o jednotku a přitom ostatní regresory zůstaly beze změny.

10.3 Odhady metodou nejmenších čtverců

Nechť $\widehat{\boldsymbol{\beta}}$ je nějaký odhad vektoru parametrů $\boldsymbol{\beta}$. Označme $\widehat{\boldsymbol{Y}} = X\widehat{\boldsymbol{\beta}}$ odhadnuté střední hodnoty odezvy, tj.

$$\widehat{Y}_i = \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \cdots + \widehat{\beta}_p X_{ip} = \boldsymbol{X}_i^\mathsf{T} \widehat{\boldsymbol{\beta}}.$$

Odhad $\widehat{\boldsymbol{\beta}}$ vybereme tak, aby vektor $\widehat{\boldsymbol{Y}}$ byl co nejblíže vektoru $\boldsymbol{Y}$ v euklidovské vzdálenosti, tj.

$$\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^\mathsf{T} \boldsymbol{\beta})^2.$$

Definice 10.3. Tento odhad $\widehat{\boldsymbol{\beta}}$ nazýváme *odhad metodou nejmenších čtverců*[*].

Funkci, kterou minimalizujeme, lze přepsat jako

$$\sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^\mathsf{T} \boldsymbol{\beta})^2 = (\boldsymbol{Y} - X\boldsymbol{\beta})^\mathsf{T} (\boldsymbol{Y} - X\boldsymbol{\beta}).$$

Vyjádření pro odhad $\widehat{\boldsymbol{\beta}}$ dostaneme snadno pomocí maticových derivací:

$$\frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{Y} - X\boldsymbol{\beta})^\mathsf{T} (\boldsymbol{Y} - X\boldsymbol{\beta}) = -X^\mathsf{T} (\boldsymbol{Y} - X\boldsymbol{\beta}) - [(\boldsymbol{Y} - X\boldsymbol{\beta})^\mathsf{T} X] = -2(X^\mathsf{T}\boldsymbol{Y} - X^\mathsf{T} X\boldsymbol{\beta}).$$

Položíme-li derivaci rovnou nule, zjistíme, že $\widehat{\boldsymbol{\beta}}$ řeší soustavu p lineárních rovnic o p neznámých

$$(X^\mathsf{T} X)\widehat{\boldsymbol{\beta}} = X^\mathsf{T} \boldsymbol{Y}.$$

Řešení musí být globálním minimem, protože minimalizovaná funkce je konvexní v $\boldsymbol{\beta}$. Jelikož matice X má plnou hodnost p, matice $X^\mathsf{T} X$ (čtvercová, $p \times p$) má také plnou hodnost p. Tudíž existuje právě jedno řešení dané soustavy

$$\widehat{\boldsymbol{\beta}} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} \boldsymbol{Y}.$$

Poznámka.

[*] Angl. *least squares estimator*

1. Vektor $\widehat{\boldsymbol{Y}} = X\widehat{\boldsymbol{\beta}} = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{Y}$ nazýváme vektorem *odhadnutých (vyrovnaných) hodnot* odezvy*. Je to lineární kombinace původních pozorování $\boldsymbol{Y}$.

2. Matice $H \overset{\mathsf{df}}{=} X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ je idempotentní. Platí $\widehat{\boldsymbol{Y}} = H\boldsymbol{Y}$ a $H\widehat{\boldsymbol{Y}} = HH\boldsymbol{Y} = \widehat{\boldsymbol{Y}}$. Matice H je čtvercová $n \times n$, její hodnost je p.

3. Čtvercová matice $I_n - H = I_n - X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ je také idempotentní. Její hodnost je $n - p$.

Definice 10.4.

- Náhodný vektor $\boldsymbol{u} \overset{\mathsf{df}}{=} \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (I_n - H)\boldsymbol{Y}$ se nazývá *vektor reziduí*†. Jeho prvek $u_i = Y_i - \widehat{Y}_i = Y_i - \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}$ se nazývá *residuum*.
- Náhodná veličina
$$SS_e \overset{\mathsf{df}}{=} \boldsymbol{u}^\mathsf{T}\boldsymbol{u} = \sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}})^2$$
se nazývá *residuální součet čtverců*‡. Je to vlastně minimalisovaný součet čtverců odchylek.

Poznámka. Pro vektor reziduí platí $\mathsf{E}\,\boldsymbol{u} = \boldsymbol{0}$, $\mathsf{var}\,\boldsymbol{u} = \sigma^2(I_n - H)$. Residua tedy nejsou nezávislá a nemají stejné rozptyly. Rozptylová matice vektoru reziduí je singulární (má hodnost $n - p$).

Věta 10.3 (Vlastnosti odhadu metodou nejmenších čtverců)**.**

1. $\widehat{\boldsymbol{\beta}}$ je nestranný odhad, $\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$

2. $\mathsf{var}\,\widehat{\boldsymbol{\beta}} = \sigma^2(X^\mathsf{T}X)^{-1}$

3. Jsou-li $(Y_i, \boldsymbol{X}_i)$ nezávislé a stejně rozdělené náhodné vektory, pak $\widehat{\boldsymbol{\beta}}$ je konsistentní odhad $\boldsymbol{\beta}$ a
$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\boldsymbol{0}, \sigma^2(\mathsf{E}\,\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T})^{-1})$$

Nyní začneme předpokládat normalitu a dokážeme několik dalších užitečných vlastností.

Věta 10.4. Nechť v modelu (10.1) navíc platí $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2 I_n)$. Pak
$$\frac{SS_e}{\sigma^2} \sim \chi^2_{n-p}.$$

Poznámka.

* Angl. *fitted values* † Angl. *residuals* ‡ Angl. *residual sum of squares*

- Z věty 10.4 plyne, že $\frac{SS_e}{n-p}$ je nestranný a konsistentní odhad rozptylu σ^2 (toto platí i bez předpokladu normality).
- Věta 10.4 může být použita i ke konstrukci intervalu spolehlivosti pro σ^2. Postupem použitým v Kapitole 2 k odvození intervalu (2.2) dostaneme přesný interval

$$\left(\frac{SS_e}{\chi^2_{n-p}(1-\alpha/2)}, \frac{SS_e}{\chi^2_{n-p}(\alpha/2)} \right). \tag{10.2}$$

- Za předpokladu normality lze dokázat, že $\widehat{\boldsymbol{\beta}}$ a SS_e jsou nezávislé.

Věta 10.5. Nechť v modelu (10.1) navíc platí $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2 I_n)$. Nechť $\boldsymbol{c}$ je libovolný p-rozměrný vektor reálných konstant. Pak

$$\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{\frac{SS_e}{n-p}\boldsymbol{c}^\mathsf{T}(X^\mathsf{T}X)^{-1}\boldsymbol{c}}} \sim t_{n-p}$$

Poznámka. Věta 10.5 se používá k testování hypotéz o parametrech a lineárních kombinacích parametrů a ke konstrukci intervalů spolehlivosti. Chceme-li například otestovat hypotézu $H_0 : \beta_j = 0$, zvolíme $\boldsymbol{c} = \boldsymbol{e}_j$ (vektor nul kromě j-tého prvku, který je 1). Dostaneme $\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} = \widehat{\beta}_j$, $\boldsymbol{c}^\mathsf{T}\boldsymbol{\beta} = \beta_j$ a $\boldsymbol{c}^\mathsf{T}(X^\mathsf{T}X)^{-1}\boldsymbol{c} = v_j$, kde v_j je j-tý diagonální prvek matice $(X^\mathsf{T}X)^{-1}$. Použijeme testovou statistiku

$$T_j = \frac{\widehat{\beta}_j}{\sqrt{\frac{SS_e}{n-p}v_j}},$$

která má za platnosti $H_0 : \beta_j = 0$ rozdělení t_{n-p}. H_0 budeme zamítat na hladině α, pokud $|T_j| \geq t_{n-p}(1-\alpha/2)$.

Chceme-li testovat hypotézu $H_0 : \beta_2 = \beta_3$, zvolíme $\boldsymbol{c} = (0, 1, -1, 0, \ldots, 0)^\mathsf{T}$ a dále postupujeme stejně.

Interval spolehlivosti pro β_j s pravděpodobností pokrytí $1-\alpha$ by vyšel

$$\widehat{\beta}_j \mp t_{n-p}(1-\alpha/2)\sqrt{\frac{SS_e}{n-p}v_j}.$$

Poznámka. Pomocí věty 10.5 lze dokázat větu 10.1 o výběrovém korelačním koeficientu.

Poznámka. Dvouvýběrový t-test je speciální případ věty 10.5 v modelu $Y_i = \beta_1 + \beta_2 X_{i2} + \varepsilon_i$, kde X_{i2} nabývá pouze hodnot $0, 1$ a $\boldsymbol{c} = (0, 1)^\mathsf{T}$.

Věta 10.5 umožňuje testovat hypotézy o jednotlivých parametrech nebo lineárních kombinacích parametrů. Následující věta zobecňuje tento výsledek na testování více parametrů najednou.

Nechť platí lineární model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, kde $X = (X_A|X_B)$ a $\boldsymbol{\beta} = (\boldsymbol{\beta}_A^\mathsf{T}, \boldsymbol{\beta}_B^\mathsf{T})^\mathsf{T}$, $\boldsymbol{\beta}_B \in \mathbb{R}_d$, $\boldsymbol{\beta}_A \in \mathbb{R}^{p-d}$. Model lze přepsat ve tvaru

$$\boldsymbol{Y} = X_A\boldsymbol{\beta}_A + X_B\boldsymbol{\beta}_B + \boldsymbol{\varepsilon}.$$

Pokud jsou všechny složky parametru $\boldsymbol{\beta}_B$ nulové, lze z modelu vynechat regresory X_B a přejít k jednodušššímu modelu s regresní maticí X_A.

Označme SS_h residuální součet čtverců v tomto jednodušším modelu. Symbolem SS_e stále označujeme residuální součet čtverců v plném modelu s regresní maticí X. Vzhledem k tomu, že metoda nejmenších čtverců minimalizuje residuální součet čtverců, musí platit nerovnost $SS_e < SS_h$.

Věta 10.6. Nechť platí hypotéza $H_0 : \boldsymbol{\beta}_B = 0$. Pak statistika

$$F = \frac{n-p}{d}\,\frac{SS_h - SS_e}{SS_e}$$

má rozdělení F s d a $n-p$ stupni volnosti.

Poznámka. F-test analýzy rozptylu je speciální případ věty 10.6 v modelu $Y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$, kde $X_{ij} = \mathbb{I}_{\{Z_i=j\}}$ pro $j = 2, \ldots, k$, a $\boldsymbol{\beta}_B = (\beta_2, \ldots, \beta_k)^\mathsf{T}$.

Poznámka. Věta 10.6 je důležitá při budování regresního modelu, tj. rozhodování, které regresory do modelu zařadit a které nikoli.

Obsah