

01MAS Matematická statistika

Martin Kovanda
(revize a korektury Václav Kůs)

14. dubna 2020

Obsah

1	Statistika - setup a základní postupy	1
1.1	Statistické bodové odhady	2
1.2	Vlastnosti bodových odhadů	3
1.3	Výběrové charakteristiky a jejich vlastnosti	6
1.4	Výběrový kvantil (a jeho vlastnosti)	8
1.5	Neparametrické (empirické) odhady distribucí	9
2	Metody pro hledání bodových odhadů	12
2.1	Metoda momentů	12
2.2	Nestranné odhady s minimálním rozptylem (UMVUE)	14
2.3	Rao-Cramérova nerovnost	16
2.4	Metoda maximální věrohodnosti (MLE)	18
3	Testování statistických hypotéz	23
3.1	Základní strategie TSH	23
3.2	UMP testy pro parametr $\theta = \theta(F)$	24
3.3	Neyman-Pearsonovo lemma (N-PL)	26
3.4	Složené hypotézy a MLR systémy	28
3.5	Nestranné UMP testy (UMPU)	30
4	Další metody testování hypotéz	33
4.1	Test poměrem věrohodností (LRT)	33
4.2	Analýza variance (ANOVA)	35
4.3	Odůvodnitelné testy (RT)	36
4.4	Dvouvýběrové testy ($2 \times \mathcal{N}_1$)	36
4.5	Test koeficientu korelace (\mathcal{N}_2)	38
5	Asymptotické testy hypotéz	41
5.1	Asymptotické testy středních hodnot <i>iid</i> \mathcal{L}_2	42
5.2	Asymptotický LRT a Waldův test v \mathbb{R}^k	43
5.3	Testy dobré shody (GoF)	44
5.4	Modifikace χ^2 -testů dobré shody	47
6	Konfidenční množiny, Intervaly spolehlivosti	49
6.1	Konstrukce $CM_{1-\alpha}$ pomocí pivotů (PQ)	49
6.2	Konstrukce $CM_{1-\alpha}$ pomocí TSH ϕ_α	51
6.3	Asymptotické konfidenční množiny	53

Předmluva

Materiál byl původně sestavován na základě přednášek Ing. Václava Kůse, Ph.D., které proběhly v letním semestru akademického roku 2019/2020 na Fakultě jaderné a fyzikálně inženýrské ČVUT v Praze. Vzhledem k probíhající pandemii Covid-19 však byla přerušena kontaktní výuka, a proto byl zbytek materiálů vytvořen přepracováním zápisků z tabule z akademického roku 2018/2019. Tímto bych chtěl poděkovat Ing. Václavu Kůsovi, Ph.D., za rozsáhlou korekturu a doplnění materiálů, bez kterého by se tato práce neobešla. Dále bych chtěl poděkovat všem, kteří v této práci našli chyby a upozornili na ně, případně je i opravili.

Zároveň tímto vyzývám čtenáře, aby podobným způsobem zpracovali další předměty. Tvorba těchto materiálů zabere hodně času a usilí, ale dle mého názoru má veliký smysl pro všechny, kteří budou tento předmět navštěvovat.

Tento učební text je určen posluchačům 3. ročníku základního studia navštěvujícím kurs 01MAS *Matematická statistika*, který je zařazen mezi předměty oborů AMSM a MM. Při sestavování textu se předpokládaly znalosti základů matematiky na úrovni absolvování kurzů 01MAB2-4, 01LAB1-2 a 01MIP.

1 Statistika - setup a základní postupy

Představme si, že máme nějakou reálnou situaci a chceme na ní vytvořit statistický matematický model, který závisí na jistých parametrech (θ). Narozdíl od jiných předmětů ale máme k dispozici výsledky měření z navrženého experimentu a zkoumáme jejich příčinu. Máme opět $(\Omega, \mathcal{A}, \mathbb{P})$. Ω zde nazveme **populace**, $\omega \in \Omega$ nazveme **individuum**.

Náhodná veličina $X : \Omega \rightarrow \mathbb{R}$ zde bude představovat **vlastnost** zkoumanou na dané populaci Ω (např. počet špatných výrobků v sérii). Tuto vlastnost zjišťujeme experimentálně, např. měřením, vážením apod. Víme také, že $X \sim \mathbb{P}^X$ má jisté rozložení na $(\Omega, \mathcal{A}, \mathbb{P})$ dané např. pomocí F_X, f_X, p_X, \dots , které ale neznáme a chceme ho zjistit.

Definice 1.1. n -tici nezávislých náhodných veličin X_1, \dots, X_n stejně rozdělených s distribuční funkcí F nazýváme **náhodný výběr** z rozdělení F . Konkrétní realizací \mathbf{X} získáme vektor čísel $\mathbf{x} = (x_1, \dots, x_n)$, který nazveme **realizace** náhodného výběru X_1, \dots, X_n , neboli naměřená data.

PŘÍKLAD 1.2. Příkladem je třeba n opakování stejných experimentů, ve kterých máme pokaždé stejné nastavení. Používáme tedy stejnou metodu měření. Dalším příkladem může být n výrobků z nějaké dodávky zboží, na základě kterých určujeme celkový počet zmetků.

Definice 1.3. Vybereme vektor individuí $\omega^{(n)} = (\omega_1, \dots, \omega_n) \in \Omega^n$ a definujeme

$$X_j(\omega^{(n)}) := X(\omega_j), \quad \forall j \in \hat{n},$$

jako **pozorování** (v tomto tvaru máme skutečná data). Zavedeme dále $\Omega^{(n)} := \Omega^n$, $\mathcal{A}^{(n)} := \sigma(\mathcal{A}^n)$, $\mathbb{P}^{(n)} := \bigotimes_{i=1}^n \mathbb{P}^X$ (součinnová míra), tedy

$$\mathbb{P}^{(n)}(A_1 \times \dots \times A_n) = \prod_{j=1}^n \mathbb{P}^X(A_j), \quad \forall A_j \in \mathcal{A}.$$

Definujeme **realizaci** náhodného výběru $\mathbf{X} = (X_1, \dots, X_n)$ vztahem

$$\mathbf{x} = \mathbf{X}(\omega^{(n)}).$$

Věta 1.4. $(X_j)_{j=1}^n$ iid \mathbb{P}^X .

Důkaz. Pro $\forall B \in \mathcal{B}$ a $\forall j \in \hat{n}$ platí, že

$$\mathbb{P}^{X_j}(B) = \mathbb{P}^{(n)} \circ X_j^{-1}(B) = \mathbb{P}^{(n)}\left(\underbrace{\{X_j \in B\}}_{\Omega \times \Omega \dots \times \{X \in B\} \times \Omega \dots}\right) = \mathbb{P}(X \in B) = \mathbb{P}^X(B).$$

Zbývá nám ještě dokázat nezávislost:

$$\mathbb{P}^{(n)}\left(\bigcap_{j=1}^n \{X_j \in B_j\}\right) = \mathbb{P}^{(n)}\left(\bigotimes_{j=1}^n \{X \in B_j\}\right) = \prod_{j=1}^n \mathbb{P}(X \in B_j) = \prod_{j=1}^n \mathbb{P}^{(n)}(X_j \in B_j),$$

protože B_j jsou pokaždé na jiné "pozici" v kartézském součinu $\Omega \times \dots \times \{X \in B_j\} \times \dots \times \Omega$. \square

Definice 1.5. Statistika je libovolná funkce náhodného výběru X_1, \dots, X_n , jejíž funkční předpis nezávisí na parametrech příslušného rozdělení.

Příkladem statistiky může být **výběrový průměr** (sample mean)

$$\overline{X}_n = \frac{1}{n} \sum_{j=1}^n X_j,$$

případně **geometrický průměr**

$$\overline{X}_n^G = \sqrt[n]{X_1 X_2 \dots X_n},$$

výběrový rozptyl

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \overline{X}_n)^2 \quad \text{nebo} \quad s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2,$$

kde $\hat{\sigma}_n$ nebo případně s_n je **výběrová směrodatná odchylka**. Dále uveďme **r -tý výběrový obecný a centrální moment** (sample moments)

$$m'_r = \frac{1}{n} \sum_{j=1}^n X_j^r, \quad m_r = \frac{1}{n} \sum_{j=1}^n (X_j - \overline{X}_n)^r,$$

medián

$$\hat{X}_{\frac{1}{2}} = \begin{cases} X_{(\frac{n+1}{2})} & n \text{ liché,} \\ \frac{1}{2} (X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}) & n \text{ sudé,} \end{cases}$$

kde $X_{(j)}$ značí j -tou zdola uspořádanou statistiku z náhodného výběru $(X_{(1)} \leq \dots \leq X_{(n)})$.

1.1 Statistické bodové odhady

Nechť θ je parametr spojený s rozdělením \mathbb{P}^X , tzn. $\theta = \theta(\mathbb{P}^X)$. Máme dvě možnosti.

a) θ může být spojený s \mathbb{P}^X , např. $\theta = \mathbb{E}X, DX, F_X(t), \dots$, nebo

b) θ může být přímo parametr rozdělení \mathbb{P}_θ^X .

V obou případech ale požadujeme tzv. identifikovatelnost rodiny $\mathcal{P} = \{\mathbb{P}^X\}$, resp. $\mathcal{P} = \{\mathbb{P}_\theta^X : \theta \in \Theta\}$, tzn., že každému parametru přísluší právě jedna pravděpodobnost X .

$$\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_1^X \neq \mathbb{P}_2^X, \quad \text{resp.} \quad \theta_1 \neq \theta_2 \Rightarrow P_{\theta_1}^X \neq P_{\theta_2}^X.$$

Opačná implikace často neplatí, třeba pro danou střední hodnotu najdeme vícero různých rozdělení. Předpokládáme, že $\theta \in \Theta \subset \mathbb{R}^k$, kde Θ se nazývá **parametrický prostor**. Dále máme $\tau(\theta) : \Theta \rightarrow \mathbb{R}^s$, kde τ je tzv. **parametrická funkce** (většinou nás ale zajímá např. jedna vybraná složka, tedy $s = 1$). Odhadnout celé rozdělení se nám většinou nepodaří (nebo to nepotřebujeme), proto hledáme odhad parametru θ .

Definice 1.6. Libovolná (borelovsky) měřitelná funkce náhodného výběru X_1, \dots, X_n (tedy statistika)

$\hat{\theta}_n(\mathbf{X}) : \Omega^n \rightarrow \mathbb{R}^k$ se nazývá **odhadem** (estimator) $\theta \in \Theta \subset \mathbb{R}^k$.

$\widehat{T}_n(\mathbf{X}) : \Omega^n \rightarrow \mathbb{R}^s$ se nazývá **odhadem** (estimator) parametrické funkce $\tau(\theta) \in \tau(\Theta) \subset \mathbb{R}^s$.

POZNÁMKA 1.7. Tato definice nám však neříká, jak tuto statistiku volit. Aby však poskytovala námi žádané výsledky, je potřeba, aby splňovala některé vlastnosti.

1.2 Vlastnosti bodových odhadů

Definice 1.8. $T_n(\mathbf{X})$ nazýváme **nestranný odhad** $\tau(\theta)$, pokud

$$\mathbb{E}_\theta T_n(\mathbf{X}) = \tau(\theta), \quad \forall \theta \in \Theta, \quad \forall n \in \mathbb{N}.$$

$T_n(\mathbf{X})$ nazýváme **asymptoticky nestranný odhad** τ , pokud

$$\mathbb{E}_\theta T_n(\mathbf{X}) \rightarrow \tau(\theta), \quad \forall \theta \in \Theta,$$

kde \mathbb{E}_θ značí střední hodnotu vzhledem k rozdělení \mathbb{P}_θ^X .

POZNÁMKA 1.9. Střední hodnotu tedy chápeme jako "průměrování" přes všechny možné výběry, které lze získat pod rozdělením \mathbb{P}_θ^X při daném fixním, ale libovolném θ .

Definice 1.10. $T_n(\mathbf{X})$ nazýváme **eficientní** (vydatný), pokud pro $\forall \tilde{T}_n, \forall \theta \in \Theta$ platí

$$\begin{aligned} s = 1 : \mathbb{E} \left[(T_n(\mathbf{X}) - \tau(\theta))^2 \right] &\leq \mathbb{E} \left[(\tilde{T}_n(\mathbf{X}) - \tau(\theta))^2 \right], \\ s > 1 : \mathbb{E} \left[\|T_n(\mathbf{X}) - \tau(\theta)\|_\varepsilon^2 \right] &\leq \mathbb{E} \left[\|\tilde{T}_n(\mathbf{X}) - \tau(\theta)\|_\varepsilon^2 \right]. \end{aligned}$$

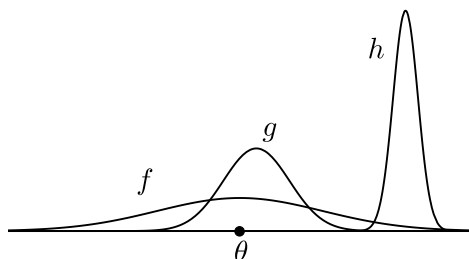
$\mathbb{E}(T_n - \tau)^2$ se nazývá střední kvadratická chyba - MSE (*mean squared error*) odhadu $T_n(\mathbf{X})$. Tedy eficientní odhad má nejnižší možnou MSE. Pro nestranné odhady ($\tau(\theta) = \mathbb{E}T_n(\mathbf{X})$) tento vztah pro $s = 1$ přechází na $DT_n(\mathbf{X}) \leq D\tilde{T}_n(\mathbf{X})$.

Definice 1.11. Máme-li $T_n^1, T_n^2 : \Omega \rightarrow \mathbb{R}^1$ jako odhady $\tau(\theta)$, pak definujeme **relativní eficientci** vztahem

$$\text{RE}_{2,1} = \frac{\mathbb{E}(T_n^1 - \tau)^2}{\mathbb{E}(T_n^2 - \tau)^2} \stackrel{\text{nestranný odhad}}{=} \frac{DT_n^1}{DT_n^2}.$$

Zkusme nyní porovnat dvě statistiky tak, abychom určili, která z nich je lepší. Pokud porovnáváme 2 nestranné statistiky, pak lepší (**vydatnější**) z nich je ta s menším rozptylem. Pokud však porovnáváme statistiky, které nejsou nestranné, pak nutně nemusí být lepší ta s nižším rozptylem. Používáme k tomu tedy právě definovanou **MSE** (střední kvadratickou odchylku, mean squared error) (viz obrázek 1.1)

$$\text{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2, \quad \text{kde } \hat{\theta} \text{ je odhadem neznámého parametru } \theta.$$



Obrázek 1.1: Ukázka toho, že statistika s nejmenším rozptylem (h) nebo s nejbližší střední hodnotou (f) nemusí být nutně ta nejvhodnější.

Definice 1.12. $T_n(\mathbf{X})$ se nazývá **konzistentní** odhad τ , pokud

$$T_n(\mathbf{X}) \xrightarrow{\mathbb{P}, s.j.} \tau(\theta), \quad \forall \theta \in \Theta$$

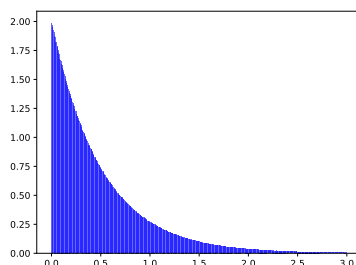
(pro $\xrightarrow{\mathbb{P}}$ slabě konzistentní, pro $\xrightarrow{s.j.}$ silně konzistentní).

Věta 1.13 (Kritérium konzistence). *Odhad $T_n(\mathbf{X})$ je slabě konzistentní ($T_n(\mathbf{X}) \xrightarrow{\mathbb{P}} \tau(\theta)$), pokud*

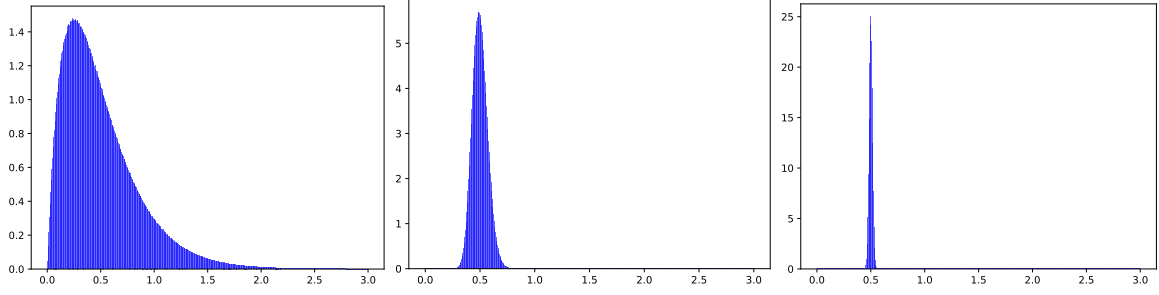
1. $T_n(\mathbf{X})$ je asymptoticky nestranný ($\mathbb{E}_\theta T_n(\mathbf{X}) \rightarrow \tau(\theta)$, pro $\forall \theta \in \Theta$) a
2. platí pro něj, že $DT_n(\mathbf{X}) \rightarrow 0$.

PŘÍKLAD 1.14. Statistika je funkce náhodných veličin, a proto se taky chová jako náhodná veličina. Má proto taky své vlastní rozdělení a například střední hodnota jejího rozdělení nám může poskytnout užitečné informace. Příkladem je zkoumání realizace exponenciálního rozdělení $\text{Exp}(\frac{1}{2})$ (viz obr. 1.2). Vezměme nyní statistiku "výběrový průměr" (výběrová střední hodnota). Na obrázku 1.3 vidíme, že střední hodnota výběrového průměru odpovídá $\mu = \frac{1}{2}$, což je důsledek toho, že je statistika nestranná. Pro vyšší n jde zároveň její rozptyl k nule. Tyto vlastnosti vyplývají z centrální limitní věty, která říká, že výběrový průměr má přibližně normální rozdělení (pro data z normálního rozdělení pak přímo normální rozdělení) dané vztahem

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$



Obrázek 1.2: Hustota pravděpodobnosti rozdělení $\text{Exp}(\frac{1}{2})$.



Obrázek 1.3: Hustota pravděpodobnosti výběrového průměru pro $n = 2$, $n = 50$ a $n = 1000$.

Vidíme tedy, že $\overline{X}_n \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right)$, kde $\mu = 0.5$.

Definice 1.15. $T_n(\mathbf{X})$ se nazývá **asymptoticky normální** odhadem $\tau(\theta)$ s asymptotickou kovarianční maticí $\mathbb{C}(\theta)$ (matice tvaru $s \times s$), pokud pro $\forall \theta \in \Theta$

$$T_n(\mathbf{X}) \sim \mathcal{AN}_s\left(\tau(\theta), \frac{1}{n}\mathbb{C}(\theta)\right), \text{ tzn. } \sqrt{n}(T_n(\mathbf{X}) - \tau(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}_s(\mathbf{0}, \mathbb{C}(\theta)) \text{ (viz CLT).}$$

Pro $s = 1$ definice přechází na $\sqrt{n}(T_n(\mathbf{X}) - \tau(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(\theta))$, kde $\sigma^2(\theta)$ nazýváme asymptotický rozptyl odhadu $T_n(\mathbf{X})$.

Definice 1.16. Máme-li T_n^1, T_n^2 jako odhady $\tau(\theta)$, které jsou oba \mathcal{AN} odhady τ s asymptotickými rozptyly $\sigma_1^2(\theta)$ a $\sigma_2^2(\theta)$, pak definujeme **asymptotickou relativní eficienci** vztahem $\text{ARE}_{2,1} = \frac{\sigma_1^2(\theta)}{\sigma_2^2(\theta)}$.

POZNÁMKA 1.17. • Odtud nutně neplyne, že $\mathbb{E}\left(\sqrt{n}(T_n(\mathbf{X}) - \tau(\theta))\right) \rightarrow 0$ (tedy asymptotická nestrannost T_n), protože $\lim_{n \rightarrow +\infty} \mathbb{E}T_n$ nemusí nutně existovat.

- Nemusí dokonce platit ani vztah $D\left(\sqrt{n}(T_n(\mathbf{X}) - \tau(\theta))\right) = D\left(\sqrt{n}(T_n(\mathbf{X}))\right) \rightarrow \sigma^2(\theta)$. (rovnítka vyplývá z posunutí o konstantu).

Věta 1.18. $T_n(\mathbf{X})$ je asymptoticky normální $\mathcal{AN}\left(\tau(\theta), \frac{1}{n}\sigma^2(\theta)\right)$. Pak $T_n(\mathbf{X}) \xrightarrow{\mathbb{P}} \tau(\theta)$, $\forall \theta \in \Theta$.

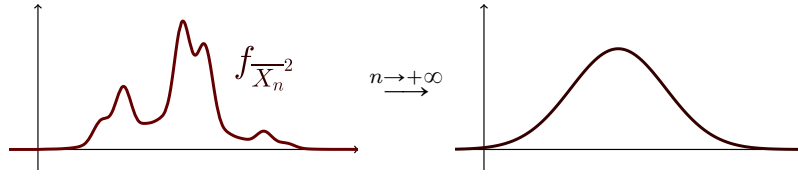
Tato slabá konzistence je řádu $o_p(n^{-\alpha})$, $\alpha < \frac{1}{2}$, tzn., že $n^\alpha(T_n(\mathbf{X}) - \tau(\theta)) \xrightarrow{\mathbb{P}} 0$, $\forall \alpha < \frac{1}{2}$. To vyplývá z věty (MIP) (Mějme $(X_n)_{n=1}^{+\infty}$, $X_n \sim \mathcal{AN}(\mu, \sigma_n^2)$ tak, že $\sigma_n \rightarrow 0$. Pak $X_n \xrightarrow{\mathbb{P}} \mu$).

Věta 1.19 (Δ -metoda). Nechť $T_n(\mathbf{X}) \sim \mathcal{AN}\left(\tau(\theta), \frac{1}{n}\sigma^2(\theta)\right)$ a $g: \mathbb{R}^1 \rightarrow \mathbb{R}^1$ spojitě diferencovatelnou v $\tau(\theta)$, $\forall \theta \in \Theta$. Pak

$$g(T_n(\mathbf{X})) \sim \mathcal{AN}_1\left(g(\tau(\theta)), \frac{\sigma^2(\theta)}{n} [g'(\tau(\theta))]^2\right).$$

PŘÍKLAD 1.20. Například pokud $T_n(\mathbf{X}) = \overline{X}_n$, $g(t) = t^2$, $g'(t) = 2t$, pak

$$(\overline{X}_n)^2 \sim \mathcal{AN}\left(\mu^2, \frac{\sigma^2}{n} \cdot 4\mu^2\right).$$



1.3 Výběrové charakteristiky a jejich vlastnosti

Věta 1.21. Mějme $X \in \mathcal{L}_1$, resp. \mathcal{L}_2 , volme $\theta = \theta(\mathbb{P}^X) = \mathbb{E}X = \mu$ a označme $\sigma^2 = DX < +\infty$ (\mathcal{L}_2). Pak **sample mean**

$$T_n(\mathbf{X}) = \overline{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$$

je *nestranným, konzistentním a $\mathcal{AN}\left(\mathbb{E}X, \frac{1}{n}\sigma^2\right)$ odhadem $\theta = \mathbb{E}X = \mu$.*

Důkaz. Víme, že X_1, \dots, X_n jsou iid podle věty 1.4. Pak

1. $\mathbb{E}\overline{X}_n = \mu, \forall \mu,$
2. $\overline{X}_n \xrightarrow{\mathbb{P}, s.j.} \mathbb{E}X$, což platí díky ZVČ (Kolmogorov 2, kde požadujeme $(X_j)_{j=1}^n$ iid \mathcal{L}_1),

□

POZNÁMKA 1.22. 1. $\text{CLT}_{L-L}: \overline{X}_n \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right)$ (zde požadujeme $(X_j)_{j=1}^n$ iid \mathcal{L}_2).

2. X zde značí vlastnost na Ω a $\mathbf{X} = (X_1, \dots, X_n)$ je náhodný výběr. Není to tedy realizace!

Věta 1.23. $X \in \mathcal{L}_2$, resp. \mathcal{L}_4 , volme $\theta = \theta(\mathbb{P}^X) = DX = \sigma^2$. Pak **výběrový rozptyl**

$$T_n(\mathbf{X}) = \hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \overline{X}_n)^2 \quad i \quad T_n(\mathbf{X}) = s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2$$

jsou oba asymptoticky nestranné, konzistentní a $\mathcal{AN}\left(\sigma^2, \frac{1}{n}(\mu_4 - \sigma^4)\right)$ odhady σ^2 , kde $\mu_4 = \mathbb{E}[(X - \mathbb{E}X)^4]$. V případě $T_n(\mathbf{X}) = s_n^2$ je $T_n(\mathbf{X})$ navíc nestranný odhad σ^2 , $\forall n \in \mathbb{N}$.

Důkaz. Konzistence:

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \overline{X}_n)^2 = \frac{1}{n} \sum_{j=1}^n X_j^2 - \frac{2}{n} \sum_{j=1}^n X_j \overline{X}_n + \overline{X}_n^2 = \underbrace{\frac{1}{n} \sum_{j=1}^n X_j^2}_{\overline{Y}_n \xrightarrow{\mathbb{P}, s.j.} \mathbb{E}Y_1 = \mathbb{E}X^2} - \underbrace{\overline{X}_n^2}_{\rightarrow (\mathbb{E}X)^2} \xrightarrow{\mathbb{P}, s.j.} DX$$

Nestrannost (pro $\hat{\sigma}_n^2$ pouze asymptotická):

$$\begin{aligned} \mathbb{E}(\hat{\sigma}_n^2) &= \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^n X_j^2 - (\overline{X}_n)^2\right) = \mathbb{E}X^2 - \mathbb{E}(\overline{X}_n^2) = \mathbb{E}X^2 - \frac{1}{n^2} \mathbb{E}\left(\sum_{j=1}^n X_j^2 - \sum_{i \neq j} X_i X_j\right) = \\ &= \mathbb{E}X^2 - \frac{1}{n} \mathbb{E}X^2 - \frac{1}{n^2} n(n-1)(\mathbb{E}X)^2 = \frac{n-1}{n} \underbrace{[\mathbb{E}(X^2) - (\mathbb{E}X)^2]}_{=DX=\sigma^2} = \frac{n-1}{n} \sigma^2 \rightarrow \sigma^2, \end{aligned}$$

1 Statistika - setup a základní postupy

$$\mathbb{E}s_n^2 = \frac{n}{n-1}\mathbb{E}\hat{\sigma}_n^2 = \sigma^2.$$

Asymptotická normalita $\hat{\sigma}_n^2$ i s_n^2 plyne z rozkladu $\hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \mu)^2 - (\bar{X}_n - \mu)^2$ a ze Slutskyho lemma. \square

POZNÁMKA 1.24. V případě dat 1.2, kde $\sigma^2 = 0.25$ je situace zachycena na obrázku 1.4.

Důsledek 1.25. Mějme $X \in \mathcal{L}_2$. Potom pro tzv. **t-statistiku** platí, že

$$t_n = t_n(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

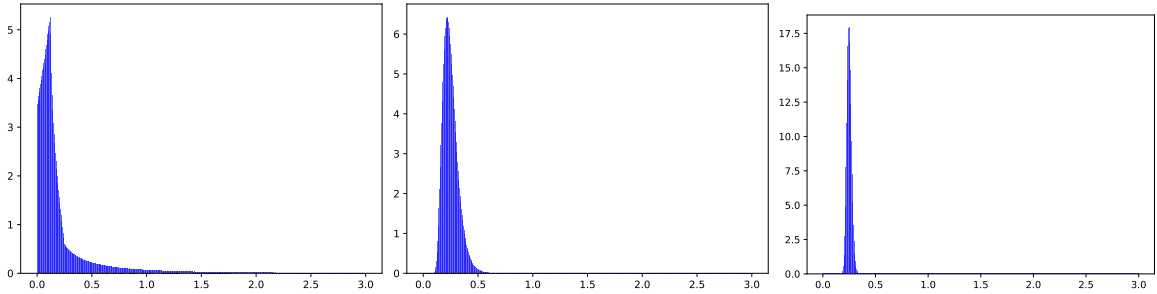
Důkaz. Z CLT_{L-L}

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2), \quad s_n \xrightarrow{\mathbb{P}, s.j.} \sigma.$$

Podíl tedy konverguje v distribuci (Slutsky),

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \xrightarrow{\mathcal{D}} \frac{\mathcal{N}(0, \sigma^2)}{\sigma} = \mathcal{N}(0, 1).$$

\square



Obrázek 1.4: Histogram výběrového rozptylu realizace 1.2 (10000000 rozptylů) pro $n = 2$, $n = 100$ a $n = 1000$.

Věta 1.26 (Chinčín). Mějme $X \in \mathcal{L}_r$, resp. $X \in \mathcal{L}_{2r}$, $r \geq 1$, volíme

$$\theta_1 = \theta_1(\mathbb{P}^X) = \mathbb{E}(X^r) = \mu'_r, \quad \theta_2 = \mathbb{E}(X - \mathbb{E}X)^r = \mu_r.$$

Pak r -tý výběrový moment

$$m'_r = m'_r(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^n X_j^r$$

je nestranným, konzistentním a \mathcal{AN} odhadem $\theta_1 = \mu'_1$ a

$$m_r = m_r(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^r$$

je konzistentním odhadem μ_r .

Věta 1.27. Speciálně nyní mějme $(X_j)_{j=1}^{n,+\infty}$ iid $\mathcal{N}(\mu, \sigma^2)$. Pak

a) $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, $\forall n \in \mathbb{N}$. Dále potom $\mathbb{E}\overline{X}_n = \mu$, $D\overline{X}_n = \sigma^2$, $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$,

b) $\hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \overline{X}_n)^2 \xrightarrow{\mathcal{N}} \frac{n\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-1)$, $\forall n \in \mathbb{N}$ (Pearsonovo rozdělení).

$s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2 \xrightarrow{\mathcal{N}} \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1)$, $\forall n \in \mathbb{N}$. Dále platí

$$D(\hat{\sigma}_n^2) = \frac{\sigma^4}{n^2} 2(n-1) = \frac{n-1}{n^2} (2\sigma^4) \rightarrow 0,$$

$$D(s_n^2) = D\left(\frac{\sigma}{n-1} \chi^2(n-1)\right) = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1} \rightarrow 0,$$

$$D(s_n^2) > D(\hat{\sigma}_n^2).$$

c) \overline{X}_n a s_n^2 jsou nezávislé a platí

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} = \frac{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}}{\frac{s_n}{\sigma}} \sim \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1), \quad \forall n \in \mathbb{N} \text{ (Studentovo rozdělení)}.$$

d) Pokud jsou X, Y na $(\Omega, \mathcal{A}, \mathbb{P})$ nezávislé, pak pro

X_1, \dots, X_n iid $\mathcal{N}(\mu_1, \sigma_1^2)$, ze kterých známe $\overline{X}_n, s_{1,n}^2$,

Y_1, \dots, Y_m iid $\mathcal{N}(\mu_2, \sigma_2^2)$, ze kterých známe $\overline{Y}_m, s_{2,m}^2$,

platí, že

$$\frac{s_{1,n}^2(n-1)}{\sigma_1^2} \sim \chi^2(n-1), \quad \frac{s_{2,m}^2(m-1)}{\sigma_2^2} \sim \chi^2(m-1).$$

Navíc lze ukázat, že $s_{1,n}^2$ a $s_{2,m}^2$ jsou nezávislé.

Důsledek 1.28.

$$T_n(\mathbf{X}) = \frac{\frac{s_{1,n}^2}{\sigma_1^2}}{\frac{s_{2,m}^2}{\sigma_2^2}} = \frac{\frac{\chi^2(n-1)}{n-1}}{\frac{\chi^2(m-1)}{m-1}} \sim F(n-1, m-1) \quad (\text{Fisherovo rozdělení}).$$

1.4 Výběrový kvantil (a jeho vlastnosti)

Definice 1.29. Mějme náhodný výběr (X_1, \dots, X_n) . Pak $(X_{(1)}, \dots, X_{(n)})$ nazveme **uspořádaný náhodný výběr** (vzestupně). **Výběrový α -kvantil** definujeme jako $\hat{X}_{\alpha,n} = X_{([\alpha n]+1)}$, pro $\alpha = \frac{1}{2}$ nazýváme $\hat{X}_{\frac{1}{2}}$ **výběrovým mediánem**, který alternativně označujeme jako \hat{X}_{med} . **Výběrové rozpětí** pak definujeme jako $d = X_{(n)} - X_{(1)}$ a **výběrové interkvartilové rozpětí** jako

$$d_{\frac{1}{4}} = X_{([\frac{3}{4}n]+1)} - X_{([\frac{1}{4}n]+1)}.$$

POZNÁMKA 1.30. Připomeňme, že $\theta = \theta(F) = \inf\{x : F(x) \geq \alpha\} \stackrel{\text{ozn}}{=} x_\alpha$ je α -kvantil rozdělení F . Pro F rostoucí a spojitou potom $F(x_\alpha) = \alpha$, $\alpha \in (0, 1)$. Výběrový kvantil $\hat{X}_{\alpha,n}$ je odhadem x_α a zjednodušeně ho označujeme jako \hat{X}_α .

PŘÍKLAD 1.31.

1. Mějme data $\{1, 2, 3, 4, 5\}$, tedy $n = 5$. Pak $\hat{x}_{1/2} = x_{[5/2]+1} = x_{(3)} = 3 = \bar{x}_5$.
2. Nyní mějme $n = 5$, ale tentokrát $\{1, 2, 3, 4, 500\}$. Pak $\hat{x}_{1/2} = 3$ (je robustní), ale $\bar{x}_5 = 102$ (není robustní, tedy neodolá jedné, či několika větším výchyilkám nebo chybám (*outliers*) v datech).

Věta 1.32. Mějme X_1, \dots, X_n iid F , $\theta = x_\alpha$, $\alpha \in (0, 1)$. Nechť x_α je jednoznačně určeno rovnicí $F(x_\alpha) = \alpha$ a existuje $F'(x_\alpha) > 0$. Pak

$$\hat{X}_\alpha \sim \mathcal{N}\left(x_\alpha, \frac{\alpha(1-\alpha)}{n[F'(x_\alpha)]^2}\right).$$

Důkaz. Dokážeme, že platí vztah $Y_n := \sqrt{n}(\hat{X}_\alpha - x_\alpha) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{[F'(x_\alpha)]^2}\right)$. Máme

$$F_{Y_n}(t) = \mathbb{P}(Y_n \leq t) = \mathbb{P}\left(\sqrt{n}(\hat{X}_\alpha - x_\alpha) \leq t\right) = \mathbb{P}\left(\hat{X}_\alpha \leq x_\alpha + \frac{t}{\sqrt{n}}\right) = \circledast$$

Nyní označme $S_n = \#\left\{j \in \hat{n} : X_j \leq x_\alpha + \frac{t}{\sqrt{n}}\right\}$. Pak $S_n \sim \text{Bi}(n, p_n)$, kde

$$\underbrace{A, \text{ iid } F}_{A, \text{ iid } F}$$

$$p_n = \mathbb{P}(A) = \mathbb{P}\left(X_j \leq x_\alpha + \frac{t}{\sqrt{n}}\right) = F\left(x_\alpha + \frac{t}{\sqrt{n}}\right).$$

Dále si připomeneme vztah $\hat{X}_\alpha = X_{([\alpha n]+1)}$, ve kterém označíme $m = [\alpha n] + 1$.

$\circledast = \mathbb{P}(S_n \geq m) = 1 - \mathbb{P}(S_n \leq m-1)$, který se dá převést na $1 - \Phi_{\mathcal{N}}(\dots)$, protože

$$S_n \stackrel{\text{CLT}}{\sim} \mathcal{N}(np_n, np_n(1-p_n)).$$

□

Důsledek 1.33. Z \mathcal{AN} plyne, že $\hat{X}_\alpha \xrightarrow{\mathbb{P}} x_\alpha$ řádu $o_p(n^{-\beta})$, $\beta < \frac{1}{2}$.

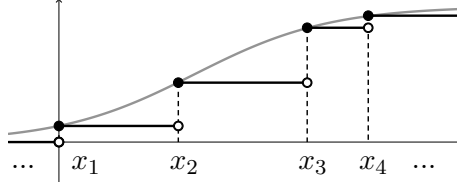
1.5 Neparametrické (empirické) odhady distribucí

Definice 1.34. Mějme X_1, \dots, X_n iid F , označme pro dané $t \in \mathbb{R}$ charakteristickou funkci na intervalu $(-\infty, t]$ jako $\mathbb{I}_{(-\infty, t]}$. Pak **empirickou distribuční funkci** (EDF) F_n definujeme vztahem

$$F_n(t) = F_n(t, \mathbf{X}) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{(-\infty, t]}(X_j), \quad \forall t \in \mathbb{R}.$$

Pro t fixní můžeme psát $F_n(t, \mathbf{X}) = T_n(\mathbf{X})$.

1 Statistika - setup a základní postupy



Věta 1.35 (ZVMS). Mějme X_1, \dots, X_n iid F . Pak pro $\forall t \in \mathbb{R}$ fixní platí, že $F_n(t)$ je ne-stranným, konzistentním a \mathcal{AN} odhadem $\theta = F(t)$, tzn.

1. $\mathbb{E}F_n(t) = F(t), \forall n,$
2. $F_n(t) \xrightarrow{\mathbb{P}, s.j.} F(t), \forall t \in \mathbb{R},$
3. $F_n(t) \sim \mathcal{AN}(F(t), \frac{1}{n}F(t)(1-F(t))).$

Navíc dokonce platí, že

1. $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{s.j.} 0, \text{ (Glivenko-Cantelliho lemma),}$
2. $\mathbb{P}\left(\sup_t |F_n(t) - F(t)| > \varepsilon\right) \leq 8(n+1)e^{-\frac{n\varepsilon^2}{32}}, \forall n, \forall \varepsilon > 0, \text{ (Glivenko-Cantelli),}$
3. $\mathbb{P}\left(\sup_t |F_n(t) - F(t)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}, \forall n, \forall \varepsilon > 0 \text{ (Massart, 1990).}$

Důkaz. Pro důkaz volme fixní libovolné t a označme $\mathbb{I}_{(-\infty, t]}(X_j) = Y_j^t$. Pak $Y_j^t = \begin{cases} 1, & X_j \leq t, \\ 0, & \text{jinak,} \end{cases}$ přičemž spočteme

$$\mathbb{P}(Y_j^t = 1) = \mathbb{P}(X_j \leq t) \stackrel{X_j \text{ iid}}{=} P(X \leq t) = F(t), \quad \mathbb{P}(Y_j^t = 0) = 1 - F(t),$$

což nám poskytuje rozdělení $Y_j^t \sim A(p = F(t))$ - alternativní (Bernoulliho) rozdělení. $(Y_j^t)_{j=1}^n$ jsou nezávislé, protože X_j iid. Dále víme, že $nF_n(t) = \sum_{j=1}^n Y_j^t \sim \text{Bi}(n, p = F(t))$. Z vlastností binomického rozdělení potom vyplývá, že

- a) $\mathbb{E}F_n(t) = \frac{1}{n}\mathbb{E}[\text{Bi}(n, p = F(t))] = F(t) = \theta, \forall t,$
- b) $\frac{1}{n} \sum_{j=1}^n Y_j^t = \bar{Y}_n^t \xrightarrow{\mathbb{P}, s.j.} \mathbb{E}Y_j^t = \mathbb{E}Y_1^t = F(t) = \theta \quad (\text{ZVČ}),$
- c) $F_n(t) = \frac{1}{n} \sum_{j=1}^n Y_j^t = \bar{Y}_n^t \sim \mathcal{AN}\left(F(t), \frac{1}{n}F(t)(1-F(t))\right) \quad \left(\text{CLT pro } (Y_j^t)_{j=1}^n \text{ iid } \mathcal{L}_2(A(p))\right).$

□

Definice 1.36. Mějme $\theta = \theta(F)$ (funkcionál na prostoru distribučních funkcí). Pak

$$T_n(\mathbf{X}) = \theta\left(\underbrace{F_n}_{\rightarrow F}\right)$$

se nazývá **statistický funkcionál**.

PŘÍKLAD 1.37.

$$\theta(F) = \mathbb{E}X \quad \Rightarrow \quad T_n(\mathbf{X}) = \theta(F_n) = \int x dF_n = \sum_{j=1}^n x_j \cdot \mathbb{P}_n(X = x_j) = \sum_{j=1}^n x_j \cdot \frac{1}{n} = \overline{X}_n,$$

$$\theta(F) = DX \quad \Rightarrow \quad T_n(\mathbf{X}) = \theta(F_n) = \int x^2 dF_n - \left(\int x dF_n \right)^2 = \frac{1}{n} \sum_{j=1}^n x_j^2 - \overline{X}_n^2 = \hat{\sigma}_n^2,$$

$$\theta(F) = x_\alpha \quad \Rightarrow \quad T_n(\mathbf{X}) = \theta(F_n) = \inf\{x : F_n(x) \geq \alpha\} = \hat{x}_\alpha, \text{ kde } x_\alpha \text{ je } \alpha\text{-kvantil } X.$$

Definice 1.38 (Histogram). Mějme $X \sim f$, $\text{supp } f = [a, b]$, resp. zde BÚNO $[0, 1]$. Zavedeme dělení intervalu $[0, 1]$

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \dots, B_m = \left[\frac{m-1}{m}, 1\right]$$

a označíme dále $h = \frac{1}{m} = \lambda(B_j)$, $Y_j = \#\{i : X_i \in B_j\}_{i=1}^n$, $\hat{p}_j = \frac{Y_j}{n}$ jako odhad $p_j = \int_{B_j} f(x)dx$.

Pak **histogramový odhad hustoty pravděpodobnosti** definujeme vztahem

$$\hat{f}_n^H(t) = \sum_{j=1}^m \frac{\hat{p}_j}{h} \mathbb{I}_{B_j}(t) = \frac{1}{nh} \sum_{j=1}^m Y_j \mathbb{I}_{B_j}(t) = \frac{1}{nh} \sum_{i=1}^n \mathbb{I}\{X_i \in B_j\}, \quad \forall t \in B_j.$$

Věta 1.39 (IMSE). Pro $\hat{f}_n^H(t)$ předpokládejme, že f' je absolutně spojitá a platí, že $\int_{\mathbb{R}} (f'(u))^2 du < +\infty$. Potom

$$R(\hat{f}_n^H, f) = \frac{h^2}{12} \int_0^1 (f'(x))^2 dx + \frac{1}{nh} + o(h^2) + o\left(\frac{1}{n}\right),$$

což při volbě $h_n = O(n^{-1/3})$ vede na řád integrované střední kvadratické chyby (Integrated Mean Square Error)

$$\text{IMSE} = R(\hat{f}_n^H, f) = \int_0^1 \mathbb{E} \left[\hat{f}_n^H(t) - f(t) \right]^2 dt = O(n^{-2/3}).$$

Definice 1.40. Mějme jádro $K(x)$ takové, že $K(x) \geq 0$, $\int_{\mathbb{R}} K(x)dx = 1$, $\int_{\mathbb{R}} xK(x)dx = 0$, $\sigma_K^2 = \int_{\mathbb{R}} x^2 K(x)dx > 0$. Označíme $h \in \mathbb{R}^+$ jako tzv. **šířku okna** (bin width), neboli vyhlazovací parametr (smoothing parameter). Pak definujeme **jádrový odhad hustoty** vztahem

$$\hat{f}_n^K(t) := \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{t - X_i}{h}\right), \quad \forall t \in \mathbb{R}.$$

POZNÁMKA 1.41. Příklady takových používaných jader mohou být:

1. $K(x) = \frac{1}{2} \mathbb{I}_{[-1,1]}(x)$ (boxcar),
2. $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ (Gaussovo),
3. $K(x) = \frac{3}{4}(1 - x^2) \mathbb{I}_{[-1,1]}(x)$ (Epanechnikovo) apod.

POZNÁMKA 1.42. Podobný výsledek řádu IMSE lze ukázat pro jádrový odhad $\hat{f}_n^K(t)$. Při volbě $h_n = O(n^{-1/5})$ lze dosáhnout pro $\hat{f}_n^K(t)$ řád $\text{IMSE} = O(n^{-4/5})$.

2 Metody pro hledání bodových odhadů

2.1 Metoda momentů

Tato metoda je založená na užití výběrových momentů, ať už centrálních nebo necentrálních, případně i z momentů rozdělení. V praxi pak využijeme tu, která se dá vypočítat nejjednodušeji. V této metodě bereme v potaz všechny realizace.

Máme tedy $(\Omega, \mathcal{A}, \mathbb{P})$, $\theta \in \Theta \subset \mathbb{R}^k$, $\tau(\theta)$ jako odhadovanou parametrickou funkci, vlastnost $X \text{ iid } f_X(x, \theta)$. Nechť $X_1, \dots, X_n \text{ iid } \mathcal{L}_k$ (aby existovaly momenty do řádu k) a označíme

$$\begin{aligned}\mu_r &= \mu_r(\theta) = \mathbb{E}X^r, \quad r \in \hat{k}, \\ \boldsymbol{\mu}(\theta) &= (\mu_1(\theta), \dots, \mu_k(\theta)) : \mathbb{R}^k \rightarrow \mathbb{R}^k.\end{aligned}$$

Dále požadujeme, aby $\exists \boldsymbol{\mu}^{-1}$, tedy například aby $\boldsymbol{\mu}$ bylo regulární a prosté.

Definice 2.1 (Momentový odhad). **Momentový odhad** θ definujeme vztahem

$$\hat{\theta}_M := \hat{\theta}_M(\mathbf{X}) = \boldsymbol{\mu}^{-1}(m'_1(\mathbf{X}), \dots, m'_k(\mathbf{X})), \quad \text{kde } m'_r(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^n X_j^r,$$

což znamená, že $\hat{\theta}_M$ je řešením *soustavy momentových rovnic* (značíme ME_q) ve tvaru

$$\mu_r(\theta) = m'_r(\mathbf{X}), \quad \forall r \in \hat{k}.$$

Definujeme dále **momentový odhad** $\tau(\theta)$ vztahem $T_M(\mathbf{X}) := \tau(\hat{\theta}_M(\mathbf{X}))$, metoda momentů je tedy invariantní na transformace parametru θ (je to jen jiné vyjádření pro vztah $T_M(\mathbf{X}) = \tau(\hat{\theta}_M) = \tau(\hat{\theta}_M)$).

POZNÁMKA 2.2. 1. Pokud soustava ME_q není jednoznačně řešitelná, případně některý z momentů μ_r nezávisí na θ , pak můžeme přidat další rovnici ve tvaru $\mu_{k+1}(\theta) = m_{k+1}(\mathbf{X})$.

2. Alternativně lze užít i centrální momenty

$$\mu_r(\theta) = \mathbb{E}(X - \mathbb{E}X)^r, \quad m_r(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^r.$$

PŘÍKLAD 2.3. Mějme $(X_j)_{j=1}^n \text{ iid } \mathcal{N}(\mu, \sigma^2)$. Pak je soustava ME_q ve tvaru

$$\begin{aligned}\mathbb{E}X &= \mu \stackrel{!}{=} \bar{X}_n, \\ \mathbb{E}(X - \mathbb{E}X)^2 &= \text{DX} = \sigma^2 \stackrel{!}{=} \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^2 = \hat{\sigma}_n^2.\end{aligned}$$

Momentové odhady $\hat{\mu}_M = \bar{X}_n$ a $\hat{\sigma}_M^2 = \hat{\sigma}_n^2$ jsou tedy konzistentní a \mathcal{AN} odhady μ a σ^2 . Obecně však momentové odhady nejsou eficientní.

2 Metody pro hledání bodových odhadů

Věta 2.4. Pokud je $\boldsymbol{\mu}^{-1}$ spojitá funkce, pak $\hat{\theta}_M(\mathbf{X})$ je **konzistentním** odhadem θ . Pokud je navíc τ spojitá, pak $T_M(\mathbf{X})$ je konzistentní.

Důkaz. Ze ZVČ (Chinchin) víme, že $m_1(\mathbf{X}) \xrightarrow{\mathbb{P}, s.j.} \mu_1, \dots, m_k(\mathbf{X}) \xrightarrow{\mathbb{P}, s.j.} \mu_k$. Pak ale

$$\hat{\theta}_M = \boldsymbol{\mu}^{-1}(m_1(\mathbf{X}), \dots, m_k(\mathbf{X})) \xrightarrow{\mathbb{P}, s.j.} \boldsymbol{\mu}^{-1}(\mu_1, \dots, \mu_k) = \boldsymbol{\mu}^{-1}(\boldsymbol{\mu}(\theta)) = \theta.$$

Podobně pak $T_M(\mathbf{X}) = \tau(\hat{\theta}_M) \xrightarrow{\mathbb{P}, s.j.} \tau(\theta)$ pro $\tau(\theta)$ spojitě. □

POZNÁMKA 2.5 (Připomínka Delta metody). Mějme $(\mathbf{X}_n)_{n=1}^{+\infty}$ do \mathbb{R}^k , $\mathbf{X}_n \sim \mathcal{N}_k(\theta, \frac{1}{n}\mathbb{C}(\theta))$. Necht' $g: \mathbb{R}^k \rightarrow \mathbb{R}$ je borelovská, $\nabla g(\theta) \neq 0$, $\exists \frac{\partial g}{\partial \theta_i}$ na H_θ (okolí θ) a jsou spojitě v bodě θ . Pak

$$g(\mathbf{X}_n) \sim \mathcal{N}_1\left(g(\theta), \frac{1}{n} \nabla g(\theta) \mathbb{C}(\theta) \nabla g(\theta)^T\right).$$

Zobecnění pro $g: \mathbb{R}^k \rightarrow \mathbb{R}^k$: Pokud Jacobiho matice \mathbb{J}_g zobrazení g existuje na okolí H_θ a je spojitá v bodě θ , $\mathbb{J}_g(\theta) \neq \mathbf{0}$, pak

$$g(\mathbf{X}_n) \sim \mathcal{N}_k\left(g(\theta), \frac{1}{n} \mathbb{J}_g(\theta) \mathbb{C}(\theta) \mathbb{J}_g(\theta)^T\right).$$

Věta 2.6. Necht' $\hat{\theta}_M$ je odhad metodou momentů, $(X_j)_{j=1}^{+\infty} \in \mathcal{L}_{2k}$, $\boldsymbol{\mu}$ je difeomorfismus $(\boldsymbol{\mu}, \boldsymbol{\mu}^{-1})$ spojitě diferencovatelné. Pak $\forall \theta \in \Theta$ je

$$\hat{\theta}_M \sim \mathcal{N}\left(\theta, \frac{1}{n} \mathbb{C}_M(\theta)\right),$$

kde $\mathbb{C}_M(\theta) = \mathbb{J}_{\boldsymbol{\mu}^{-1}}(\theta) \mathbb{C}(\theta) \mathbb{J}_{\boldsymbol{\mu}^{-1}}(\theta)^T$ a $\mathbb{C} = \text{Cov}(X, X^2, \dots, X^k)$. Pokud je navíc $\tau(\theta)$ spojitě diferencovatelné a $\nabla \tau(\theta) \neq 0$, pak

$$T_M(\mathbf{X}) \sim \mathcal{N}\left(\tau(\theta), \frac{1}{n} \nabla \tau(\theta) \mathbb{C}_M(\theta) \nabla \tau(\theta)^T\right).$$

Důkaz. Definujeme $\mathbb{Z}_j := (X_j, X_j^2, \dots, X_j^k)$. Potom

$$\mathbb{E}\mathbb{Z}_j = (\mathbb{E}X_j, \mathbb{E}X_j^2, \dots, \mathbb{E}X_j^k) = (\mathbb{E}X, \mathbb{E}X^2, \dots, \mathbb{E}X^k) = \boldsymbol{\mu}(\theta).$$

Označme nyní $\text{Cov}(\mathbb{Z}_j) = \text{Cov}(X, X^2, \dots, X^k) = \mathbb{C}(\theta)$. Víme, že $\mathbb{Z}_j \text{ iid } \mathcal{L}_2$, tudíž dle CLT v \mathbb{R}^k platí, že $\overline{\mathbb{Z}}_n = (m_r(\mathbf{X}))_{r=1}^k \sim \mathcal{N}_k(\boldsymbol{\mu}(\theta), \frac{1}{n} \mathbb{C}(\theta))$. Pak pro $g = \boldsymbol{\mu}^{-1}$ spojitě diferencovatelné dostáváme z Δ -metody vztah

$$\hat{\theta}_M = \boldsymbol{\mu}^{-1}\left((m_r(\mathbf{X}))_{r=1}^k\right) \stackrel{1.19}{\sim} \mathcal{N}\left(\theta, \frac{1}{n} \underbrace{\mathbb{J}_{\boldsymbol{\mu}^{-1}}(\theta) \mathbb{C}(\theta) \mathbb{J}_{\boldsymbol{\mu}^{-1}}(\theta)^T}_{\mathbb{C}_M(\theta)}\right).$$

Následně pro τ spojitě diferencovatelné opět z Δ -metody dostaneme, že

$$T_M(\mathbf{X}) = \tau(\hat{\theta}_M) \sim \mathcal{N}\left(\tau(\theta), \frac{1}{n} \nabla \tau(\theta) \mathbb{C}_M(\theta) \nabla \tau(\theta)^T\right).$$

□

2 Metody pro hledání bodových odhadů

PŘÍKLAD 2.7. Mějme $X_1, \dots, X_n \text{ iid } \mathcal{N}(\mu, \sigma^2)$, kde neznáme hodnotu parametrů μ a σ^2 . Víme ale, že

$$\mathbb{E}X_i = \mu, \quad \text{D}X_i = \sigma^2 \quad \text{a} \quad \mathbb{E}(X_i)^2 = \text{D}X_i + (\mathbb{E}X_i)^2 = \sigma^2 + \mu^2.$$

Odhady parametrů metodou momentů dostaneme z rovnic

$$\mu = m'_1 = \overline{X_n} \quad \text{a} \quad \sigma^2 = m_2 = \frac{1}{n} \left(\sum_{j=1}^n X_j - \overline{X_n} \right)^2 = \hat{\sigma}_n^2.$$

Tím získáme odhady $\hat{\mu}_M, \hat{\sigma}_M^2$.

2.2 Nestranné odhady s minimálním rozptylem (UMVUE)

Definice 2.8. Mějme $X_1, \dots, X_n \text{ iid } F$, $\theta \in \Theta \subset \mathbb{R}^k$, $\tau(\theta) \in \mathbb{R}^1$, $T(\mathbf{X})$ jako odhad $\tau(\theta)$. Definujeme

$$T_{\text{UMR}} = \underset{T}{\operatorname{argmin}} \mathbb{E}(T(\mathbf{X}) - \tau(\theta))^2, \quad \forall \theta \in \Theta.$$

Definujeme dále kvadratickou **ztrátovou funkci** (*loss function*) jako

$$\mathcal{L}_2(T, \theta) := (T(\mathbf{X}) - \tau(\theta))^2$$

a příslušnou **rizikovou funkci** (*risk function*) vztahem

$$R(T, \theta) := \mathbb{E}\mathcal{L}_2(T, \theta).$$

Potom $T_{\text{UMR}} = \underset{T}{\operatorname{argmin}} R(T, \theta)$, kde UMR je zkratka pro *uniformly minimum risk*. T_{UMR} je tedy odhad, který minimalizuje hodnotu rizikové funkce R .

Definice 2.9. $S(\mathbf{X})$ se nazývá **postačující statistika** (*sufficiency*) pro θ , pokud rozdělení \mathbf{X} podmíněné hodnotou $S(\mathbf{X}) = s$ nezávisí na parametru θ . Pro diskrétní případy tedy $\mathbb{P}(\mathbf{X} = \mathbf{x} | S(\mathbf{X}) = s)$ nezávisí na θ , případně $f_{\mathbf{X}|S}(\mathbf{x}|s)$ nezávisí na θ .

Postačující statistika je tedy taková funkce náhodného výběru (statistika), která umí sama o sobě nahradit původní výběr bez ztráty informace o parametru θ .

PŘÍKLAD 2.10. Házíme mincí, $X_i \sim \text{Be}(\theta)$, $\mathbf{x} = (0, 1, 1, 0, 1, \dots)$. Představme si, že chceme vypočítat tzv. MLE (bude definováno později) ve tvaru

$$L = \prod_{i=1}^N \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{x_1+x_2+\dots} (1 - \theta)^{1-x_1+1-x_2+\dots} = \theta^{\sum_{i=1}^N x_i} (1 - \theta)^{N - \sum_{i=1}^N x_i}.$$

Pak ale ve výsledku nezáleží na samotných datech, ale pouze na jejich součtu, tedy pokud označíme $S(\mathbf{X}) := \sum_{i=1}^N X_i$, pak $S(\mathbf{X})$ je postačující statistika (podle Neymannova fatrorizačního kritéria).

Definice 2.11. Mějme X, Y náhodné veličiny na $(\Omega, \mathcal{A}, \mathbb{P})$. Pak

$$\mathbb{E}(X|Y = y) = \int_{\mathbb{R}} x d\mathbb{P}^{X|Y=y} = \left[\text{pro ASR } f_{X|Y}(x|y) := \frac{f_{X,Y}(x, y)}{f_Y(y)} \right] = \dots$$

Z toho vyplývá, že $\mathbb{E}(X|Y) : \Omega \rightarrow \mathbb{R}$ je náhodná veličina.

2 Metody pro hledání bodových odhadů

Věta 2.12. Pro náhodné veličiny (X, Y) s ASR $f_{X,Y}$ platí vztah

$$\mathbb{E}[\mathbb{E}(X|Y)] \stackrel{\text{ASR}}{=} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} x f_{X|Y} dx \right) f_Y dy \stackrel{\text{F.V.}}{=} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f_{X|Y} f_Y dy \right) x dx = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f_{X,Y} dy \right) x dx = \int_{\mathbb{R}} x f_X dx = \mathbb{E}X.$$

Pro účely následující Rao-Blackwellovy věty označme

$$T_{\text{RB}}(\mathbf{X}) = T_{\text{RB}}(S(\mathbf{X})) := \mathbb{E}[T(\mathbf{X})|S(\mathbf{X}) = s]$$

za předpokladu existence \mathbb{E} jako odhad zkonstruovaný v Rao-Blackwellově větě. $T_{\text{RB}}(\mathbf{X})$ je tedy opět statistika, pro kterou platí, že

$$T_{\text{RB}}(\mathbf{X}) := \mathbb{E}[T(\mathbf{X})|S(\mathbf{X}) = s] \stackrel{\text{ASR}}{=} \int T(\mathbf{x}) f_{T(\mathbf{X})|S(\mathbf{X})=s} d\mathbf{x}.$$

Na $T_{\text{RB}}(\mathbf{X})$ pohlížíme jako na funkci \mathbf{X} , která vznikne ve dvou krocích:

1. spočítá se podmíněná střední hodnota $\mathbb{E}(T(\mathbf{X})|S(\mathbf{X}) = s) = T(s)$ při libovolně daném pevném s ,
2. za s se dosadí vazba $s = S(\mathbf{X})$, čímž vznikne $T_{\text{RB}}(S(\mathbf{X}))$.

Věta 2.13 (Rao-Blackwell). Mějme X_1, \dots, X_n iid F , $\theta \in \Theta \subset \mathbb{R}^k$, $\tau(\theta) \in \mathbb{R}^1$, $T(\mathbf{X})$ jako odhad $\tau(\theta)$, nechť $S(\mathbf{X})$ je postačující statistika pro θ a nechť $\mathcal{L}(T, \theta)$ je konvexní funkcí v T pro $\forall \theta \in \Theta$. Pak

$$R(T_{\text{RB}}(\mathbf{X}), \theta) \leq R(T(\mathbf{X}), \theta), \quad \forall \theta \in \Theta,$$

přičemž T_{RB} nezávisí na θ .

Důkaz. Větu ukážeme obecněji pro $\mathcal{L}(T, \theta)$ konvexní v T , která zahrnuje i naši kvadratickou ztrátovou funkci \mathcal{L}_2 . Pro $\forall \theta \in \Theta$ z Jensenovy nerovnosti platí, že

$$R(T_{\text{RB}}, \theta) = \mathbb{E} \mathcal{L}(\mathbb{E}[T|S = s], \theta) \leq \mathbb{E}[\mathbb{E}[\mathcal{L}(T, \theta)|S = s]] = \mathbb{E} \mathcal{L}(T, \theta) = R(T, \theta).$$

□

Definice 2.14. Systém hustot $\mathcal{F} = \mathcal{F}_X = \{f(x, \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$ se nazývá **úplný**, pokud pro $\forall h : \mathbb{R} \rightarrow \mathbb{R}$, $\mathbb{E}_{\theta} h(X) = 0$, $\forall \theta \in \Theta$ platí, že

$$h(X) = 0 \text{ s.j. } \forall \theta \in \Theta, \quad \text{neboli} \quad \mathbb{P}_{\theta}(h(X) = 0) = 1.$$

Definice 2.15. Postačující statistika S se nazývá **úplná postačující statistika**, pokud systém rozdělení \mathcal{F}_S je úplný, tzn. pro libovolnou borelovskou funkci $g : \mathbb{R} \rightarrow \mathbb{R}$ platí, že pokud

$$\mathbb{E}[g(S(X))] = 0, \quad \forall \theta \in \Theta, \quad \text{pak} \quad g(S(X)) = 0 \text{ s.j.}, \quad \forall \theta \in \Theta.$$

Věta 2.16 (Lehmann-Scheffé). Nechť jsou splněny předpoklady R.-B. věty a navíc $T(\mathbf{X})$ je nestranný odhad $\tau(\theta)$ (tzn. $\mathbb{E}T = \tau$, $\forall \theta \in \Theta$), a dále nechť $S(\mathbf{X})$ je úplná postačující statistika. Pak $T_{\text{RB}} = T_{\text{UMR}}$, což pro volbu ztrátové funkce \mathcal{L}_2 označíme jako T_{UMVUE} . (uniformly minimum variance unbiased estimator)

2 Metody pro hledání bodových odhadů

Důkaz. • Z předpokladu víme, že T je nestranné, tedy i T_{RB} je nestranné, protože $\mathbb{E}T_{\text{RB}} = \mathbb{E}[\mathbb{E}(T|S = s)] = \mathbb{E}T = \tau(\theta), \forall \theta \in \Theta$.

- T_{RB} nestranné $\Rightarrow R(T_{\text{RB}}, \theta) = \mathbb{E}\mathcal{L}_2(T_{\text{RB}}, \theta) = \mathbb{E}(T_{\text{RB}}(\mathbf{X}) - \underbrace{\tau(\theta)}_{\mathbb{E}T_{\text{RB}}(\mathbf{X})})^2 = D[T_{\text{RB}}(\mathbf{X})]$.
- $T_{\text{RB}} = T_{\text{UMVUE}} : T^{(1)}, T^{(2)}$ oba nestranné odhady $\tau(\theta)$. Z toho vyplývá, že $T_{\text{RB}}^{(1)}$ a $T_{\text{RB}}^{(2)}$ jsou také nestranné odhady $\tau(\theta)$. Potom ale

$$\mathbb{E}(T_{\text{RB}}^{(1)} - T_{\text{RB}}^{(2)}) = 0 \xRightarrow{\text{S je úplná}} T_{\text{RB}}^{(1)} - T_{\text{RB}}^{(2)} = 0 \quad \text{s.j. } \forall \theta \in \Theta,$$

protože na $T_{\text{RB}}^{(1)} - T_{\text{RB}}^{(2)}$ aplikujeme úplnou postačující statistiku $S(\mathbf{X}) = s$.

□

POZNÁMKA 2.17. T_{RB} jen stejnoměrně vylepšuje, ale nemusí dosahovat na ten stejnoměrně nejlepší. Že z R-B věty vyleze ten úplně nejlepší, který už nelze stejnoměrně vylepšit, právě zajišťují předpoklady Lehmann-Scheffé.

Pozor, UMVUE není invariantní na transformace, tzn. $\widehat{\tau(\theta)}_{\text{UMVUE}} \neq \tau(\hat{\theta}_{\text{UMVUE}})$.

2.3 Rao-Cramérova nerovnost

Definice 2.18. Mějme $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$ a označíme

$$l(\theta) = \ln f(x, \theta), \quad l'_i(\theta) := \frac{\partial}{\partial \theta_i} \ln f(x, \theta), \quad \nabla_{\theta} l = \nabla l(\theta) = (l'_1(\theta), \dots, l'_k(\theta)).$$

Systém \mathcal{F} se nazývá **regulární systém hustot**, ozn. \mathcal{F}_{reg} , pokud

- 1) $\text{supp } f$ nezávisí na θ a Θ je otevřená množina,
- 2) pro všechna $\forall \theta \in \Theta$ existuje $\nabla l(\theta)$ na $\text{supp } f$,
- 3) $\mathbb{E}[\nabla l(\theta)] = 0, \forall \theta \in \Theta$, což je zajištěno předpokladem následující záměny

$$\mathbb{E}[l'_i(\theta)] = \int_{\mathbb{R}} \frac{f'_i}{f} \cdot f dx = \int_{\mathbb{R}} \frac{\partial}{\partial \theta_i} f dx = \frac{\partial}{\partial \theta_i} \underbrace{\int_{\text{supp } f} f dx}_{=1} = 0,$$

- 4) $\text{Cov}(\nabla l(\theta))$ je konečná a PD matice rozměru $(k \times k)$.

Systém \mathcal{F} označíme $\mathcal{F}_{\text{reg}}^+$, pokud navíc splňuje podmínku

$$5) \mathbb{E}\left[\frac{f''_{i,j}(\mathbf{X}, \theta)}{f(\mathbf{X}, \theta)}\right] = 0, \forall \theta \in \Theta, \quad \text{neboli } \int_{\mathbb{R}^n} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(\mathbf{x}, \theta) d\mathbf{x} = 0, \forall \theta \in \Theta.$$

Definice 2.19. $\mathbb{I}(\theta) = \text{Cov}(\nabla l(\theta))$ se nazývá **Fisherova informační matice** a platí, že

$$\mathbb{I}_{i,j}(\theta) := \text{Cov}(l'_i, l'_j) = \mathbb{E}(l'_i \cdot l'_j) - \underbrace{\mathbb{E}l'_i}_{=0} \underbrace{\mathbb{E}l'_j}_{=0} = \mathbb{E}\left(\frac{\partial \ln f}{\partial \theta_i} \cdot \frac{\partial \ln f}{\partial \theta_j}\right) \stackrel{5)}{=} -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial \theta_i \partial \theta_j}\right].$$

2 Metody pro hledání bodových odhadů

Lemma 2.20. Mějme $\mathbf{X} = (X_1, \dots, X_n)$ nezávislé, $X_i \sim f_{X_i}(x_i, \theta)$. Pak $\mathbb{I}_{\mathbf{X}}(\theta) = \sum_{j=1}^n \mathbb{I}_{X_j}(\theta)$. Pokud navíc X_1, \dots, X_n jsou iid f_X , pak $\mathbb{I}_{\mathbf{X}}(\theta) = n\mathbb{I}_X(\theta)$.

Důkaz. Speciálně pro systém \mathcal{F}_{reg}^+ dostaneme

$$(\mathbb{I}_{\mathbf{X}})_{ij}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial \theta_i \partial \theta_j}\right] = -\mathbb{E}\left[\frac{\partial^2 \ln \prod_{l=1}^n f_{X_l}}{\partial \theta_i \partial \theta_j}\right] = -\sum_{l=1}^n \mathbb{E}\left[\frac{\partial^2 \ln f_{X_l}}{\partial \theta_i \partial \theta_j}\right] = \sum_{l=1}^n (\mathbb{I}_{X_l})_{ij}(\theta).$$

□

Věta 2.21 (Rao-Cramérova nerovnost). Mějme $T(\mathbf{X})$ jako nestranný odhad $\tau(\theta)$, \mathcal{F}_{reg} , nechť dále $(\forall \theta \in \Theta)(\exists \nabla \tau(\theta))$ a $\mathbb{E}[T(\mathbf{X})]$ lze derivovat podle θ_i pod znakem \mathbb{E} (tj. derivace pod integrálem) pro $\forall i \in \hat{k}$. Pak

$$D(T(\mathbf{X})) \geq \underbrace{\nabla \tau(\theta) \mathbb{I}_{\mathbf{X}}^{-1}(\theta) \nabla \tau(\theta)^T}_{\text{RCLB}_{\tau}(\theta)}, \quad \forall \theta \in \Theta. \quad (\text{RCLB} = \text{Rao-Cramérova spodní hranice}).$$

Náznak důkazu. Označme $\mathbb{D} := \begin{pmatrix} DT(\mathbf{X}) & \nabla \tau \\ \nabla \tau^T & \mathbb{I}_{\mathbf{X}}(\theta) \end{pmatrix}$. Protože každá kovarianční matice je PSD, stačí dokázat, že se jedná o Cov matici vektoru (T, l'_1, \dots, l'_k) , tedy že $\mathbb{D} = \text{Cov}(T, l'_1, \dots, l'_k)$. To plyne z následujícího výpočtu:

$$\frac{\partial \tau}{\partial \theta_i}(\theta) = \frac{\partial}{\partial \theta_i} \int T(\mathbf{X}) f_{\mathbf{X}}(\mathbf{x}, \theta) d\mathbf{x} = \int T \frac{f'_i}{f} d\mathbf{x} = \int T \frac{\partial \ln f}{\partial \theta_i} f d\mathbf{x} = \mathbb{E}\left(T \frac{\partial \ln f}{\partial \theta_i}\right) = \text{Cov}(T, l'_i),$$

protože střední hodnota $\mathbb{E}l'_i$ je nulová, což víme z regularity systému \mathcal{F} . Pak tedy $|\mathbb{D}| \geq 0$ a R.-C. nerovnost získáme z rozvoje $|\mathbb{D}|$ podle 1. řádku a poté podle 1. sloupce. Tím se získá dvojnásobná suma, ze které se vyjádří $DT(\mathbf{X})$. Následně se použije Cramerovo pravidlo pro $\mathbb{I}_{ij}^{-1}(\theta)$. □

Definice 2.22. Mějme \mathcal{F}_{reg} . Pak definujeme **eficienci** nestranného odhadu $T_n(\mathbf{X})$ funkce $\tau(\theta)$ vztahem

$$e_n := \frac{\text{RCLB}_{\tau}(\theta)}{D(T_n(\mathbf{X}))}.$$

Pokud $e_n = 1$, $\forall n \in \mathbb{N}$, případně $\lim_{n \rightarrow +\infty} e_n = 1$, pak $T(\mathbf{X})$ nazýváme **(asymptoticky) eficientní**.

Speciálně volme do definice 2.18 jednorozměrný $\theta \in \Theta \subset \mathbb{R}^1$. Pak \mathcal{F}_{reg}^+ je definován následovně:

- 1), 2), 3) $\mathbb{E}\left[\frac{\partial \ln f}{\partial \theta}(x, \theta)\right] = 0$,
- 4) $\mathbb{I}(\theta) = D\left(\frac{\partial \ln f}{\partial \theta}\right) \stackrel{5)}{=} -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial \theta^2}\right]$, což se nazývá **Fisherova informace** o θ .

Pak R.-C.N. přechází do tvaru $D(T(\mathbf{X})) \geq \frac{[\tau'(\theta)]^2}{\mathbb{I}_{\mathbf{X}}(\theta)}$, $\forall \theta \in \Theta$, přičemž rovnost nastane právě tehdy, když $|\mathbb{D}| = \begin{vmatrix} DT & \tau' \\ \tau' & \mathbb{I}_{\mathbf{X}}(\theta) \end{vmatrix} = 0$. Z toho vyplývá, že

$$DT \cdot \mathbb{I}_{\mathbf{X}}(\theta) = \tau'(\theta)^2 = \text{Cov}^2\left(T, \frac{\partial \ln f}{\partial \theta}\right) = DT \cdot D\left(\frac{\partial \ln f}{\partial \theta}\right).$$

2 Metody pro hledání bodových odhadů

Po odmocnění pak dostaneme rovnost ve Schwarzově nerovnosti, která nastává právě tehdy, když veličiny $(T - \tau)$ a $\frac{\partial \ln f}{\partial \theta}$ jsou lineárně závislé s.j., tzn.

$$\frac{\partial \ln f_{\mathbf{X}}}{\partial \theta} = \sum_{j=1}^n \frac{\partial \ln f_{X_j}}{\partial \theta} = K(\theta, n)(T(\mathbf{X}) - \tau(\theta)), \quad \text{s.j.} \quad (2.1)$$

PŘÍKLAD 2.23. Necht' $f_{\mathbf{X}}(\mathbf{x}, \theta) = c(\theta)h(\mathbf{x})e^{Q(\theta)T(\mathbf{x})}$, tzn. \mathcal{F} je exponenciální třída hustot. Pak

$$\begin{aligned} \ln f_{\mathbf{X}}(\mathbf{x}, \theta) &= \ln c(\theta) + \ln h(\mathbf{x}) + Q(\theta)T(\mathbf{x}) \text{ a dále} \\ \frac{\partial \ln f_{\mathbf{X}}}{\partial \theta} &= \frac{c'(\theta)}{c(\theta)} + Q'(\theta)T(\mathbf{x}) = Q'(\theta) \left[T(\mathbf{x}) - \underbrace{\left(-\frac{1}{c(\theta)} \cdot \frac{c'(\theta)}{Q'(\theta)} \right)}_{\tau(\theta)} \right]. \end{aligned}$$

Tedy pokud $T(\mathbf{X})$ je nestranný a $\tau(\theta) = -\frac{1}{c(\theta)} \frac{c'(\theta)}{Q'(\theta)}$ (předpokládáme existenci $c'(\theta)$ a $Q'(\theta)$), pak z rovnice (2.1) získáváme UMVUE $T(\mathbf{X})$ pro $\tau(\theta)$.

Věta 2.24 (Bhattacharya). Mějme $\theta \in \Theta \subset \mathbb{R}^1$, $T(\mathbf{X})$ jako odhad $\tau(\theta)$. Necht' dále existuje vektor derivací podle θ do řádu m : $\tilde{\tau}' = (\tau', \tau'', \dots, \tau^{(m)})$. Pak za podobných dodatečných předpokladů jako v R.-C.N., platí, že

$$D(T(\mathbf{X})) \geq \tilde{\tau}'(\theta) \tilde{\mathbb{I}}_{\mathbf{X}}^{-1}(\theta) (\tilde{\tau}'(\theta))^T, \quad (\text{Bhattacharyova spodní hranice, BLB}_{\tau}(\theta)),$$

$$\text{kde } \tilde{\mathbb{I}}_{\mathbf{X}} = \text{Cov} \left(\left(\frac{\partial^i f}{\partial \theta^i} / f \right)_{i=1}^m \right).$$

POZNÁMKA 2.25. Pro $m = 1$ přechází $\text{BLB}_{\tau}(\theta)$ na $\text{RCLB}_{\tau}(\theta)$. Pokud se ani pomocí této BLB nedosáhne hranice, je třeba hledat UMVUE prostřednictvím R-B věty: $T_{\text{UMVE}} = \mathbb{E}[T|S = s]$.

2.4 Metoda maximální věrohodnosti (MLE)

MLE je ve statistice velice často používaná metoda, pomocí které hledáme bodové odhady parametrů. Snažíme se maximalizovat sdruženou hustotu experimentu vzhledem k parametru, který je obsažen v navrženém rozdělení. Předpokládáme totiž, že známe toto rozdělení, jen neznáme jeho parametr. Tímto způsobem získáme tzv. **věrohodnostní funkci**.

Definice 2.26 (Věrohodnostní funkce). Mějme X_1, \dots, X_n s odpovídajícím systémem hustot $\mathcal{F} = \{f_{\mathbf{X}}(\mathbf{x}, \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$. Pak definujeme **věrohodnostní funkci** vztahem

$$L(\theta) = f(\mathbf{x}, \theta), \quad \text{resp.} \quad L(\theta, \mathbf{x}) = h(\mathbf{x})f(\mathbf{x}, \theta), \quad \forall \theta \in \Theta, \quad \forall \mathbf{x} \in \mathbb{R}^n$$

a **logaritmickou věrohodnostní funkci** vztahem

$$l(\theta) = \ln L(\theta, \mathbf{x}).$$

Mějme nezávislý náhodný výběr X_1, \dots, X_n id. Pak věrohodnostní funkci můžeme zavést jako sdruženou hustotu tohoto výběru při dané realizaci $\mathbf{x} = (x_1, \dots, x_n)$, tedy

$$L(\theta) = \prod_{i=1}^n f_{X_i}(x_i, \theta), \quad \left(L(\theta) = \prod_{i=1}^n \mathbb{P}_{\theta}(X_i = x_i) - \text{pro diskrétní případ} \right).$$

2 Metody pro hledání bodových odhadů

Definice 2.27 (Maximálně věrohodný odhad). Definujeme **maximálně věrohodný odhad** vztahem

$$\hat{\theta}_{\text{ML}}(\mathbf{X}) = \underset{\theta \in \Theta}{\operatorname{argsup}} L(\theta)$$

za předpokladu, že $\hat{\theta}_{\text{ML}}$ je borelovsky měřitelná, jednoznačná a závisí na \mathbf{X} . Dále pak definujeme **maximálně věrohodný odhad** parametrické funkce $\tau(\theta)$ jako

$$T_{\text{ML}}(\mathbf{X}) = \tau(\hat{\theta}_{\text{ML}}), \quad \text{tedy MLE je invariantní na transformace } \tau.$$

POZNÁMKA 2.28. Postup v praxi:

maximalizujeme $L(\theta_1, \dots, \theta_k)$ přes všechny možné hodnoty $\theta_1, \dots, \theta_k$. Vzhledem ke tvaru L je ale vhodné ji nejprve zlogaritmovat (logaritmus je monotonní a nemění proto polohu maxima). To je právě důvod, proč se zavádí $l(\theta_1, \dots, \theta_k) = \ln L(\theta_1, \dots, \theta_k)$. Potom získáme

$$\frac{\partial \ln L(\theta_1, \dots, \theta_k)}{\partial \theta_i} = \frac{\partial l(\theta_1, \dots, \theta_k)}{\partial \theta_i} = 0, \quad i \in \hat{k},$$

který nazveme **systém věrohodnostních rovnic**, ozn LE_q .

Věta 2.29. Mějme $(X_j)_{j=1}^{+\infty}$ iid $f(x, \theta_0)$, kde $\operatorname{supp} f_{\mathbf{X}}$ nezávisí na θ , $\ln \frac{f(X, \theta)}{f(X, \theta_0)} \in \mathcal{L}_1$. Dále předpokládáme identifikovatelnost rodiny hustot, tzn.

$$\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2} \quad (\text{různé parametry definují různá rozdělení}).$$

Pak $\mathbb{P}_{\theta_0}(L(\theta_0, \mathbf{x}) > L(\theta, \mathbf{x})) \xrightarrow{n \rightarrow +\infty} 1, \quad \forall \theta \neq \theta_0$.

Důkaz.

$$\mathbb{P}_{\theta_0}\left(\frac{L(\theta, \mathbf{x})}{L(\theta_0, \mathbf{x})} < 1\right) = \mathbb{P}_{\theta_0}\left(\frac{1}{n} \ln \frac{L(\theta, \mathbf{x})}{L(\theta_0, \mathbf{x})} < 0\right) \stackrel{\text{iid}}{=} \mathbb{P}_{\theta_0}\left(\underbrace{\frac{1}{n} \sum_{j=1}^n \ln \frac{f(x_j, \theta)}{f(x_j, \theta_0)}}_{\xrightarrow{\mathbb{P}} a < 0} < 0\right)$$

Podle ZVČ aplikovaného na $Y_j = \ln \frac{f(X_j, \theta)}{f(X_j, \theta_0)}$ dostaneme $\frac{1}{n} \sum_{j=1}^n Y_j \xrightarrow{\mathbb{P}_{\theta_0}} \mathbb{E}_{\theta_0} Y_1$. Podle Jensenovy nerovnosti pro ryze konkávní transformaci $\ln(u)$ potom plyne, že

$$\mathbb{E}_{\theta_0} Y_1 = \mathbb{E}_{\theta_0} \left(\ln \frac{f(X_1, \theta)}{f(X_1, \theta_0)} \right) < \ln \mathbb{E}_{\theta_0} \left(\frac{f(X_1, \theta)}{f(X_1, \theta_0)} \right) = \ln \int_{\mathbb{R}} \frac{f(x, \theta)}{f(x, \theta_0)} f(x, \theta_0) dx = \ln \int_{\mathbb{R}} f(x, \theta) dx = 0.$$

Nyní využijeme implikaci (důkaz ponechán čtenáři) $(Y_n \xrightarrow{\mathbb{P}} a < 0 \Rightarrow \mathbb{P}(Y_n < 0) \rightarrow 1)$, díky kterému plyne tvrzení věty, tzn. $\mathbb{P}_{\theta_0}\left(\frac{L(\theta, \mathbf{x})}{L(\theta_0, \mathbf{x})} < 1\right) \xrightarrow{n \rightarrow +\infty} 1$. \square

Definice 2.30. Systém hustot $\mathcal{F} := \{f(x, \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$ se nazývá **ML-regulární**, ozn. $\mathcal{F}_{\text{reg}}^{\text{ML}}$, pokud pro něj platí, že

1. Θ je otevřená množina, $\operatorname{supp} f_{\mathbf{X}}$ nezávisí na θ ,
2. $f(x, \theta) \in \mathcal{C}^{(3)}$ vzhledem k θ pro $\forall x \in \mathbb{R}$,

2 Metody pro hledání bodových odhadů

$$3. \int_{\mathbb{R}} \frac{\partial f}{\partial \theta_r} dx = 0 \quad \text{a} \quad \int_{\mathbb{R}} \frac{\partial^2 f}{\partial \theta_r \partial \theta_j} dx = 0, \quad \forall \theta \in \Theta, \text{ tzn. záměna derivací a integrálu je možná,}$$

4. Fisherova informační matice $\mathbb{I}_X(\theta)$ je PD a konečná,

$$5. (\forall \theta_0)(\exists H_{\theta_0})(\exists M(X) \in \mathcal{L}_1(\mathbb{P}_{\theta_0}))(\forall \theta \in H_{\theta_0}) \left(\|\partial_{\theta}^3 \ln f\| \leq M(X), \text{ přičemž } \mathbb{E}_{\theta_0} M(X) < +\infty \right).$$

Definice 2.31. Mějme $\hat{\theta}_n \sim \mathcal{AN}(\theta_0, \frac{1}{n} \mathbb{C}(\theta))$. $\hat{\theta}_n$ se nazývá **asymptoticky eficientní**, pokud $\mathbb{C}(\theta) = \mathbb{I}_X^{-1}(\theta_0)$.

Věta 2.32. Mějme X_1, \dots, X_n iid $f(x, \theta_0) \in \mathcal{F}_{reg}^{ML}$. Pak pro každé konzistentní řešení $\hat{\theta}_n(\mathbf{X})$ soustavy věrohodnostních rovnic LE_q platí, že

$$\hat{\theta}_n \sim \mathcal{AN}_k\left(\theta_0, \frac{1}{n} \mathbb{I}_X^{-1}(\theta_0)\right), \quad \text{neboli} \quad \sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}_k\left(0, \frac{1}{\mathbb{I}_X(\theta_0)}\right).$$

Takové konzistentní řešení je tedy \mathcal{AN} a dle definice 2.31 i asymptoticky eficientním odhadem θ_0 , ozn. $\hat{\theta}_{ELE}$ ($ELE = \text{eficient likelihood estimator}$).

Důkaz pro $k = 1$. $\hat{\theta}_n$ je konzistentní řešení rovnice $\frac{\partial l_n}{\partial \theta}(\theta, x) = 0$, tedy $l'_n(\hat{\theta}_n) = 0$, kde $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$

$l_n(\theta) = \ln \prod_{i=1}^n f_{X_i}(x, \theta)$. Nyní uděláme Taylorův rozvoj $l'_n(\hat{\theta}_n)$ do 2. řádu, tedy

$$l'_n(\hat{\theta}_n) = l'_n(\theta_0) + (\hat{\theta}_n - \theta_0) l''_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 l'''_n(\theta_n^*) = 0, \quad \text{kde } \theta_n^* \in [\hat{\theta}_n, \theta_0].$$

Z této rovnice vyjádříme $\sqrt{n}(\hat{\theta}_n - \theta_0)$ v následujícím tvaru:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\overbrace{\frac{1}{\sqrt{n}} l'_n(\theta_0)}^{4) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbb{I}_X(\theta))}}{\underbrace{\frac{1}{n} l''_n(\theta_0)}_{1) \xrightarrow{\mathbb{P}} -\mathbb{I}_X(\theta_0)} + \underbrace{\frac{1}{2}(\hat{\theta}_n - \theta_0)}_{2) \xrightarrow{\mathbb{P}} 0} \underbrace{\frac{1}{n} l'''_n(\theta_n^*)}_{3) \xrightarrow{\mathbb{P}} C < +\infty}} \xrightarrow{\mathcal{D}} \frac{1}{\mathbb{I}_X(\theta)} \mathcal{N}(0, \mathbb{I}_X(\theta)) \quad (\text{Stutskyho lemma.})$$

Nyní rozebereme jednotlivé části vzniklého výrazu:

1)

$$\begin{aligned} \frac{1}{n} l''_n(\theta_0) &= \frac{1}{n} \left(\ln \prod_{j=1}^n f_{X_j} \right)'' = \frac{1}{n} \sum_{j=1}^n (\ln f_{X_j})'' = \frac{1}{n} \sum_{j=1}^n l''_j(\theta_0) \xrightarrow{\mathbb{P} \text{ (ZVČ)}} \mathbb{E}[l''_1(\theta_0)] = \\ &= \mathbb{E} \left[\frac{\partial^2 \ln f_{X_1}}{\partial \theta^2} \right] = -\mathbb{I}_X(\theta_0) \end{aligned}$$

2) $\hat{\theta}_n - \theta_0 \xrightarrow{\mathbb{P}} 0$, protože $\hat{\theta}_n$ je konzistentním řešením LE_q .

3) Zde využijeme předpoklad č. 5 z definice, \mathcal{F}_{reg}^{ML} .

$$\left| \frac{1}{n} l'''_n(\theta_n^*) \right| = \left| \frac{1}{n} \sum_{j=1}^n l'''_j(\theta_n^*) \right| \leq \frac{1}{n} \sum_{j=1}^n \left| l'''_j(\theta_n^*) \right| \leq \frac{1}{n} \sum_{j=1}^n M(X_j) \xrightarrow{\mathbb{P} \text{ (ZVČ)}} \mathbb{E} M(X) = C < +\infty.$$

2 Metody pro hledání bodových odhadů

4) CLT říká, že pro Y_1, \dots, Y_n iid platí, že $\sqrt{n}(\bar{Y}_n - \mu) = \sqrt{n}\left(\frac{1}{n} \sum_{j=1}^n Y_j - \mathbb{E}Y_1\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \text{DY}_1)$.

Zkusíme se tedy dostat do tohoto tvaru.

$$\frac{1}{\sqrt{n}}l'_n(\theta_0) \stackrel{iid}{=} \sqrt{n}\left(\frac{1}{n} \sum_{j=1}^n l'_1(\theta_0) - \underbrace{\mathbb{E}l'_1(\theta_0)}_{=0}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \text{D}(l'_1(\theta_0))\right) = \mathcal{N}\left(0, \underbrace{\mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(X_j, \theta_0)\right]}_{\mathbb{I}_X(\theta_0)}\right),$$

kde člen $\mathbb{E}l'_1(\theta_0)$ je nulový, protože

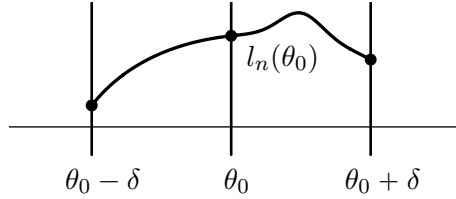
$$\mathbb{E}_{\theta_0}[\partial_{\theta} \ln f(x_1, \theta_0)] = \mathbb{E}\left[\frac{f'}{f}\right] = \int_{\mathbb{R}} \frac{f'}{f} f dx = \int_{\mathbb{R}} f' dx = \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f dx = 0.$$

□

Věta 2.33. Mějme $\mathcal{F}_{reg}^{\text{ML}}$, $\theta \in \mathbb{R}^1$. Pak s pravděpodobností $n \rightarrow +\infty$ 1 existuje konzistentní řešení LE_q .

Důkaz. Z věty 2.29 víme, že $\mathbb{P}(L(\theta_0, \mathbf{x}) > L(\theta, \mathbf{x})) \rightarrow 1$, $\forall \theta \neq \theta_0$. Zvolme libovolně $\delta > 0$. Pak dosazením $\theta = \theta_0 \pm \delta$ získáme vztah

$$\mathbb{P}(l_n(\theta_0) - l_n(\theta_0 \pm \delta) > 0) \rightarrow 1, \quad \forall \delta > 0.$$



Protože $l_n(\theta)$ je na intervalu $(\theta_0 - \delta, \theta_0 + \delta)$ diferencovatelná (tedy i spojitá), pak pro $\forall \delta$ s pravděpodobností $\mathbb{P}_{\theta_0} \rightarrow 1$ existuje $\hat{\theta}_n$ jako řešení $l'_n(\hat{\theta}_n) = 0$, které se nachází v intervalu $(\theta_0 - \delta, \theta_0 + \delta)$ a je tedy konzistentní pro θ_0 . □

PŘÍKLAD 2.34 (Původně ze SME). Mějme X_1, \dots, X_n iid $\mathcal{N}(\mu, \sigma^2)$, $\hat{\mu}_{\text{ML}} = ?$, $\hat{\sigma}_{\text{ML}}^2 = ?$

$$L(\mu, \sigma^2) = \prod_{i=1}^n f_{X_i}(x_i, \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2}.$$

$$l(\mu, \sigma^2) = \ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2.$$

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{j=1}^n (x_j - \mu) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{n} \sum_{j=1}^n x_j = \bar{x}_n,$$

$$\frac{\partial l}{\partial (\sigma^2)} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^n (x_j - \mu)^2 = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x}_n)^2.$$

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2}, \quad \frac{\partial^2 l}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{j=1}^n (x_j - \mu), \quad \frac{\partial^2 l}{\partial (\sigma^2)^2} = \frac{n}{2} \frac{1}{\sigma^4} - \frac{1}{\sigma^6} \sum_{j=1}^n (x_j - \mu)^2,$$

2 Metody pro hledání bodových odhadů

$$\mathbb{I}_{12} = \mathbb{I}_{21} = -\mathbb{E} \left[-\frac{1}{\sigma^4} \sum_{j=1}^n (X_j - \mu) \right] = \frac{1}{\sigma^4} \left(\sum_{j=1}^n \underbrace{(\mathbb{E} X_j - \mu)}_{=\mu} \right) = 0,$$

$$\mathbb{I}_{11} = \frac{n}{\sigma^2}, \quad \mathbb{I}_{22} = -\frac{n}{2} \frac{1}{\sigma^4} + \frac{1}{\sigma^6} \sum_{j=1}^n \underbrace{\mathbb{E} (X_j - \mu)^2}_{\sigma^2} = \frac{n}{2\sigma^4}.$$

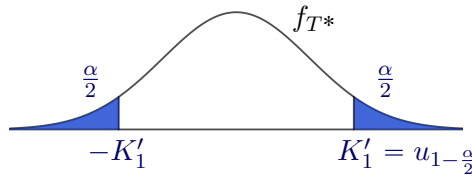
Fisherova informační matice má tedy tvar $\mathbb{I}_n = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} = n\mathbb{I}_1(\mu, \sigma^2)$.

Speciálně pro odhad jednorozměrného parametru μ získáme odhad $\hat{\mu}_{\text{ML}} = \overline{X}_n$, pro který plyne asymptotická normalita:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \underbrace{\sigma^2}_{\mathbb{I}_1^{-1}(\mu)}).$$

V případě náhodného výběru z normálního rozdělení se tedy jedná o přesné rozdělení \overline{X}_n (nejenom asymptotické), protože víme, že

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \overline{X}_n - \mu \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right), \quad \sqrt{n}(\overline{X}_n - \mu) \sim \mathcal{N}(0, \sigma^2).$$



3 Testování statistických hypotéz

3.1 Základní strategie TSH

Cílem testování statistických hypotéz je rozhodnout, zda jsme na základě nějakého experimentu schopni ověřit platnost určitého vysloveného tvrzení (hypotézy) na úrovni celé populace, případně parametru θ s touto populací spojeného. Můžeme tak například posoudit, jestli výsledky maturitních testů z matematiky závisí na pohlaví a na místě narození studentů. Experiment provádíme na jednotlivých jedincích, přičemž často máme k dispozici tzv. pokusnou a kontrolní skupinu. Příkladem toho může být například dvojité zaslepený experiment (*double-blind experiment*), pomocí něhož zkoumáme účinky daného léku. Zde máme dvě skupiny pacientů - první skupině podáváme lék, který chceme otestovat, a té druhé placebo, případně jiný medikament. Zároveň ale ani pacienti, ani lékaři, neví, které skupině jaký typ léku aplikujeme. Výsledky se pak zpracují právě užitím testování statistických hypotéz.

Mějme nějaký objekt/subjekt O a jeho stav $S \in \mathcal{S} = \mathcal{S}_0 + \mathcal{S}_1$ (disjunktní sjednocení $\mathcal{S}_0 \cup \mathcal{S}_1$). Testujeme hypotézu $H_0 : O$ je ve stavu z \mathcal{S}_0 oproti $H_1 : O$ je ve stavu z \mathcal{S}_1 . Máme-li náhodnou veličinu $X \sim f \in \mathcal{F}$ spojenou se stavy objektu O způsobem, že $\mathcal{S} \leftrightarrow \mathcal{F}$ jsou ve vzájemně jednoznačném vztahu, potom úlohu reformulujeme na

$$H_0 : X \sim f \in \mathcal{F}_0 \subset \mathcal{F} \quad \text{vs.} \quad H_1 : X \sim f \in \mathcal{F}_1 = \mathcal{F} \setminus \mathcal{F}_0.$$

Při případné parametrizaci modelu musíme dbát na identifikovatelnost rodiny \mathcal{F}

Definice 3.1. Mějme populaci Ω a na ní vlastnost $X \sim F$, kde $F \in \mathcal{F}$. Označme $\theta = \theta(F)$ parametr modelu, který nás zajímá, kde $\theta \in \Theta \subset \mathbb{R}^k$. Označme $H_0 : \theta \in \Theta_0$ jako **základní nulovou hypotézu** (*null hypothesis*) a $H_1 : \theta \in \Theta_1$, kde $\Theta_1 = \Theta \setminus \Theta_0$, jako **alternativní hypotézu**.

Nulová hypotéza H_0 může označovat "původní stav", tedy že zkoumaná věc se nezměnila, nebo že je lepší, než nějaký alternativní stav z hypotézy H_1 . Ta naproti tomu většinou doplňkově vyvrací platnost nulové hypotézy H_0 , např. že nový lék funguje lépe než starý, nebo že alternativní rozdělení odpovídá naměřeným datům více, než distribuce deklarovaná v H_0 . O zamítnutí H_0 , resp. přijetí, rozhodujeme na základě dostupného náhodného výběru $\mathbf{X} = (X_j)_{j=1}^n$ pořízeného v rámci zvoleného experimentu.

PŘÍKLAD 3.2. Hypotézou H_0 může být například to, že náhodný výběr pochází z normálního rozdělení, nebo že 2 náhodné výběry pochází ze stejného rozdělení, případně že mají alespoň stejnou střední hodnotu nebo rozptyl. Máme-li dvě výrobní metody a k nim příslušné náhodné výběry $X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ a $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$, můžeme zkoumat například to, jestli platí $H_0 : \mu_1 = \mu_2$ (výrobní metody jsou ve své střední hodnotě stejné), versus $H_1 : \mu_1 < \mu_2$ (nová metoda č.1 (X_j) je lepší, než původní metoda č.2 (Y_j)).

Definice 3.3. Definujeme R_{H_0} jako jev představující **zamítnutí** H_0 (*rejection*) a \bar{R}_{H_0} jako **přijetí** H_0 (*acceptation*). Pak **kritickou funkci testu** definujeme jako pravděpodobnost, že

3 Testování statistických hypotéz

zamítneme H_0 na základě naměřených dat \mathbf{x} , tzn.

$$\phi(\mathbf{x}) := \mathbb{P}(R_{H_0} | \mathbf{X} = \mathbf{x}) \in [0, 1], \quad \text{pro } \forall \mathbf{x} \in \mathcal{X},$$

kde \mathcal{X} je tzv. **výběrový prostor**, $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \exists \omega \in \Omega, \mathbf{X}(\omega) = \mathbf{x}\}$. Dále definujeme pro test $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, $\Theta = \Theta_0 + \Theta_1$, funkci

$$\begin{aligned} \beta_\phi(\theta) &:= \mathbb{E}_\theta[\phi(\mathbf{X})] = \mathbb{E}_\theta[\mathbb{P}(R_{H_0} | \mathbf{X} = \mathbf{x})] = \left| \begin{array}{l} \mathbb{P}(A) = \int_A 1 d\mathbb{P} = \int_\Omega \mathbb{I}_A d\mathbb{P} = \mathbb{E}[\mathbb{I}_A] \\ A = \{R_{H_0} | \mathbf{X} = \mathbf{x}\} \end{array} \right| = \\ &= \mathbb{E}_\theta[\mathbb{E}[\mathbb{I}_{R_{H_0}} | \mathbf{X} = \mathbf{x}]] \stackrel{\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}X}{=} \mathbb{E}_\theta[\mathbb{I}_{R_{H_0}}] = \mathbb{P}_\theta(R_{H_0}), \quad \text{pro } \forall \theta \in \Theta. \end{aligned}$$

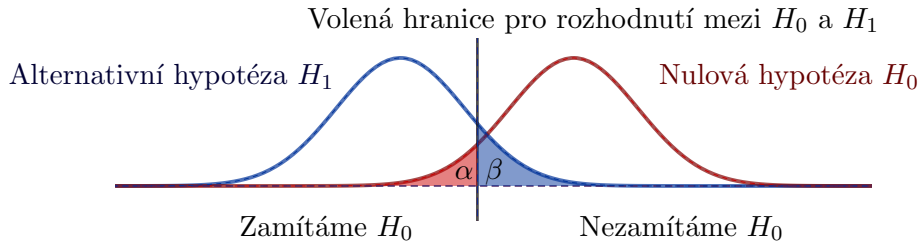
Funkce $\beta_\phi|_{\Theta_1}$, tedy zúžení β_ϕ z Θ na obor Θ_1 , se nazývá **silofunkce testu** ϕ .

Pokud testujeme hypotézu H_0 oproti H_1 , mohou nastat 4 navzájem se vylučující situace:

	Zamítáme H_0	Nezamítáme H_0
H_0 platí	Nastává chyba I. druhu	Správný výsledek
H_0 neplatí	Správný výsledek	Nastává chyba II. druhu

Chybu I. druhu považujeme za kritickou (tu horší) chybu. Pravděpodobnost chyby I. druhu vyjadřuje právě funkce $\beta_\phi|_{\Theta_0}$, což je zúžení β_ϕ z Θ na obor Θ_0 . Právě tuto pravděpodobnost budeme chtít mít pod kontrolou, tzn. pro vhodně malé zvolené číslo $\alpha \in (0, 1)$ požadujeme, aby celé zúžení $\beta_\phi|_{\Theta_0}$ bylo stejnoměrně pod zadanou hranicí α . Číslo α nazýváme **hladina významnosti** testu H_0 versus H_1 a požadujeme tedy, aby $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha$.

Pravděpodobnost chyby II. druhu vyjadřuje funkce $1 - \beta_\phi|_{\Theta_1}$ a budeme se ji snažit minimalizovat za vazební podmínky $\beta_\phi|_{\Theta_0} \leq \alpha$ na chybu I. druhu.



Obrázek 3.1: Porovnání hypotéz pro případ jednoduché H_0 (jeden stav) oproti jednoduché H_1 (druhý alternativní stav).

Shrňme strategii testování $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ v následující sekci.

3.2 UMP testy pro parametr $\theta = \theta(F)$

Definice 3.4 (UMP strategie testování H_0 vs. H_1). Hledáme takovou optimální kritickou funkci testu ϕ^* , aby při zvolené hladině významnosti $\alpha \in (0, 1)$ byla pravděpodobnost chyby II. druhu minimální, tzn. aby pro $\forall \theta \in \Theta_1$ bylo $\beta_{\phi^*}(\theta)$ stejnoměrně na Θ_1 **maximální silou**

3 Testování statistických hypotéz

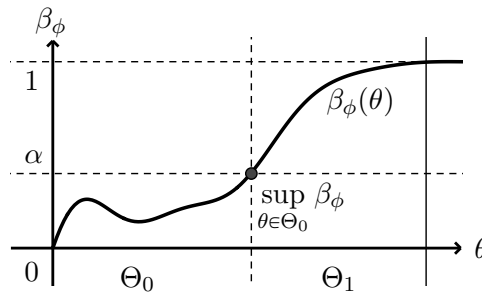
testu, za podmínky, že pravděpodobnost chyby I. druhu bude stále (stejněměrně na Θ_0) pod hranicí α , tzn.

$$\forall \theta \in \Theta_0, \beta_{\phi^*}(\theta) \leq \alpha.$$

Číslo $\sup_{\theta \in \Theta_0} \beta_{\phi^*}(\theta)$ se nazývá **hladina testu** (*size of test ϕ^**) a v praxi může být ostře pod nastavenou hranicí **hladiny významnosti** testu α (*significance level α*).

Konkrétní hodnotu $\beta_{\phi^*}(\theta)$ pro $\theta \in \Theta_1$ nazýváme **síla testu ϕ^*** pro dané $\theta \in \Theta_1$, celé zúžení $\beta_{\phi^*}|_{\Theta_1}$ pak nazýváme **silofunkce** testu ϕ^* .

Pokud takový test ϕ^* splňující uvedené podmínky existuje, nazýváme ho stejněměrně nejsilnějším testem H_0 oproti H_1 , ozn. **UMP test** (*uniformly most powerful test*). Situaci UMP ilustruje obrázek 3.2.



Obrázek 3.2: UMP model testování $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ na hladině α .

POZNÁMKA 3.5. Ještě lepší strategií by bylo hledat test ϕ , který minimalizuje stejněměrně oba druhy chyb I. a II. najednou. To však obecně nelze splnit, protože bohužel platí, že pokud se snažíme snížit pravděpodobnost chyby jednoho druhu, pak pravděpodobnost druhé chyby roste. Tedy chyby I. a II. druhu jsou komplementární, a proto musíme volit určitou formu kompromisu (viz. definice UMP testu ϕ^*).

Definice 3.6. Pokud Θ_0 je jednoprvková (1 stav), pak H_0 nazveme **jednoduchá** hypotéza (*simple*), v opačném případě je H_0 **složená** hypotéza (*composite*). Totéž platí pro Θ_1 a H_1 alternativu.

UMP je tedy test, který má nejvyšší statistickou sílu β na celém Θ_1 (H_1) mezi všemi možnými testy pod zadanou hranicí α pro chybu I. druhu na Θ_0 (H_0). Je taky dobré si uvědomit, že β je pravděpodobnost, že nenastává chyba II. druhu. Takovýto optimální test nutně nemusí existovat. Pokud však existuje, je možné ho pro speciální případ jednoduché H_0 a jednoduché H_1 nalézt pomocí Neyman-Pearsonova lemmatu, které bude uvedeno dále v sekci 3.3.

Definice 3.7. Pokud uvažujeme kritickou funkci ve tvaru

$$\phi(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in W \subset \mathbb{R}^n, \\ 0 & \text{jinak,} \end{cases}$$

pak W nazveme **kritický obor testu** (*critical region*). Je to tedy obor naměřených hodnot, při kterém zamítáme H_0 . Tomuto tvaru testu se říká **neznáhodněný test** a o přijetí H_0

3 Testování statistických hypotéz

rozhodujeme následovně:

$$\begin{aligned} \text{nastává jev } \{\mathbf{X} \in W\} &\Rightarrow \text{zamítáme } H_0, \\ \text{nastává opačný jev } \{\mathbf{X} \notin W\} &\Rightarrow \text{nezamítáme } H_0. \end{aligned}$$

Kritický obor W musí opět splňovat podmínku omezenosti pravděpodobnosti chyby I. druhu ve tvaru

$$\mathbb{P}_\theta((X_1, \dots, X_n) \in W) \leq \alpha, \quad \forall \theta \in \Theta_0,$$

pro zadanou hladinu významnosti testu $\alpha \in (0, 1)$. Příležitostně budeme pro kritický obor proto užívat označení W_α .

Pro stejnoměrně optimální UMP test ϕ^* pak odpovídající kritickou oblast značíme W^* a nazýváme ji **UMP kritickou oblastí testu** (*UMP critical region - UMPCR*). Pokud tedy $(x_1, \dots, x_n) \in W^*$, pak zamítneme H_0 . Opět příležitostně označíme UMP_α , ϕ_α^* , nebo W_α^* .

3.3 Neyman-Pearsonovo lemma (N-PL)

Nyní už přichází na řadu **Neyman-Pearsonovo lemma**, které nám umožní najít nejlepší možný test ϕ^* pro případ jednoduché hypotézy H_0 (1 stav) oproti jednoduché alternativě H_1 (také pouze 1 stav).

Věta 3.8 (Neyman-Pearsonovo lemma). *Mějme dvě hypotézy $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ a číslo $\alpha \in (0, 1)$ jako hladinu významnosti testu. Označme nyní hustotu pravděpodobnosti $f_0 := f(\mathbf{x}, \theta_0)$ a $f_1 := f(\mathbf{x}, \theta_1)$ obě vzhledem ke vhodné dominující míře λ . Pak existuje $K > 0$ a UMP test ϕ^* ve tvaru*

$$\phi^*(\mathbf{x}) = \begin{cases} 1 & f_1 > K f_0, \\ \gamma & f_1 = K f_0, \\ 0 & f_1 < K f_0, \end{cases} \quad \text{tak, že } \beta_{\phi^*}(\theta_0) = \alpha.$$

Pokud ϕ je nějaký jiný UMP test na hladině α , pak ϕ je nutně stejného tvaru jako ϕ^ na množině $\{f_1 \neq K f_0\}$. Výjimkou je situace, kdy existuje test ϕ s $\beta_\phi(\theta_1) = 1$, přičemž $\beta_\phi(\theta_0) < \alpha$, což znamená, že test ϕ nemůže dosáhnout zadané hranice α pro svou pravděpodobnost chyby I. druhu tak, jako ji dosahuje test ϕ^* .*

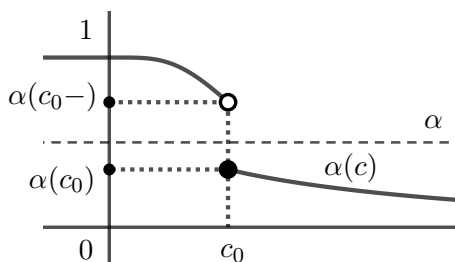
Důkaz. Konstruktivní důkaz (nutno znát ke zkoušce).

a) Nejprve zkonstruujeme nějaký test ϕ^* požadovaného tvaru. Definujeme

$$\begin{aligned} \alpha(c) &:= \mathbb{P}_{\theta_0}(f_1 > c f_0) = \mathbb{P}_{\theta_0}(f_1 > c f_0 \wedge f_0 > 0) = \mathbb{P}_{\theta_0}\left(\frac{f_1}{f_0} > c\right) = 1 - \mathbb{P}_{\theta_0}\left(\underbrace{\frac{f_1}{f_0}}_{Y \geq 0} \leq c\right) = \\ &= 1 - F_Y(c). \end{aligned}$$

Protože F_Y je nějaká distribuční funkce jisté náhodné veličiny Y , pak $\alpha(c)$ je nerostoucí, zprava spojitá a limitně se chová jako $\lim_{c \rightarrow -\infty} \alpha(c) = 1$, $\lim_{c \rightarrow +\infty} \alpha(c) = 0$.

3 Testování statistických hypotéz



Pro c_0 takové, že $\alpha(c_0-) \geq \alpha \geq \alpha(c_0)$, definujeme

$$\phi^*(\mathbf{x}) := \begin{cases} 1 & f_1 > c_0 f_0, \\ \gamma = \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} & f_1 = c_0 f_0, \\ 0 & f_1 < c_0 f_0, \end{cases}$$

kde $\alpha(c_0-)$ značí $\lim_{c \rightarrow c_0-} \alpha(c)$. V případě, že $\alpha(c)$ je spojitá v c_0 , pak číslo γ je nedefinováno ($\frac{0}{0}$). To však nevadí, protože v tomto případě platí

$$0 = \alpha(c_0-) - \alpha(c_0) = 1 - \alpha(c_0) - (1 - \alpha(c_0-)) = \underbrace{\mathbb{P}_{\theta_0}\left(\frac{f_1}{f_0} \leq c_0\right)}_{F_Y(c_0)} - \underbrace{\mathbb{P}_{\theta_0}\left(\frac{f_1}{f_0} < c_0\right)}_{F_Y(c_0-)} = \mathbb{P}_{\theta_0}\left(\frac{f_1}{f_0} = c_0\right),$$

a tedy vidíme, že množina $\{f_1 = c_0 f_0\}$ má nulovou pravděpodobnostní míru. To znamená, že test ϕ^* je definován jednoznačně s.j. \mathbb{P}_{θ_0} . Hladina tohoto testu ϕ^* je

$$\begin{aligned} \beta_{\phi^*}(\theta_0) &= \mathbb{E}_{\theta_0} \phi(\mathbf{X}) = 1 \cdot \mathbb{P}(\phi^* = 1) + \gamma \mathbb{P}(\phi^* = \gamma) + 0 \cdot \mathbb{P}(\phi^* = 0) = \\ &= \underbrace{\mathbb{P}(f_1 > c_0 f_0)}_{\alpha(c_0)} + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} \cdot \underbrace{\mathbb{P}(f_1 = c_0 f_0)}_{\alpha(c_0-) - \alpha(c_0)} = \alpha. \end{aligned}$$

- b) Nyní ukážeme, že zkonstruovaný test ϕ^* je UMP testem. Necht' ϕ^* je test z předchozí části důkazu a ϕ je libovolný jiný test na hladině významnosti α , tzn. $\beta_{\phi}(\theta_0) \leq \alpha$. Chceme ukázat, že $\beta_{\phi^*}(\theta_1) - \beta_{\phi}(\theta_1) \geq 0$.

$$\begin{aligned} \beta_{\phi^*}(\theta_1) - \beta_{\phi}(\theta_1) &= \mathbb{E}_{\theta_1}[\phi^*(\mathbf{X}) - \phi(\mathbf{X})] = \int_{\mathbb{R}^n} (\phi^* - \phi) f_1(\mathbf{x}) d\mathbf{x} = \\ &= \underbrace{\int_{S^+ := \{\phi^* - \phi > 0\}} (\phi^* - \phi) f_1 d\mathbf{x}}_{\mathbf{x} \in S^+ \Rightarrow \phi^* > 0 \Rightarrow f_1 \geq c_0 f_0} + \underbrace{\int_{S^- := \{\phi^* - \phi < 0\}} (\phi^* - \phi) f_1 d\mathbf{x}}_{\mathbf{x} \in S^- \Rightarrow \phi^* < 1 \Rightarrow f_1 \leq c_0 f_0} + \underbrace{\int_{\{\phi^* - \phi = 0\}} (\phi^* - \phi) f_1 d\mathbf{x}}_{=0} \geq \\ &\geq \int_{S^+} (\phi^* - \phi) c_0 f_0 d\mathbf{x} + \int_{S^-} (\phi^* - \phi) c_0 f_0 d\mathbf{x} = c_0 \underbrace{\int_{\mathbb{R}^n} (\phi^* - \phi) f_0 d\mathbf{x}}_{\mathbb{E}_{\theta_0}(\phi^* - \phi)} = c_0 (\underbrace{\beta_{\phi^*}(\theta_0)}_{=\alpha} - \underbrace{\beta_{\phi}(\theta_0)}_{\leq \alpha}) \geq 0. \end{aligned}$$

Z toho vyplývá, že síla testu ϕ^* je větší, než síla testu ϕ . Konstantu c_0 z důkazu ztotožňujeme s $K > 0$ ze znění věty.

□

Důsledek 3.9. Pokud platí, že $\mathbb{P}_{\theta_0}(f_1 = Kf_0) = 0$, pak můžeme psát

$$\phi^* = \begin{cases} 1 & x \in W^* = \{f_1 \geq Kf_0\}, \\ 0 & x \in (W^*)^c = \{f_1 < Kf_0\}. \end{cases}$$

Tedy v případě, že hranice $\{f_1 = Kf_0\}$ je nulové \mathbb{P}_{θ_0} míry, potom existuje neznáhodněný test s UMP kritickou oblastí W^* pro testování H_0 versus H_1 , přičemž pravděpodobnost chyby I. druhu je přímo rovna požadované signifikanci α , $\beta_{W^*}(\theta_0) = \alpha$, zatímco síla testu $\beta_{W^*}(\theta_1)$ je maximální možná.

PŘÍKLAD 3.10. Neyman-Pearsonovo lemma se zpravidla používá pro tzv. jednoduché testy, což znamená, že je testovaný parametr zadaný konkrétní jednou hodnotou. Příkladem může být třeba testování hypotéz pro parametry $\mathcal{N}(\mu, \sigma^2)$ typu

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1 \neq \mu_0 \text{ (resp. } \mu \geq \mu_0 \text{)}.$$

Pro složené testy, např. typu

$$H_0 : \sigma^2 \geq 7 \text{ vs. } H_1 : \sigma^2 < 7, \quad \text{nebo } H_0 : \mu \leq \mu_0 \text{ vs. } H_1 : \mu > \mu_0,$$

ani jiné podobně zadané testy, optimální UMP_α test nemusí obecně existovat a jsou nutné dodatečné podmínky (viz následující sekce).

3.4 Složené hypotézy a MLR systémy

Postup použití N-PL v praxi pro test z důsledku 3.9.

Hledáme takový test ϕ^* tvaru

$$\phi^*(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in \{f_1 \geq Kf_0\} = W^* \dots \text{UMP CR}, \\ 0 & \mathbf{x} \in \{f_1 < Kf_0\} = (W^*)^c, \end{cases}$$

pro který je dosažena hladina testu $\beta_{\phi^*}(\theta_0) = \alpha$, přičemž síla (sílofunkce) testu β je optimální.

- 1) Nejdříve najdeme **tvar** W^* jako řešení nerovnice $f_1 \geq Kf_0$. Získáme ho v nějakém tvaru $W^* = \{T(\mathbf{x}) \geq K'\}$, resp. $W^* = \{T(\mathbf{x}) \leq K'\}$, s blíže nespecifikovanou volnou konstantou K' , tedy

$$\{f_1 \geq Kf_0\} \sim \{T(\mathbf{X}) \geq K'\},$$

resp.

$$\{f_1 \geq Kf_0\} \sim \{T(\mathbf{X}) \leq K'\},$$

kde $T(\mathbf{X})$ se nazývá **testovací statistika**.

- 2) Konkrétní hodnotu K' pak určíme z rovnice $\mathbb{P}_{\theta_0}(T(\mathbf{X}) \geq K') = \alpha$, resp. $\mathbb{P}_{\theta_0}(T(\mathbf{X}) \leq K') = \alpha$. K vyřešení této nerovnosti však nutně potřebujeme umět vyjádřit $\mathbb{P}_{\theta_0}(T \geq K')$, resp. $\mathbb{P}_{\theta_0}(T \leq K')$ za předpokladu platnosti hypotézy H_0 , tzn. při θ_0 . Odvození rozdělení $T(\mathbf{X})$ při θ_0 se říká "distributional problem" testování hypotéz a jde o stěžejní část úspěšné aplikace.

3 Testování statistických hypotéz

- 3) V případě, že H_1 je složená, tzn. $H_1 : \theta \in \Theta_1$, postupujeme takto: volíme $\theta_1 \in \Theta_1$ libovolně pevně a testujeme hypotézu $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ na hladině α . Z Neyman-Pearsonova lemmatu existuje UMP_α test ϕ^* , případně UMPCR W^* . Pokud tento ϕ^* , případně W^* , nezávisí na volbě θ_1 , máme finální UMP_α test při složené alternativě H_1 .
- 4) Pokud i H_0 je složená, tzn. $H_0 : \theta \in \Theta_0$, pak, pokud to lze, ještě navíc ukážeme, že $\sup_{\theta \in \Theta_0} \beta_{\phi^*}(\theta) \leq \alpha$, tzn., že $\forall \theta'_0 \in \Theta$, platí, že $\beta_{\phi^*}(\theta'_0) \leq \alpha$. Průchodnost bodů 3) a 4) zajišťuje například následující koncept MLR.

Definice 3.11. Systém hustot \mathcal{F} se nazývá **MLR** (*Monotone likelihood ratio*), pokud $\exists T(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^1$ tak, že pro $\forall \theta_0 < \theta_1$ platí, že $\frac{f_1}{f_0}$ je monotonní funkcí statistiky $T(\mathbf{x})$, tzn. $\frac{f_1}{f_0} = g(T(\mathbf{x}))$, kde g je monotonní. Podíl $\frac{f_1}{f_0} = \frac{L(\theta_1)}{L(\theta_0)}$ se nazývá **věrohodnostním poměrem** (*likelihood ratio*), ozn. $\text{LR}(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$.

POZNÁMKA 3.12. Pokládáme $\frac{f_1}{f_0} = +\infty$, pokud $f_1 > 0$, $f_0 = 0$. Pro $f_1 = 0$ a $f_0 = 0$ výraz nedefinujeme. Monotonii vyžadujeme pouze tam, kde je výraz $\text{LR}(\mathbf{x})$ definován.

Věta 3.13. Mějme rostoucí MLR systém hustot \mathcal{F} se statistikou $T(\mathbf{X})$. Testujeme hypotézu $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$, $\theta \in \mathbb{R}^1$, tzv. jednostrannou hypotézu oproti jednostranné alternativě, na zadané hladině $\alpha \in (0, 1)$. Pak existuje UMP_α test ϕ^* ve tvaru

$$\phi^*(\mathbf{x}) = \begin{cases} 1 & T(\mathbf{x}) > K, \\ \gamma & T(\mathbf{x}) = K, \\ 0 & T(\mathbf{x}) < K, \end{cases}$$

přičemž K a γ jsou určeny podmínkou $\beta_{\phi^*}(\theta_0) = \alpha$, tzn. $\mathbb{E}_{\theta_0}[\phi^*(\mathbf{X})] = \alpha$, tedy

$$\mathbb{P}_{\theta_0}(T(\mathbf{X}) > K) + \gamma \mathbb{P}_{\theta_0}(T(\mathbf{X}) = K) + 0 \cdot \mathbb{P}_{\theta_0}(T(\mathbf{X}) < K) = \alpha.$$

Pro případ klesajícího MLR systému \mathcal{F} stačí v tvrzení zaměnit nerovnosti za opačné.

Důkaz. Důkaz provedeme speciálně pro ostře rostoucí MLR systém \mathcal{F} .

- a) Nejprve testujeme $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$. Pro tento účel zvolíme libovolně pevně $\theta_1 > \theta_0$ a testujeme $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$. Tím splníme předpoklady Neyman-Pearsonova lemmatu, podle něhož pak existuje UMP test

$$\phi^*(\mathbf{x}) = \begin{cases} 1 & f_1 > K' f_0, \\ \gamma & f_1 = K' f_0, \\ 0 & f_1 < K' f_0, \end{cases}$$

tak, že $\beta_{\phi^*}(\theta_0) = \alpha$. Nechť $\frac{f_1}{f_0} = g(T(\mathbf{x}))$, kde g je ostře rostoucí funkcí. Nyní upravíme podmínky do tvaru

$$\left\{ \begin{array}{l} f_1 > K' f_0 \\ f_1 = K' f_0 \\ f_1 < K' f_0 \end{array} \right\} \sim \left\{ \begin{array}{l} f_1/f_0 > K' \\ f_1/f_0 = K' \\ f_1/f_0 < K' \end{array} \right\} \sim \left\{ \begin{array}{l} T(\mathbf{x}) > g^{-1}(K') \\ T(\mathbf{x}) = g^{-1}(K') \\ T(\mathbf{x}) < g^{-1}(K') \end{array} \right\} \sim \left\{ \begin{array}{l} T(\mathbf{x}) > K \\ T(\mathbf{x}) = K \\ T(\mathbf{x}) < K \end{array} \right\},$$

kde $K := g^{-1}(K')$. Tvar testu ϕ^* je stejný nezávisle na volbě $\theta_1 > \theta_0$. Konstanty K a γ jsou pak určeny z rovnice $\beta_{\phi^*}(\theta_0) = \alpha$, a proto také nezávisí na volbě $\theta_1 > \theta_0$.

3 Testování statistických hypotéz

- b) Vezmeme právě zkonstruované ϕ^* a ukážeme, že pro $\forall \theta'_0 < \theta_0$ platí, že $\beta_{\phi^*}(\theta'_0) \leq \alpha$. Definujeme pomocný (čárkovaný) test

$$H'_0 : \theta = \theta'_0 \text{ vs. } H'_1 : \theta = \theta_0.$$

Příslušný UMP test H'_0 z N-PL má stejný tvar jako UMP ϕ^* zkonstruovaný v předchozím bodě, ale na hladině $\beta_{\phi^*}(\theta'_0) = \alpha'$. Jeho síla je pak rovna $\beta_{\phi^*}(\theta_0) = \beta'$.

- c) Ukážeme, že pro UMP test ϕ^* platí $\alpha' \leq \beta'$. Volme test $\phi(\mathbf{x}) = \alpha'$ pro $\forall \mathbf{x} \in \mathcal{X}$. Pak ϕ je test H'_0 vs. H'_1 na hladině α' se silou α' , která nemůže překročit sílu β' UMP testu ϕ^* , tzn. $\alpha' \leq \beta'$.

□

PŘÍKLAD 3.14. Testujeme hypotézu $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ pro exponenciální třídu hustot

$$\mathcal{F} = \{f(x, \theta) = c(\theta)h(x)e^{Q(\theta)T(x)} : \theta \in \Theta \subset \mathbb{R}^1\}.$$

Pokud $Q(\theta)$ je ryze rostoucí, resp. ryze klesající, pak $\mathcal{F}_n \stackrel{iid}{=} \left\{f(\mathbf{x}, \theta) = \prod_{i=1}^n f(x_i, \theta)\right\}$ je MLR systém hustot se statistikou $T(\mathbf{x}) = \sum_{i=1}^n T(x_i)$. Z věty 3.13 pak plyne konkrétní tvar UMP_α testu ϕ^* pro testování H_0 vs. H_1 . Díky této MLR exponenciální třídě hustot umíme najít stejnoměrně optimální UMP testy například pro následující parametry specifických rozdělení:

$$\begin{aligned} H_0 : p \leq p_0 \text{ vs. } H_1 : p > p_0, & \quad \text{v případě } X \sim \text{Bi}(n, p), \\ H_0 : \lambda \leq \lambda_0 \text{ vs. } H_1 : \lambda > \lambda_0, & \quad \text{v případě } X \sim \text{Po}(\lambda), \\ H_0 : \theta \leq \theta_0 \text{ vs. } H_1 : \theta > \theta_0, & \quad \text{pro případ } X \sim \text{Exp}(\theta), \\ H_0 : \mu \leq \mu_0 \text{ vs. } H_1 : \mu > \mu_0, & \quad \text{při modelu } X \sim \mathcal{N}(\mu, \sigma^2 \text{ známé}), \text{ atp.} \end{aligned}$$

3.5 Nestranné UMP testy (UMPU)

V aplikacích TSH v praxi vyvstává nutnost testovat další složitější hypotézy, jako například

$$\begin{aligned} H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0, & \quad \text{nebo} \\ H_0 : \theta_1 \leq \theta \leq \theta_2 \text{ vs. } H_1 : \theta \notin [\theta_1, \theta_2]. \end{aligned}$$

Pro takovéto případy, kdy alternativní hypotéza H_1 je tzv. oboustranná ($\theta < \theta_1$ nebo $\theta > \theta_2$), zpravidla neexistuje stejnoměrně nejsilnější UMP_α test ϕ^* na hladině α a jsme nuceni z optimality UMP testu slevit. Zavedeme proto UMPU testy.

Definice 3.15. Testujme $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$. Test ϕ se nazývá nestranný, pokud $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \inf_{\theta \in \Theta_1} \beta_\phi(\theta)$.

Věta 3.16. Každý UMP test ϕ^* je nestranný, tzn. $\beta_{\phi^*}|_{\Theta_0} \leq \beta_{\phi^*}|_{\Theta_1}$, tj.

$$\sup_{\theta \in \Theta_0} \beta_{\phi^*}(\theta) \leq \inf_{\theta \in \Theta_1} \beta_{\phi^*}(\theta).$$

3 Testování statistických hypotéz

Důkaz. Volme $\phi(\mathbf{x}) = \tilde{\alpha} = \sup_{\theta \in \Theta_0} \beta_{\phi^*}(\theta)$ pro $\forall \mathbf{x} \in \mathcal{X}$. Pak

$$\begin{aligned} \beta_{\phi}|_{\Theta_0} &= \mathbb{E}_{\theta_0} \phi(\mathbf{X}) = \tilde{\alpha}, \quad \forall \theta_0 \in \Theta_0, \quad (\text{hladina testu } \phi) \\ \beta_{\phi}|_{\Theta_1} &= \mathbb{E}_{\theta_1} \phi(\mathbf{X}) = \tilde{\alpha}, \quad \forall \theta_1 \in \Theta_1. \quad (\text{silofunkce testu } \phi) \end{aligned}$$

Podle předpokladu je ϕ^* UMP, a tedy jeho silofunkce je stejnoměrně na Θ_1 vyšší, než u všech ostatních testů včetně testu ϕ . Proto platí, že $\beta_{\phi^*}(\theta_1) \geq \tilde{\alpha}$ pro $\forall \theta_1 \in \Theta_1$. Věta 3.16 zobecňuje výsledek z bodu c) z důkazu věty 3.13. \square

Definice 3.17. Stejnoměrně nejsilnější test mezi všemi nestrannými testy se nazývá UMPU. (*UMP Unbiased*).

Věta 3.18. Testujeme hypotézu $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, kde $\theta \in \Theta \subset \mathbb{R}^1$ a θ_0 je vnitřním bodem Θ , tedy $\theta_0 \in \Theta^\circ$. Nechť \mathcal{F} je exponenciální třída hustot z příkladu 3.14 s $Q(\theta)$ ryze rostoucí a diferencovatelnou, tedy $\mathcal{F}_n = \left\{ f(\mathbf{x}, \theta) \stackrel{iid}{=} \prod_{i=1}^n f(x_i, \theta) \right\}$ je MLR systém se statistikou $T(\mathbf{x}) = \sum_{i=1}^n T(x_i)$. Pak existuje UMP_{U_α} test tvaru

$$\phi_u^*(\mathbf{x}) = \begin{cases} 1 & T(\mathbf{x}) < K_1 \vee T(\mathbf{x}) > K_2, \\ \gamma_1 & T(\mathbf{x}) = K_1, \\ \gamma_2 & T(\mathbf{x}) = K_2, \\ 0 & T(\mathbf{x}) \in (K_1, K_2), \end{cases}$$

takový, že $\beta_{\phi_u^*}(\theta_0) = \alpha$. Konstanty $K_1, K_2, \gamma_1, \gamma_2$ určíme tak, aby byla splněna podmínka

$$\mathbb{P}(T(\mathbf{X}) < K_1) + \mathbb{P}(T(\mathbf{X}) > K_2) + \gamma_1 \mathbb{P}(T(\mathbf{X}) = K_1) + \gamma_2 \mathbb{P}(T(\mathbf{X}) = K_2) = \alpha.$$

Důkaz. Bez důkazu. (ze zobecněného N-PL) \square

Na základě věty 3.18 umíme nalézt alespoň UMP_{U_α} optimální testy mezi všemi nestrannými testy, např. pro hypotézy ve tvaru

$$\begin{aligned} H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0, \quad & \text{při Gaussovském modelu } X \sim \mathcal{N}(\mu, \sigma^2 \text{ známé}), \\ H_0 : \sigma^2 = \sigma_0^2 \text{ vs. } H_1 : \sigma^2 \neq \sigma_0^2, \quad & \text{při Gaussovském modelu } X \sim \mathcal{N}(\mu \text{ známé}, \sigma^2). \end{aligned}$$

PŘÍKLAD 3.19 (Podrobněji viz MASC). Mějme $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, kde σ^2 známe. Testujeme hypotézu $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. Za účelem nalezení UMP_α volíme $\mu_1 \neq \mu_0$ a testujeme $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$. Z N-PL určíme tvar kritické oblasti

$$W^* = \{\mathbf{x} \in \mathbb{R}^n : f_1(\mathbf{x}) \geq K f_0(\mathbf{x})\} = \left\{ (\mu_1 - \mu_0) \sum_{j=1}^n x_j \geq K' \right\},$$

kde do K' byly zahrnuty všechny konstanty nezávislé na \mathbf{x} . Nyní, pokud $\mu_1 > \mu_0$, pak $W^* = \left\{ \sum_{j=1}^n x_j \geq K'' \right\}$, je-li $\mu_1 < \mu_0$, pak $W^* = \left\{ \sum_{j=1}^n x_j \leq K'' \right\}$. Protože se tvar W^* takto mění v závislosti na volbě $\mu_1 \geq \mu_0$, nelze najít stejnoměrně univerzální kritickou oblast W^* pro celý obor $H_1 : \mu \neq \mu_0$. Z toho vyplývá, že UMP_α test neexistuje.

3 Testování statistických hypotéz

Umíme však nalézt UMPU_α z věty 3.18, protože Gaussovská rodina $\mathcal{N}(\mu, \sigma^2 \text{ známé})$ je z exponenciální třídy hustot s příslušnými $Q(\mu) = \frac{\mu}{\sigma^2}$ a $T(\mathbf{X}) = \sum_{j=1}^n X_j$ (vzhledem k \mathcal{F}_n).

Rozdělení testovací statistiky $T(\mathbf{X})$ za platnosti $H_0 : \mu = \mu_0$ umíme vyřešit:

$T(\mathbf{X})|_{H_0} \sim \mathcal{N}(n\mu_0, n\sigma^2)$. Volíme-li ve větě 3.18 konstatny K_1 a K_2 symetricky, dostáváme

UMPU_α CR ve tvaru $W^* = \left\{ \frac{\sqrt{n}|\bar{X}_n - \mu_0|}{\sigma} \geq K'_1 \right\}$. Pro tuto novou testovací statistiku

$$T^*(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma} \sim \mathcal{N}(0, 1)$$

můžeme snadno určit $K'_1 = u_{1-\frac{\alpha}{2}}$ kvantil $\mathcal{N}(0, 1)$ tak, že platí

$$\beta_{W^*}(\mu_0) = \mathbb{P}(\text{chyby I. druhu}) = \mathbb{P}\left(\frac{\sqrt{n}|\bar{X}_n - \mu_0|}{\sigma} \geq u_{1-\frac{\alpha}{2}}\right) = \alpha.$$

4 Další metody testování hypotéz

PŘÍKLAD 4.1. Mějme $X \sim \text{Po}(\lambda)$, tedy $\mathbb{E}X = \lambda$, a testujme hypotézu $H_0 : \lambda = 1$ vs. $H_1 : \lambda = 10$.

Použijeme-li optimální test $\text{UMP}_{\alpha=0.05} : \phi^* = \begin{cases} 1 & x \geq 4, \\ 0.5058 & x = 3, \\ 0 & x \leq 2, \end{cases}$

dostaneme sílu tohoto UMP testu $\beta = \beta_{\phi^*}(10) = 1 - 0.0065$, tzn., že $\mathbb{P}(\text{chyby II. druhu}) = 0.0065$ je ještě o řád nižší, než $\mathbb{P}(\text{chyby I. druhu}) = \alpha = 0.05$, kterou považujeme za kritickou (vážnější) chybu. Kritickou chybu tak máme pod horší kontrolou, než nekritickou chybu.

Pokud v praxi použijeme **neoptimální** test $\phi = \begin{cases} 1 & x \geq 6, \\ 0.4516 & x = 5, \\ 0 & x \leq 4, \end{cases}$

pro který $\alpha = 0.0019$ se sílou testu $\beta = 0.95$, dostaneme test s lepší kontrolou kritické $\mathbb{P}(\text{chyby I. druhu}) = 0.0019$, při zachování rozumné velikosti $1 - \beta = 0.05$ pro pravděpodobnost nekritické chyby II. druhu.

Zabývejme se dále i dalšími potenciálně neoptimálními testy, jako je například LRT test.

4.1 Test poměrem věrohodností (LRT)

Definice 4.2. Mějme rodinu $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$ a testujme obecnou hypotézu $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ na zadané hladině významnosti $\alpha \in (0, 1)$. Zavedme funkci

$$\Lambda(\mathbf{x}) := \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)}, \quad \text{kde } L(\theta) = f(\mathbf{x}, \theta)$$

je věrohodnostní funkcí testovaného modelu, založenou na vzorku $\mathbf{x} \in \mathcal{X}$ z náhodného výběru $\mathbf{X} = (X_j)_{j=1}^n \text{ iid } f(x, \theta)$. Definujme test tvaru

$$\phi_\Lambda(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in W_\Lambda \subset \mathbb{R}^n, \\ 0 & \mathbf{x} \notin W_\Lambda, \end{cases}$$

kde $W_\Lambda = \{\mathbf{x} \in \mathbb{R}^n : \Lambda(\mathbf{x}) \leq K\}$ je taková, že pro nějakou konstantu $K \in [0, 1]$ platí $\beta_{\phi_\Lambda}|_{\Theta_0} \leq \alpha$, tzn.

$$\beta_{W_\Lambda}(\theta) := \beta_{\phi_\Lambda}(\theta) = \mathbb{E}_\theta \phi_\Lambda(\mathbf{X}) = \mathbb{P}_\theta(\phi_\Lambda(\mathbf{X}) = 1) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

Takové ϕ_Λ , pokud existuje, se nazývá **LRT test** pro testování $H_0 \times H_1$ na hladině významnosti α , ozn. LRT_α . W_Λ je odpovídající LRT kritická oblast tohoto testu (LRT_α CR). Pokud nastal jev $\{\mathbf{X} \in W_\Lambda\}$, zamítáme H_0 , pokud nastává opačný jev, pak H_0 nezamítáme.

4 Další metody testování hypotéz

Jde o test založený na limitních vlastnostech statistického modelu, přičemž smysluplnost zavedení tohoto LRT testů vyplývá z lemmatu, dokázaného v sekci 2 MLE odhadů, které říká, že pro $(X_j)_{j=1}^{+\infty} \text{ iid } f(x, \theta_0)$, kde $\text{supp } f$ je nezávislý na θ , platí, že

$$\mathbb{P}_{\theta_0}(L(\theta_0) > L(\theta)) \xrightarrow{n \rightarrow +\infty} 1, \quad \forall \theta \neq \theta_0.$$

a) Pokud H_0 platí, a tedy skutečná hodnota parametru θ_0 leží jak v Θ_0 , tak v $\Theta_0 \cup \Theta_1$, pak

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)} = 1 \text{ s pravděpodobností } \mathbb{P}_{\theta_0} \text{ jdoucí k } 1 \text{ při } n \rightarrow +\infty.$$

b) Pokud H_0 neplatí, a tedy skutečná hodnota parametru θ_0 neleží v Θ_0 , ale stále je obsažena ve $\Theta_0 \cup \Theta_1$, pak $\Lambda(\mathbf{x}) \leq K < 1$ je ostře odraženo od 1 s pravděpodobností \mathbb{P}_{θ_0} jdoucí k 1 při $n \rightarrow +\infty$. Právě tak jsme nastavili v definici ϕ_Λ kritický obor W_Λ pro přijetí/zamítnutí H_0 .

POZNÁMKA 4.3. LRT test nemusí být obecně optimální, je založen pouze na asymptotické vlastnosti a tedy pro konečné n nemusí dosahovat uspokojivých kvalit ve své síle testu β . Dokonce lze nalézt příklady LRT testů, které mají sílu testu β nižší, než zadaná hranice α pro chybu I. druhu, tzn. chyba II. druhu je velmi častá!

LRT test je v praxi často využíván pro svou obecnost ($\theta \in \mathbb{R}^k$), někdy za cenu složitější implementace při vyhodnocování funkce $\Lambda(\mathbf{x})$ nebo při řešení distribučního problému rozdělení testovací statistiky $T_\Lambda(\mathbf{X})$ za platnosti H_0 .

PŘÍKLAD 4.4 (jednovýběrový t-test: podrobněji v MASC). Mějme $X_1, \dots, X_n \text{ iid } \mathcal{N}(\mu, \sigma^2)$, kde σ^2 **neznáme**. Testujeme hypotézu $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ na hladině α pro nějaké vybrané $\mu_0 \in \mathbb{R}$ fixní. Pak zde máme

$$\theta = (\mu, \sigma^2), \quad \Theta_0 = \mu_0 \times \mathbb{R}^+, \quad \Theta = \Theta_0 \cup \Theta_1 = \mathbb{R} \times \mathbb{R}^+,$$

$$\Lambda(\mathbf{x}) = \frac{\sup L(\mu_0, \sigma^2) : \sigma > 0}{\sup L(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0},$$

kde $L(\mu, \sigma^2)$ je věrohodnostní funkce odpovídajícího systému $\mathcal{F}_n = \left\{ \prod_{j=1}^n f_{X_j} \right\}$. Z teorie ML odhadů vyplývá, že supréma L se nabývá právě v bodech maximálně věrohodných odhadů v čitateli i jmenovateli, tzn.

$$\hat{\sigma}_{0n}^2 = \underset{\sigma^2}{\operatorname{argmax}} L(\mu_0, \sigma^2) = \frac{1}{n} \sum_{j=1}^n (X_j - \mu_0)^2 \text{ a}$$

$$(\hat{\mu}_n, \hat{\sigma}_n^2) = \underset{(\mu, \sigma^2)}{\operatorname{argmax}} L(\mu, \sigma^2) = \left(\bar{X}_n, \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^2 \right).$$

Dosažením do $\Lambda(\mathbf{x})$ a upravením nerovnosti $\Lambda(\mathbf{x}) < K$ získáme LRT CR ve tvaru

$$W_\Lambda = \left\{ \mathbf{x} \in \mathbb{R}^n : \frac{\sqrt{n}|\bar{X}_n - \mu|}{s_n} \geq K' \right\}, \text{ kde } s_n = \frac{n}{n-1} \hat{\sigma}_n \text{ je výběrová směrodatná odchylka. Zave-}$$

dením LRT testovací statistiky $T_\Lambda(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \sim t(n-1)$ určíme $K' = t_{1-\frac{\alpha}{2}}(n-1)$ -kvantil Studentova rozdělení s $(n-1)$ stupni volnosti, aby bylo naplněno, že pro $\forall \sigma^2 > 0$

$$\beta_{W_\Lambda}(\mu_0, \sigma^2) = \mathbb{P}_{H_0}(\text{chyby I. druhu}) = \mathbb{P}_{H_0}(T_\Lambda(\mathbf{X}) \geq t_{1-\frac{\alpha}{2}}(n-1)) = \alpha.$$

Upozornění: v případě tohoto obecně neoptimálního LRT testu je velmi žádoucí výpočet, resp. aproximace, síly testu v okolí $H_{\mu_0}^+$ (pravé okolí bodu μ_0).

4.2 Analýza variance (ANOVA)

Analýza rozptylu (*analysis of variance*, ANOVA) je metoda, která umožňuje zjistit, jestli má na Gaussovskou náhodnou veličinu vliv některý ze znaků u jednotlivých jedinců, např. zda na plat zaměstnanců má vliv dosažené vzdělání, pohlaví, věk apod.

Mějme nezávislé náhodné výběry $X_{i1}, \dots, X_{in_i} \sim \mathcal{N}(\mu_i, \sigma^2)$, $i \in I$, $N = \sum_{i=1}^I n_i$. Potom sdružená hustota z \mathcal{F}_N je tvaru

$$f(\mathbf{x}|\mu_1, \dots, \mu_I, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^I \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 \right\}.$$

Testujeme hypotézu

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_I (= \mu) \quad \text{vs.} \quad H_1 : \text{alespoň jedna nerovnost}$$

na hladině $\alpha \in (0, 1)$ za dodatečného předpokladu $\sigma_1^2 = \dots = \sigma_I^2 = \sigma^2$ neznámé, tzn. předpokládáme homogenitu rozptylů jednotlivých testovaných podskupin $i \in I$.

Odvození ANOVA LRT $_{\alpha}$ testu

Mějme

$$\Lambda(\mathbf{x}) = \frac{\sup\{f(\mathbf{x}|\mu, \mu, \dots, \mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}}{\sup\{f(\mathbf{x}|\mu_1, \mu_2, \dots, \mu_I, \sigma^2) : \mu_i \in \mathbb{R}, \sigma^2 > 0\}}.$$

Při řešení extrémů prostřednictvím diferenciálního počtu $\partial_r f = 0$ v čitateli získáme 2 rovnice a ve jmenovateli $I + 1$ rovnic, které vyřešíme a příslušné hodnoty maximálně věrohodných odhadů $\hat{\mu}, \hat{\sigma}^2$, resp. $\hat{\mu}_i, \hat{\sigma}^2$, $i \in I$, zpětně dosadíme do $\Lambda(\mathbf{x})$. Dále potom nalezneme tvar LRT kritické oblasti $W_{\Lambda} = \{\mathbf{x} : \Lambda(\mathbf{x}) \leq K\}$, kdy platí

$$\Lambda(\mathbf{x}) \leq K \quad \Leftrightarrow \quad \mathcal{F}_{\Lambda}(\mathbf{x}) = \frac{(N - I)S_A}{(I - 1)S_e} \geq C,$$

kde

$$S_A = \sum_{i=1}^I n_i (\bar{x}_i - \bar{\bar{x}})^2, \quad S_e = \sum_{i=1}^I \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \quad \bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad \text{a} \quad \bar{\bar{x}} = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} x_{ij}.$$

PŘÍKLAD 4.5. Distribuční problém tohoto LRT testu ANOVA spočívá v odvození rozdělení testovací statistiky $\mathcal{F}_{\Lambda}(\mathbf{x})$ za předpokladu platnosti $H_0 : \mu_1 = \dots = \mu_I = \mu$. Postupujeme následovně:

$$\begin{aligned} \sum_{i=1}^I \sum_{j=1}^{n_i} X_{ij}^2 &= \sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i + \bar{X}_i)^2 = \sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + \underbrace{2 \sum_{i=1}^I \sum_{j=1}^{n_i} \bar{X}_i (X_{ij} - \bar{X}_i)}_0 + \sum_{i=1}^I n_i \bar{X}_i^2 = \\ &= \underbrace{\sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}_{Q_1 = S_e} + \sum_{i=1}^I n_i (\bar{X}_i - \bar{\bar{X}} + \bar{\bar{X}})^2 = \\ &= S_e + \underbrace{\sum_{i=1}^I n_i (\bar{X}_i - \bar{\bar{X}})^2}_{Q_2 = S_A} + \underbrace{2 \sum_{i=1}^I n_i \bar{\bar{X}} (\bar{X}_i - \bar{\bar{X}})}_0 + \underbrace{\sum_{i=1}^I \bar{\bar{X}}^2 n_i}_{N \cdot \bar{\bar{X}}^2 =: Q_3} = S_e + S_A + Q_3 = \sum_{i=1}^3 Q_i, \end{aligned}$$

4 Další metody testování hypotéz

což je součet tří kvadratických forem. Dá se ukázat (viz lineární algebra), že součet hodnotí těchto tří kvadratických forem dává plnou dimenzi úlohy N ,

$$\sum_{i=1}^3 h(Q_i) = h(S_e) + h(S_A) + h(Q_3) = (N - I) + (I - 1) + 1 = N.$$

Z Cochranovy věty pak vyplývá, že Q_i jsou **nezávislé** a $Q_i(\mathbf{X}) \sim \chi^2(h(Q_i))$, důsledkem čehož

$$\mathcal{F}_\Lambda(\mathbf{x})|_{H_0} = \frac{S_A/(I-1)}{S_e/(N-I)} \sim \frac{\chi^2(I-1)/(I-1)}{\chi^2(N-I)/(N-I)} \sim F(I-1, N-I),$$

tedy $\mathcal{F}_\Lambda(\mathbf{x})$ má za platnosti H_0 Fisherovo rozdělení s $(I-1)$ a $(N-I)$ stupni volnosti.

Nyní hledáme konstantu C tak, aby platilo $\mathbb{P}_{H_0}(\mathcal{F}_\Lambda(\mathbf{X}) \geq C) = \alpha$, což vede na LRT kritický obor $W_\Lambda = \{\mathbf{x} : \mathcal{F}_\Lambda(\mathbf{x}) \geq F_{1-\alpha}(I-1, N-I)\}$, kde $F_{1-\alpha}$ značí $(1-\alpha)$ -kvantil příslušného Fisherova rozdělení F . Skončí-li experiment v tomto kritickém oboru, pak zamítáme H_0 .

4.3 Odůvodnitelné testy (RT)

Doted' jsme při TSH postupovali podle následujícího schématu:

- stanovení principu testu (UMP, UMPU, LRT,...), nastavení $\alpha \in (0, 1)$,
- odvození $\phi = 1/\gamma/0$, tzn. nalezení W_α ve vhodném tvaru $\{T(\mathbf{X}) \geq K_\alpha\}$,
- řešení distribučního problému, tj. nalezení rozdělení testovací statistiky $T(\mathbf{X})|_{H_0} \sim F_T$ při platnosti H_0 ,
- určení K_α z podmínky $\mathbb{P}(T(\mathbf{X}) \geq K_\alpha) \leq \alpha$ při platnosti H_0 , tzn. $\beta_\phi|_{\Theta_0} \leq \alpha$,
- výpočet síly β nebo silofunkce testu $\beta_\phi|_{\Theta_1}$, tzn. $\mathbb{P}(T(\mathbf{X}) \geq K_\alpha)$ při platnosti H_1 .

Odůvodnitelné testy (*reasonable*, RT) **přímo** využívají nějakou vhodnou ("uhádnutou") statistiku $T(\mathbf{X})$, pro kterou lze nalézt její rozdělení $T(\mathbf{X})|_{H_0} \sim F_T$ takové, že F_T nezávisí na neznámých testovaných parametrech θ . Poté navrhne logicky vhodný (zdůvodnitelný) test ϕ_α , např. ve tvaru $W_\alpha = \{T(\mathbf{X}) \geq K_\alpha\}$, doladíme hodnotu K_α a nakonec spočteme silofunkci testu, což je v tomto případě velmi žádoucí, pokud to lze. Pokud ne, prověřujeme sílu odvozeného testu numerickou simulací Monte-Carlo nebo použijeme různé aproximace (např. pomocí CLT). Příklady RT testů si ukážeme v následující sekci!

4.4 Dvouvýběrové testy ($2 \times \mathcal{N}_1$)

Dvouvýběrový nepárový t-test porovnává střední hodnoty dvou Gaussovských výběrů. Příkladem toho může být třeba střední hodnota tlaku krve u kuřáků a nekuřáků, atp.

PŘÍKLAD 4.6 (Dvouvýběrový t-test). Uvažujme dva náhodné výběry ze dvou různých Gaussovských distribucí

$$\begin{aligned} X_1, \dots, X_{n_1} \text{ iid } \mathcal{N}(\mu_1, \sigma_1^2) &\Rightarrow \bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}, \\ Y_1, \dots, Y_{n_2} \text{ iid } \mathcal{N}(\mu_2, \sigma_2^2) &\Rightarrow \bar{Y}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right), \quad \frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}. \end{aligned}$$

4 Další metody testování hypotéz

Budeme testovat hypotézu shodnosti středních hodnot obou souborů $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ na hladině α . Rozlišujeme tři případy:

a) Známe-li σ_1^2, σ_2^2 , potom

$$\frac{\bar{X}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1),$$

protože z reprodukcí vlastnosti \mathcal{N} víme, že $(\bar{X}_1 - \bar{Y}_2) \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

Při $H_0 : \mu_1 = \mu_2$ pak nalezneme rozdělení testovací statistiky

$$U = U(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_1 - \bar{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

Vyřešením rovnice $\mathbb{P}(|U| \geq K_\alpha) = \alpha$ dostaneme $K_\alpha = u_{1-\frac{\alpha}{2}}$ s následným RT kritickým oborem $W_\alpha = \left\{(\mathbf{x}, \mathbf{y}) : |U(\mathbf{x}, \mathbf{y})| \geq u_{1-\frac{\alpha}{2}}\right\}$, kde $u_{1-\frac{\alpha}{2}}$ značí příslušný kvantil $\mathcal{N}(0, 1)$ rozdělení.

b) Pokud neznáme σ_1^2, σ_2^2 , ale víme, že $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (analogie ANOVA pro $I = 2$, kde σ^2 neznáme), pak volíme testovací statistiku jako

$$T = T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_1 - \bar{Y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \text{ při platnosti } H_0,$$

kde $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ se nazývá *pooled sample variance*. Studentovo $t(n_1 + n_2 - 2)$ rozdělení plyne z faktu, že $T = \frac{U}{s/\sigma}$, přičemž

$$(n_1 + n_2 - 2) \frac{s^2}{\sigma^2} = \left[\frac{(n_1 - 1)s_1^2}{\sigma_1^2} + \frac{(n_2 - 1)s_2^2}{\sigma_2^2} \right] \sim \chi^2(n_1 - 1) + \chi^2(n_2 - 1) \stackrel{id}{\sim} \chi^2(n_1 + n_2 - 2),$$

což plyne z reprodukcí vlastnosti χ^2 rozdělení. Podobně jako v a) dostáváme RT kritický obor $W_\alpha = \left\{|T(\mathbf{x}, \mathbf{y})| \geq t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)\right\}$, kde $t_{1-\frac{\alpha}{2}}$ značí kvantil příslušného $t(n_1 + n_2 - 2)$ Studentova rozdělení.

c) Pokud σ_1^2, σ_2^2 neznáme, ale víme, že $\sigma_1^2 \neq \sigma_2^2$, pak užíváme testovací statistiku

$$T_\nu = T_\nu(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\nu) \quad (\text{Welchova aproximace}),$$

kde

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}.$$

Následný kritický obor $W_\alpha = \left\{|T_\nu| \geq t_{1-\frac{\alpha}{2}}(\nu)\right\}$ definuje dvouvýběrový **t-test**. Pro ne-celočíselné ν interpolujeme $t(\nu)$ z hodnot sousedních $t([\nu])$ a $t([\nu] + 1)$.

4 Další metody testování hypotéz

PŘÍKLAD 4.7 (Test homogenity rozptylů = F-test). Za stejných předpokladů jako u dvouvýběrového t-testu z příkladu 4.6 testujeme hypotézu homogenity rozptylů dvou Gaussovských výběrů

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 \neq \sigma_2^2 \quad \text{na hladině } \alpha \in (0, 1).$$

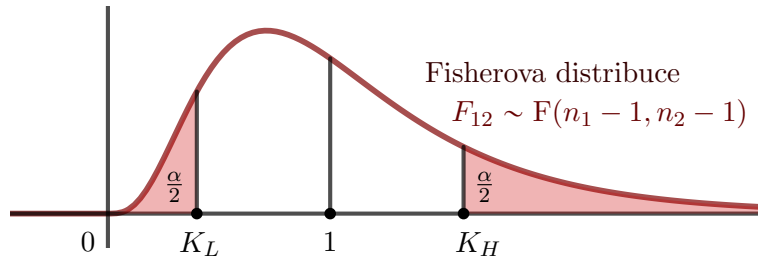
Testovací statistiku volíme

$$F_{12} = F(\mathbf{X}, \mathbf{Y}) = \frac{s_1^2}{s_2^2} = \left| H_0 : \sigma_1^2 = \sigma_2^2 \right| \stackrel{H_0}{=} \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}}{\frac{(n_2-1)s_2^2}{\sigma_2^2}} \cdot \frac{n_2-1}{n_1-1} \sim \frac{\frac{\chi^2(n_1-1)}{n_1-1}}{\frac{\chi^2(n_2-1)}{n_2-1}} \stackrel{id}{\sim} F(n_1-1, n_2-1),$$

za platnosti H_0 . Pak RT kritický obor F-testu je při symetrické volbě kvantilů Fisherova F rozdělení

$$W_\alpha = \left\{ (\mathbf{x}, \mathbf{y}) : F(\mathbf{x}, \mathbf{y}) \geq F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1) \text{ nebo } F(\mathbf{x}, \mathbf{y}) \leq F_{\frac{\alpha}{2}}(n_1-1, n_2-1) \right\},$$

viz obrázek 4.1. Tuto volbu odůvodňuje fakt, že $s_{1,2}^2 \xrightarrow{s.j.} \sigma_{1,2}^2$ a $\mathbb{E}s_{1,2}^2 = \sigma_{1,2}^2$.



Obrázek 4.1: Kritický obor F-testu homogenity rozptylů.

POZNÁMKA 4.8. V praxi lze použít testovací statistiku

$$\tilde{F}_{12} = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} \sim F(n' - 1, n'' - 1), \text{ kde } \begin{aligned} n' &= \max(n_1, n_2), \\ n'' &= \min(n_1, n_2), \end{aligned}$$

kterou pak porovnáváme pouze s horním $F_{1-\frac{\alpha}{2}}$ příslušným kvantilem Fisherova rozdělení.

4.5 Test koeficientu korelace (\mathcal{N}_2)

Předpokládejme $(X_j, Y_j)_{j=1}^n \text{ iid } \mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho)$ z dvourozměrného nede degenerovaného Gaussova rozdělení při $\sigma_1 > 0$, $\sigma_2 > 0$, $|\varrho| < 1$. Testujeme nekorelovanost X, Y , tzn. nulovou hodnotu korelačního koeficientu $\varrho = \varrho(X, Y)$:

$$H_0 : \varrho = 0 \text{ vs. } H_1 : \varrho \neq 0 \quad (\text{tzn. test nezávislosti } X \text{ a } Y \text{ v } \mathcal{N}_2 \text{ modelu})$$

na hladině významnosti $\alpha \in (0, 1)$.

Odvodíme LRT test H_0 :

4 Další metody testování hypotéz

Logaritmická věrohodnostní funkce modelu $\mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho)$ při označení $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho)$ je

$$l(\theta) = \ln L(\theta) = \frac{-1}{2(1-\varrho^2)} \left[\sum_{j=1}^n \frac{(X_j - \mu_1)^2}{\sigma_1^2} + \sum_{j=1}^n \frac{(Y_j - \mu_2)^2}{\sigma_2^2} - 2\varrho \sum_{j=1}^n \frac{(X_j - \mu_1)(Y_j - \mu_2)}{\sigma_1 \sigma_2} \right] - n \ln(2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}).$$

Následně vyhodnotíme výraz

$$\Lambda(\mathbf{x}, \mathbf{y}) = \frac{\sup\{L(\theta) : \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho = 0\}}{\sup\{L(\theta) : \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, |\varrho| < 1\}} \sim \frac{4 \text{ rovnice typu } \partial_r \ln L = 0}{5 \text{ rovnic typu } \partial_l \ln L = 0}.$$

Řešením obou extrémů jsou MLE odhady: $\hat{\mu}_1 = \bar{X}_n$, $\hat{\mu}_2 = \bar{Y}_n$, $\hat{\sigma}_1^2 = \hat{\sigma}_{n,X}^2$ a $\hat{\sigma}_2^2 = \hat{\sigma}_{n,Y}^2$ a dále

$$\hat{\varrho}_{XY} = \hat{\varrho}_n(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{j=1}^n (X_j - \bar{X}_n)(Y_j - \bar{Y}_n)}{\sqrt{\sum_{j=1}^n (X_j - \bar{X}_n)^2 \sum_{j=1}^n (Y_j - \bar{Y}_n)^2}} = \frac{\widehat{\text{Cov}}_{XY}}{\hat{\sigma}_1 \cdot \hat{\sigma}_2},$$

který se nazývá **Pearsonův výběrový koeficient korelace**. Dosazením těchto odhadů do $\Lambda(\mathbf{x}, \mathbf{y})$ získáme

$$\begin{aligned} \ln \Lambda(\mathbf{x}, \mathbf{y}) &= \ln L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \varrho = 0) - \ln L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\varrho}_{XY}) = \\ &= -n \ln(2\pi\hat{\sigma}_1\hat{\sigma}_2) + n \ln(2\pi\hat{\sigma}_1\hat{\sigma}_2\sqrt{1-\hat{\varrho}_{XY}^2}) = \\ &= \frac{n}{2} \ln(1-\hat{\varrho}_{XY}^2) \leq K \Leftrightarrow |\hat{\varrho}_{XY}| \geq K' \Leftrightarrow \frac{|\hat{\varrho}_{XY}|\sqrt{n-2}}{\sqrt{1-\hat{\varrho}_{XY}^2}} \geq K'', \end{aligned}$$

kde K, K', K'' jsou vhodné konstanty nezávislé na (\mathbf{x}, \mathbf{y}) . Lze ukázat, že při platnosti $H_0 : \varrho = 0$ má testovací LRT statistika rozdělení

$$T = T(\mathbf{X}, \mathbf{Y}) = \frac{\hat{\varrho}_{XY}\sqrt{n-2}}{\sqrt{1-\hat{\varrho}_{XY}^2}} \sim t(n-2),$$

tedy Studentovo rozdělení s $(n-2)$ stupni volnosti. To vede na kritický obor testu $H_0 : \varrho = 0$ ve tvaru

$$W_\alpha = \left\{ (\mathbf{x}, \mathbf{y}) : |T(\mathbf{x}, \mathbf{y})| \geq t_{1-\frac{\alpha}{2}}(n-2) \right\},$$

opět s použitím $(1 - \frac{\alpha}{2})$ -kvantilu příslušného Studentova rozdělení.

PŘÍKLAD 4.9. Mějme následující dvourozměrná pozorování (data) o rozsahu $n = 10$

x_j	94	98	127	88	85	95	111	75	102	82
y_j	2.1	1.9	3.5	1.5	3.2	1.6	1.9	2.5	2.6	1.9

pocházející z náhodného výběru $(X_j, Y_j)_{j=1}^n \stackrel{iid}{\sim} \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho)$. Pro test nekorelovanosti (zde i test nezávislosti) X a Y na hladině $\alpha = 0.05$ spočteme pro $n = 10$ následující hodnoty

$$\hat{\varrho}_{X,Y} = 0.3425, \quad T = \frac{\hat{\varrho}_{XY}\sqrt{n-2}}{\sqrt{1-\hat{\varrho}_{XY}^2}} = 1.031, \quad t_{1-\alpha/2}(n-2) = t_{0.975}(8) = 2.306.$$

4 Další metody testování hypotéz

Nyní tedy testujeme $W_\alpha : |T| \geq t_{0.975}(8)$? Nerovnost ale neplatí, a proto H_0 nezamítáme na hladině $\alpha = 0.05$. Pokud síla testu β (sílofunkce) je dostatečně vysoká, např. $\beta > \beta_1$ na intervalu $\varrho \in (-1, -\delta) \cup (\delta, 1)$, pak H_0 přijímáme, tzn. s danou signifikací $1 - \beta_1$ deklarujeme, že mezi Gaussovskými náhodnými veličinami X a Y není korelační vztah, tedy X a Y jsou nezávislé.

5 Asymptotické testy hypotéz

Doposud jsme pracovali převážně s formou testů statistických hypotéz, které využívaly odvozené (UMP, UMPU, LRT) nebo jinak odůvodněné (RT) testovací statistiky $T_n = T_n(\mathbf{X})$, pro které bylo možné odvodit konkrétní **přesné** rozdělení $T_n|_{H_0} \sim F_{T_n}$ za platnosti H_0 pro daný fixní rozsah n náhodného výběru $(X_j)_{j=1}^n$. K tomu byla vyžadována specifická znalost statistického modelu, např. Gaussovskost pro t-test, F-test, ANOVA,...

V mnoha případech však v praxi, po naměření nebo obdržení dat z nějakého komplikovanějšího experimentu, narážíme na dva problémy:

- Statistický model, ze kterého pochází naše data, není znám (technologie měření není dostupná) nebo je znám pouze přibližně na základě statistických testů shody dat s předpokládaným rozdělením. Tyto testy však opět fungují pouze na určité hladině spolehlivosti (signifikanci), jsou velmi často navíc založené pouze na limitních větách, a proto předpoklad statistického modelu pak může být zavádějící či dokonce chybný.
- Statistický model sice umíme více méně přesně odhalit, ale je nepříznivý v tom smyslu, že pro něj nedokážeme explicitně dovodit rozdělení vhodné testovací statistiky $T_n|_{H_0}$ pro dané fixní n . To často nastává, pokud data pochází z nestandardních distribucí nebo z více-komponentních distribučních směsí.

Obě tyto komplikace se dají překonat, pokud máme k dispozici střední či vyšší rozsahy n souborů naměřených dat. To umožňuje aproximovat rozdělení vhodné zvolené testovací statistiky $T_n|_{H_0}$ limitním rozdělením při $n \rightarrow +\infty$ ve smyslu (slabé) limity v distribuci, tzn. $T_n|_{H_0} \xrightarrow{\mathcal{D}} G$, kde limitní distribuční funkce G je nezávislá na neznámých parametrech modelu. Aby se toho dalo dosáhnout, je někdy potřeba nalézt navíc posloupnosti $(a_n)_{n=1}^{+\infty} \in \mathbb{R}$, $(b_n)_{n=1}^{+\infty} > 0$, pro které

$$T'_n(\mathbf{X}) = \frac{T_n(\mathbf{X}) - a_n}{b_n} \Big|_{H_0} \xrightarrow{\mathcal{D}} G \quad \left(\text{tzn. } T'_n \sim AG(a_n, b_n^2) \right).$$

Následně použijeme přibližné rozdělení G ke konstrukci tzv. **asymptotického testu** ϕ_α , resp. jeho příslušné kritické oblasti

$$W_\alpha = \{T_n(\mathbf{x}) \geq K_\alpha\} \stackrel{\text{resp.}}{=} \{T'_n(\mathbf{x}) \geq K'_\alpha\} \quad \text{tak, aby}$$

$$\lim_{n \rightarrow +\infty} \mathbb{P}(\text{chyby I. druhu}) = \lim_{n \rightarrow +\infty} \mathbb{P}(T_n(\mathbf{X}) \geq K_\alpha) \stackrel{\text{resp.}}{=} \lim_{n \rightarrow +\infty} \mathbb{P}(T'_n(\mathbf{X}) \geq K'_\alpha) = \alpha \quad \text{za platnosti } H_0.$$

K doladění konstanty K'_α opět použijeme vhodné typy kvantilů limitního rozdělení G , například $G_{1-\alpha}$, G_α , $G_{1-\frac{\alpha}{2}}$, podle povahy nerovností \geq charakterizující W_α . Dosažené signifikanci testu α skrze toto limitní rozdělení G pak říkáme **asymptotická hladina** (*size*) testu a test založený na takové W_α se nazývá **asymptotický (přibližný) test** hypotézy H_0 vs. H_1 .

5.1 Asymptotické testy středních hodnot iid \mathcal{L}_2

Věta 5.1 (Jednovýběrový asymptotický test $\mu = \mu_0$). *Mějme náhodný výběr X_1, \dots, X_n iid \mathcal{L}_2 pocházející z libovolného rozdělení s $\mathbb{E}X_j = \mu$ a s konečným rozptylem $\mathrm{D}X_j = \sigma^2 > 0$, který je neznámý. Testujeme hypotézu $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ (resp. $\mu \geq \mu_0$ apod.) na hladině $\alpha \in (0, 1)$. Pak testovací statistika*

$$T_n = T_n(\mathbf{X}) = \sqrt{n} \frac{\bar{X}_n - \mu_0}{s_n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

za platnosti H_0 . Následně test $H_0 : \mu = \mu_0$, založený na kritické oblasti $W_\alpha = \{|T_n(\mathbf{x})| \geq u_{1-\frac{\alpha}{2}}\}$, kde $u_{1-\frac{\alpha}{2}}$ značí kvantil $\mathcal{N}(0, 1)$, zamítá H_0 na **asymptotické hladině** α .

Důkaz. Plyne triviálně z asymptotických vlastností \bar{X}_n a s_n v \mathcal{L}_2 , viz kapitola 1. \square

POZNÁMKA 5.2. Test je asymptotický a tedy vyžaduje dostupnost dostatečně velkého počtu experimentálních dat $(x_j)_{j=1}^n$, avšak to je vyváženo tím, že statistický (apriorní) model pro tyto realizace může být zcela neznámého typu, splňující pouze předpoklad konečného $\sigma^2 > 0$.

Věta 5.3. *Nechť X a Y jsou nezávislé z \mathcal{L}_2 a mějme dva náhodné výběry (např. testovací a kontrolní) $(X_i)_{i=1}^{n_1}$ iid $\mathcal{L}_2(\mu_1, \sigma_1^2 > 0)$ a $(Y_j)_{j=1}^{n_2}$ iid $\mathcal{L}_2(\mu_2, \sigma_2^2 > 0)$. Pak*

$$T_{12} = T_{12}(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{při } n_1, n_2 \rightarrow +\infty.$$

Schéma důkazu. Zavedeme

$$U_n = \sum_{i=1}^{n_1} \underbrace{\frac{1}{n_1} \frac{X_i - \mu_1}{\sigma_{12}}}_{\xi_i} + \sum_{j=1}^{n_2} \underbrace{\frac{-1}{n_2} \frac{Y_j - \mu_2}{\sigma_{12}}}_{\eta_j} = \sum_{i=1}^{n_1} \xi_i + \sum_{j=1}^{n_2} \eta_j,$$

kde $\sigma_{12} := \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ a $n = n_1 + n_2$. Nemůžeme použít standardní Lindeberg-Lévyho CLT, protože v U_{12} nemáme součty stejně rozdělených náhodných veličin. Přímým výpočtem však ověříme, že pro $\forall i \in \widehat{n}_1, \forall j \in \widehat{n}_2$, platí

$\mathbb{E}\xi_i = 0, \mathbb{E}\eta_j = 0$ a $\mathrm{D}\xi_i < +\infty, \mathrm{D}\eta_j < +\infty$, přičemž $B_n^2 := \sum_{i=1}^{n_1} \mathrm{D}\xi_i + \sum_{j=1}^{n_2} \mathrm{D}\eta_j = 1$, kde jsme označili $n := n_1 + n_2$. Nyní budeme aplikovat obecnější CLT Lindeberg-Fellerův, tzn. je potřeba ověřit Lindebergovu podmínku $LP_n^\varepsilon \rightarrow 0, \forall \varepsilon > 0$ (viz 01MIP):

$$LP_n^\varepsilon = \sum_{i=1}^{n_1} \mathbb{E}[\xi_i^2 \mathbb{I}_{|\xi_i| > \varepsilon}] + \sum_{j=1}^{n_2} \mathbb{E}[\eta_j^2 \mathbb{I}_{|\eta_j| > \varepsilon}] \rightarrow 0, \quad \forall \varepsilon > 0.$$

Následně z CLT_{L-F} postupně dostáváme

$$\bar{U}_n = \frac{1}{n} U_n \sim \mathcal{N}\left(\bar{\mu}_n, \frac{\bar{\sigma}_n^2}{n}\right) = \mathcal{N}\left(0, \frac{1}{(n_1 + n_2)^2}\right), \quad \text{tzn. } U_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

5 Asymptotické testy hypotéz

Protože víme, že

$$s_1^2 \xrightarrow{s.j.} \sigma_1^2 \wedge s_2^2 \xrightarrow{s.j.} \sigma_2^2 \Rightarrow \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \xrightarrow{s.j.} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

dostaneme ze Slutskyho lemmatu finální výsledek $T_{12} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$. □

Důsledek 5.4 (Dvouvýběrový asymptotický test $\mu_1 = \mu_2$). *Díky větě 5.3 lze zkonstruovat asymptotický test pro testování hypotézy $H_0 : \mu_1 = \mu_2$ v obecném \mathcal{L}_2 modelu, kdy máme k dispozici dva nezávislé náhodné výběry ze dvou potenciálně zcela typově odlišných libovolných distribucí F_X a F_Y , o kterých víme pouze to, že obě distribuce mají konečné neznámé rozptyly $\sigma_1^2 > 0$, $\sigma_2^2 > 0$. Test $H_0 : \mu_1 = \mu_2$, založený na kritické oblasti*

$$W_\alpha = \left\{ |T_{12}(\mathbf{x}, \mathbf{y})| \geq u_{1-\frac{\alpha}{2}} \right\}, \quad \text{kde } T_{12}(\mathbf{x}, \mathbf{y}) = \frac{\bar{x}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

zamítá H_0 na asymptotické hladině α . Hranice zamítnutí $u_{1-\frac{\alpha}{2}}$ zde opět označuje příslušný kvantil Gaussova rozdělení $\mathcal{N}(0, 1)$.

5.2 Asymptotický LRT a Waldův test v \mathbb{R}^k

Věta 5.5. *Mějme $\theta \in \Theta \subset \mathbb{R}^k$ a testujeme hypotézu $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ na základě náhodného výběru X_1, \dots, X_n iid $f \in \mathcal{F}$. Nechť jsou dále splněny předpoklady z věty 2.33 o asymptotické normalitě MLE odhadů, tedy $\mathcal{F} = \mathcal{F}_{reg}^{ML}$, a nechť $\mathbb{I}(\theta)$ je spojitá ($k \times k$ Fisherova informační matice) v bodě θ_0 . Pak za platnosti H_0 platí*

$$\lambda_n(\mathbf{X}) = -2 \ln \Lambda(\mathbf{X}) \xrightarrow{\mathcal{D}} \chi^2(k).$$

Důkaz. Provedeme důkaz pro dimenzi $k = 1$. Předpokládejme, že $H_0 : \theta = \theta_0$ platí a že $\hat{\theta}_n = \hat{\theta}_{ML}$ je konzistentním řešením LE_q z věty 2.33 o AN MLE odhadů. Pak Taylorem dostaneme

$$\begin{aligned} \lambda_n(\mathbf{X}) &= -2 \ln \Lambda(\mathbf{X}) = -2 \ln \frac{\sup\{L(\theta) : \theta = \theta_0\}}{\sup\{L(\theta) : \theta \in \Theta\}} = 2(l(\hat{\theta}_n) - l(\theta_0)) = \\ &= 2l(\hat{\theta}_n) - 2l(\hat{\theta}_n) - 2(\theta_0 - \hat{\theta}_n) \underbrace{l'(\hat{\theta}_n)}_0 - (\theta_0 - \hat{\theta}_n)^2 l''(\hat{\theta}_n) + \frac{1}{3}(\hat{\theta}_n - \theta_0)^3 l'''(\theta_n^*), \end{aligned}$$

kde $\theta_n^* \in [\theta_0, \hat{\theta}_n]$. Předchozí vztah upravíme na

$$\lambda_n(\mathbf{X}) = \left\{ \sqrt{n}(\hat{\theta}_n - \theta_0) \left[-\frac{1}{n} l''(\hat{\theta}_n) \right]^{\frac{1}{2}} \right\}^2 + \frac{1}{3}(\hat{\theta}_n - \theta_0) [\sqrt{n}(\hat{\theta}_n - \theta_0)]^2 \cdot \frac{1}{n} l'''(\theta_n^*),$$

$$\begin{aligned} \text{kde } \sqrt{n}(\hat{\theta}_n - \theta_0) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{\mathbb{I}(\theta_0)}\right), & (\text{viz MLE } \hat{\theta}_n \text{ věta 2.33}), \\ -\frac{1}{n} l''(\hat{\theta}_n) &\xrightarrow{\mathbb{P}} \mathbb{I}(\theta_0) > 0, & (\text{ze spojitosti } \mathbb{I}(\theta) \text{ v } \theta_0), \\ (\hat{\theta}_n - \theta_0) &\xrightarrow{\mathbb{P}} 0, & (\text{z konzistence } \hat{\theta}_n), \\ [\sqrt{n}(\hat{\theta}_n - \theta_0)]^2 &\xrightarrow{\mathcal{D}} \left[\mathcal{N}\left(0, \frac{1}{\mathbb{I}(\theta_0)}\right) \right]^2, & (g(t) = t^2 \text{ spojitá}), \\ \left| \frac{1}{n} l'''(\theta_n^*) \right| &< M \text{ s pravděpodobností jdoucí k } 1 & (\text{stejně jako ve větě 2.33 o MLE}). \end{aligned}$$

5 Asymptotické testy hypotéz

Pak celkově ze Slutskyho lemma dostáváme

$$\lambda_n(\mathbf{X}) \xrightarrow{\mathcal{D}} [\mathcal{N}(0, 1)]^2 = \chi^2(1).$$

□

Důsledek 5.6 (Asymptotický LRT). *Za předpokladů věty 5.5 je test hypotézy $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, založený na kritické oblasti*

$$W_\alpha = \left\{ \mathbf{x} : \lambda_n(\mathbf{x}) \geq \chi_{1-\alpha}^2(k) \right\},$$

kde $\chi_{1-\alpha}^2(k)$ značí $(1 - \alpha)$ -kvantil příslušného $\chi^2(k)$ rozdělení, je tzv. **asymptotickým LRT testem**, který zamítá H_0 na asymptotické hladině α .

Věta 5.7 (Waldův test). *Nechť $\theta \in \Theta \subset \mathbb{R}^k$ a testujeme $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ na hladině $\alpha \in (0, 1)$. Definujeme **Waldovu testovací statistiku** předpisem*

$$W_{\theta_0}(\mathbf{X}) = n(\hat{\theta}_n - \theta_0)^T \mathbb{I}(\theta_0)(\hat{\theta}_n - \theta_0),$$

kde $\hat{\theta}_n$ je MLE. Pak při stejných předpokladech jako ve větě 5.5 platí, že $W_{\theta_0}(\mathbf{X}) \xrightarrow{\mathcal{D}} \chi^2(k)$ za platnosti H_0 a **Waldův test** H_0 , založený na CR

$$W_\alpha = \left\{ \mathbf{x} : W_{\theta_0}(\mathbf{x}) \geq \chi_{1-\alpha}^2(k) \right\}$$

zamítá H_0 na asymptotické hladině α .

Důkaz. Důkaz ponechán čtenáři (triviální). □

POZNÁMKA 5.8. Oba testy, asymptotický LRT i Waldův test jsou konzistentní, tzn. jejich síla/silofunkce splňuje podmínku, že

$$\beta_{W_\alpha}(\theta) \rightarrow 1 \text{ při } n \rightarrow +\infty, \quad \forall \theta \neq \theta_0.$$

5.3 Testy dobré shody (GoF)

Testy dobré shody jsou jedním z velmi užitečných nástrojů při identifikaci konkrétního typu statistického modelu, který stojí za naměřenými realizacemi sledované náhodné veličiny (vlastnosti) X . Shoda těchto dostupných dat $\mathbf{x} = (x_i)_{i=1}^n$ s apriorně předpokládaným modelem $F \in \mathcal{F}$ by měla být vždy ověřena ještě předtím než přistoupíme ke statistickým testům nad konkrétními parametry θ tohoto modelu. Tedy chceme-li po provedeném experimentálním měření nějaké fyzikální či ekonomické veličiny X , začít testovat např. hypotézu $H_0 : \mu = \mu_0$ prostřednictvím klasického t -testu, měli bychom nejdříve ověřit fakt, že experimentální data \mathbf{x} skutečně mohou pocházet z nějaké Gaussovske distribuce $\mathcal{N}(\mu, \sigma^2)$ s potenciálně neznámými parametry $\mu, \sigma^2 > 0$. Pokud se takové potvrzení předpokládaného funkčního tvaru statistického modelu nezdaří, je potřeba buď prověřit kvalitu naměřených dat (přeměřit/doměřit data, upravit experiment) nebo přejít k jinému vhodnějšímu statistickému modelu jiného typu využívající odlišný parametr $\tilde{\theta}$ co do jeho významu nebo dimenze. A znovu testovat shodu dat s tímto novým modelem.

Je patrné, že tato úloha prověření typu modelu je zcela zásadní a rozhoduje o kvalitě celého statistického šetření nebo predikce. Testy dobré shody jsou pouze jedním z mnoha takových detekčních nástrojů pro identifikaci relevantního statistického modelu, další možné přístupy viz 01MEX, 01SKE, a jiné.

5 Asymptotické testy hypotéz

Definice 5.9. Mějme náhodnou veličinu (vlastnost) X na $(\Omega, \mathcal{A}, \mathbb{P})$ a příslušný náhodný výběr X_1, \dots, X_n pocházející z nějaké neznámé distribuce F . Volme jednu konkrétní distribuci $F_0 \in \mathcal{F}$. Pak test hypotézy

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0 \quad (\text{resp. } F = F_1)$$

se nazývá **testem dobré shody** (*GoF - Goodness-of-Fit*) modelu F_0 . Opět volíme hladinu významnosti $\alpha \in (0, 1)$ a testujeme H_0 na této signifikantní hranici pro chybu I.druhu. Tedy, za kritickou chybu považujeme rozhodnutí o zamítnutí modelu F_0 , přestože ten je správný.

χ^2 -testy GoF

Za jeden z nejznámějších GoF testů lze považovat následující χ^2 -test, který převádí celou úlohu na specifický *asymptotický* test v Multinomickém parametrickém modelu za cenu jistého *binování* (diskretizace) dostupného náhodného výběru \mathbf{X} . Označme H_X obor hodnot náhodné veličiny $X \sim F$ a vytvořme dělení $\{A_1, \dots, A_k\}$ oboru hodnot H_X na k disjunktních boxů či tříd (binů). Dále zavedeme

$$p_j = \mathbb{P}_F(X \in A_j), \quad p_{0j} = \mathbb{P}_{F_0}(X \in A_j), \quad \forall j \in \hat{k}.$$

Mějme nyní k dispozici náhodný výběr $\mathbf{X} = (X_i)_{i=1}^n \text{ iid } F$ a nechť

$$Y_j = \# \{i : X_i \in A_j\} = \sum_{i=1}^n \mathbb{I}\{X_i \in A_j\} = \sum_{i=1}^n \mathbb{I}_{A_j}(X_i)$$

je počet těch pozorování X_i z $\{X_1, \dots, X_n\}$, která se vyskytují v j -tém binu A_j , $j \in \hat{k}$. Vzhledem k *iid* předpokladu pro jednotlivá X_i pak plyne, že každé $Y_j \sim \text{Bi}(n, p_j)$, a proto i celý vektor $\mathbf{Y} = (Y_1, \dots, Y_k)$ má Multinomické rozdělení rozdělení $\mathbf{Y} \sim \text{Mult}(n, \mathbf{p})$, při označení $\mathbf{p} = (p_1, \dots, p_k)$. Namísto testu $H_0 : F = F_0$ vs. $H_1 : F \neq F_0$ pak testujeme parametrickou hypotézu na hladině $\alpha > 0$

$$H_0 : \mathbf{p} = \mathbf{p}_0 \quad \text{vs.} \quad H_1 : \mathbf{p} \neq \mathbf{p}_0 \quad \text{kde } \mathbf{p}_0 = (p_{01}, \dots, p_{0k})$$

v Multinomickém modelu $\mathbf{Y} \sim \text{Mult}(n, \mathbf{p})$, přičemž $\sum_1^k p_j = \sum_1^k p_{0j} = 1$.

Věta 5.10. Nechť $\hat{p}_j = Y_j/n$ značí MLE odhady parametrů p_j v binomickém modelu $\text{Bi}(n, p_j)$, $\forall j \in \hat{k}$. Následující tři testovací statistiky v $\text{Mult}(n, \mathbf{p})$ modelu

$$\begin{aligned} \chi^2(\mathbf{Y}) &= \sum_{j=1}^k \frac{n(\hat{p}_j - p_{0j})^2}{p_{0j}} = \sum_{j=1}^k \frac{(Y_j - np_{0j})^2}{np_{0j}} && (\text{Pearsonova}), \\ \tilde{\chi}^2(\mathbf{Y}) &= \sum_{j=1}^k \frac{n(\hat{p}_j - p_{0j})^2}{\hat{p}_j} = \sum_{j=1}^k \frac{(Y_j - np_{0j})^2}{Y_j} && (\text{Neymanova}), \\ \lambda_n(\mathbf{Y}) &= -2 \ln \Lambda(\mathbf{Y}) = -2 \ln \prod_{j=1}^k \left(\frac{p_{0j}}{\hat{p}_j} \right)^{n\hat{p}_j} = -2 \sum_{j=1}^k Y_j \ln \left(\frac{np_{0j}}{Y_j} \right) && (LRT \text{ stat.}), \end{aligned}$$

jsou za platnosti $H_0 : \mathbf{p} = \mathbf{p}_0$ asymptoticky ekvivalentní a všechny konvergují v distribuci k limitnímu $\chi^2(k-1)$ rozdělení.

5 Asymptotické testy hypotéz

Důkaz. Protože platí $\sum_1^k p_j = \sum_1^k p_{0j} = 1$, je skutečná dimenze testovaného parametrického prostoru $\Theta = \{(p_1, \dots, p_k) : p_j \in (0, 1)\}$ rovna $k - 1$. Pak z věty 5.5 o asymptotickém LRT testu víme, že za platnosti H_0 platí $\lambda_n(\mathbf{Y}) \xrightarrow{\mathcal{D}} \chi^2(k - 1)$, protože jsou naplněny předpoklady na ML-regularitu $\text{Mult}(n, \mathbf{p})$ systému. Dále lze ukázat pomocí Taylorova rozvoje funkce $\ln(1 + u)$ a s využitím asymptotických vlastností MLE odhadů \hat{p}_j , že $\lambda_n(\mathbf{Y}) = \chi^2(\mathbf{Y}) + o_p(1)$. Pak ze Slutskyho lemma mají $\lambda_n(\mathbf{Y})$ a $\chi^2(\mathbf{Y})$ shodné asymptotické rozdělení. Podobně pro $\tilde{\chi}^2(\mathbf{Y})$. \square

Důsledek 5.11 (Pearson χ^2 GoF). *Test $H_0 : \mathbf{p} = \mathbf{p}_0$ založený na kritické oblasti*

$$W_\alpha = \left\{ \mathbf{y} : \chi^2(\mathbf{y}) \geq \chi_{1-\alpha}^2(k - 1) \right\} \quad (\chi_{1-\alpha}^2 \text{ značí kvantil } \chi^2 \text{ rozdělení})$$

je asymptotickým Pearsonovým χ^2 -testem dobré shody dosahujícím asymptotické significance α . Podobně pro $\tilde{\chi}^2(\mathbf{Y})$ a $\lambda_n(\mathbf{Y})$ testovací statistiky.

Jde o asymptotický test, tedy je vyžadován jednak dostatečný počet n pozorování $(x_i)_1^n$, obvykle alespoň $n \geq 50$, a současně by měla být splněna podmínka $np_{0j} \geq 5$, $\forall j \in \hat{k}$. Hodnoty testovacích statistik $\chi^2(\mathbf{y})$ a $\tilde{\chi}^2(\mathbf{y})$ představují součet vážených kvadratických odchylek tzv. pozorovaných četností y_j od teoretických četností np_{0j} přes všechny biny A_j , $j \in \hat{k}$.

Kolmogorov-Smirnovův GoF test

Test dobré shody dat s modelem $H_0 : F = F_0$ vs. $H_1 : F \neq F_0$ na hladině α lze také provést čistě neparametrickým způsobem s využitím empirické distribuční funkce

$$F_n(t) = F_n(t, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, t]}(X_i),$$

jejíž statistické asymptotické vlastnosti byly uvedeny v kapitole 2.

Věta 5.12. *Nechť $\mathbf{X} = (X_i)_1^n$ jsou iid F a F_n značí příslušnou empirickou distribuční funkci. Definujeme **Kolmogorov-Smirnovovu** statistiku*

$$D_n(F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| = K(F_n, F) \quad (\text{Kolmogorovova vzdálenost}).$$

Pak rozdělení $D_n(F)$ nezávisí na modelu F , tzn. $D_n(F)$ je PQ pivotem pro F . Dále $D_n(F) \xrightarrow{s.j.} 0$ a asymptotické rozdělení $\sqrt{n}D_n(F)$ je dáno vztahem

$$\lim_{n \rightarrow +\infty} \mathbb{P}(\sqrt{n}D_n(F) \leq t) = 1 - 2 \sum_{m=1}^{\infty} (-1)^{m-1} e^{-2m^2 t^2} =: G^{\text{KS}}(t), \quad \forall t > 0.$$

Důkaz. Konvergence $D_n(F) \xrightarrow{s.j.} 0$ byla ukázána již v kap.2 (Glivenko-Cantelli lemma). Asymptotické rozdělení G^{KS} ukázal Kolmogorov (1933). Za DCv zbývá pouze prověřit fakt, že $D_n(F)$ je pivotální veličinou. \square

Důsledek 5.13 (K-S GoF test). *Test $H_0 : F = F_0$ založený na kritické oblasti*

$$W_\alpha = \left\{ \mathbf{x} : D_n(F_0) \geq K_\alpha \right\} = \left\{ \mathbf{x} : \sqrt{n}D_n(F_0) \geq K'_\alpha \right\},$$

kde $K'_\alpha = G_{1-\alpha}^{\text{KS}}$ je $(1 - \alpha)$ -kvantil asymptotického Kolmogorov-Smirnovova rozdělení G^{KS} , je testem H_0 na asymptotické hladině α . Tento test se nazývá Kolmogorov-Smirnovův (K-S) neparametrický test dobré shody.

5 Asymptotické testy hypotéz

Příslušné hodnoty kvantilů $G_{1-\alpha}^{\text{KS}}$ limitního rozdělení $\sqrt{n}D_n(F_0)$, resp. $D_n(F_0)$, jsou tabulovány. Pro jejich přibližný výpočet lze užít aproximaci $G^{\text{KS}}(t)$ konečnou řadou. Použijeme-li takto pouze první člen řady pro $m = 1$, tzn. $G^{\text{KS}}(t) \doteq 1 - 2e^{-2t^2}$, dostaneme jednoduchý předpis pro přibližnou hodnotu $(1-\alpha)$ -kvantilu K-S rozdělení $K'_\alpha = G_{1-\alpha}^{\text{KS}} \doteq \frac{\sqrt{2}}{2} [\ln 2 - \ln \alpha]^{1/2}$.

5.4 Modifikace χ^2 -testů dobré shody

Pokud potřebujeme testovat shodu dat s modelem ve složené podobě, tzn.

$$H_0 : F = F_\theta \quad \text{vs.} \quad H_1 : F \neq F_\theta \quad \text{na hladině } \alpha \in (0, 1),$$

kde $\theta \in \Theta \subset \mathbb{R}^s$ je neznámý parametr takový, že $\dim(\Theta) = s$, pak postup testování upravíme následovně. Označíme opět

$$p_j = \mathbb{P}_F(X \in A_j), \quad p_{0j} = p_{0j}(\theta) = \mathbb{P}_{F_\theta}(X \in A_j), \quad \forall j \in \hat{k},$$

kde nyní $\mathbf{p} = (p_1, \dots, p_k)$ a $\mathbf{p}_0 = \mathbf{p}_0(\theta) = (p_{01}(\theta), \dots, p_{0k}(\theta))$. Na základě binovaných náhodných veličin $(Y_j)_{j=1}^k$ testujeme v $\text{Mult}(n, \mathbf{p})$ modelu parametrickou hypotézu

$$H_0 : \mathbf{p} = \mathbf{p}_0(\theta) \quad \text{vs.} \quad H_1 : \mathbf{p} \neq \mathbf{p}_0(\theta) \quad \text{na hladině } \alpha > 0.$$

Protože však nyní vektor \mathbf{p}_0 je funkcí neznámého parametru θ , musíme tento parametr odhadnout za platnosti H_0 . Takový vhodný odhad $\hat{\theta}_n = \hat{\theta}_n(\mathbf{Y})$ by měl disponovat dostatečně dobrými asymptotickými vlastnostmi při $n \rightarrow +\infty$. Dostupnou realizaci odhadu $\hat{\theta}_n(\mathbf{y})$ pak dosadíme do funkce $\mathbf{p}_0(\theta)$ a testujeme již jednoduchou bodovou hypotézu v Multinomickém modelu na hladině α

$$H_0 : \mathbf{p} = \hat{\mathbf{p}}_0 \quad \text{vs.} \quad H_1 : \mathbf{p} \neq \hat{\mathbf{p}}_0, \quad \text{kde } \hat{\mathbf{p}}_0 = \hat{\mathbf{p}}_0(\hat{\theta}_n(\mathbf{y})).$$

Toto zanesení odhadu $\hat{\theta}_n$ do testované hypotézy H_0 však modifikuje asymptotické vlastnosti použitých testovacích statistik.

Věta 5.14. *Mějme test $H_0 : \mathbf{p} = \hat{\mathbf{p}}_0$ vs. $H_1 : \mathbf{p} \neq \hat{\mathbf{p}}_0$ za výše uvedených podmínek, kdy $\dim(\Theta) = s$. Nechť $\hat{\theta}_n$ značí maximálně věrohodný odhad $\hat{\theta}_{\text{ML}}(\mathbf{Y})$ za platnosti H_0 . Pak testovací statistika*

$$\chi^2(\mathbf{Y}) = \sum_{j=1}^k \frac{(Y_j - n\hat{p}_{0j})^2}{n\hat{p}_{0j}} \xrightarrow{\mathcal{D}} \chi^2(k - s - 1)$$

a test založený na kritické oblasti

$$W_\alpha = \left\{ \mathbf{y} : \chi^2(\mathbf{y}) \geq \chi_{1-\alpha}^2(k - s - 1) \right\} \quad (\chi_{1-\alpha}^2 \text{ značí kvantil } \chi^2 \text{ rozdělení})$$

je Pearsonovým χ^2 -testem dobré shody dosahujícím asymptotické signifikance α . Podobně pro testovací statistiky $\tilde{\chi}^2(\mathbf{Y})$ a $\lambda_n(\mathbf{Y})$.

Důkaz. Podobně jako u věty 5.10 s využitím asymptotiky (zobecněného) LRT testu (nedělali jsme). □

5 Asymptotické testy hypotéz

POZNÁMKA 5.15. Takto upraveným χ^2 -testem dobré shody tedy umíme otestovat například hypotézu H_0 , zda naše data \mathbf{y} mohou pocházet z nějakého libovolného rozdělení ze systému Gaussovských distribucí, tzn. hypotézu

$$H_0 : \mathbf{F} \in \mathcal{F} = \{\mathcal{N}(\mu, \sigma^2) : \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}.$$

Avšak pozor, dosazované MLE odhady $\hat{\mu}_n, \hat{\sigma}_n^2$ nejsou běžné MLE odhady parametrů μ, σ^2 vytvořené na základě původně naměřených dat $(x_i)_1^n$, ale na základě výhradně binovaných dat $(y_j)_1^k$ podléhajících Multinomickému modelu $\text{Mult}(n, \mathbf{p}_0(\mu, \sigma^2))$. Obecně nesprávný, ale přesto užívaný postup testování normality dat $H_0 : \mathbf{F} = \mathcal{N}(\mu, \sigma^2)$ je ten, kdy na základě původních dat $(x_i)_1^n$ jsou spočteny standardní MLE odhady $\hat{\mu} = \bar{x}_n$ a $s_n^2 = \sum_1^n (x_i - \bar{x}_n)^2 / (n - 1)$, poté je náhodný výběr $(X_i)_1^n$ tzv. 'standardizován' na $U_i = (X_i - \bar{x}_n) / s_n$ a nakonec je testována hypotéza $H_0 : \mathbf{F}_U = \mathcal{N}(0, 1)$ na hladině α pomocí běžné Pearsonovy testovací statistiky aplikované na 'standardizovaná' data $(u_i)_1^n$.

6 Konfidenční množiny, Intervaly spolehlivosti

Bodové odhady $\hat{\theta}_n(\mathbf{X}) = \hat{\theta}_n$, resp. $T_n(\mathbf{X}) = T_n$, sice poskytují odhad hodnoty daného parametru, nicméně například ve spojitém statistickém modelu (ASR_λ) vždy platí, že $\mathbb{P}(\hat{\theta}_n(\mathbf{X}) = \theta_0) = 0$, $\forall \theta_0$, resp. $\mathbb{P}(T_n(\mathbf{X}) = \tau(\theta)) = 0$, $\forall \theta \in \Theta$. Z tohoto důvodu se zavádí tzv. interval spolehlivosti, což je interval, ve kterém se neznámý parametr θ nachází s námi zadanou pravděpodobností (například 95%). Přesnost takového odhadu je pak dán šířkou tohoto intervalu.

Definice 6.1. Necht' $\mathbf{X} = (X_1, \dots, X_n) \sim \mathbb{P} \in \mathcal{P}$ na (Ω, \mathcal{A}) , označme $\theta = \theta(\mathbb{P}) \in \Theta \subset \mathbb{R}^1$ funkcionál na \mathcal{P} , resp. $\theta \in \Theta \subset \mathbb{R}^k$. Označíme \mathcal{B}_Θ systém borelovských podmnožin Θ a zvolíme číslo $\alpha \in (0, 1)$. Pak $C(\mathbf{X}) \in \mathcal{B}_\Theta$ se nazývá **konfidenční množina** ($\text{CM}_{1-\alpha}$) pro θ na hladině $(1 - \alpha)$, pokud

$$\inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\theta \in C(\mathbf{X})) \geq 1 - \alpha.$$

Číslo $\inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\theta \in C(\mathbf{X}))$ se pak nazývá **konfidenční koeficient** (koeficient spolehlivosti). Pokud speciálně $C(\mathbf{X}) = [\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})] \subset \Theta$, nazveme ji **konfidenční interval** ($\text{CI}_{1-\alpha}$) (interval spolehlivosti) na hladině $1 - \alpha$.

POZNÁMKA 6.2. $C(\mathbf{X})$ je tedy "náhodná" množina (založená na náhodném výběru \mathbf{X}) **pokrývající** skutečnou hodnotu parametru θ s pravděpodobností alespoň $1 - \alpha$, ať je tento skutečný neznámý parametr θ kdekoli ve svém parametrickém prostoru Θ .

Označíme-li $\mathbb{E}[\lambda(C(\mathbf{X}))]$ jako střední λ -objem konfidenční množiny, resp. střední λ -délku intervalu spolehlivosti, pak hledáme takovou $C(\mathbf{X})$, která minimalizuje $\mathbb{E}[\lambda(C(\mathbf{X}))]$ stejnoměrně na Θ při podmínce, že konfidenční koeficient $\geq 1 - \alpha$. Tato optimalizační úloha je však často neřešitelná, proto se v praxi většinou snažíme naplnit alespoň kritérium kladené na konfidenční koeficient a přitom dosáhnout rozumného objemu/délky $\lambda(C(\mathbf{x}))$. Při náhodném opakování celého experimentu pak víme, že $\text{CM}_{1-\alpha}$ pokrývá skutečnou hodnotu parametru θ v průměru ve $(1 - \alpha) \cdot 100\%$ případech.

6.1 Konstrukce $\text{CM}_{1-\alpha}$ pomocí pivotů (PQ)

Definice 6.3. Borelovsky měřitelná funkce $\mathcal{R}(\mathbf{X}, \theta)$ se nazývá **pivotální veličina** (*pivot*) (PQ) pro parametr θ , pokud **rozdělení** $\mathcal{R}(\mathbf{X}, \theta)$ nezávisí na volbě $\mathbb{P} \in \mathcal{P}$.

Metoda konstrukce $\text{CM}_{1-\alpha}$ pro θ :

- 1) Nalezneme vhodnou **pivotální veličinu** $\mathcal{R}(\mathbf{X}, \theta)$, pokud taková existuje.
- 2) Volíme pevně $\mathbb{P} \in \mathcal{P}$ a najdeme konstanty c_1, c_2 takové, že platí

$$\mathbb{P}(c_1 \leq \mathcal{R}(\mathbf{X}, \theta) \leq c_2) \geq 1 - \alpha, \quad \text{případně} \quad \mathbb{P}(c_1 \leq \mathcal{R}(\mathbf{X}, \theta) \leq c_2) = 1 - \alpha.$$

3) Pak $C(\mathbf{X}) = \left\{ \theta \in \Theta : c_1 \leq \mathcal{R}(\mathbf{X}, \theta) \leq c_2 \right\}$ je $\text{CI}_{1-\alpha}$.

Důkaz. Zřejmý. □

Problémy konstrukce $C(\mathbf{X})$:

- a) existence alespoň jednoho pivotu $\mathcal{R}(\mathbf{X}, \theta)$,
- b) existence mnoha různých $\mathcal{R}(\mathbf{X}, \theta)$: tedy nevíme, kterou z nich v konstrukci $C(\mathbf{X})$ použít, tzn. která poskytuje nejmenší střední objem/délku $C(\mathbf{X})$,
- c) volba c_1, c_2 : kritériem může být například $\alpha/2$ -symetrie, tzn. volba c_1, c_2 tak, aby $\mathbb{P}(\mathcal{R}(\mathbf{X}, \theta) > c_2) = \frac{\alpha}{2}$ a $\mathbb{P}(\mathcal{R}(\mathbf{X}, \theta) < c_1) = \frac{\alpha}{2}$.
- d) výpočet $C(\mathbf{X})$ ze soustavy nerovnic $c_1 \leq \mathcal{R}(\mathbf{X}, \theta) \leq c_2$: pokud je $\mathcal{R}(\mathbf{X}, \theta)$ např. ryze rostoucí v proměnné $\theta \in \mathbb{R}^1$, pak $C(\mathbf{X}) = \left\{ \theta : \mathcal{R}^{-1}(c_1, \mathbf{X}) \leq \theta \leq \mathcal{R}^{-1}(c_2, \mathbf{X}) \right\}$. Podobně pro funkci $\mathcal{R}(\mathbf{X}, \theta)$ regulární a prostou vzhledem k $\theta \in \mathbb{R}^k$.

PŘÍKLAD 6.4. Mějme Gaussovský náhodný výběr $X_1, \dots, X_n \text{ iid } \mathcal{N}(\mu, \sigma^2)$.

- a) Chceme najít $\text{CI}_{1-\alpha}$ pro parametr $\theta = \mu$, přičemž předpokládáme, že σ^2 neznáme. Víme, že t-statistika

$$T(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \sim t(n-1) \quad (\text{Studentovo rozdělení})$$

a tedy $T(\mathbf{X})$ je pivotální veličina pro parametr μ . Volme konstanty c_1, c_2 z podmínky $\alpha/2$ -symetrie následovně

$$\mathbb{P}\left(\underbrace{t_{\frac{\alpha}{2}}(n-1)}_{c_1} \leq T(\mathbf{X}) \leq \underbrace{t_{1-\frac{\alpha}{2}}(n-1)}_{c_2}\right) = 1 - \alpha.$$

Ze soustavy nerovnic $t_{\frac{\alpha}{2}} \leq T(\mathbf{X}) \leq t_{1-\frac{\alpha}{2}}$ po dosazení za $T(\mathbf{X})$ vyjádříme μ , tzn. dostáváme $\text{CI}_{1-\alpha}$ pro μ v Gaussovském modelu:

$$C(\mathbf{X}) = \left[\bar{X}_n - t_{1-\frac{\alpha}{2}}(n-1) \frac{s_n}{\sqrt{n}}, \quad \bar{X}_n + t_{1-\frac{\alpha}{2}}(n-1) \frac{s_n}{\sqrt{n}} \right],$$

kde jsme použili vlastnost symetrie kvantilů $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$ Studentova rozdělení. V praxi se někdy (nesprávně) tento $\text{CI}_{1-\alpha}$ zapisuje jako

$$\mu = \bar{X}_n \pm t_{1-\frac{\alpha}{2}}(n-1) \frac{s_n}{\sqrt{n}}.$$

- b) Podobně lze získat $\text{CI}_{1-\alpha}$ pro parametr σ^2 při neznámém μ prostřednictvím χ^2 -statistiky (PQ pro σ^2):

$$\chi^2(\mathbf{X}) = \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1) \quad (\text{Pearsonovo } \chi^2\text{-rozdělení}).$$

Protože zde $\chi^2(n-1)$ není symetrické rozdělení, nelze $\text{CI}_{1-\alpha}$ psát ve tvaru $\sigma^2 = \pm c_{\alpha,n} s_n^2$ ani $\sigma^2 = s_n^2 \pm c_{\alpha,n}$.

6 Konfidenční množiny, Intervaly spolehlivosti

PŘÍKLAD 6.5 (Simultální $\text{CM}_{1-\alpha}$ pro \mathcal{N}). Mějme $(X_j)_{j=1}^n \text{ iid } \mathcal{N}(\mu, \sigma^2)$ a uvažujme dvou-
rozměrný parametr $\theta = (\mu, \sigma^2)$. Volme dvourozměrnou pivotální veličinu,
 $\mathcal{R}(\mathbf{X}, \theta) = (U(\mathbf{X}), \chi^2(\mathbf{X}))$, kde

$$U(\mathbf{X}) = \sqrt{n} \frac{(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0, 1) \quad (PQ \text{ pro } \mu \text{ při fixním } \sigma^2),$$

$$\chi^2(\mathbf{X}) = \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1) \quad (PQ \text{ pro } \sigma^2 \text{ nezávisle na } \mu).$$

Volíme konstanty c_1, c_2 tak, aby $\mathbb{P}(c_1 \leq U(\mathbf{X}) \leq c_2) = \sqrt{1-\alpha}$, tzn. např. $c_1 = -u_{1-\frac{\alpha}{2}}$ a $c_2 = u_{1-\frac{\alpha}{2}}$ kvantily $\mathcal{N}(0, 1)$. Dále volíme konstanty d_1, d_2 tak, aby $\mathbb{P}(d_1 \leq \chi^2(\mathbf{X}) \leq d_2) = \sqrt{1-\alpha}$, tzn. např. $d_1 = \chi_{\frac{\alpha}{2}}^2(n-1)$ a $d_2 = \chi_{1-\frac{\alpha}{2}}^2(n-1)$ kvantily rozdělení $\chi^2(n-1)$. Ze soustavy nerovností $\{c_1 \leq U(\mathbf{X}) \leq c_2 \wedge d_1 \leq \chi^2(\mathbf{X}) \leq d_2\}$ pak vyjádříme dvourozměrnou konfidenční množinu pro $\theta = (\mu, \sigma^2)$

$$C(\mathbf{X}) = \left\{ (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ : \frac{n(\bar{X}_n - \mu)^2}{u_{1-\frac{\alpha}{2}}^2} \leq \sigma^2, \frac{(n-1)s_n^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{\chi_{\frac{\alpha}{2}}^2(n-1)} \right\},$$

která dosahuje konfidenčního koeficientu přesně na hladině $1-\alpha$. (nakreslete $C(\mathbf{X}) \subset \mathbb{R} \times \mathbb{R}^+$ za DCv).

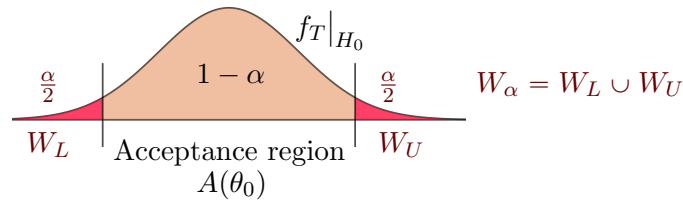
6.2 Konstrukce $\text{CM}_{1-\alpha}$ pomocí TSH ϕ_α

Definice 6.6. Nechť $\phi(\mathbf{x})$ je test hypotézy H_0 vs. H_1 . Pak množinu $A(H_0) = \{\mathbf{x} : \phi(\mathbf{x}) < 1\}$ nazýváme **přípustná oblast testu ϕ** (ozn. AR - acceptance region). Speciálně pro test $H_0 : \theta = \theta_0$ označíme přípustnou oblast AR jako $A(\theta_0)$.

POZNÁMKA 6.7. Nechť je test ϕ pro $H_0 : \theta = \theta_0$ **neznáhodněný** a tedy je charakterizován pro dané θ_0 příslušnou kritickou oblastí typu

$$W_\alpha = \{T(\mathbf{x}) \leq K_1 \vee T(\mathbf{x}) \geq K_2\} = \{T(\mathbf{x}) \leq K_1\} \cup \{T(\mathbf{x}) \geq K_2\} \stackrel{\text{ozn.}}{=} W_L \cup W_U$$

na hladině testu α , tzn. $\mathbb{P}_{\theta_0}(\mathbf{X} \in W_\alpha) = \mathbb{P}_{\theta_0}(\mathbf{X} \in W_L) + \mathbb{P}_{\theta_0}(\mathbf{X} \in W_U) = \alpha$. V tomto případě pak platí, že $A(\theta_0) = \{\mathbf{x} : T(\mathbf{x}) \in (K_1, K_2)\} = W_\alpha^c$, kde $\mathbb{P}(\mathbf{X} \in A(\theta_0)) = 1 - \alpha$. Tuto situaci ilustruje obrázek 6.1.



Obrázek 6.1: Souvislost $\text{CI}_{1-\alpha}$ a testování hypotézy $H_0 : \theta = \theta_0$ skrze CR W_α .

Mějme nyní náš statistický model $\mathbb{P} \in \mathcal{P}$, parametr zájmu $\theta = \theta(\mathbb{P})$ a náhodný výběr $X_1, \dots, X_n \text{ iid } \mathbb{P}_\theta$.

Metoda konstrukce $CM_{1-\alpha}$ pro θ :

1. Volíme $\theta_0 \in \Theta$ libovolně pevně.
2. Testujeme $H_0 : \theta = \theta_0$ na hladině významnosti α . Příslušný test označíme ϕ_α , resp. W_α pro případ neznáhodněného testu.
3. Vyjádříme $A(\theta_0)$ pro test ϕ_α , resp. W_α , při libovolném $\theta_0 \in \Theta$. Tím získáme $A(\theta)$ pro $\forall \theta \in \Theta$.
4. Pak $C(\mathbf{X}) = \{\theta \in \Theta : \mathbf{X} \in A(\theta)\}$ je $CM_{1-\alpha}$ pro θ . Navíc, pokud je test ϕ_α neznáhodněný a je na hladině α , $\beta_{\phi_\alpha}(\theta_0) = \alpha$ pro $\forall \theta_0 \in \Theta$, pak uvedená $C(\mathbf{X})$ je $CM_{1-\alpha}$ s **konfidenčním koeficientem** rovným $(1 - \alpha)$.

Důkaz. Ukážeme tvrzení z bodu 4 pro případ identifikovatelné rodiny $\mathcal{P} = \{\mathbb{P}_\theta\}$, ve které je θ jediným parametrem modelu (nejsou zde tzv. rušivé parametry):

$$\begin{aligned} \inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\theta \in C(\mathbf{X})) &= \inf_{\theta \in \Theta} \mathbb{P}_\theta(\theta \in C(\mathbf{X})) = \inf_{\theta \in \Theta} \mathbb{P}_\theta(\mathbf{X} \in A(\theta)) = 1 - \sup_{\theta \in \Theta} \mathbb{P}_\theta(\mathbf{X} \notin A(\theta)) = \\ &= 1 - \underbrace{\sup_{\theta \in \Theta} \mathbb{P}_\theta(\phi_\alpha(\mathbf{X}, \theta) = 1)}_{\leq \alpha} \geq 1 - \alpha. \end{aligned}$$

□

POZNÁMKA 6.8. Pokud $C(\mathbf{X})$ vznikne prostřednictvím **neznáhodněného** UMP testu ϕ_α na hladině α pro $\forall \theta_0 \in \Theta$, pak $C(\mathbf{X})$ nazveme UMA (*uniformly most accurate*) $CM_{1-\alpha}$ pro θ . Podobně použitím UMPU testu na hladině α získáme UMAU $_{1-\alpha}$ (*UMA Unbiased, $CM_{1-\alpha}$*) pro θ .

PŘÍKLAD 6.9. Uvažujme statistický model $X_1, \dots, X_n \text{ iid } \mathcal{N}(\mu, \sigma^2)$, kde $\sigma^2 > 0$ je neznámé. Jde nám o $CI_{1-\alpha}$ pro parametr $\theta = \mu$ na hladině $\alpha \in (0, 1)$. Testujeme pro $\forall \mu_0 \in \mathbb{R}$ hypotézu

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0 \quad \text{na hladině } \alpha.$$

Použijeme LRT_α test s příslušnou LRT kritickou oblastí

$$W_\alpha = \left\{ \mathbf{x} : \left| \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{s_n} \right| \geq \underbrace{t_{1-\frac{\alpha}{2}}(n-1)}_{K_\alpha} \right\}, \quad \text{pro } \forall \mu_0 \in \mathbb{R}.$$

Pak přípustná oblast pro $\forall \mu_0 \in \mathbb{R}$ je rovna

$$A(\mu_0) = W_\alpha^c = \left\{ \mathbf{x} : \left| \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{s_n} \right| < t_{1-\frac{\alpha}{2}}(n-1) \right\}$$

nezávisle na neznámém (rušivém) parametru $\sigma^2 > 0$. Pak podle bodu 4 konstrukce,

$$C(\mathbf{X}) = \{\mu \in \mathbb{R} : \mathbf{X} \in A(\mu)\} = \left| \begin{array}{c} \text{řešení příslušné} \\ \text{nerovnosti v } A(\mu) \end{array} \right| = \left(\bar{X}_n - t_{1-\frac{\alpha}{2}}(n-1) \frac{s_n}{\sqrt{n}}, \bar{X}_n + t_{1-\frac{\alpha}{2}}(n-1) \frac{s_n}{\sqrt{n}} \right)$$

je CI pro μ s konfidenčním koeficientem rovným $1 - \alpha$. Tento CI je shodný s $CI_{1-\alpha}$ získaným v příkladu 6.4 prostřednictvím pivotální náhodné veličiny $T(\mathbf{X})$ (t-statistika).

POZNÁMKA 6.10. Všimněte si, že v obou konstrukcích $CM_{1-\alpha}$ pomocí PQ pivotů nebo ϕ_α testů jsme neřešili velikost středního objemu/délky tohoto $C(\mathbf{X})$, tzn. stejnoměrnou minimalizaci $\mathbb{E}[\lambda(C(\mathbf{X}))]$. Většinou je to úloha obtížná, mnohdy (explicitně) neřešitelná, na stejné úrovni jako je úloha nalezení testu ϕ_α se stejnoměrně maximální silou (silou) testu $\beta_{\phi_\alpha}|_{H_1}$.

6.3 Asymptotické konfidenční množiny

Definice 6.11. Mějme $\mathbb{P} \in \mathcal{P}$ na (Ω, \mathcal{A}) , $\theta = \theta(\mathbb{P}) \in \Theta \subset \mathbb{R}^1$, resp. $\Theta \subset \mathbb{R}^k$ a \mathcal{B}_Θ Borelovské. Mějme $X_1, \dots, X_n \sim \mathbb{P} \in \mathcal{P}$ a $\alpha \in (0, 1)$. Pak $C(\mathbf{X}) \in \mathcal{B}_\Theta$ se nazývá **asymptotická** $\text{CM}_{1-\alpha}$, ozn $\text{ACM}_{1-\alpha}$, pokud pro

$$\forall \mathbb{P} \in \mathcal{P}, \quad \lim_{n \rightarrow +\infty} \mathbb{P}(\theta \in C(\mathbf{X})) \geq 1 - \alpha.$$

Metody konstrukce:

- I) Najdeme takovou vhodnou náhodnou veličinu $\mathcal{R}_n(\mathbf{X}, \theta)$, která je **asymptoticky pivotální** veličinou (ozn. APQ), tzn. její *limitní* rozdělení nezávisí na $\mathbb{P} \in \mathcal{P}$:

$$\mathcal{R}_n(\mathbf{X}, \theta) \xrightarrow{\mathcal{D}} G, \quad \text{kde } G \text{ nezávisí na } \mathbb{P} \in \mathcal{P}.$$

Toto limitní G použijeme pro konstrukci $\text{ACM}_{1-\alpha}$ stejně jako v případě neasymptotických $\text{CM}_{1-\alpha}$ (viz sekce 6.1).

- II) Stejně jako v sekci 6.2, pro konstrukci $\text{ACM}_{1-\alpha}$ použijeme přípustnou oblast $A(\theta_0)$ založenou na asymptotickém testu ϕ_α dosahujícím asymptotické hladiny α pro testování $H_0 : \theta = \theta_0$. Tedy $C(\mathbf{X}) = \{\theta \in \Theta : \mathbf{X} \in A(\theta)\}$, kde $A(\theta)$ je AR asymptotického testu ϕ_α , je $\text{ACM}_{1-\alpha}$ pro parametr θ .

PŘÍKLAD 6.12 (I). Mějme $X \sim F \in \mathcal{F}$ a volme $\theta = \theta(F) = F(t)$ pro nějaké fixní $t \in \mathbb{R}$. Najdeme $\text{ACI}_{1-\alpha}$ pro $F(t)$ založený na $X_1, \dots, X_n \text{ iid } F$ náhodného výběru. Víme, že

$$F_n(t) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{(-\infty, t]}(X_j) \quad (\text{empirická c.d.f.})$$

je náhodnou veličinou, která je \mathcal{AN} odhadem $F(t)$ pro $\forall t \in \mathbb{R}$, konkrétně

$F_n(t) \sim \mathcal{AN}\left(F(t), \frac{1}{n}F(t)(1-F(t))\right)$. Pak

$$\mathcal{R}_n(\mathbf{X}) = \frac{\sqrt{n}(F_n(t) - F(t))}{\sqrt{F(t)(1-F(t))}} \xrightarrow{\mathcal{D}} U \sim \mathcal{N}(0, 1),$$

a proto $\mathcal{R}_n(\mathbf{X})$ je APQ veličinou. Volíme $c_2 = -c_1 = u_{1-\frac{\alpha}{2}}$ kvantil $\mathcal{N}(0, 1)$, pro který platí, že

$$\lim_{n \rightarrow +\infty} \mathbb{P}_F(|\mathcal{R}_n(\mathbf{X})| \leq u_{1-\frac{\alpha}{2}}) = \mathbb{P}_{\mathcal{N}(0,1)}(|U| \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

Vyřešením nerovnosti $|\mathcal{R}_n(\mathbf{X})| \leq u_{1-\frac{\alpha}{2}}$ vzhledem k $F(t)$ pak snadno dostaneme $\text{ACI}_{1-\alpha}$ pro $F(t)$ při fixním $t \in \mathbb{R}$.

PŘÍKLAD 6.13 (II). Mějme opět systém distribucí \mathcal{P} , $\theta = \theta(\mathbb{P}) \in \Theta \subset \mathbb{R}^k$ a $X_1, \dots, X_n \text{ iid } \mathbb{P} \in \mathcal{P}$. Pro konstrukci $\text{ACM}_{1-\alpha}$ využijeme **asymptotický LRT** test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ na hladině α pro $\forall \theta_0 \in \Theta$. Víme, že za platnosti H_0 a předpokladu věty 5. má veličina $\lambda_n(\mathbf{X}) = -2 \ln \Lambda(\mathbf{x})$ asymptotické rozdělení $\chi^2(k)$, tzn.

$$\lambda_n(\mathbf{X}) = -2 \ln \frac{L(\theta_0)}{L(\hat{\theta}_n)} = 2(l(\hat{\theta}_n) - l(\theta_0)) \xrightarrow{\mathcal{D}} \chi^2(k) \quad \text{pro } \forall \theta_0 \in \Theta,$$

6 Konfidenční množiny, Intervaly spolehlivosti

kde $\hat{\theta}_n$ je MLE odhad parametru θ . Příslušná LRT kritická oblast $W_\alpha = \{\mathbf{x} : \Lambda(\mathbf{x}) \leq K_\alpha\}$ vede na přípustnou oblast

$$A(\theta_0) = \left\{ \mathbf{x} : \lambda_n(\mathbf{x}) < -2 \ln K_\alpha \stackrel{\text{ozn}}{=} K'_\alpha \right\} \quad \text{pro } \forall \theta_0,$$

pro kterou však platí limitní vztah

$$\lim_{n \rightarrow +\infty} \mathbb{P}_{\theta_0}(\mathbf{X} \in A(\theta_0)) = \lim_{n \rightarrow +\infty} \mathbb{P}_{\theta_0}(\underbrace{\lambda_n(\mathbf{X})}_{\xrightarrow{\mathcal{D}} \chi^2(k)} < \chi_{1-\alpha}^2(k)) = 1 - \alpha,$$

kde jsme použili volbu $K'_\alpha = \chi_{1-\alpha}^2(k)$ kvantil $\chi^2(k)$ rozdělení. Máme tedy $\text{ACM}_{1-\alpha}$ pro parametr θ ve tvaru

$$C(\mathbf{X}) = \{\theta \in \Theta : \mathbf{X} \in A(\theta)\} = \left\{ \theta \in \Theta : l(\hat{\theta}_n) < \frac{\chi_{1-\alpha}^2(k)}{2} + l(\theta) \right\}.$$

POZNÁMKA 6.14. $\text{ACM}_{1-\alpha} C(\mathbf{X})$ získaná prostřednictvím asymptotického LRT testu může být podstatně odlišná od $\text{ACM}_{1-\alpha}$ získaná metodou APQ nebo PQ, viz např. **simultální** $\text{ACM}_{1-\alpha}$ oproti $\text{CM}_{1-\alpha}$ pro parametr $\theta = (\mu, \sigma^2)$ v Gaussovském modelu $\mathcal{N}(\mu, \sigma^2)$.