KAPITOLA 1.

Opakování základů teorie testování hypotéz

1.1. Formulace problému

Nechť $\underset{\sim}{X} = (X_1,\dots,X_N)$ je náhodný vektor (vektor pozorování) a nechť H a K jsou 2 disjunktní množiny rozdělení pravděpodobností na $(R^N,\mathcal{B}^N)$. Řekneme, že vektor $\underset{\sim}{X}$ splňuje hypotézu, jestliže rozdělení pravděpodobností $\underset{\sim}{X}$ patří do H, a že splňuje alternativu, jestliže jeho rozdělení patří do K. Pro hypotézu použijeme rovněž symbolu H a pro alternativu symbolu K ; tedy H i K označují jak výrok, tak množinu rozdělení náhodných vektorů výrok splňujících - to jistě nepovede k omylu a zjednoduší se zápis.

Hypotézu obvykle formulujeme tak, že je to množina rozdělení majících určitou vlastnost homogenity, symetrie, nezávislosti apod. - proto často užíváme přívlastku "nulová hypotéza". Rozdělení patřící do alternativy naopak vykazují nehomogenitu, nesymetrii, závislost, apod.

Problém spočívá v tom, že na základě pozorovaných dat $x_1,\dots,x_N$ máme rozhodnout, zda platí hypotéza H nebo alternativa K. Každé pravidlo, které každému bodu $x_1,\dots,x_N$ přiřadí právě jedno ze 2 možných rozhodnutí: "přijetí H " - "zamítnutí H " , nazveme (nerandomizovaným) testem hypotézy H proti alternativě K . Takový test rozděluje výběrový prostor na 2 komplementární části : kritický obor (obor zamítnutí) A_K a obor přijetí A_H. Jestliže pak $(x_1,\dots,x_N) \in A_K$, test

hypotézu zamítá a v případě $(x_1,\dots,x_N) \in A_H$ ji nezamítá.

Jestliže na základě pozorování $x_1,\dots,x_N$ provádíme nějaký test, může být naše rozhodnutí správné, nebo se můžeme dopustit jedné ze 2 druhů chyb :

(1) zamítneme H_1 i když H platí (chyba I.druhu),

(2) přijmeme H_1 i když H neplatí (chyba II.druhu).

Je žádoucí použít takový test, který má co nejnižší pravděpodobnosti obou druhů chyb. Pravděpodobnost chyby I.druhu, je-li $P \in H$ skutečne rozdělení vektoru $\underset{\sim}{X}$, je rovna

$$P(\underset{\sim}{X} \in A_K). \tag{1.1}$$

Číslo $\sup_{P \in H} P(\underset{\sim}{X} \in A_K)$ nazýváme velikostí testu s kritickým oborem A_K.

Pravděpodobnost chyby II.druhu v případě, že $Q \in K$ je skutečné rozdělení $\underset{\sim}{X}$, je rovna

$$Q(\underset{\sim}{X} \in A_H) = 1 - Q(\underset{\sim}{X} \in A_K). \tag{1.2}$$

Hodnotu pravděpodobnosti

$$\beta(Q) = Q(\underset{\sim}{X} \in A_K), \quad Q \in K \tag{1.3}$$

nazýváme silou testu proti alternativě Q, $Q \in K$. Funkci $\beta(Q) : K \to [0,1]$ pak nazýváme silofunkcí příslušného testu.

Žádoucí je pak test, jehož silofunkce je maximální stejnoměrně přes celou alternativu a jehož pravděpodobnost chyby I. druhu je malá pro všecka rozdělení splňující hypotézu.

Teorie testování a hledání optima se značně zjednoduší, rozšíříme-li množinu testů o tzv. randomizované testy. Randomizovaný test zamítá H při daném $\underset{\sim}{x}$ s pravděpodobností $\Phi(\underset{\sim}{x})$

a přijímá s pravděpodobností $1-\varphi(\underset{\sim}{x})$, $0 \le \varphi(\underset{\sim}{x}) \le 1$ pro vš. $\underset{\sim}{x}$. Každý test je tak charakterizován funkcí φ, zvanou testová (nebo kritická) funkce takovou, že $0 \le \varphi \le 1$. Kritický obor testu je pak $\{\underset{\sim}{x} : \varphi(\underset{\sim}{x}) = 1\}$, takže u nerandomizovaných testů je φ indikátorem kritického oboru.

Zavedením randomizovaných testů se množina všech testů ztotožňuje s množinou všech funkcí $\{\varphi(\underset{\sim}{x}) : 0 \le \varphi \le 1\}$, a je tedy konvexní a slabě kompaktní.

Jestliže P je skutečné rozdělení pravděpodobností $\underset{\sim}{X}$, pak test s kritickou funkcí φ zamítá H s pravděpodobností

$$\beta_\varphi(P) = E_P(\varphi(\underset{\sim}{X})) = \int \varphi(\underset{\sim}{x}) dP(\underset{\sim}{x}). \tag{1.4}$$

Je zřejmé, že mezi všemi testy H proti K by byl nejlepší ten, který by splňoval

$$\beta_\varphi(Q) = E_Q(\varphi(\underset{\sim}{X})) := \max \quad \text{pro vš. } Q \in K \tag{1.5}$$

a zároveň

$$\beta_\varphi(P) = E_P(\varphi(\underset{\sim}{X})) := \min \quad \text{pro vš. } P \in H . \tag{1.6}$$

Protože nelze oběma podmínkám současně vyhovět, formulujeme úlohu optimálního testu takto: zvolíme malé číslo α, $0 < \alpha < 1$, tzv. hladinu významnosti, a mezi všemi testy vyhovujícími

$$\beta_\varphi(P) \le \alpha \quad \text{pro vš. } P \in H \tag{1.7}$$

hledáme test vyhovující (1.5). Pokud takový test existuje, nazýváme jej stejnoměrně nejsilnější (SN) test H proti K velikosti $\le \alpha$, krátce stejnoměrně nejsilnější α-test.

Podle toho, zdali H (resp. K) obsahuje jeden nebo více prvků, mluvíme o jednoduché nebo složené hypotéze (resp. alter-

nativě).

V případě jednoduché hypotézy i alternativy je problém maximalizace (1.5) za podmínky (1.6) plně řešen Neyman-Pearsonovým lemmatem, které uvedeme bez důkazu.

Neyman-Pearsonovo lemma. Nechť P a Q jsou dvě rozdělení pravděpodobností, která mají hustoty p a q vzhledem k míře μ (může být i $\mu = P+Q$). Pak pro test jednoduché hypotézy $H:\{P\}$ proti jednoduché alternativě $K:\{Q\}$ existuje test Φ a konstanta k tak, že

$$E_P(\Phi(\underset{\sim}{X})) = \alpha \tag{1.8}$$

a

$$\Phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{jestliže } q(\underset{\sim}{x}) > k.p(\underset{\sim}{x}) \\ 0 & \text{jestliže } q(\underset{\sim}{x}) < k.p(\underset{\sim}{x}). \end{cases} \tag{1.9}$$

Tento test je nejsilnějším α-testem H proti K.

Chceme-li nalézt nejsilnější test složené hypotézy proti jednoduché alternativě, můžeme někdy použít následující větu:

VĚTA 1. Uvažujme problém testování složené hypotézy H proti jednoduché alternativě $K:\{Q\}$. Nechť všechna rozdělení patřící do H i K mají hustoty vzhledem k σ-konečné míře μ. Nechť $P_0 \in H$ a nechť Φ_0 je test takový, že

$$\Phi_0(\underset{\sim}{x}) = \begin{cases} 1 \ldots & \text{jestliže } q(\underset{\sim}{x}) > kp_0(\underset{\sim}{x}) \\ 0 \ldots & \text{jestliže } q(\underset{\sim}{x}) < kp_0(\underset{\sim}{x}) \end{cases} \tag{1.10}$$

pro některou konstantu k ; $p_0 = \frac{d P_0}{d \mu}$, $q = \frac{dQ}{d\mu}$.
Pak, jestliže platí

$$E_{P_0}(\Phi_0(\underset{\sim}{X})) = \sup_{P \in H} E_P(\Phi_0(\underset{\sim}{X})) = \alpha \tag{1.11}$$

kde $0<\alpha<1$, je test Φ_0 nejsilnějším α- testem H proti K.

Rozdělení P_0 pak nazveme nejméně příznivé rozdělení hypotézy H vzhledem k alternativě $K : \{Q\}$.

Důkaz věty 1 jakož i Neyman-Pearsonova lemmatu lze nalézt v knize Lehmann (1959).

Na závěr ještě připomeneme definici nestranného testu:

DEFINICE 1 : Řekneme, že test Φ hypotézy H proti alternativě K je nestranný, jestliže platí

$$\begin{aligned} &\beta_\Phi(P) \leq \alpha && \text{pro vš. } P \in H \\ &\beta_\Phi(Q) \geq \alpha && \text{pro vš. } Q \in K . \end{aligned} \tag{1.12}$$

Pokud existuje SM α-test H proti K, je nutně nestranný, protože jeho silofunkce nemůže být menší než silofunkce testu $\Phi(x) \equiv \alpha$.

Příklad 1. Nechť X je charakteristika výrobku v hromadné výrobě. Výrobek je dobrý, jestliže $X > u$, kde u je daná konstanta. Chceme testovat hypotézu $H : p \geq p_0$ proti alternativě $K : p < p_0$, kde $p = P(X \leq u)$ je pravděpodobnost, že výrobek je vadný.

Nechť $X_1, \ldots, X_N$ jsou měření provedená na náhodném výběru N výrobků. X_i jsou tedy nezávislé náhodné veličiny se stejným rozdělením P, které neznáme. Každé takové rozdělení lze jednoznačně charakterizovat trojicí (p, P^-, P^+), kde P^- je rozdělení X podmíněné jevem $X \leq u$ a P^+ je rozdělení X podmíněné jevem $X > u$. Jestliže p^- a p^+ jsou hustoty P^- a P^+ vzhledem k nějaké míře μ (např. $\mu = P^- + P^+$) a jestliže $\underset{\sim}{x} = (x_1, \ldots, x_N)$ je bod výběrového prostoru takový, že

$$x_{i_1}, \ldots, x_{i_m} \leq u < x_{j_1}, \ldots, x_{j_{N-m}}$$

pak hustota sdruženého rozdělení $X_1,\dots,X_N$ v bodě $\underset{\sim}{x}$ je rovna

$$p^m(1-p)^{N-m}\, p^-(x_{i_1}) \dots p^-(x_{i_m})\, p^+(x_{j_1}) \dots p^+(x_{j_{N-m}}).$$

Uvažujme nyní pevnou alternativu, řekněme (p_1, P^-, P^+), kde $p_1 < p_0$. Očekáváme, že nejméně příznivé rozdělení hypotézy H proti této alternativě bude (p_0, P^-, P^+), protože je nejblíže zvolené alternativě. Ověříme tuto skutečnost pomocí věty 1. Test Φ_0 příslušný (p_0, P^-, P^+) má tvar

$$\Phi_0(\underset{\sim}{x}) = \begin{cases} 1 & \dots \text{ jestliže } \left(\frac{p_1}{p_0}\right)^m \left(\frac{1-p_1}{1-p_0}\right)^{N-m} > C_1 \\ \gamma & \dots \text{ jestliže } = C_1 \\ 0 & \dots \text{ jestliže } \left(\frac{p_1}{p_0}\right)^m \left(\frac{1-p_1}{1-p_0}\right)^{N-m} < C_1 \end{cases}$$

kde C_1 je konstanta. Protože p_0 a p_1 jsou pevné, $p_1 < p_0$, můžeme $\Phi_0(\underset{\sim}{x})$ psát také ve tvaru

$$\Phi_0(\underset{\sim}{x}) = \begin{cases} 1 & \dots \quad M < C_2 \\ \gamma & \dots \quad M = C_2 \\ 0 & \dots \quad M > C_2 \end{cases}$$

kde M je počet vadných výrobků ve výběru a konstanty C_2 a γ jsou určeny podmínkou $P(M < C_2) + \gamma P(M = C_2) = \alpha$ pro $p = p_0$.

Rozdělení náhodné veličiny M je binomické $b(p,N)$ a je tedy nezávislé na P^- a P^+. Odtud plyne, že silofunkce testu Φ_0 závisí jen na p a je klesající v p, takže za platnosti H dosahuje maxima pro $p = p_0$. Tím je dokázáno, že (p_0, P^-, P^+) je <u>nejméně příznivé</u> rozdělení H vzhledem k (p_1, P^-, P^+) a že Φ_0 je nejsilnějším α-testem proti

této alternativě. Zároveň vidíme, že test ϕ_0 nezávisí na speciálně zvolené alternativě a je tedy stejnoměrně nejsilnější pro H proti K .

Jestliže označíme $Z_i = X_i - u$, pak testová statistika je počet Z_i mezi $Z_1,\dots,Z_N$ pro která platí $Z_i \leq 0$. Později uvidíme, že se jedná o tzv. znaménkový test.

Tím jsme také získali první příklad testu optimálního pro neparametrický model. Kdybychom předpokládali nějaký specifický tvar rozdělení X , mohli bychom dostat jiný test.

1.2. Princip invariance v testování hypotéz

Mnohé statistické problémy jsou z určitého hlediska symetrické. Pak je přirozené, omezíme-li se při řešení těchto problémů na statistické postupy, které jsou podobně symetrické. Např. předpokládejme, že $X_1,\dots,X_n$ jsou nezávislé náhodné veličiny s rozděleními pravděpodobností $P_{\theta_1},\dots,P_{\theta_n}$. Pak od vhodného testu hypotézy $H : \theta_1 = \dots = \theta_n$ proti alternativě, že alespoň 2 hodnoty z $\theta_1,\dots,\theta_n$ jsou různé, očekáváme, že bude symetrický vzhledem k uspořádání dat $x_1,\dots,x_n$.

Matematickým vyjádřením symetrie problému je jeho invariance ke vhodné grupě zobrazení. V našem příkladě je touto grupou množina všech permutací dat $x_1,\dots,x_n$.

Obecně nechť g je prosté zobrazení výběrového prostoru $\mathcal{X}$ na sebe. Řekneme, že problém testu H proti K je invariantní vzhledem ke g, jestliže g zachovává hypotézu i alternativu, tj. :

(1.13) $\quad P \in H \Rightarrow P\,g^{-1} \in H,\ Q \in K \Rightarrow Q\,g^{-1} \in K$

a jestliže ke každému rozdělení $P' \in H$ (resp. $Q' \in K$) existuje $P \in H$ (resp. $Q \in K$) tak, že $P = P'g^{-1}$ (resp. $Q = Q'g^{-1}$). Jinými slovy, problém testu H proti K zůstává invariantní vzhledem k zobrazení g, jestliže rozdělení $g\underset{\sim}{X}$ splňuje H právě tehdy, splňuje-li rozdělení $\underset{\sim}{X}$ hypotézu H, a podobně pro alternativu K.

Představme si, že problém testu H proti K je invariantní vzhledem ke grupě G zobrazení výběrového prostoru $\mathcal{X}$ na sebe. Pak je přirozené, omezíme-li se na testy, které jsou také invariantní vzhledem ke G, tj. které splňují

(1.14) $$\Phi(g\,\underset{\sim}{x}) = \Phi(\underset{\sim}{x}) \quad \text{pro vš. } \underset{\sim}{x} \in \mathcal{X} \text{ a } g \in G.$$

Mezi všemi testy, invariantními vzhledem ke G, se pak snažíme nalézt stejnoměrně nejsilnější invariantní α-test. K tomu musíme prozkoumat strukturu množiny všech testů, invariantních vzhledem ke grupě G. Ukáže se, že v některých případech lze nalézt vhodnou statistiku, zvanou maximální invarianta takovou, že každý invariantní test je funkcí této statistiky.

DEFINICE. Statistiku $T = T(\underset{\sim}{x})$ nazveme maximální invariantou vzhledem ke grupě zobrazení G, jestliže platí

(1.15) $T(g\,\underset{\sim}{x}) = T(\underset{\sim}{x})$ pro vš. $\underset{\sim}{x} \in \mathcal{X}$ a $g \in G$,
tj. T je invariantní,
a

(1.16) $T(\underset{\sim}{x}_1) = T(\underset{\sim}{x}_2)$ implikuje, že existuje $g \in G$ tak, že $\underset{\sim}{x}_2 = g\,\underset{\sim}{x}_1$.

VĚTA 2. Nechť $T(\underset{\sim}{x})$ je maximální invarianta vzhledem ke grupě zobrazení G. Pak test $\Phi(\underset{\sim}{x})$ je invariantní vzhledem

ke G právě tehdy, jestliže existuje funkce h taková, že

$$\Phi(\underset{\sim}{x}) = h(T(\underset{\sim}{x})) \quad \text{pro vš.} \quad \underset{\sim}{x} \in \mathcal{X}.$$

Důkaz. (1) Jestliže $\Phi(\underset{\sim}{x}) = h(T(\underset{\sim}{x}))$ pro vš. $\underset{\sim}{x}$, pak $\Phi(g\,\underset{\sim}{x}) = h(T(g\,\underset{\sim}{x})) = h(T\,\underset{\sim}{x})) = \Phi(\underset{\sim}{x})$ pro lib. $g \in G$, a tedy Φ je invariantní.

(2) Nechť Φ je invariantní a nechť $T(\underset{\sim}{x}_1) = T(\underset{\sim}{x}_2)$. Pak podle definice je $\underset{\sim}{x}_2 = g\,\underset{\sim}{x}_1$ pro nějaké $g \in G$ a tedy $\Phi(\underset{\sim}{x}_2) = \Phi(\underset{\sim}{x}_1)$.

Příklad 1

Příklady maximálních invariant

(1) Nechť $\underset{\sim}{x} = (x_1,\ldots,x_n)$ a nechť G je grupa posunutí

$$g\,\underset{\sim}{x} = (x_1 + c,\ldots, x_n + c), \quad c \in R^1 .$$

Pak $T(\underset{\sim}{x}) = \underset{\sim}{y} = (x_1 - x_n,\ldots,x_{n-1} - x_n)$ je maximální invarianta. Skutečně, $T(\underset{\sim}{x})$ je invariantní vzhledem k posunutí; na druhé straně, nechť $T(\underset{\sim}{x}) = T(\underset{\sim}{x}')$; položíme-li $x_n' - x_n = c$, dostaneme $x_i' = x_i + c$, $i = 1,\ldots,n$.

(2) Nechť $\underset{\sim}{x} = (x_1,\ldots,x_n)$ a nechť G je grupa všech ortogonálních transformací $R^n \to R^n$. Pak $T(\underset{\sim}{x}) = \sum_{i=1}^{n} x_i^2$ je maximální invarianta, protože 2 body $\underset{\sim}{x}$ a $\underset{\sim}{x}^*$ se dají vzájemně zobrazit ortogonální transformací právě tehdy, mají-li stejné vzdálenosti od počátku.

(3) Nechť $\underset{\sim}{x} = (x_1,\ldots,x_n)$ a nechť G je množina n! permutací složek $\underset{\sim}{x}$. Pak maximální invarianta

$$T(\underset{\sim}{x}) = \left(x^{(1)} \leq x^{(2)} \leq \ldots \leq x^{(n)}\right)$$

je vektor uspořádaných složek $\underset{\sim}{x}$, tzv. vektor pořádkových statistik.

(4) Nechť G je množina všech zobrazení tvaru

$$x_i' = f(x_i), \qquad i=1,\dots,n$$

takových, že f je spojitá a ryze rostoucí funkce. Uvažujme jen ty body výběrového prostoru $\mathcal{X}$, jejichž všechny složky jsou různé.

Nechť R_i je pořadí x_i mezi $x_1,\dots,x_n$, tj.

$$R_i = \sum_{j=1}^{n} u(x_i - x_j), \qquad i=1,\dots,n,$$

kde $u(x) = 1$ pro $x \geq 0$ a $u(x) = 0$ pro $x < 0$. Pak $T(\underset{\sim}{x}) = (R_1,\dots,R_n)$ je maximální invarianta vzhledem ke G.

Skutečně, spojitá rostoucí funkce nemění pořadí složek $\underset{\sim}{x}$, což znamená, že $T(\underset{\sim}{x})$ je invariantní.

Abychom dokázali vlastnost (1.16) statistiky $T(\underset{\sim}{x})$, předpokládejme, že 2 různé vektory $\underset{\sim}{x}$ a $\underset{\sim}{x}'$ mají stejný vektor pořadí $R_1,\dots,R_n$. To znamená, že je-li $x_{i_1} < x_{i_2} < \dots < \dots < x_{i_n}$ vektor $\underset{\sim}{x}$ uspořádaný podle velikosti složek, platí i $x'_{i_1} < \dots < x'_{i_n}$ pro vektor $\underset{\sim}{x}'$. Definujme funkci $f : R^1 \to R^1$ takto :

$$f(x_i) = x_i', \qquad i=1,\dots,n,$$

f je lineární na intervalech $\langle x_{i_1}, x_{i_2}\rangle, \dots, \langle x_{i_{n-1}}, x_{i_n}\rangle$,
f je spojitá a rostoucí.
Alespoň jedna taková funkce existuje, a tedy $T(\underset{\sim}{x})$ je maximální invarianta.

Příklad 2. Uvažujme náhodný vektor $\underset{\sim}{X} = (X_1,\dots,X_n)$. Chceme testovat hypotézu H_0 proti alternativě H_1 :

H_i : " $\underset{\sim}{X}$ má hustotu $f_i(x_i-\theta,\dots,x_n-\theta)$, $\theta \in R^1$ "
$i=0,1$

kde f_0, f_1 jsou dané hustoty, Θ je neznámý (rušivý) parametr. Problém testu H_0 proti H_1 je invariantní vzhledem ke grupě posunutí

$$g\,\underset{\sim}{x} = (x_1+c,\ldots,x_n+c), \qquad c\in R^1.$$

Maximální invarianta je $T(\underset{\sim}{x}) = \underset{\sim}{y} = (x_1-x_n,\ldots,x_{n-1}-x_n)$.

$\underset{\sim}{Y}$ má za H_i rozdělení s hustotou

$$f_i^*(y_1,\ldots,y_{n-1}) = \int f_i(y_i+z,\ldots,y_{n-1}+z,z)dz\,, \qquad i=0,1.$$

Omezíme-li se tedy na testy invariantní vzhledem k posunutí, redukuje se problém na test jednoduché hypotézy $H_0 : \underset{\sim}{Y}$ má hustotu $f_0^*(y_1,\ldots,y_{n-1})$ proti alternativě

$H_1 : \underset{\sim}{Y}$ má hustotu $f_1^*(y_1,\ldots,y_{n-1})$.

Podle Neyman-Pearsonova lemmatu stejnoměrně nejsilnější invariantní test má tedy tvar

$$\Phi(\underset{\sim}{y}) = \begin{cases} 1 \ \ldots & \dfrac{\int f_1(y_1+z,\ldots,y_{n-1}+z,z)dz}{\int f_0(y_1+z,\ldots,y_{n-1}+z,z)dz} > k \\ 0 \ \ldots & \dfrac{\int f_1(y_1+z,\ldots,y_{n-1}+z,z)dz}{\int f_0(y_1+z,\ldots,y_{n-1}+z,z)dz} < k \end{cases}$$

kde k je zvoleno tak, aby test měl danou velikost α.

Vyjádřeno pomocí původních pozorování $x_1,\ldots,x_n$,

$$\Phi(\underset{\sim}{x}) = 1 \ \ldots \ \frac{\int f_1(x_1+t,\ldots,x_n+t)dt}{\int f_0(x_1+t,\ldots,x_n+t)dt} > k\,.$$

Test nezávisí na Θ, a je tedy stejnoměrně nejsilnějším invariantním testem H_0 proti H_1.

Příklad 3. Nechť $\underset{\sim}{X} = (X_1,\ldots,X_n)$ je náhodný výběr z rozdělení $N(\mu,\sigma^2)$. Testujeme

$$H : \sigma \geq \sigma_0 \quad \text{proti} \quad K : \sigma < \sigma_0 .$$

Problém je invariantní vzhledem ke grupě G posunutí $g\,\underset{\sim}{x} = (x_1+c,\ldots,x_n+c)$, $c \in R^1$. Protože $Y = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ a $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$ jsou postačující statistiky pro μ a σ^2, stačí omezit se na testy, které jsou funkcemi těchto postačujících statistik. Posunutí $g\,\underset{\sim}{x} = (x_1+c,\ldots,x_n+c)$ indukuje v prostoru postačujících statistik zobrazení

$$y' = y + c \,, \quad (S^2)' = S^2 .$$

Maximální invarianta vzhledem ke grupě takovýchto zobrazení je zřejmě S^2, a tedy stejnoměrně nejsilnější invariantní test bude záviset jen na S^2 : přesněji, bude mít tvar

$$\Phi(\underset{\sim}{x}) = \Phi(y, s^2) = \begin{cases} 1 & \ldots \ \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 \leq C \\ 0 & \ldots \ \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 > C , \end{cases}$$

kde kritickou hodnotu C stanovíme z tabulek rozdělení χ^2_{n-1}.

KAPITOLA 2.

Základní výběrové statistiky : pořadí a pořádkové statistiky

Nechť $\underset{\sim}{X} = (X_1,\dots,X_n)$ je vektor pozorování, tj. měřitelné zobrazení základního prostoru $(\Omega,\mathcal{A})$ do výběrového prostoru $(\mathcal{X},\mathcal{B})$, kde $\mathcal{X}$ je borelovská množina z R^n a $\mathcal{B}$ je systém borelovských podmnožin $\mathcal{X}$. Nechť $X^{(1)} \leq X^{(2)} \leq \dots \leq X^{(n)}$ značí tentýž vektor uspořádaný podle velikosti. Pak statistiku $X^{(i)}$ nazveme i-tou pořádkovou statistikou a vektor $X^{(\cdot)} = (X^{(1)},\dots,X^{(n)})$ vektorem pořádkových statistik.

Nechť $\underset{\sim}{x} = (x_1,\dots,x_n)$ je realizace vektoru pozorování, jejíž žádné 2 složky nejsou shodné. Označme $r_i(x)$ počet složek vektoru $\underset{\sim}{x}$, které jsou $\leqslant x_i$, tj. pořadí složky x_i v posloupnosti $x^{(1)} < \dots < x^{(n)}$. Pořadí X_i můžeme též definovat zápisem :

$$R_i = \sum_{j=1}^{n} u(X_i - X_j), \qquad i=1,\dots,n \tag{2.1}$$

kde $u(t) = 1$ pro $t \geqslant 0$ a $u(t) = 0$ pro $t < 0$.

Vidíme, že realizace náhodného pokusu vede k úplnému uspořádání dat pouze tehdy, jsou-li všecka pozorování $x_1,\dots,x_n$ různá. Jen tehdy je vektor pořadí definován jednoznačně. Má-li rozdělení pravděpodobností náhodného vektoru $\underset{\sim}{X}$ spojitou distribuční funkci, nastane shoda dvou pozorování s pravděpodobností 0 . Protože v dalším testu budeme uvažovat jen spojitá rozdělení pravděpodobností, zdálo by se, že problémem, zda je vektor po-

řadí dobře definován, se nemusíme zabývat.

Bohužel v praxi se shody dvou nebo více pozorování často vyskytnou, i když jsou měřené veličiny spojitého typu, protože měření vždy zaznamenáváme jen na konečný počet desetinných míst. Pro takové případy můžeme definici pořadí u testů na nich založených vhodně modifikovat; k této otázce se později vrátíme.

Vektor pořadí R a vektor pořádkových statistik $X^{(\cdot)}$ budou východiskem našich dalších úvah, proto věnujeme tuto kapitolu jejich vlastnostem. V kapitole 1 jsme viděli, že obě tyto statistiky jsou maximálními invariantami vzhledem k určitým grupám zobrazení výběrového prostoru $\mathcal{X}$ na sebe. Nyní budeme uvažovat jejich postačitelnost a chování jejich rozdělení pravděpodobností.

Nechť $\mathcal{R}$ značí množinu všech permutací $r = (r_1,\ldots,r_n)$ posloupnosti $(1,\ldots,n)$. $\mathcal{R}$ obsahuje $n!$ bodů a je-li vektor pořadí R dobře definován, nabývá hodnot právě z množiny $\mathcal{R}$.

Nechť $\mathcal{X}^{(\cdot)}$ značí podprostor výběrového prostoru $\mathcal{X}$ obsahující body $x^{(\cdot)} = (x^{(1)},\ldots,x^{(n)})$ splňující $x^{(1)} < \ldots < \ldots < x^{(n)}$. Nechť $\mathcal{B}^{(\cdot)}$ je σ-algebra borelovských podmnožin $\mathcal{X}^{(\cdot)}$.

VĚTA 1. Dvojice $(X^{(\cdot)},R)$ je postačující statistikou pro libovolný systém rozdělení $\underset{\sim}{X}$ absolutně spojitého typu.

Důkaz. Jestliže je dáno $X^{(\cdot)} = x^{(\cdot)}$ a $R = r$, pak při libovolném absolutně spojitém rozdělení $\underset{\sim}{X}$ a pro libovolné $A \in \mathcal{B}$ platí

$$P(\underset{\sim}{X} \in A \mid X^{(\cdot)} = x^{(\cdot)},\ R = r) =$$

$= P\left\{(x^{(r_1)},\ldots,x^{(r_n)}) \in A \,\middle|\, X^{(\cdot)} = x^{(\cdot)},\ R=r\right\} = 0$ nebo 1 podle toho, zda $(x^{(r_1)},\ldots,x^{(r_n)})$ je nebo není prvkem A; tato pravděpodobnost zřejmě nezávisí na základním rozdělení P.

DEFINICE 1. Nechť $\mathcal{P}$ je systém rozdělení náhodného vektoru $\underset{\sim}{X}$ a nechť $T = T(\underset{\sim}{X})$ je statistika. Řekneme, že statistika T je úplná, jestliže pro libovolnou funkci $h(t)$ takovou, že

(2.2) $\quad E_P\, h(T(\underset{\sim}{X})) = 0 \quad$ pro vš. $P \in \mathcal{P}$,

platí $h(T(\underset{\sim}{x})) = 0$ sj. $[\mathcal{P}]$.

DEFINICE 2. Řekneme, že náhodný vektor $\underset{\sim}{X} = (X_1,\ldots,X_n)$ splňuje hypotézu H_0, jestliže má rozdělení pravděpodobnosti s hustotou tvaru

$$(2.3) \qquad p(\underset{\sim}{x}) = \prod_{i=1}^{n} f(x_i), \qquad \underset{\sim}{x} \in R^n$$

kde f je libovolná jednorozměrná hustota; tj. jestliže složky $\underset{\sim}{X}$ jsou vzájemně nezávislé náhodné veličiny se stejným rozdělením absolutně spojitého typu (hypotéza náhodnosti).

VĚTA 2. Vektor $X^{(\cdot)}$ je úplnou postačující statistikou pro hypotézu H_0, tj. pro systém $\mathcal{P}$ všech rozdělení tvaru (2.3).

Důkaz.

(1) Postačitelnost $X^{(\cdot)}$. Jestliže je dáno $X^{(\cdot)} = x^{(\cdot)}$, může náhodný vektor $\underset{\sim}{X}$ nabývat jen hodnot $(x^{(r_1)},\ldots,x^{(r_n)})$ kde $r = (r_1,\ldots,r_n) \in \mathcal{R}$ a vzhledem ke (2.3) platí

$$(2.4) \quad P\left\{\underset{\sim}{X} = (x^{(r_1)},\ldots,x^{(r_n)}) \,\middle|\, X^{(\cdot)} = (x^{(1)},\ldots,x^{(n)})\right\} = \frac{1}{n!}$$

pro vš. $r \in \mathcal{R}$. Podmíněné rozdělení (2.4) je tedy stejné pro všecka rozdělení $\underset{\sim}{X}$ patřící k H_0.

(2) Úplnost $X^{(\cdot)}$. Důkaz úplnosti zde neuvádíme, protože je značně složitější a rozsáhlejší. Podrobný důkaz lze nalézt v Lehmannově knize (1959). Myšlenka důkazu je taková, že se nejprve dokáže, že statistika $X^{(\cdot)}$ je ekvivalentní statistice

$$T(\underset{\sim}{X}) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \ldots, \sum_{i=1}^{n} X_i^n \right)$$

v tom smyslu, že $X^{(\cdot)}$ i $T(\underset{\sim}{X})$ indukují stejnou pod-σ-algebru $\mathcal{B}$, a pak se dokáže úplnost statistiky $T(\underset{\sim}{X})$ vzhledem ke vhodnému podsystému rozdělení $\underset{\sim}{X}$ splňujících H_0. Odtud pak plyne úplnost $T(\underset{\sim}{X})$, a tedy i úplnost $X^{(\cdot)}$, vzhledem k celé hypotéze H_0.

VĚTA 3. Nechť náhodný vektor $\underset{\sim}{X}$ má rozdělení pravděpodobností s hustotou $p(x_1,\ldots,x_n)$. Pak (1) $X^{(\cdot)}$ má rozdělení pravděpodobností s hustotou

$$\bar{p}(x^{(1)},\ldots,x^{(n)}) = \begin{cases} \sum_{r \in \mathcal{R}} p(x^{(r_1)},\ldots,x^{(r_n)}) & \ldots (x^{(1)},\ldots,x^{(n)}) \in \mathcal{X}^{(\cdot)} \\ 0 & \ldots \text{jinde;} \end{cases} \tag{2.5}$$

(2) rozdělení R podmíněné jevem $X^{(\cdot)} = x^{(\cdot)}$ má tvar

$$P(R=r \mid X^{(\cdot)} = x^{(\cdot)}) = \frac{p(x^{(r_1)},\ldots,x^{(r_n)})}{\bar{p}(x^{(1)},\ldots,x^{(n)})} \tag{2.6}$$

pro lib. $r \in \mathcal{R}$ a $x^{(\cdot)} \in \mathcal{X}^{(\cdot)}$.

Důkaz.

(1) Má-li $\bar{p}$ být hustotou $X^{(\cdot)}$, musí splňovat vztah

$$P(X^{(\cdot)} \in B) = \int_B \ldots \int \bar{p}(x^{(1)},\ldots,x^{(n)}) dx^{(1)} \ldots dx^{(n)} \tag{2.7}$$

pro lib. $B \in \mathcal{B}^{(\cdot)}$.

Platí

$$P(X^{(\cdot)} \in B) = \sum_{r \in \mathcal{R}} P(X^{(\cdot)} \in B,\ R=r) =$$

$$= \sum_{r \in \mathcal{R}} \int_{\substack{x^{(\cdot)} \in B \\ R=r}} \cdots \int p(x_1,\ldots,x_n)\, dx_1 \ldots dx_n =$$

$$= \sum_{r \in \mathcal{R}} \int_B \cdots \int p(x^{(r_1)},\ldots,x^{(r_n)})\, dx^{(1)} \ldots dx^{(n)} =$$

$$= \int_B \cdots \int \bar{p}(x^{(1)},\ldots,x^{(n)})\, dx^{(1)} \ldots dx^{(n)}, \quad \text{což dokazuje (1).}$$

(2) Pro $B \in \mathcal{B}^{(\cdot)}$ a $r \in \mathcal{R}$ platí

$$P(X^{(\cdot)} \in B,\ R=r) = \int_B \cdots \int P(R=r \mid X^{(\cdot)} = x^{(\cdot)})\, \bar{p}(x^{(1)},\ldots,$$

$$\ldots,x^{(n)})\, dx^{(1)} \ldots dx^{(n)}.$$

Podobně jako u (1) můžeme psát

$$P(X^{(\cdot)} \in B, R=r) = \int_B \cdots \int p(x^{(r_1)},\ldots,x^{(r_n)})\, dx^{(1)} \ldots dx^{(n)} =$$

(2.8)

$$= \int_B \cdots \int \frac{p(x^{(r_1)},\ldots,x^{(r_n)})}{\bar{p}(x^{(1)},\ldots,x^{(x)})}\, \bar{p}(x^{(1)},\ldots,x^{(x)})\, dx^{(1)} \ldots dx^{(x)};$$

úprava integrandu je přípustná vzhledem k tomu, že z $\bar{p}(x^{(1)},\ldots,x^{(n)}) = 0$ plyne $p(x^{(r_1)},\ldots,x^{(r_n)}) = 0$ pro vš. $r \in \mathcal{R}$. Z (2.8) plyne (2.6).

VĚTA 4. Nechť H_* je systém hustot p takových, že $p(x_1,\ldots,x_n) = p(x_{i_1},\ldots,x_{i_n})$ pro libovolnou permutaci $i_1,\ldots,i_n$ posloupnosti $1,\ldots,n$. Jestliže $\underset{\sim}{X}$ má hustotu $p \in H_*$,

jsou vektor pořadí R a vektor pořádkových statistik $X^{(\cdot)}$ vzájemně nezávislé; R má rozdělení pravděpodobností

$$(2.9) \qquad P(R=r) = \frac{1}{n!} \quad , \qquad r \in \mathcal{R}$$

a $X^{(\cdot)}$ má rozdělení pravděpodobností s hustotou

$$(2.10) \qquad \bar{p}(x^{(1)},\dots,x^{(n)}) = \begin{cases} n!\, p(x^{(1)},\dots,x^{(n)}) \dots x^{(\cdot)} \in \mathcal{X}^{(\cdot)} \\ 0 \dots \text{jinde.} \end{cases}$$

Důkaz. (2.10) plyne přímo z (2.5). Dále pro lib. $r \in \mathcal{R}$ platí

$$P(R=r) = \int_{R=r} \dots \int p(x_1,\dots,x_n)\,dx_1,\dots dx_n =$$

$$= \int_{(x^{(1)},\dots x^{(n)}) \in \mathcal{X}^{(\cdot)}} \dots \int p(x^{r_1)},\dots,x^{(r_n)})\,dx^{(1)} \dots dx^{(n)} =$$

$$= \int_{\mathcal{X}^{(\cdot)}} \dots \int p(x^{(1)},\dots,x^{(n)})\,dx^{(1)} \dots dx^{(n)} =$$

$$= P\{R = (1,\dots,n)\} \quad ,$$

odtud vyplývá (2.9).

Abychom dokázali nezávislost R a $X^{(\cdot)}$, stačí, když ověříme

$$(2.11) \qquad P(R=r \mid X^{(\cdot)}=x^{(\cdot)}) = P(R=r) = \frac{1}{n!}$$

pro lib. $x^{(\cdot)} \in \mathcal{X}^{(\cdot)}$ a $r \in \mathcal{R}$.

Z (2.5) však plyne

$$\bar{p}(x^{(1)},\dots,x^{(n)}) = n!\, p(x^{(r_1)},\dots,x^{(r_n)})$$

a tedy vzhledem k (2.6) platí

$$P(R=r \mid X^{(\cdot)}=x^{(\cdot)}) = \frac{1}{n!} \quad , \qquad r \in \mathcal{R},\ x^{(\cdot)} \in \mathcal{X}^{(\cdot)}.$$

Důsledek věty 4. Nechť $\underset{\sim}{X}$ splňuje hypotézu H_0. Pak R a $X^{(\cdot)}$ jsou nezávislé a jejich rozdělení jsou dána vztahy (2.9) a (2.10).

VĚTA 5. Nechť $\underset{\sim}{X}$ má rozdělení s hustotou $p \in H$ a nechť $T = T(\underset{\sim}{X})$ je statistika. Pak platí

$$(2.12) \qquad E\left[T(X_1,\dots,X_n) \,\middle|\, R=r\right] = E\left[T(X^{(r_1)},\dots,X^{(r_n)})\right] .$$

Jestliže Q je rozdělení pravděpodobnostní s hustotou q takovou, že $q(x_1,\dots,x_n)=0$ implikuje $p(x_1,\dots,x_n) = 0$, platí pro lib. $r \in \mathcal{R}$

$$(2.13) \qquad Q(R=r) = \frac{1}{n!}\, E_p\left[\frac{q(X^{(r_1)},\dots,X^{(r_n)})}{p(X^{(r_1)},\dots,X^{(r_n)})}\right] .$$

Důkaz. Podle věty 4 jsou $X^{(\cdot)}$ a R nezávislé, proto platí

$$E\left[T(X_1,\dots,X_n) \,\middle|\, R=r\right] = E\left[T(X^{(r_1)},\dots,X^{(r_n)}) \,\middle|\, R=r\right] =$$

$$= E\left[T(X^{(r_1)},\dots,X^{(r_n)})\right] .$$

Na druhé straně z (2.10) plyne

$$Q(R=r) = \int_{R=r}\cdots\int q(x_1,\dots,x_n)\,dx_1 \cdots dx_n =$$

$$= \frac{1}{n!}\int_{\mathcal{X}^{(\cdot)}}\cdots\int \frac{q(x^{(r_1)},\dots,x^{(r_n)})}{p(x^{(r_1)},\dots,x^{(r_n)})}\; n!\; p(x^{(1)},\dots,x^{(n)})\,dx^{(1)} \cdots$$

$$\cdots dx^{(n)} = \frac{1}{n!}\, E_p\left[\frac{q(X^{(r_1)},\dots,X^{(r_n)})}{p(X^{(r_1)},\dots,X^{(r_n)})}\right] .$$

Na závěr kapitoly ještě uvedeme některé vlastnosti marginálních rozdělení vektorů R a $X^{(\cdot)}$ za platnosti hypotézy H_0.

VĚTA 6. Nechť X splňuje hypotézu H_0. Pak platí

(i) $P(R_i=j) = \frac{1}{n}$ pro vš. $j=1,\dots,n$; $i=1,\dots,n$.

(ii) $P(R_i=k, R_j=m) = \frac{1}{n(n-1)}$ pro $1 \leq i,j,k,m \leq n$, $i \neq j$, $k \neq m$

(iii) $E\,R_i = \frac{n+1}{2}$, $i=1,\dots,n$

(iv) $\mathrm{Var}\,R_i = \frac{n^2-1}{12}$, $i=1,\dots,n$

(v) $\mathrm{Cov}(R_i,R_j) = -\frac{n+1}{12}$, $1 \leq i,j \leq n$, $i \neq j$.

Důkaz.

(i) $$P(R_1=j) = \sum_{(j_2,\dots,j_n)} P(R_1=j, R_2=j_2,\dots,R_n=j_n) =$$

$$= \sum_{(j_2,\dots,j_n)} \frac{1}{n!} = \frac{(n-1)!}{n!},$$

kde sčítáme přes množinu permutací $(j_2,\dots,j_n)$ čísel $(1,\dots,j-1, j+1,\dots,n)$. Úvaha pro $i=2,\dots,n$ je analogická. Podobným způsobem dokážeme i (ii).

(iii) $$E\,R_i = \sum_{j=1}^{n} \frac{j}{n} = \frac{n+1}{2}.$$

(iv) $$\mathrm{Var}\,R_i = \frac{1}{n} \sum_{j=1}^{n} j^2 - \left(\frac{n+1}{2}\right)^2 = \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 =$$

$$= \frac{n^2-1}{12}.$$

(v) $$\mathrm{Cov}(R_i,R_j) = \sum_{\substack{k=1 \\ k\neq m}}^{n} \sum_{m=1}^{n} k.m \frac{1}{n(n-1)} - \left(\frac{n+1}{2}\right)^2 =$$

$$= \frac{1}{n(n-1)} \left[\left(\sum_{k=1}^{n} k \right)^2 - \sum_{k=1}^{n} k^2 \right] - \left(\frac{n+1}{2}\right)^2 = -\frac{n+1}{12}. \square$$

VĚTA 7. Nechť $\underset{\sim}{X}$ má rozdělení pravděpodobností s hustotou

$$p(x_1,\dots,x_n) = \prod_{i=1}^{n} f(x_i).$$

Pak $X^{(i)}$ má rozdělení s hustotou

$$(2.15) \qquad f_n^{(i)}(x) = n\binom{n-1}{i-1}(F(x))^{i-1}(1-F(x))^{n-i} f(x), \quad x \in R^1$$

kde $F(x) = \int_{-\infty}^{x} f(y)dy$, $i=1,\ldots,n$.

Důkaz. Pro distribuční funkci $X^{(i)}$ platí

$$P(X^{(i)} < x) = \sum_{j=1}^{n-1} P(X^{(j)} < x \leq X^{(j+1)}) + P(X^{(n)} < x) =$$

$$= \sum_{j=1}^{n} \binom{n}{j}(F(x))^{j}(1-F(x))^{n-j} ,$$

odkud dostaneme hustotu $f_n^{(i)}$ jako derivaci.

Speciálně, jestliže $\underset{\sim}{X}$ je náhodný výběr z rovnoměrného rozdělení $R[0,1]$, řídí se $X^{(i)}$ beta rozdělením $B(i,n-i+1)$ a má střední hodnotu a rozptyl

$$E\, X^{(i)} = \frac{i}{n+1}, \quad \mathrm{Var}\, X^{(i)} = \frac{i(n-i+1)}{(n+1)^2(n+2)} .$$

Problémy a cvičení

(1) Nechť $X_1,\ldots,X_n$ je náhodný výběr z $R(0,1)$.
Pak platí

$$\mathrm{Cov}(X^{(i)},X^{(j)}) = \frac{i(n-j+1)}{(n+1)^2(n+2)}$$

(2) Nechť $X_1,\ldots,X_n$ je náhodný výběr z rozdělení s distribuční funkcí F a hustotou f. Pak sdružené rozdělení $X^{(i_1)},\ldots,X^{(i_k)}$; $i_1 < i_2 < \ldots < i_k$; $k \leq n$, má hustotu

$$\frac{n!\, f(y_1)\ldots f(y_k)}{(i_1-1)(i_2-i_1-1)!\ldots(n-i_k)!} \cdot (F(y_1))^{i_1-1}(F(y_2)-F(y_1))^{i_2-i_1-1}\ldots$$

$$\ldots (1-F(y_k))^{n-i_k} \qquad \text{pro } y_1 < y_2 < \ldots < y_k.$$

(3) Nechť $X_1,\dots,X_n$ je náhodný výběr z rozdělení s distribuční funkcí F a hustotou f. Pak <u>výběrové rozpětí</u> $D = X^{(n)} - X^{(1)}$ má hustotu

$$f_D(y) = n(n-1)\int_{-\infty}^{\infty}\left[F(x)-F(x-y)\right]^{n-2} f(x-y)f(x)dx.$$

Speciálně, je-li f hustota $R(0,1)$, má D beta rozdělení $B(n-1,2)$.