KAPITOLA 3.

Testy hypotézy o shodnosti dvou populací

V této kapitole budeme uvažovat problém srovnání dvou populací s distribučními funkcemi F a G na základě dvou nezávislých náhodných výběrů, $X_1,\dots,X_m$ z první populace a $Y_1,\dots,Y_n$ z druhé populace. Experimentální pozadí problému se diskutovalo v příkladu 2 kapitoly 1.

V celé kapitole budeme uvažovat základní hypotézu H_0:$F=G$. Chceme-li nalézt vhodné testy této hypotézy, nemůžeme uvažovat hypotézu samostatně, ale vždy jen ve spojení s příslušnou alternativou. Volba testu pak úzce závisí na příslušné alternativě a volba alternativy závisí na tom, co chceme a můžeme o rozděleních F a G předpokládat. Jednotlivé části této kapitoly budou věnovány jednotlivým alternativám; budeme podle možnosti postupovat od speciálnějších (užších) k obecnějším (širším) alternativám.

Během celé kapitoly budeme předpokládat, že F a G jsou absolutně spojité s hustotami f a g.

3.1. Dvouvýběrový t-test

Nechť F a G jsou distribuční funkce normálního rozdělení $N(\xi,\sigma^2)$ a $N(\nu,\sigma^2)$, kde ξ, ν a σ^2 jsou neznámé. Uvažujme hypotézu H_0, kterou lze v tomto případě psát ve tvaru

(3.1) $$H_0 : \xi = \nu$$

proti tzv. jednostranné alternativě

$$K_1 : \xi < \nu .$$

Jak je známo ze základního kursu matematické statistiky, v tomto případě existuje stejnoměrně nejsilnější nestranný α-test H_0 proti K_1 a je jím dvouvýběrový t-test s kritickou funkcí

$$\varphi_1(\underset{\sim}{X},\underset{\sim}{Y}) = \begin{cases} 1 & \dots \quad t(\underset{\sim}{X},\underset{\sim}{Y}) \geq C_0 \\ 0 & \dots \quad t(\underset{\sim}{X},\underset{\sim}{Y}) < C_0 , \end{cases}$$

kde

$$t(\underset{\sim}{X},\underset{\sim}{Y}) = \frac{(\bar{Y}-\bar{X})\sqrt{\frac{1}{m}+\frac{1}{n}}}{\left\{\left[\sum_{i=1}^{m}(X_i-\bar{X})^2 + \sum_{j=1}^{n}(Y_j-\bar{Y})^2\right] / (m+n-2)\right\}^{1/2}}$$

a $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$, $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$.

Jestliže platí $\xi = \nu$, má $t(\underset{\sim}{X},\underset{\sim}{Y})$ Studentovo t-rozdělení o $(m+n-2)$ stupních volnosti a konstanta C_0 je určena vztahem

$$\int_{C_0}^{\infty} t_{m+n-2}(y)\,dy = \alpha .$$

Jestliže uvažujeme hypotézu H_0, tj. (3.1), proti oboustranné alternativě

$$K_1' : \xi \neq \nu$$

má nejsilnější nestranný α-test tvar

$$\varphi_1(\underset{\sim}{X},\underset{\sim}{Y}) = \begin{cases} 1 & \dots \quad |t(\underset{\sim}{X},\underset{\sim}{Y})| \geq C \\ 0 & \dots \quad |t(\underset{\sim}{X},\underset{\sim}{Y})| < C, \end{cases}$$

kde C je určeno vztahem

$$\int_{C}^{\infty} t_{m+n-2}(y)dy = \frac{\alpha}{2} \quad .$$

3.2. F - test

Nechť F a G jsou distribuční funkce normálního rozdělení $N(\xi, \sigma_1^2)$, $N(\xi, \sigma_2^2)$ kde ξ, σ_1^2, σ_2^2 jsou neznámé parametry. Uvažujeme hypotézu H_0, která má zde tvar

$$H_0 : \sigma_1^2 = \sigma_2^2$$

proti alternativě

$$K_2 : \sigma_1^2 < \sigma_2^2 \ .$$

Opět z klasické matematické statistiky je známo, že existuje stejnoměrně nejsilnější nestranný α-test H_0 proti K_2 s testovou funkcí

$$\Phi_2(\underset{\sim}{X},\underset{\sim}{Y}) = \begin{cases} 1 & \dots \quad a(\underset{\sim}{X},\underset{\sim}{Y}) \geq C_0 \\ 0 & \dots \quad a(\underset{\sim}{X},\underset{\sim}{Y}) < C_0 \end{cases},$$

kde

$$a(\underset{\sim}{X},\underset{\sim}{Y}) = \frac{\sum_{j=1}^{n} (Y_j - \overline{Y})^2 / (n-1)}{\sum_{i=1}^{m} (X_i - \overline{X})^2 / (m-1)} \quad .$$

Jestliže platí $\sigma_1^2 = \sigma_2^2$, má $a(\underset{\sim}{X},\underset{\sim}{Y})$ rozdělení F o n-1 a m-1 stupních volnosti. Konstanta C_0 je pak určena vztahem

$$\int_{C_0}^{\infty} F^*_{n-1,m-1}(y)dy = \alpha,$$

kde F^* značí hustotu F-rozdělení.

3.3. Permutační t-test

Chceme testovat hypotézu H_o proti alternativě

(3.2) $K_3 : G(x) = F(x-\Delta)\ ; \quad x \in R^1\ , \quad \Delta > 0,$

a přitom máme podezření, že F je distribuční funkce normálního rozdělení, ale nejsme si tím stoprocentně jisti. Je zřejmé, že taková situace se často vyskytuje.

Kdybychom použili t-test a skutečné rozdělení by nebylo normální, byla by pravděpodobnost chyby I.druhu obecně odlišná od α , a tedy mohla by být i větší než α . Abychom stabilizovali velikost testu, je vhodné použít tzv. permutačního testu, který nyní odvodíme.

Předpokládejme, že $X_1,\dots,X_m$ je náhodný výběr z rozdělení s hustotou $f(x)$ a $Y_1,\dots,Y_n$ náhodný výběr z rozdělení s hustotou $f(y-\Delta)$ a předpokládejme, že hustota f je skoro všude spojitá, ale jinak neznámá. Označme $\mathcal{F}$ systém všech takových hustot.

Sdružená hustota náhodného vektoru $(X_1,\dots,X_m, Y_1,\dots, \dots,Y_n) = (Z_1,\dots,Z_N)$, $N=m+n$, pak je

(3.3) $$q_\Delta(\underset{\sim}{z}) = f(x_1) \dots f(x_m) f(y_1-\Delta) \dots f(y_n-\Delta).$$

Chceme, aby použitý test $\Phi(\underset{\sim}{z})$ byl nestranný (viz def. 1 kap.1); odtud plyne, že musí platit

(3.4) $$\int \dots \int \Phi(z_1,\dots,z_N)\, f(z_1)\dots f(z_N)\, dz_1 \dots dz_N = \alpha$$

pro všecka $f \in \mathcal{F}$. Následující věta dává jednoduchou charakteristiku testů, splňujících (3.4).

VĚTA 1. Nechť $\mathcal{F}$ je systém všech sk.vš. spojitých hustot f. Nechť $(Z^{(1)},\dots,Z^{(N)})$ je vektor pořádkových statistik odpovídající $Z_1,\dots,Z_N$. Pak (3.4) platí pro vš. $f\in\mathcal{F}$ právě tehdy, když

$$\frac{1}{N!}\sum_{r\in\mathcal{R}}\Phi(Z^{(r_1)},\dots,Z^{(r_N)}) = \alpha \tag{3.5}$$

platí s pravděpodobností 1, kde $\mathcal{R}$ je množina všech permutací r čísel $(1,\dots,N)$.

Důkaz. Podle věty 2 kapitoly 1 je vektor $Z^{(\cdot)}$ pořádkových statistik úplnou postačující statistikou pro $\mathcal{F}$. Rozdělení náhodného vektoru $\underset{\sim}{Z}$ podmíněné $Z^{(\cdot)} = z^{(\cdot)}$ je tedy za platnosti H_0 stejné pro vš. $f\in\mathcal{F}$ a tudíž i $E\left[\Phi(\underset{\sim}{Z})\,\middle|\, Z^{(\cdot)} = z^{(\cdot)}\right]$ je za H_0 stejná pro vš. $f\in\mathcal{F}$ a sk.vš. $z^{(\cdot)}$. Má-li platit (3.4), musí být i

$$E\left[\Phi(\underset{\sim}{Z})\,\middle|\, Z^{(\cdot)} = z^{(\cdot)}\right] = \alpha \quad \text{pro sk.vš.} \quad z^{(\cdot)}$$

a ze tvaru podmíněného rozdělení $\underset{\sim}{Z}$ při daném $Z^{(\cdot)} = z^{(\cdot)}$ (viz (2.4)) vyplývá

$$\alpha = E\left[\Phi(\underset{\sim}{Z})\,\middle|\, Z^{(\cdot)} = z^{(\cdot)}\right] = \frac{1}{N!}\sum_{r\in\mathcal{R}}\Phi(z^{(r_1)},\dots,z^{(r_N)}) . \qquad \square$$

Testy splňující (3.5) se nazývají permutační testy. Stanovme sílu permutačního testu Φ nejprve proti jednoduché alternativě, že sdružená hustota vektoru $\underset{\sim}{Z}$ je rovna $q(z_1,\dots,z_N)$. Dostaneme

$$\int\cdots\int\Phi(\underset{\sim}{z})q(\underset{\sim}{z})dz_1\cdots dz_N = \sum_{r\in\mathcal{R}}\int_{R=r}\cdots\int\Phi(\underset{\sim}{z})q(\underset{\sim}{z})dz_1\cdots dz_N =$$

$$= \sum_{r\in\mathcal{R}}\int\cdots\int\Phi(z^{(r_1)},\dots,z^{(r_N)})q(z^{(r_1)},\dots,z^{(r_N)})dz^{(1)}\cdots z^{(N)} \tag{3.6}$$

$$= \int \dots \int \frac{\sum_{r \in \mathcal{R}} \overline{\Phi}(z^{(r_1)}, \dots, z^{(r_N)}) q(z^{(r_1)}, \dots, z^{(r_N)})}{\overline{q}(z^{(r_1)}, \dots, z^{(r_N)})} \cdot$$

$$\cdot \overline{q}(z^{(r_1)}, \dots, z^{(r_N)}) \, dz^{(1)} \dots dz^{(N)} ,$$

kde $\overline{q}$ je hustota $(Z^{(1)}, \dots, Z^{(N)})$ příslušná hustotě q vektoru $(Z_1, \dots, Z_N)$ (viz kap.2, věta 2). Má-li test maximalizovat silofunkci (3.6) za podmínky, že splňuje (3.5) pro sk.vš. $(z^{(1)}, \dots, z^{(N)})$, musí pro sk.vš. $(z^{(1)}, \dots, z^{(N)})$ maximalizovat výraz

$$\frac{\sum_{r \in \mathcal{R}} \overline{\Phi}(z^{(r_1)}, \dots, z^{(r_N)}) q(z^{(r_1)}, \dots, z^{(r_N)})}{\overline{q}(z^{(1)}, \dots, z^{(N)})} =$$

(3.7)

$$= \frac{\sum_{r \in \mathcal{R}} \overline{\Phi}(z^{(r_1)}, \dots, z^{(r_N)}) q(z^{(r_1)}, \dots, z^{(r_N)})}{\sum_{r' \in \mathcal{R}} q(z^{(r'_1)}, \dots, z^{(r'_N)})} = \max$$

mezi všemi testy splňujícími

$$(3.8) \qquad \frac{1}{N!} \sum_{r \in \mathcal{R}} \overline{\Phi}(z^{(r_1)}, \dots, z^{(r_N)}) = \alpha .$$

Při daném $(z^{(1)}, \dots, z^{(N)})$ test, optimální ve smyslu (3.7) a (3.8), podle Neyman-Pearsonova fundamentálního lemmatu zamítá hypotézu H_0 pro ty permutace $(z^{(r_1)}, \dots, z^{(r_N)})$, pro které podíl

$$\frac{q(z^{(r_1)}, \dots, z^{(r_N)}) \, N!}{\sum_{r \in \mathcal{R}} q(z^{(r'_1)}, \dots, z^{(r'_N)})}$$

nabývá velkých hodnot; nejsilnější test má tedy testovou funkci

$$(3.9)\qquad \Phi(z^{(r_1)},\dots,z^{(r_N)}) = \begin{cases} 1 & \dots\ q(z^{(r_1)},\dots,z^{(r_N)}) > C(z^{(\cdot)}) \\ \gamma & \dots\ q(z^{(r_1)},\dots,z^{(r_N)}) = C(z^{(\cdot)}) \\ 0 & \dots\ q(z^{(r_1)},\dots,z^{(r_N)}) < C(z^{(\cdot)}) \end{cases}$$

pro funkci $C(z^{(\cdot)})$ takovou, že platí (3.8). Prakticky to znamená, že $N!$ permutací hodnot $z^{(1)},\dots,z^{(N)}$ uspořádáme podle rostoucích hodnot $q(z^{(r_1)},\dots,z^{(r_N)})$ a zamítneme H_o pro k permutací s největšími hodnotami $q(z^{(r_1)},\dots,r^{(r_N)})$; pro $(k+1)$-ní největší hodnotu $q(z^{(r_1)},\dots,z^{(r_N)})$ zamítneme H_o s pravděpodobností γ, kde k a γ jsou určena vztahem

$$k + \gamma = \alpha N!.$$

Speciálně nás zajímá třída normálních alternativ tvaru (3.3), kde f probíhá hustoty $N(\mu,\sigma^2)$, $\mu\in R^1$, $\sigma^2 > 0$. Uvidíme, že nejsilnější permutační test proti těmto alternativám nezávisí na μ, σ^2 a Δ. Tento test je vhodný zvláště tehdy, když předpokládáme, že náhodné výběry pocházejí z normálního rozdělení, ale nejsme si tím zcela jisti; použitím permutačního testu máme zaručeno, že velikost nepřekročí předepsané α při žádné alternativě tvaru (3.3).

Při normální hustotě f nabývá alternativa (3.3) tvaru

$$(3.10)\qquad q(z_1,\dots,z_N) =$$

$$= (\sqrt{2\pi}\,\sigma)^{-1} \exp\left\{ -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{m} (z_i-\mu)^2 + \sum_{i=m+1}^{N} (z_i-\mu-\Delta)^2 \right]\right\}.$$

Protože faktor $\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Z_i-\mu)^2\right\}$ nezávisí na uspořádání $(z_1,\dots,z_N)$, test (3.9) zamítá H_o ve prospěch alternativy (3.10), jestliže $\exp\left\{\Delta\cdot\sum_{i=m+1}^{N} z_i\right\} > C(z^{(\cdot)})$, tedy

$$(3.11)\quad \varphi(z_1,\dots,z_N)=\begin{cases} 1 \dots \sum_{j=1}^{n} y_j = \sum_{i=m+1}^{N} z_i > C'(z^{(\cdot)}) \\ \gamma \dots \sum_{j=1}^{n} y_j = \sum_{i=m+1}^{N} z_i = C'(z^{(\cdot)}) \\ 0 \dots \sum_{j=1}^{n} y_j = \sum_{i=m+1}^{N} z_i < C'(z^{(\cdot)}). \end{cases}$$

Z hodnot, kterých testová statistika $\sum_{i=m+1}^{N} z_i$ nabývá na množině $N!$ permutací hodnot $(z^{(1)},\dots,z^{(N)})$, je zřejmě jen $\binom{N}{n}$ různých a stačí mezi sebou porovnat jen tyto různé hodnoty; hypotézu H_o zamítneme pro k největších hodnot $\sum_{i=m+1}^{N} z_i$ a s pravděpodobností γ pro $(k+1)$-ní hodnotu, kde

$$k+\gamma=\alpha\binom{N}{n}.$$

Shrneme postup, jak provádíme permutační test (3.11) proti alternativě 2 normálních výběrů lišících se posunutím:

(i) pozorujeme hodnoty $(x_1,\dots,x_m, y_1,\dots,y_n) = (z_1,\dots,z_N)$;

(ii) stanovíme $k\geq 0$, celé a γ, $0\leq\gamma<1$ tak, že $k+\gamma=\alpha\binom{N}{n}$.

(iii) vypočteme hodnoty $\sum_{i=m+1}^{N} z^{(r_i)}$ pro všecky permutace $r\in\mathcal{R}$; mezi těmito hodnotami je $\binom{N}{n}$ různých. Z těchto hodnot najdeme $(\binom{N}{n}-k+1)$-ní podle velikosti.

(iv) Zamítneme H_o, jestliže $\sum_{j=1}^{n} y_j$ je alespoň rovno této $(\binom{N}{n}-k+1)$-ní hodnotě a zamítneme H_o s pravděpodobností γ, jestliže $\sum_{j=1}^{n} y_j$ je rovno $(\binom{N}{n}-k)$-té hodnotě.

Nakonec uvedeme jinou formu permutačního testu (3.11), která ukáže souvislost tohoto testu s klasickým t-testem. Vynásobíme-li nerovnost

$$\sum_{j=1}^{n} y_j > C(z^{(\cdot)})$$

výrazem $(\frac{1}{m} + \frac{1}{n})$ a odečteme $\frac{1}{m}(\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} y_j)$, dostaneme

$$\bar{y} - \bar{x} > C'(z^{(\cdot)})(\frac{1}{m} + \frac{1}{n}) - \frac{1}{m}\sum_{i=1}^{N} z_i = C''(z^{(\cdot)})$$

a další úpravou dostaneme jinou ekvivalentní nerovnost

$$(3.12)\quad \frac{(\bar{y}-\bar{x})\sqrt{\frac{1}{m}+\frac{1}{n}}}{\left\{\left[\sum_{i=1}^{m}(x_i-\bar{x})^2+\sum_{j=1}^{n}(y_j-\bar{y})^2\right]/(m+n-2)\right\}^{\frac{1}{2}}} > C'''(z^{(\cdot)}).$$

Kritický obor permutačního testu má tedy tvar t-testu, ve kterém konstantní mez C_o je nahrazena náhodnou mezí $C'''(z^{(\cdot)})$.

3.4. Pořadové testy rozdílu v poloze dvou populací

Uvažujeme nejprve problém testu hypotézy H_o proti obecné alternativě

$$(3.13)\quad K_4 : G(x) \leq F(x) \text{ pro vš. } x \in R^1, \quad G \neq F,$$
$$G,F \text{ spojité, jinak neznámé.}$$

tj. proti jednostranné alternativě, která tvrdí, že náhodná veličina s rozdělením G je stochasticky větší než náhodná veličina s rozdělením F.

Problém testu H_o proti alternativě K_4 je invariantní vzhledem ke grupě $\mathcal{G}$ zobrazení tvaru

$$z_i' = f(z_i), \quad i=1,\dots,N$$

kde f je libovolná spojitá rostoucí funkce, $(z_1,\dots,z_N) = (x_1,\dots,x_m, y_1,\dots,y_n)$, $N = m+n$. Z příkladu 1, část 4, kapitola 1 vyplývá, že maximální invariantou vzhledem ke G je vektor pořadí $(R_1,\dots,R_N)$ náhodných veličin $(Z_1,\dots,Z_N) = (X_1,\dots,X_m, Y_1,\dots,Y_n)$.

Protože problém testu H_o proti K_4 je invariantní vzhledem ke $\mathcal{G}$, bude vhodné omezit své úvahy na testy, které jsou také invariantní vzhledem ke $\mathcal{G}$. Z věty 2 kapitoly 1 víme, že každý invariantní test je funkcí maximální invarianty. V našem případě bude každý invariantní test funkcí pouze pořadí $(R_1,\dots,R_N)$, tedy bude <u>pořadovým testem.</u>

Vzhledem k invarianci omezujeme tedy své úvahy na pořadové testy. Třídu testů však můžeme omezit ještě dále. Najdeme-li vhodné zobrazení vektoru $(R_1,\dots,R_N)$, které je postačující statistikou pro $(R_1,\dots,R_N)$ za hypotézy i za alternativy, stačí hledat vhodný test mezi testy, závislými jen na této postačující statistice. Protože rozdělení náhodného vektoru $(R_1,\dots,R_m, R_{m+1},\dots,R_N)$ je symetrické v prvních m a posledních n proměnných při všech rozděleních F a G, vidíme, že postačující statistikou jsou pořadí

$$R_1' < \dots < R_m' \quad \text{a} \quad R_{m+1}' < \dots < R_N'$$

veličin $X_1,\dots,X_m$ a $Y_1,\dots,Y_n$ ve spojeném výběru, uspořádaná podle velikosti. Protože kterákoli z množin ve (3.14) jednoznačně určuje druhou, vidíme, že množina invariantních

testů se nakonec redukuje na testy, které jsou funkcemi uspořádaných pořadí jednoho z výběrů, např. výběru $Y_1,\dots,Y_n$.

Vektor $(R'_{m+1},\dots,R'_N)$ může být roven všem kombinacím n-té třídy z čísel $1,\dots,N$; může tedy nabývat celkem $\binom{N}{n}$ různých hodnot. Za platnosti H_0 jsou všecky tyto hodnoty stejně pravděpodobné a kritický obor libovolného pořadového testu velikosti $\alpha = k/\binom{N}{n}$ se skládá z právě k bodů $(s_1,\dots,s_n)$, kde $1 \leq s_1 < \dots < s_n \leq N$. Podle toho, které body zahrneme do kritického oboru, dostaneme různé pořadové testy. Bohužel mezi těmito testy neexistuje stejnoměrně nejsilnější pořadový test a tedy ani stejnoměrně nejsilnější invariantní test. O tom se přesvědčíme, až se seznámíme s různými standardními pořadovými testy: různé testy budou nejsilnější proti různým podmnožinám alternativy K_4.

3.4.1. Wilcoxonův test

Uvažujme opět alternativu posunutí v poloze $K_3: G(x) = F(x-\Delta)$, $x \in R^1$, $\Delta > 0$. Kdybychom věděli, že rozdělení F je normální, použili bychom t-testu uvedeného v 3.1.

Nyní zkonstruujeme jinou testovou statistiku tak, že dosadíme do výrazu pro t-statistiku $t(\underset{\sim}{X},\underset{\sim}{Y})$ místo hodnot $X_1,\dots,X_m, Y_1,\dots,Y_n$ příslušná pořadí $R_1,\dots,R_m, R_{m+1},\dots,R_N$. Dostaneme výraz

$$(3.14)\quad \left(\frac{mn}{N}\right)^{1/2}\left(\frac{1}{n}\sum_{j=m+1}^{N} R_j - \frac{1}{m}\sum_{i=1}^{m} R_i\right)\left[\frac{1}{N-2}\sum_{i=1}^{N}(R_i-\bar{R})^2\right]^{-1/2}$$

$$\text{kde } \bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i.$$

Tento výraz lze dále zjednodušit: z toho, že $(R_1,\dots,R_N)$ je

vlastně permutací $(1,\dots,N)$, plynou jednoduché identity:

$$\sum_{i=1}^{m} R_i = \sum_{i=1}^{N} R_i - \sum_{j=m+1}^{N} R_j = \frac{N(N+1)}{2} - \sum_{j=m+1}^{N} R_j$$

$$\bar{R} = \frac{N+1}{2}$$

$$\sum_{i=1}^{N} (R_i - \bar{R})^2 = \sum_{j=1}^{N} (j - \frac{N+1}{2})^2 = \frac{N(N^2-1)}{12} .$$

Využijeme-li těchto identit, vidíme, že statistika (3.14) je, až na lineární transformaci, ekvivalentní statistice

$$W = \sum_{i=m+1}^{N} R_i , \tag{3.15}$$

která je rovna součtu pořadí druhého výběru. Příslušná testová funkce má tvar

$$\Phi(R_1,\dots,R_N) = \begin{cases} 1 & \dots \quad W > C_\alpha \\ \gamma & \dots \quad W = C_\alpha \\ 0 & \dots \quad W < C_\alpha , \end{cases} \tag{3.16}$$

kde C_α je určeno tak, aby $P(W>C_\alpha) + \gamma P(W=C_\alpha) = \alpha$.

Test daný vztahy (3.15) a (3.16) se nazývá <u>Wilcoxonův test</u>. Patří do množiny invariantních testů, kterou jsme výše vymezili, tj. testů, závislých jen na uspořádaných pořadích druhého výběru.

Později uvidíme, že Wilcoxonův test je lokálně nejsilnějším pořadovým testem proti alternativám posunutí K_3, kde F je distribuční funkce logistického rozdělení s hustotou

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2} , \qquad x \in R^1 .$$

Vektor uspořádaných pořadí $R'_{m+1},\dots,R'_N$ nabývá za H_o každé kombinace n prvků z čísel $1,\dots,N$ s pravděpodobností $\frac{1}{\binom{N}{n}}$. Při malých hodnotách N můžeme stanovit rozdělení pravděpodobností W tak, že pro každou možnou hodnotu $W = w$ stanovíme počet kombinací (řekněme k_w) k ní vedoucích; pak $P(W=w|H_o) = \frac{k_w}{\binom{N}{n}}$. Odtud můžeme stanovit kritickou hodnotu C_α. Tento systematický způsob se však stává velmi pracný při větších N a n. Naštěstí existuje řada tabulek kritických hodnot Wilcoxonova testu (malou tabulku viz též na konci skript); dále např. Likeš-Laga: Základní statistické tabulky, SNTL 1978.

Rozdělení Wilcoxonovy statistiky bývá tabelováno různým způsobem: buď jsou tabelovány přímo kritické hodnoty nebo distribuční funkce W_N za platnosti H_o pro různá m a n. Často bývá tabelováno rozdělení statistiky

$$U_N = \sum_{i=m+1}^{N} \sum_{j=1}^{m} u(Z_i - Z_j)$$

kde $u(t) = 1$ pro $t \geq 0$ a $u(t)=0$ pro $t < 0$. Statistiky U_N a W_N jsou ve vzájemném vztahu

$$W_N = U_N + \frac{1}{2} n(n+1).$$

Pro výpočet Wilcoxonovy statistiky bývá vhodné její <u>duální vyjádření</u>: nechť $Z^{(1)} < \dots < Z^{(N)}$ jsou pořádkové statistiky příslušné náhodnému vektoru $(Z_1,\dots,Z_N) = (X_1,\dots,X_m, Y_1,\dots,Y_N)$. Definujme náhodné veličiny $V_1,\dots,V_N$ takto: $V_i = 0$ jestliže $Z^{(i)}$ pochází z 1.výběru a $V_i = 1$ jestliže $Z^{(i)}$ pochází z 2.výběru, $i=1,\dots,N$. Pak platí

(3.17) $$W_N = \sum_{i=1}^{N} i\, V_i.$$

Při velkých hodnotách m a n již nejsou k dispozici tabulky kritických hodnot W_N; v tom případě lze však kritické hodnoty stanovit pomocí normální aproximace rozdělení W_N. Statistika W_N totiž má za platnosti H_0 při $\min(m,n) \to \infty$ asymptoticky normální rozdělení v tom smyslu, že platí

$$(3.18) \qquad \lim_{\min(m,n)\to\infty} P\left\{ \frac{W_N - E\, W_N}{(\operatorname{var} W_N)^{1/2}} < x \,\middle|\, H_0 \right\} = \Phi(x), \quad x \in R^1,$$

kde Φ je distribuční funkce normálního rozdělení $N(0,1)$. Abychom mohli aproximace (3.18) použít ke stanovení kritických hodnot, potřebujeme střední hodnotu a rozptyl W_N za H_0. V následující větě odvodíme střední hodnotu a rozptyl poněkud obecnější statistiky.

VĚTA 2. Nechť náhodný vektor $(R_1,\dots,R_N)$ má diskrétní rovnoměrné rozdělení na množině všech permutací $\mathcal{R}$ čísel $1,\dots,N$; tj. $P(R=r) = \frac{1}{N!}$ pro lib. $r \in \mathcal{R}$; nechť $c_1,\dots,c_N$ a $a_1 = a(1),\dots,a_N = a(N)$ jsou libovolné konstanty. Pak střední hodnota a rozptyl statistiky

$$(3.19) \qquad S_N = \sum_{i=1}^{N} c_i\, a(R_i)$$

jsou rovny

$$(3.20) \qquad E\, S_N = \frac{1}{N} \sum_{i=1}^{N} c_i \sum_{j=1}^{N} a_j,$$

$$(3.21) \qquad \operatorname{var} S_N = \frac{1}{N-1} \sum_{i=1}^{N} (c_i - \bar{c})^2 \sum_{j=1}^{N} (a_j - \bar{a})^2,$$

kde $\bar{c} = \frac{1}{N} \sum_{i=1}^{N} c_i$, $\bar{a} = \frac{1}{N} \sum_{j=1}^{N} a_j$.

Důkaz. Pro střední hodnotu platí

$$E\,S_N = \sum_{i=1}^{N} c_i \cdot E\,a(R_i) = \sum_{i=1}^{N} c_i \cdot \frac{1}{N} \sum_{j=1}^{N} a_j \cdot$$

Rozptyl můžeme psát ve tvaru

$$\operatorname{var} S_N = \sum_{i=1}^{N} c_i^2 \operatorname{var}\ a(R_i) + \sum_{i=j} \sum c_i c_j \operatorname{cov}(a(R_i), a(R_j)) =$$

$$= \operatorname{var} a(R_1) \sum_{i=1}^{N} c_i^2 + \operatorname{cov}(a(R_1), a(R_2)) \sum_{i \neq j} \sum c_i c_j =$$

$$= \operatorname{cov}(a(R_1), a(R_2)).N^2\bar{c}^2 + \left[\operatorname{var} a(R_1) - \operatorname{cov}(a(R_1), a(R_2))\right] \sum_{i=1}^{N} c_i^2 .$$

Podle věty 6 kapitoly 2 dostaneme

$$\operatorname{var} a(R_1) = \frac{1}{N} \sum_{i=1}^{N} (a_i - \bar{a})^2 ,$$

$$\operatorname{cov}(a(R_1), a(R_2)) = -\frac{1}{N(N-1)} \sum_{i=1}^{N} (a_i - \bar{a})^2 ,$$

a tedy

$$\operatorname{var} S_N = -\frac{1}{N-1} \sum_{i=1}^{N} (a_i - \bar{a})^2 . N\bar{c}^2 + \frac{1}{N-1} \sum_{i=1}^{N} (a_i - \bar{a})^2 . \sum_{j=1}^{N} c_j^2 .$$

Speciálně, střední hodnota a rozptyl Wilcoxonovy statistiky za H_0 jsou rovny

$$E\,W_N = \frac{n(N+1)}{2}$$

(3.22)

$$\operatorname{var} W_N = \frac{mn(N+1)}{12} .$$

Rozdělení W_N je za platnosti H_0 symetrické kolem $E\,W_N$. Této vlastnosti využijeme při práci s tabulkami kritických hodnot: testujeme-li např. H_0 proti alternativě, že druhý výběr je posunut vlevo vzhledem k prvnímu, zamítáme H_0, jestliže $W_N < 2EW_N - C_\alpha$, kde C_α je kritická hodnota ze vztahu (3.16).

Symetrie rozdělení W_N plyne z následující věty:

VĚTA 3. Nechť náhodný vektor $(R_1,\dots,R_N)$ má diskrétní rovnoměrné rozdělení na množině všech permutací $\mathcal{R}$ čísel $1,\dots,N$, nechť $c_1,\dots,c_N$ a $a_1 = a(1),\dots,a_N = a(N)$ jsou konstanty takové, že platí buď

(a) $a_i + a_{N-i+1} = \text{konst}$, $i=1,\dots,N$

nebo

(b) $c_i + c_{N-i+1} = \text{konst}$, $i=1,\dots,N$.

Pak rozdělení pravděpodobností statistiky $S_N = \sum_{i=1}^{N} c_i\, a(R_i)$ je symetrické kolem ES_N, tj. $S_N - ES_N$ má stejné rozdělení jako $-(S_N - ES_N)$.

Důkaz.

Jestliže platí (a), pak

$$2N\bar{a} = \sum_{i=1}^{N} a_i + \sum_{i=1}^{N} a_{N-i+1} = N.\text{konst} \text{ , a tedy}$$

$$a_i + a_{N-i+1} = 2\bar{a} \text{ , } i=1,\dots,N.$$

Označme $S'_N = \sum_{i=1}^{N} c_i\, a(N-R_i+1)$. Protože $(N-R_1+1,\dots,N-R_N+1)$ má stejné rozdělení jako $(R_1,\dots,R_N)$, má S'_N stejné rozdělení jako S_N a platí

$$S'_N = \sum_{i=1}^{N} c_i\, a(N-R_i+1) = 2\bar{a} \sum_{i=1}^{N} c_i - S_N = 2ES_N - S_N$$

$$\Leftrightarrow \quad S'_N - ES'_N = ES_N - S_N$$

$$\Leftrightarrow \quad P(S_N - ES_N = s) = P(S'_N - ES'_N = s) = P(ES_N - S_N = s)$$ platí pro libovolné s. Tím je dokázána 1.část.

Jestliže platí (b), pak podobně $c_i + c_{N-i+1} = 2\bar{c}$, $i=1,\dots,N$; $(R_N,\dots,R_1)$ má stejné rozdělení jako $(R_1,\dots,R_N)$, a tedy

$S''_N = \sum_{i=1}^{N} c_{N-i+1}\, a(R_i) = \sum_{i=1}^{N} c_i\, a(R_{N-i+1})$ má stejné rozdělení jako S_N; dále platí

$$S''_N = 2\bar{c} \sum_{i=1}^{N} a_i - S_N = 2ES_N - S_N$$

$$\Longrightarrow \quad S''_N - ES_N = ES_N - S_N$$

a důkaz dokončíme podobně jako v případě (a). □

3.4.2. Van der Waerdenův test

Obecnou třídu testových statistik pro testování H_0 proti K_3 : $G(x) = F(x-\Delta)$, $\Delta>0$, tvoří statistiky tvaru

$$S_N = \sum_{i=m+1}^{N} h\left(\frac{R_i}{N+1}\right) = \sum_{i=1}^{N} h\left(\frac{i}{N+1}\right) V_i \,, \tag{3.23}$$

kde h je vhodná neklesající funkce definovaná na $(0,1)$. Tato třída v sobě zahrnuje i Wilcoxonův test ($h(t) = t$, $0<t<1$). Podrobněji si všimneme 2 testů typu (3.23), které se v praxi často používají: van der Waerdenova a mediánového testu. Van der Waerdenův test je vytvořen funkcí

$$h(t) = \Phi^{-1}(t), \quad 0<t<1, \tag{3.24}$$

kde Φ^{-1} je inverzní distribuční funkce normálního rozdělení $N(0,1)$. Tento test je vhodný pro alternativy typu K_3, je-li F distribuční funkce přibližně normálního typu.

Kritické hodnoty testu tabelovali např. B.L.van der Waerden a E.Nievergelt (1956): Tables for Comparing Two Samples by X-test and Sign Test (Springer, Berlín).

Při velkých hodnotách m a n můžeme kritické hodnoty stanovit pomocí normální aproximace rozdělení testové statis-

tiky. Při $\min(m,n)\to\infty$ je rozdělení S_N přibližně normální s parametry ES_N, var S_N. Podle věty 2 dostaneme

$$ES_N = 0, \quad \text{var } S_N = \frac{mn}{N(N-1)} \sum_{i=1}^{N} \left[\Phi^{-1}\left(\frac{i}{N+1}\right) \right]^2 .$$

Z věty 3 dále plyne, že rozdělení van der Waerdenovy statistiky je symetrické kolem nuly.

3.4.3. Mediánový test

Položíme-li ve (3.23)

$$h(t) = \begin{cases} 0 & \dots \quad 0<t<\frac{1}{2} \\ \frac{1}{2} & \dots \quad t=\frac{1}{2} \\ 1 & \dots \quad \frac{1}{2}<t<1, \end{cases}$$

dostaneme testovou statistiku tzv. mediánového testu. Testová statistika má jednoduchou interpretaci: nechť μ je medián spojeného výběru $X_1,\dots,X_m$, $Y_1,\dots,Y_n$. Mediánová statistika je rovna počtu pozorování 2.výběru, která leží nad μ, zvětšená o $\frac{1}{2}$ v případě, že N je liché.

Jestliže N je sudé a $M=\frac{N}{2}$, má mediánová statistika za platnosti H_o hypergeometrické rozdělení pravděpodobností:

$$(3.25)\quad P(S_N=k) = \begin{cases} \dfrac{\binom{M}{k}\binom{M}{n-k}}{\binom{N}{n}} & \dots \quad \max(0,n-M) \le k \le \\ & \qquad \le \min(M,n) \\ 0 & \dots \quad \text{jinak.} \end{cases}$$

Skutečně, vektor uspořádaných pořadí $R'_{m+1}<\dots<R'_N$ veličin $Y_1,\dots,Y_n$ nabývá $\binom{N}{n}$ možných kombinací, všech se

stejnou pravděpodobností. Jestliže $S_N = k$, je k prvků příslušné kombinace větších než μ a (n-k) prvků je menších než μ -takových kombinací je celkem $\binom{M}{k}\binom{M}{n-k}$.

Kritické hodnoty mediánového testu tedy můžeme získat z tabulek hypergeometrického rozdělení (např. G.J.Lieberman, D.B.Owen (1961): "Tables of the Hypergeometric Probability Distribution" a D.B.Owen:"Handbook of Statistical Tables" - ruský překlad 1966).

Pro $m,n \to \infty$ má mediánová statistika S_N přibližně normální rozdělení s parametry

$$ES_N = \frac{n}{2}, \qquad \text{var } S_N = \frac{1}{4}\frac{mn}{N-1}.$$

Při velkých počtech pozorování tedy používáme kritických hodnot založených na normální aproximaci.

Mediánový test používáme zvláště tehdy, má-li hustota příslušná distribuční funkci F protáhlé konce, tj. jestliže $\lim_{x \to \pm\infty} f(x) = 0$, ale konvergence je mnohem pomalejší než u normálního rozdělení. Např. pro Cauchyho rozdělení je mediánový test lepší než Wilcoxonův i van der Waerdenův test.

3.5. Pořadové testy rozdílu v rozptýlenosti dvou populací

Nechť $X_1,\dots,X_m$ je náhodný výběr z rozdělení se spojitou distribuční funkcí $F(x-\mu)$ a $Y_1,\dots,Y_n$ náhodný výběr z rozdělení $G(x-\mu)$, přičemž platí

(3.26) $K_5 : G(x-\mu) = F(\frac{x-\mu}{\sigma})$ pro vš. $x \in R^1$, $\sigma > 1$;

chceme testovat hypotézu o parametru σ, zatímco μ je nezná-

mý (rušivý) parametr.

Hypotéza H_0, že oba výběry pocházejí ze stejného rozdělení, v tomto případě říká, že $\sigma = 1$. Hledejme vhodný test H_0 proti alternativám K_5.

Alternativa K_5 je speciální případ následující obecnější alternativy: je-li $F(x)$ distribuční funkce $X_1,\dots,X_m$, $G(x)$ distribuční funkce $Y_1,\dots,Y_n$, pak platí pro nějaký bod μ:

$$(3.27)\qquad K_6 : F(x) \leq G(x) \quad \dots \quad -\infty < x \leq \mu$$
$$F(x) \geq G(x) \quad \dots \quad \mu \leq x < \infty .$$

Hypotéza H_0 i alternativa K_6 jsou invariantní vzhledem ke grupě $\mathcal{G}$ zobrazení tvaru $z_i' = f(z_i)$, $i=1,\dots,N$, kde f je libovolná spojitá rostoucí funkce, $(z_1,\dots,z_N) = (x_1,\dots,x_m;\ y_1,\dots,y_n)$, $N = m+n$. Invariance a postačitelnost opět redukují množinu testů na ty, které jsou funkcí uspořádaných pořadí $R'_{m+1},\dots,R'_N$ druhého výběru. Mezi těmito testy neexistuje α-test, stejnoměrně nejsilnější proti alternativě K_6. Obvykle opět volíme testové statistiky tvaru

$$(3.28)\qquad S = \sum_{i=m+1}^{N} h\left(\frac{R_i}{N+1}\right) = \sum_{i=1}^{N} h\left(\frac{i}{N+1}\right) V_i ,$$

ale s jinou funkcí h než u testů rozdílu v poloze. Z úvah o lokálně nejsilnějších testech vyplývá, že vhodné jsou funkce h konvexní (nebo konkávní) na $(0,1)$ a dosahující minima (nebo maxima) v $\frac{1}{2}$.

Pořadové testy rozptylu lze vytvořit z pořadových testů polohy vhodnou transformací, která spočívá v tom, že pořadí pozorování definujeme odlišným způsobem. Výhodou takto získa-

ných testů je, že rozdělení testových statistik je za platnosti H_0 stejné jako rozdělení příslušných testů polohy (a tedy kritické hodnoty najdeme v tabulkách kritických hodnot příslušného testu polohy).

Jako příklad uveďme Siegel-Tukeyho test, který je analogií Wilcoxonova testu. Vektor pořádkových statistik $Z^{(1)} < Z^{(2)} < \ldots < Z^{(N)}$, příslušný náhodnému vektoru $X_1,\ldots,X_m; Y_1,\ldots,Y_n$, přeuspořádáme takto:

$$(3.29)\quad Z^{(1)}, Z^{(N)}, Z^{(N-1)}, Z^{(2)}, Z^{(3)}, Z^{(N-2)}, Z^{(N-3)}, Z^{(4)}, Z^{(5)}, \ldots$$

a tuto novou posloupnost označme $\tilde{Z}^{(1)},\tilde{Z}^{(2)},\ldots,\tilde{Z}^{(N)}$; tedy

$$\tilde{Z}^{(1)}=Z^{(1)};\quad \tilde{Z}^{(4j-2)}=Z^{(N+2-2j)};\quad \tilde{Z}^{(4j-1)}=Z^{(N+1-2j)};$$

$$\tilde{Z}^{(4j)}=Z^{(2j)};\quad \tilde{Z}^{(4j+1)}=Z^{(2j+1)} \qquad \text{pro } j = 1,2,\ldots .$$

Nyní přiřadíme pořadí $1,2,\ldots,N$ posloupnosti (3.29) v tom sledu, jak je napsána. Označíme-li "nové pořadí" Z_i jako $\tilde{R}_i$, dostaneme následující vzájemně jednoznačné zobrazení mezi vektory $(\tilde{R}_1,\ldots,\tilde{R}_N)$ a $(R_1,\ldots,R_N)$:

$$\begin{aligned}
\tilde{R}_i &= 1 && \ldots && R_i = 1\\
\tilde{R}_i &= 4j-2 && \ldots && R_i = N+2-2j\\
\tilde{R}_i &= 4j-1 && \ldots && R_i = N+1-2j\\
\tilde{R}_i &= 4j && \ldots && R_i = 2j\\
\tilde{R}_i &= 4j+1 && \ldots && R_i = 2j+1,
\end{aligned}$$

$j = 1,2, \ldots$.

Kritický obor Siegel-Tukeyho testu pak je

$$\sum_{i=m+1}^{N} \tilde{R}_i > C$$

kde kritickou hodnotu C najdeme v tabulkách Wilcoxonova testu.

Nevýhodou tohoto testu je jeho nízká asymptotická vydatnost vzhledem k F-testu proti normálním alternativám ($\frac{6}{\pi^2} = 0{,}608$). Na druhé straně, pokud máme pochybnosti o tom, že základní rozdělení je normální, není vhodné používat F-testu; rozdělení jeho testové statistiky je velmi citlivé na odchylky od normality.

Místo Siegel-Tukeyho testu se v praxi častěji používají jiné pořadové testy rozptylu typu (3.28), z nichž dva zde uvedeme.

3.5.1. Klotzův test

Ve (3.28) položme $h(t) = \left[\Phi^{-1}(t)\right]^2$, $0 < t < 1$, kde Φ^{-1} je inverzní distribuční funkce standardního normálního rozdělení. Kritický obor Klotzova testu je

$$S = \sum_{i=m+1}^{N} \left[\Phi^{-1}\left(\frac{R_i}{N+1}\right)\right]^2 = \sum_{i=1}^{N} \left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2 V_i > C \ .$$

Kritické hodnoty testu tabeloval Klotz (1962) v práci "Nonparametric tests for scale", Ann.Math.Statist. 33. Pro velké hodnoty m a n lze kritické hodnoty stanovit pomocí normální aproximace $N(ES, \operatorname{var} S)$, kde

$$ES = \frac{n}{N} \sum_{i=1}^{N} \left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2$$

$$\operatorname{var} S = \frac{mn}{N(N+1)} \sum_{i=1}^{N} \left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^4 - \frac{m}{n(N+1)} (ES)^2 \ .$$

Klotzův test je zvláště vhodný v případě, že F je distribuční funkce přibližně normálního typu.

3.5.2. Kvartilový test

Jestliže ve (3.28) položíme

$$h(t) = \begin{cases} 0 & \dots \quad \frac{1}{4} < t < \frac{3}{4} \\ \frac{1}{2} & \dots \quad t = \frac{1}{4}, \quad t = \frac{3}{4} \\ 1 & \dots \quad 0 < t < \frac{1}{4} \quad \text{a} \quad \frac{3}{4} < t < 1, \end{cases}$$

dostaneme

$$S_N = \frac{1}{2} \sum_{i=m+1}^{N} \left[\operatorname{sign} \left(\left| \frac{R_i}{N+1} - \frac{1}{2} \right| - \frac{1}{4} \right) + 1 \right] ,$$

což je testová statistika tzv. kvartilového testu; příslušný kritický obor má tvar $\left[S_N > C \right]$. Pokud není (N+1) dělitelné 4, je statistika S_N rovna počtu pozorování druhého výběru, pro která je $\frac{R_i}{N+1} < \frac{1}{4}$ nebo $\frac{R_i}{N+1} > \frac{3}{4}$. Jestliže je (N+1) dělitelné 4, zvětšíme tento počet o $\frac{1}{2}$ nebo o 1, pokud pro jedno nebo dvě pozorování druhého výběru platí $\frac{R_i}{N+1} = \frac{1}{4}$ nebo $\frac{R_j}{N+1} = \frac{3}{4}$.

Jestliže $N = 4k$, má kvartilová statistika S hypergeometrické rozdělení (3.25), stejně jako statistika mediánového testu; kritické hodnoty tedy najdeme v tabulkách citovaných u mediánového testu nebo pomocí normální aproximace.

Kvartilový test je vhodný pro rozdělení pravděpodobností s „těžkými chvosty", např. klesá-li hustota $f(x)$ k nule pro $x \to \pm\infty$ stejně rychle jako $\frac{1}{x^2}$.

Poznámka. Alternativy K_5 a K_6 předpokládají, že rozdělení obou výběrů se liší jen rozptýleností, nikoli polohou. Tento případ nastane např. tehdy, jsou-li $X_1, \dots, X_m$ a $Y_1, \dots, Y_n$ výsledky měření nějaké kvantity μ různými metodami nebo v různých laboratořích. Jestliže se měří 2 kvantity,

řekněme ξ a η, dvěma různými metodami, pak obvyklý postup je, že nahradíme ξ a η vhodnými odhady $\hat{\xi}$ a $\hat{\eta}$ založenými na $X_1,\dots,X_m$ resp. $Y_1,\dots,Y_n$ a pak aplikujeme pořadové testy na pseudo-pozorování $X_1-\hat{\xi}, \dots, X_m-\hat{\xi}; Y_1-\hat{\eta},\dots, \dots,Y_n-\hat{\eta}$. Obecně se rozdělení pravděpodobností testové statistiky za H_0 touto úpravou mění, ale asymptotické rozdělení zůstane zachováno, pokud je F symetrické.

3.6. Testy založené na empirických distribučních funkcích

Nechť $X_1,\dots,X_m$ a $Y_1,\dots,Y_n$ jsou 2 nezávislé náhodné výběry ze dvou populací se spojitými distribučními funkcemi F a G. Chceme testovat hypotézu $H_0 : F = G$ proti obecné alternativě

$$K_7 : F \neq G .$$

Tato alternativa znamená, že dvě ošetření, postupy apod. se liší, ale nelze přijmout žádné zvláštní předpoklady o způsobu, jakým se odlišují.

Problém testu H_0 proti K_7 je invariantní vzhledem ke všem zobrazením tvaru $x_i' = f(x_i)$, $y_j' = f(y_j)$, $i=1,\dots,m$; $j=1,\dots,m$, kde f je libovolná spojitá funkce. Protože neexistuje statistika, invariantní vzhledem k této grupě, je jediným invariantním testem velikosti α test $\Phi(\underset{\sim}{x},\underset{\sim}{y}) \equiv \alpha$. Tento test však není přípustný, protože lze najít testy velikosti α, které mají sílu $> \alpha$ proti všem alternativám $G \neq F$ (viz cvič.(3)).

Pro testování H_0 proti obecné alternativě K_7, jakož i proti jednostranným alternativám typu $\{ G(x) \leq F(x) \ \forall x,$

$G(x) \neq F(x)$ } , u kterých se nejedná ani o rozdíl v poloze, ani v měřítku, často užíváme testy, založené na empirických distribučních funkcích.

Definice. Nechť $X_1,\dots,X_m$ jsou náhodné veličiny. Empirickou distribuční funkcí náhodného vektoru $(X_1,\dots,X_m)$ nazveme funkci $\hat{F}_m : R^1 \longrightarrow \langle 0,1\rangle$ definovanou vztahem

$$\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} u(x-X_i), \qquad x\in R^1 \tag{3.30}$$

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

Hodnota $m\hat{F}_m(x)$ je rovna počtu veličin X_i, které jsou nejvýše rovny x. Jestliže $X_1,\dots,X_m$ je náhodný výběr z rozdělení s distribuční funkcí F, má $m\hat{F}_m(x)$ binomické rozdělení $B(m;p)$, kde $p = F(x)$; tedy $\hat{F}_m(x)$ je nestranným odhadem $F(x)$ s rozptylem $\frac{1}{m}F(x)(1-F(x))$. Ze zákona velkých čísel plyne, že tento odhad je konzistentní, tj. $\hat{F}_m(x) \xrightarrow{P} F(x)$ pro $m\to\infty$, a z Moivre-Laplaceovy věty plyne, že tento odhad má asymptoticky normální rozdělení pro $m\to\infty$ a každé pevné x. Kromě toho, že $\hat{F}_m(x)$ je konzistentním odhadem $F(x)$, dále platí $\sup_{x\in R^1} |\hat{F}_m(x)-F(x)| \xrightarrow{P} 0$ pro $m\to\infty$, což je tvrzení Cantelli-Glivenkovy věty. Dá se dokázat řada dalších a silnějších limitních vlastností empirických distribučních funkcí; ale i ty, které jsme zde vyjmenovali, ukazují, že empirická distribuční funkce je dobrou aproximací skutečné distribuční funkce náhodného výběru. Proto byla navržena řada testů hypotézy H_0, které měří odchylky distribučních funkcí pomocí odchylek empirických distribučních funkcí. Popíšeme zde dva z těchto testů.

Vztah mezi pořadími, pořádkovými statistikami a empirickou distribuční funkcí veličin $X_1,\dots,X_m$ ihned vyplyne z následujících rovností (které platí, jsou-li $X_1,\dots,X_m$ různá)

$$\begin{aligned} m.\hat{F}_m(X_i) &= R_i \\ m.\hat{F}_m(X^{(i)}) &= i \end{aligned} \qquad i = 1,\dots,m. \tag{3.31}$$

3.6.1. Kolmogorov-Smirnovův test

Nechť $\hat{F}_m(x)$ je empirická distribuční funkce výběru $X_1,\dots,X_m$ a $\hat{G}_n(x)$ empirická distribuční funkce výběru $Y_1,\dots,Y_n$. Kolmogorov-Smirnovův test má tvar

$$\Phi(\underset{\sim}{X},\underset{\sim}{Y}) = \begin{cases} 1 & \dots \quad D_{mn} > C_\alpha \\ \gamma & \dots \quad D_{mn} = C_\alpha \\ 0 & \dots \quad D_{mn} < C_\alpha \end{cases} \tag{3.32}$$

kde

$$D_{mn} = \max_{x \in R^1} \left| \hat{F}_m(x) - \hat{G}_n(x) \right| \tag{3.33}$$

a C_α je kritická hodnota.

Nejprve musíme zodpovědět otázku, zda kritické hodnoty C_α závisí na základním rozdělení F a zda Kolmogorov-Smirnovův test patří k pořadovým testům. Z tvrzení následující věty vyplývá, že kritické hodnoty jsou společné pro celou hypotézu H_0 a že test je skutečně pořadovým testem.

VĚTA 4. Nechť $X_1,\dots,X_m$ a $Y_1,\dots,Y_n$ jsou 2 nezávislé náhodné výběry z rozdělení se spojitou distribuční funkcí $F(x)$. Pak statistika D_{mn} závisí jen na pořadích $R_1,\dots,R_m$, $R_{m+1},\dots,R_N$ $(N=m+n)$ pozorování a rozdělení statistiky D_{mn}

nezávisí na F.

Důkaz. Označme $Z_1,\dots,Z_N$ spojený náhodný výběr a $Z^{(1)}<Z^{(2)}<\dots<Z^{(N)}$ příslušné pořádkové statistiky.

Protože funkce $\hat{F}_m(x)$ a $\hat{G}_n(x)$ jsou neklesající, schodovité a skoky nastávají pouze v některých z bodů $Z^{(1)},\dots,Z^{(N)}$, stačí hledat maximum rozdílu $|\hat{F}_m(x) - \hat{G}_n(x)|$ v bodech $Z^{(1)},\dots,Z^{(N)}$. Nechť $V_j = 0$, jestliže $Z^{(j)}$ pochází z prvního výběru a $V_j = 1$, jestliže $Z^{(j)}$ pochází z druhého výběru, $j=1,\dots,N$. Pak platí

$$\hat{F}_m(Z^{(j)}) = \frac{1}{m}\left[(1-V_1) + (1-V_2) + \dots + (1-V_j)\right]$$

a

$$\hat{G}_n(Z^{(j)}) = \frac{1}{n}(V_1 + \dots + V_j), \qquad j=1,\dots,N,$$

tedy

$$(3.34)\quad \hat{F}_m(Z^{(j)}) - \hat{G}_n(Z^{(j)}) = \frac{m+n}{mn}\left[j\frac{n}{m+n} - V_1 - \dots - V_j\right], \qquad j=1,\dots,N$$

odkud vyplývá vyjádření testové statistiky

$$(3.35)\quad D_{mn} = \max_{1\le j\le N}\left|j\frac{n}{m+n} - V_1 - \dots - V_j\right| \cdot \frac{m+n}{mn}.$$

Nechť $R_1,\dots,R_m$, $R_{m+1},\dots,R_N$ je vektor pořadí pozorování $Z_1,\dots,Z_N$. Náhodné veličiny $V_1,\dots,V_N$ lze ekvivalentním způsobem vyjádřit takto:

$[V_i = 1] \iff$ [právě jedno z pořadí $R_{m+1},\dots,R_N$ je rovno i] a

$[V_i = 0] \iff$ [právě jedno z pořadí $R_1,\dots,R_m$ je rovno i].

Veličiny $V_1,\dots,V_N$ tedy závisí jen na pořadích, a z vyjádření (3.35) plyne, že i D_{mn} je funkcí pouze pořadí. Protože rozdělení vektoru pořadí za H_0 nezávisí na F, plyne odtud i nezávislost kritických hodnot na F. □

Poznámka. Vztahu (3.35) lze s výhodou použít ke stanovení hodnoty testové statistiky.

Kolmogorov-Smirnova testu se užívá i proti jednostranným alternativám, že druhý výběr je stochasticky větší než první výběr. Pak zamítneme H_0 ve prospěch alternativy při velkých hodnotách testové statistiky, která je v tomto případě rovna

$$(3.36)\quad D_{mn}^{+} = \max_{x \in R^1} (\hat{F}_m(x) - \hat{G}_n(x)) = \frac{m+n}{mn} \max_{1 \le j \le N} \left[j \frac{m}{m+n} - V_1 - \ldots - V_j \right] .$$

Tabulky kritických hodnot Kolmogorov-Smirnovova testu najdeme např. v pracech:
P.J.Kim a R.I.Jennrich (1970): "Tables of the Exact Sampling Distribution of Two-Sample Kolmogorov-Smirnov Criterion D_{mn}, m=n" (v serii Selected Tables in Mathematical Statistics, editoři Harter a Owen);
Jaroslav Janko: Statistické tabulky.

Pro velké hodnoty m a n lze použít limitních kritických hodnot, avšak aproximace rozdělení statistiky D_{mn} při $m,n \to \infty$ není dána normálním rozdělením, jak je vidět z následující věty, kterou uvedeme bez důkazu.

VĚTA 5. Nechť $F \equiv G$ je libovolná spojitá distribuční funkce. Pak pro libovolné $\lambda > 0$ platí

$$(3.37)\quad \lim_{m,n\to\infty} P\left\{ \left(\frac{mn}{m+n}\right)^{1/2} D_{mn} < \lambda \right\} = 1 - e^{-2\lambda^2}$$

a

$$(3.38)\quad \lim_{m,n\to\infty} P\left\{ \left(\frac{mn}{m+n}\right)^{1/2} D_{mn}^{+} < \lambda \right\} =$$

$$= 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} \; e^{-2k^2\lambda^2} \doteq 1 - 2e^{-2\lambda^2} \; .$$

Důkaz: viz Hájek-Šidák (1967), kapitola V, důsledek věty 3.6.

Limitní hodnoty (3.37) jsou tabelovány v knize Lehmann (1975).

3.6.2. Cramér-von Misesův test

Uvažujme stejnou situaci jako v 3.6.1. Testová statistika Cramér-von Misesova testu má tvar

$$(3.39) \quad M_{mn} = \frac{mn}{(m+n)^2} \left[\sum_{i=1}^{m} (\hat{F}_m(X_i) - \hat{G}_n(X_i))^2 + \sum_{j=1}^{n} (\hat{F}_m(Y_j) - \hat{G}_n(Y_j))^2 \right] .$$

Test je určen pro testování H_o proti oboustranným alternativám $K_7 : F \neq G$ a zamítá H_o pro velké hodnoty statistiky (3.39).

S použitím vztahu (3.34) můžeme testovou statistiku vyjádřit v jednodušším tvaru

$$(3.40) \quad M_{mn} = \frac{1}{mn} \sum_{j=1}^{N} (j \frac{n}{m+n} - V_1 - \ldots - V_j)^2$$

a odtud stejně jako v důkazu věty 4 vyplývá, že Cramér-von Misesův test je pořadovým testem. Tabulky kritických hodnot lze nalézt v pracech :

T.W.Anderson (1962) "On the distribution of the two-sample Cramér-von Mises criterion", Ann.Math.Statist.33, 1148-59;

E.J.Burr(1963)."Distribution of the two-sample Cramér-von Mises criterion for small equal samples",Ann.Math.Statist.34, 95-101.

Pro velké hodnoty m a n můžeme použít limitních kritických hodnot, které vyplývají v následující věty.

VĚTA 6. Nechť $F \equiv G$ je libovolná spojitá distribuční funkce. Pak pro libovoné $\lambda > 0$ platí

$$(3.41) \qquad \lim_{m,n\to\infty} P\{M_{mn} < \lambda\} = P\left\{\sum_{j=1}^{\infty} \frac{W_j^2}{j^2\pi^2} < \lambda\right\}$$

kde $W_1, W_2, \ldots$ jsou nezávislé náhodné veličiny s normálním rozdělením $N(0,1)$.

Důkaz. viz Hájek-Šidák (1967), kap.V, věta 3.8.

Limitní kritické hodnoty testu tabelovali T.W.Anderson a D.A.Darling (1952) : "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes". Ann.Math.Statist.23, 193-212.

3.7. Pořadové testy při výskytu shodných pozorování

Dosud jsme předpokládali, že distribuční funkce F a G jsou spojité. Za tohoto předpokladu nastane shoda dvou pozorování s pravděpodobností 0 a pořadí pozorování jsou s pravděpodobností 1 dobře definována.

Přesto však v praxi často narazíme na shodu dvou nebo více pozorování i když se jedná o měření spojitého typu; data totiž zaznamenáváme na konečný počet desetinných míst a hodnoty zaokrouhlujeme, takže všecka data vlastně vyjadřujeme pomocí spočetné sítě. Možnost shody dvou nebo více pozorování nelze tedy ignorovat a je třeba modifikovat pro tento případ i pořadové testy.

Existuje několik metod úpravy pořadových testů při výskytu shodných pozorování. Zde stručně popíšeme dvě z nich: metoda znáhodnění a metoda průměrných pořadí, a to s ohledem na použití v testech hypotézy o shodnosti dvou populací. Nejprve však uveďme několik obecných poznámek.

Jestliže shodná pozorování patří ke stejnému výběru, můžeme je mezi sebou uspořádat jakkoli, aniž by to mělo vliv na hodnotu testové statistiky. Proto si všímáme shodných pozorování hlavně tehdy, patří-li k různým výběrům.

Jestliže je počet shodných pozorování malý, můžeme je z výběru zcela vypustit, ovšem za cenu určité ztráty informace.

Testová statistika některých testů je dobře definována i při výskytu shod; pouze se jejich vlivem mohou měnit pravděpodobnosti chyby 1. a 2. druhu. Jako příklad uvažujme Kolmogorov-Smirnovův test: definici empirické distribuční funkce i vztahu (3.33) lze beze změny použít i při výskytu shod. Použijeme-li kritických hodnot Kolmogorov-Smirnovova testu, dostaneme kritický obor, jehož velikost je o něco nižší než nominální hladina významnosti. Tento fakt dokážeme např. tak, že chápeme pozorování $X_1,\dots,X_m$, $Y_1,\dots,Y_n$ (mezi nimiž mohou být shody) jako data vzniklá zaokrouhlením dat $X_1^*,\dots,X_m^*$, $Y_1^*,\dots,Y_n^*$ se spojitým rozdělením. Pak možné hodnoty rozdílu $\hat{F}_m(x)-\hat{G}_n(x)$, $x\in R^1$, tvoří podmnožinu hodnot $\hat{F}_m^*(x)-\hat{G}_n^*(x)$, kde $\hat{F}_m^*$ a $\hat{G}_n^*$ jsou empirické distribuční funkce X_i^* a Y_j^*; tedy

$$\max_{x\in R^1}\left[\hat{F}_m(x) - \hat{G}_n(x)\right] \leq \max_{x\in R^1}\left[\hat{F}_m^*(x) - \hat{G}_n^*(x)\right]$$

a podobně pro maxima absolutních hodnot.

3.7.1. Metoda znáhodnění

Nechť $Z_1,\dots,Z_N$ je spojený náhodný výběr. Nechť $U_1,\dots,U_N$ jsou náhodné veličiny s rovnoměrným rozdělením na $\langle 0,1\rangle$, vzájemně nezávislé a nezávislé na $Z_1,\dots,Z_N$. Uvažujme dvojice $(Z_1,U_1),\dots,(Z_N,U_N)$ a uspořádejme je podle pravidla

$$(3.42)\qquad (Z_i,U_i) < (Z_j,U_j) \iff \text{buď } Z_i < Z_j \text{ nebo } Z_i = Z_j \text{ a } U_i < U_j.$$

Pořadí R_i^* dvojice (Z_i,U_i) v tomto uspořádání prohlásíme za pořadí Z_i.

Řekneme, že $Z_1,\dots,Z_N$ splňují hypotézu $\bar{H}$, jestliže jsou nezávislé a mají stejné rozdělení (nikoli nutně spojité). Za platnosti $\bar{H}$ má vektor $R^* = (R_1^*,\dots,R_N^*)$ rovnoměrné rozdělení na $\mathcal{R}$. Pro jednoduchost to dokážeme na důležitém speciálním případě, kdy $Z_1,\dots,Z_N$ mohou nabývat spočetně mnoha ekvidistantních hodnot (což pokrývá případ, že zaznamenáváme data na k desetinných míst, kde k je celé konečné číslo).

VĚTA 7. Nechť $Z_1,\dots,Z_N$ jsou náhodné veličiny splňující hypotézu $\bar{H}$, které mohou s kladnou pravděpodobností nabývat jen hodnot z množiny

$$a + kd;\quad k = 0, \pm 1, \pm 2, \dots\ ,\quad a \in R^1,\quad d > 0.$$

Pak vektor $(R_1^*,\dots,R_N^*)$ má rozdělení pravděpodobností

$$(3.43)\qquad P(R^* = r) = \frac{1}{N!}\ ,\qquad r \in \mathcal{R},$$

kde $\mathcal{R}$ je množina všech permutací $(1,\dots,N)$.

Důkaz. Bez újmy obecnosti můžeme předpokládat, že $Z_1,\dots,Z_N$

nabývají celočíselných hodnot (jinak můžeme místo Z_i uvažovat $Z_i' = \frac{1}{d}(Z_i - a)$, $i=1,\dots,N$, čímž se $R_1^*,\dots,R_N^*$ nemění). Pak náhodná veličina $T_i = Z_i + U_i$ je ekvivalentní dvojici (Z_i, U_i), protože s pravděpodobností 1 je $Z_i = [T_i]$, $U_i = T_i - [T_i]$, kde $[T]$ označuje největší celé číslo $\leq T$; $i=1,\dots,N$. T_i má spojitou distribuční funkci, neboť $P(T_i = t) = 0$ pro lib. $t \in R^1$. Dále platí

$$(Z_i, U_i) < (Z_j, U_j) \iff T_i < T_j,$$

což znamená, že $R_1^*,\dots,R_N^*$ jsou pořadí náhodných veličin $T_1,\dots,T_N$ splňujících U_0. (3.43) pak plyne z věty 4 kapitoly 2. □

3.7.2. Metoda průměrných pořadí

Tato metoda vychází z myšlenky, že shodným pozorováním je třeba přiřadit shodná pořadí; společné "pořadí" se pak volí jako průměr všech pořadí připadajících na skupinu shodných pozorování.

Uvažujme tuto metodu ve spojitosti s Wilcoxonovým testem, pro který se nejčastěji užívá. Dá se použít podobně i pro jiné testy, jejichž skóry $a(.)$ mají smysl i pro necelé argumenty.

Předpokládejme, že mezi N pozorováními je e různých hodnot, a že d_1 z těchto pozorování je rovno nejmenší z těchto hodnot, d_2 druhé nejmenší hodnotě, ..., d_e pozorování je rovno největší hodnotě $\sum_{i=1}^{e} d_i = N$. Pak průměrná pořadí jednotlivých skupin shodných pozorování jsou

$$v_1 = \dots = v_{d_1} = \frac{1}{2}(d_1 + 1)$$

$$v_{d_1+1} = \dots v_{d_1+d_2} = d_1 + \tfrac{1}{2}(d_2+1)$$

$$v_{d_1+d_2+1} = \dots = v_{d_1+d_2+d_3} = d_1+d_2+\tfrac{1}{2}(d_3+1)$$

$$\dots\dots\dots\dots\dots\dots\dots\dots$$

$$v_{d_1+d_2+\dots+d_{e-1}+1} = \dots v_N = d_1+d_2+\dots d_{e-1}+\tfrac{1}{2}(d_e+1)$$

Označme $(R_1',\dots,R_N')$ průměrná pořadí pozorování $Z_1,\dots,Z_N$. Modifikovaná Wilcoxonova statistika má tvar

$$W_N^* = \sum_{i=m+1}^{N} R_i' \ .$$

Protože rozdělení $(R_1',\dots,R_N')$ za $\bar{H}$ už není rovnoměrné na $\mathcal{R}$ (vektor ani obecně nenabývá hodnot z $\mathcal{R}$), nelze použít standardních tabulek kritických hodnot. Obvykle používáme kritických hodnot založených na normální aproximaci, což je oprávněné pro dostatečně velká m,n a není-li rozsah žádné skupiny shodných pozorování blízký N. Abychom mohli použít normální aproximace, musíme stanovit střední hodnotu a rozptyl W_N^*.

$(R_1',\dots,R_N')$ je závislý na náhodných veličinách $d_1,\dots,d_e$. Budeme hledat střední hodnotu a rozptyl W_N^* podmíněné $d_1,\dots,d_e$ za předpokladu platnosti $\bar{H}$.

$$E(W_N^* | d_1,\dots,d_e) = \sum_{i=m+1}^{N} E(R_i' | d_1,\dots,d_e) =$$

$$= \sum_{i=m+1}^{N} \frac{1}{N}\left[d_1 \cdot \tfrac{1}{2}(d_1+1) + d_2(d_1+\tfrac{1}{2}(d_2+1)) + \dots + d_e(d_1+\dots d_{e-1}+\tfrac{1}{2}d_e)\right] =$$

$$= \frac{1}{N} \sum_{i=m+1}^{N} (1+\dots+N) = n\,\frac{N+1}{2} \ ,$$

a tedy

(3.44) $$E(W_N^* \mid d_1,\dots,d_e) = n\frac{N+1}{2} = E\,W_N ,$$

podmíněná střední hodnota W_N^* je shodná se střední hodnotou standardní Wilcoxonovy statistiky.

Rozptyl W_N^* je roven

(3.45) $$\operatorname{var} W_N^* = \frac{mn(N+1)}{12} - \frac{mn\sum_{i=1}^{e}(d_i^3 - d_i)}{12N(N-1)}$$

(vztah (3.45) je dokázán v knize Lehmann (1975), příklad 1 a 3 dodatku).

První člen v (3.45) je rozptyl standardní Wilcoxonovy statistiky, druhý člen je korekcí vzhledem ke shodám, který vymizí, nejsou-li žádná shodná pozorování.

3.8. Problémy a cvičení

(1) (Dixon a Massey(1969): Introduction to Statistical Analysis, Mc Graw Hill). Následující data udávají hladinu cholesterolu v krvi mužů dvou různých věkových skupin (20-30 letých a 40-50 letých). Užijete-li Wilcoxonův test na hladině =0,05, můžete prohlásit, že hladina cholesterolu v krvi starších mužů je významně (stochasticky) větší, než v krvi mladších mužů? Užijte normální aproximace kritických hodnot.

x(20-30 let)	135	222	251	260	269	235	386	252	352	173	156
y(40-50 let)	294	311	286	264	277	336	208	346	239	172	254

(2) Nechť $X_1,\dots,X_m$ a $Y_1,\dots,Y_n$ jsou nezávislé náhodné výběry z rozdělení se spojitými rostoucími distribučními funkcemi F a G; nechť $W = W(\underset{\sim}{X},\underset{\sim}{Y})$ je Wilcoxonova statistika.