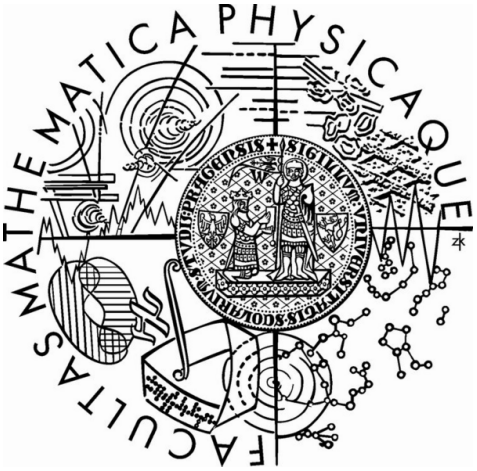
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



JITKA SETÍKOVSKÁ

Odhady založené na pořadových testech

Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Prof. RNDr. Jana Jurečková, DrSc.
Studijní program: Matematika, Matematická statistika

Ráda bych poděkovala Prof. RNDr. Janě Jurečkové, DrSc., za její cenné rady a připomínky a za její ochotu a laskavý přístup po celou dobu vedení mé diplomové práce.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 11. 4. 2006 Jitka Setíkovská

Obsah

Název práce: Odhady založené na pořadových testech
Autor: Jitka Setíkovská
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Prof. RNDr. Jana Jurečková, DrSc.
e-mail vedoucího: jurecko@karlin.mff.cuni.cz

Abstrakt: Práce se zabývá takzvanými R-odhady, což jsou odhady parametru polohy nebo posunutí navržené Hodgesem a Lehmannem (viz [7]). Jedná se o odhady vzniklé inverzí pořadových testů. Zformulována je jak definice používaná Hodgesem a Lehmannem, tak definice používaná v novější literatuře. U vybraných pořadových testů jsou ukázány základní vlastnosti rozdělení jejich statistik, které jsou pak využity při odvození explicitního vyjádření několika R-odhadů a intervalů spolehlivosti pro odpovídající parametr. Dále se práce zabývá základními vlastnostmi R-odhadů jako je nestrannost, ekvivariance vzhledem k posunutí, vydatnost, robustnost apod. Vlastnosti R-odhadů jsou zkoumány také na základě numerických výpočtů na simulovaných datech a jejich použití je ilustrováno na několika příkladech.

Klíčová slova: pořadové testy, R-odhady, Hodges-Lehmannovy odhady, odhady parametru polohy a posunutí, Wilcoxonův a Galtonův test

Title: Estimates based on rank tests
Author: Jitka Setíkovská
Department: Department of Probability and Mathematical Statistics
Supervisor: Prof. RNDr. Jana Jurečková, DrSc.
Supervisor's e-mail address: jurecko@karlin.mff.cuni.cz

Abstract: The thesis deals with R-estimators, estimators based on ranks. They were originally proposed by Hodges and Lehmann [7] as inversions of the rank tests. Not only the definition used by Hodges and Lehmann, but also the one used in later literature is formulated. Basic characteristics of some rank statistics are described and the explicit forms of the corresponding R-estimators and confidence intervals are derived. Their basic properties as unbiasedness, translation equivariance, efficiency, robustness are studied. The behavior of R-estimators is then illustrated on simulated data and on several examples.

Keywords: rank tests, R-estimators, Hodges-Lehmann estimators, estimates of location, Wilcoxon and Galton tests

Kapitola 1

Úvod

V praxi se často setkáváme s problémem porovnání účinků dvou různých postupů, například chceme zjistit účinnost nové metody léčby, srovnat dva různé způsoby výroby apod. Tyto postupy obvykle nazýváme ošetření. Provedeme tedy $m+n$ nezávislých pokusů, kdy m-krát aplikujeme první ošetření, n-krát druhé ošetření, a pozorujeme určitou hodnotu, podle které můžeme měřit účinnost příslušného postupu. My se budeme zabývat případem, kdy předpokládáme, že účinky obou ošetření se liší konstantně, to znamená o nějaké konstantní Δ. Δ se pak nazývá parametr posunutí.

Nechť tedy $X_1 \dots X_m$, $Y_1 \dots Y_n$ jsou nezávislé náhodné veličiny se spojitými distribučními funkcemi

$$\begin{array}{lll} P(X_i \leq u) & = F(u) & \text{pro } i = 1, \dots, m \qquad \text{a} \\ P(Y_j \leq u) & = F(u - \Delta) & \text{pro } j = 1, \dots, n \ . \end{array}$$

U příkladů tohoto typu řešíme většinou dva základní problémy: testujeme hypotézu $\Delta = 0$ (to znamená účinnost ošetření se neliší) proti alternativě $\Delta > 0$ (to znamená druhé ošetření má větší účinnost) nebo se pokoušíme odhadnout parametr posunutí Δ. Můžeme-li reálně předpokládat, že F je distribuční funkce normálního rozdělení, testuje se uvedená nulová hypotéza obvykle pomocí t-statistiky

$$t = \frac{(\overline{Y} - \overline{X})\sqrt{\frac{1}{m} + \frac{1}{n}}}{\left[\frac{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_i - \overline{Y})^2}{m+n-2}\right]^{\frac{1}{2}}}, \tag{1.1}$$

$$\left(\text{kde } \overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j,\ \overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i\right)$$

o které víme, že má za platnosti nulové hypotézy Studentovo t-rozdělení o $(m+n-2)$ stupních volnosti. Podobně, klasickým odhadem pro Δ je rozdíl průměrů $\hat{\Delta} = \overline{Y} - \overline{X}$.

Obě tyto metody, jak odhad, tak i test založený na statistice t, jsou však známé náchylností k hrubým chybám, například pokud se v náhodných výběrech vyskytují odlehlá pozorování. V případě testování hypotéz se tento problém řeší použitím robustnějších pořadových testů, jako například Wilcoxonova testu nebo testu s normální skórovou funkcí. Ze stejného principu vycházeli Hodges a Lehmann (viz Hodges a Lehmann [7]) a odvodili robustnější alternativu pro odhady Δ založenou na pořadových testech. Tyto odhady se nazývají R-odhady a vznikají inverzí pořadových testů hypotézy $\Delta = 0$ proti alternativě $\Delta > 0$.

K porovnání účinnosti dvou různých ošetření můžeme zvolit i jiný postup. Abychom co nejvíce omezili vlivy, které nesouvisí s ošetřeními, rozdělíme subjekty, na kterých pozorujeme účinky ošetření, do dvojic. Tentokrát provedeme $2n$ nezávislých pokusů. Každé ošetření aplikujeme n-krát, přičemž člena dvojice pro jedno ošetření vybíráme náhodně, na druhého z dvojice pak použijeme zbývající ošetření. Získáme tak dvojice pozorování $(X_1, Y_1), \ldots, (X_n, Y_n)$, které můžeme považovat za náhodný výběr ze spojitého dvourozměrného rozdělení. Položíme-li $Z_i = Y_i - X_i$, tvoří náhodné veličiny $Z_1, \ldots, Z_n$ náhodný výběr ze spojitého jednorozměrného rozdělení. Za určitých podmínek se tedy na porovnávání dvou postupů můžeme dívat také jako na jednovýběrový problém.

Budeme opět předpokládat, že účinnost obou ošetření se liší konstantně. Nechť $Z_1, \ldots, Z_n$ jsou náhodné veličiny s distribuční funkcí

$$P(Z_i \leq u) = F(u - \theta) \quad \text{pro } i = 1, \ldots, n \ ,$$

která je spojitá a kde pro F platí $F(x) + F(-x) = 1 \ \forall x$, to znamená rozdělení je symetrické kolem 0. V případě jednoho výběru se θ nazývá parametr polohy.

Nejčastějšími úkoly pak je, podobně jako v případě dvou výběrů, testovat hypotézu $\theta = 0$ (rozdělení $Z_1, \ldots, Z_n$ je symetrické kolem 0 a ošetření mají stejnou účinnost) proti alternativě $\theta > 0$ (účinnost druhého ošetření je větší) a odhadnout parametr θ. Při platnosti předpokladu normality bychom k testování zmíněné hypotézy použili párový t-test a k odhadu θ průměr $\overline{Z}$. Vůči těmto postupům máme ale stejné výhrady jako vůči klasickým metodám pro dva výběry. R-odhady se tedy uplatní i u jednovýběrových problémů, kde je odvozujeme od jednovýběrových testů pro hypotézu $\theta = 0$ proti $\theta > 0$.

V Kapitole 1.1 uvedeme základní definice a tvrzení potřebná v dalším textu. Kapitola 2 se zabývá vybranými pořadovými testy a základními

vlastnostmi rozdělení jejich statistik. Rozdělení statistiky jednovýběrového Galtonova testu je odvozeno včetně důkazu uvedeného v literatuře, která byla v době psaní diplomové práce nedostupná. V Kapitole 3 zformulujeme definici R-odhadů a to jak definici používanou Hodgesem a Lehmannem, tak ekvivalentní definici používanou v současné literatuře. Dále je zde odvozeno explicitní vyjádření několika odhadů, vedle nejznámějších odhadů založených na Wilcoxonově a znaménkovém testu je odvozen také vzorec pro odhad založený na jednovýběrovém i dvouvýběrovém Galtonově testu. Na základě podobných úvah jako R-odhady odvodíme intervaly spolehlivosti pro parametr polohy i posunutí.

Základní vlastnosti R-odhadů jsou ukázány v Kapitole 5. Dokážeme, že R-odhady jsou ekvivariantní vzhledem k posunutí, nestranné nebo alespoň mediánově nestranné a jejich rozdělení je za obecných předpokladů absolutně spojité. Na základě článku Zuo [13] jsou odvozeny meze dvou charakteristik robustnosti (míry chvostů a bodu selhání) a výsledky článku jsou rozšířeny na odhady založené na jednovýběrovém Galtonově testu. Dále se podíváme na asymptotické vlastnosti R-odhadů.

Kapitola 6 zkoumá vlastnosti R-odhadů na základě numerických výpočtů na simulovaných datech. Nakonec Kapitola 7 ilustruje použití R-odhadů na několika příkladech s reálnými daty.

1.1 Základní definice a tvrzení

Definice: *Mějme náhodný výběr $X_1, \ldots, X_n$, v němž žádné dvě náhodné veličiny nejsou shodné. Označme R_i počet náhodných veličin ve výběru, které jsou menší nebo rovny X_i, tedy $R_i = \sum_{j=1}^{n} I[X_j \leq X_i]$, $i = 1, \ldots, n$, kde I je indikátorová funkce. Pak R_i se nazývá pořadí náhodné veličiny X_i ve výběru.*

Definice: *Mějme náhodný výběr $X_1, \ldots, X_n$. Tyto veličiny uspořádáme podle velikosti. Označme $X_{(i)}$ i-tou nejmenší hodnotu z $X_1, \ldots, X_n$. $X_{(i)}$ se nazývá i-tá pořádková statistika. Platí $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ a veličinám $X_{(1)}, \ldots, X_{(n)}$ se říká uspořádaný náhodný výběr.*

Definice: *Test, jehož statistika je funkcí pouze pořadí, se nazývá pořadový test.*

Mějme dva nezávislé náhodné výběry $X_1, \ldots, X_m$ a $Y_1, \ldots, Y_n$, pro které platí

$$\begin{array}{lll} P(X_i \leq u) & = F(u) & \text{pro } i = 1, \ldots, m \qquad \text{a} \\ P(Y_j \leq u) & = F(u - \Delta) & \text{pro } j = 1, \ldots, n\ . \end{array}$$

Uvažujme testovou statistiku $h(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ pro test hypotézy $H_0 : \Delta = 0$ proti alternativě $H_1 : \Delta > 0$. Nechť pro $i = 1, \ldots, n$ je R_i pořadí Y_i ve sdruženém výběru $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. Označme $X = (X_1, \ldots, X_m)$ a $Y = (Y_1, \ldots, Y_n)$.

Věta 1 *Nechť platí H_0. Rozdělení statistiky $h(X, Y)$ je symetrické kolem μ, jestliže platí jedna z následujících podmínek:*
(i) h je funkcí pouze pořadí $R_1, \ldots, R_n$ a splňuje

$$h(x, y) + h(-x, -y) = 2\mu \tag{1.2}$$

(ii) $m = n$ a h splňuje

$$h(x, y) + h(y, x) = 2\mu \tag{1.3}$$

(iii) rozdělení s distribuční funkcí F je symetrické kolem 0 a h splňuje (1.2).

Důkaz: (i) Nechť platí H_0.
Statistika h je funkcí pouze pořadí, proto $h(x, y) = g(R_1, \ldots, R_n)$ a $h(-x, -y) = g(m+n-R_1+1, \ldots, m+n-R_n+1)$. Za platnosti $H_0 : \Delta = 0$ mají vektory $(R_1, \ldots, R_n)$ a $(m + n - R_1 + 1, \ldots, m + n - R_n + 1)$ stejné sdružené rozdělení. Z toho a vztahu (1.2) dostáváme

$$\begin{aligned} P(h(X, Y) < \mu - a) &= P(h(-X, -Y) > \mu + a) \\ &= P(g(m + n - R_1 + 1, \ldots, m + n - R_n + 1) > \mu + a) \\ &= P(g(R_1, \ldots, R_n) > \mu + a) \\ &= P(h(X, Y) > \mu + a) \ . \end{aligned}$$

(ii) Když $m = n$ a $\Delta = 0$, mají vektory (X, Y), (Y, X) stejné sdružené rozdělení. Z toho a vztahu (1.3) vyplývá

$$P(h(X, Y) < \mu - a) = P(h(Y, X) < \mu - a) = P(h(X, Y) > \mu + a) \ .$$

(iii) Z platnosti H_0 a symetrie rozdělení kolem 0 vyplývá, že vektory (X, Y) a $(-X, -Y)$ jsou stejně rozdělené. Z toho a vztahu (1.2) dostáváme

$$P(h(X, Y) < \mu - a) = P(h(-X, -Y) < \mu + a) = P(h(X, Y) > \mu + a) \ . \qquad \square$$

Podobná věta platí i pro jeden náhodný výběr. Nechť $Z_1, \ldots, Z_N$ jsou nezávislé náhodné veličiny s distribuční funkcí

$$P(Z_i \leq u) = F(u - \theta) \quad \text{pro } i = 1, \ldots, N \ ,$$

která je spojitá a kde pro F platí $F(x) + F(-x) = 1 \ \forall x$. Uvažujme testovou statistiku $h(Z_1, \ldots, Z_N)$ pro test hypotézy $H_0 : \theta = 0$ proti alternativě $H_1 : \theta > 0$. Označme $Z = (Z_1, \ldots, Z_N)$ a pro $i = 1, \ldots, N$ R_i^+ pořadí $|Z_i|$ mezi $|Z_1|, \ldots, |Z_N|$.

Věta 2 *Nechť platí H_0. Rozdělení statistiky $h(Z)$ je symetrické kolem μ, jestliže platí jedna z následujících podmínek:*
(i) h je funkcí pouze pořadí $R_1^+, \ldots, R_N^+$ a splňuje

$$h(z) + h(-z) = 2\mu \tag{1.4}$$

(ii) rozdělení s distribuční funkcí F je symetrické kolem 0 a h splňuje (1.4).

Důkaz: Důkaz je analogický jako u Věty 1. □

Kapitola 2

Vybrané pořadové testy

Dále budeme v celé kapitole při odvozování testů a jejich vlastností předpokládat, že všechna pozorování, na nichž je test založen, jsou různá (respektive mají různé absolutní hodnoty u některých jednovýběrových testů). Teoreticky toto platí vzhledem k tomu, že předpokládáme, že pozorování tvoří náhodný výběr ze spojitého rozdělení. V praxi se však například následkem zaokrouhlování shodná pozorování občas vyskytují. Není-li takových pozorování mnoho, je jejich vliv na test zanedbatelný. Případným modifikacím testů při výskytu shodných pozorování se budeme věnovat v poznámkách.

2.1 Wilcoxonův test

2.1.1 Dvouvýběrový Wilcoxonův test

Nechť $X_1, \ldots, X_m$ je náhodný výběr z rozdělení se spojitou distribuční funkcí F a nechť $Y_1, \ldots, Y_n$ je na něm nezávislý náhodný výběr z rozdělení se spojitou distribuční funkcí G. Chceme testovat hypotézu

$$\begin{aligned} H_0 : F(u) &= G(u),\ u \in \mathbf{R} \quad \text{proti alternativě} \\ H_1 : G(u) &= F(u - \Delta),\ u \in \mathbf{R},\ \Delta > 0. \end{aligned}$$

Alternativě H_1 říkáme alternativa posunutí. Všimněme si, že testování této hypotézy je ekvivalentní testování hypotézy $\Delta = 0$ proti alternativě $\Delta > 0$, o kterém jsme mluvili v Úvodu. Kdybychom mohli předpokládat, že rozdělení s distribuční funkcí F je normální, použili bychom t-test založený na statistice (1.1). Pro případy, kdy tento předpoklad splněn není, použijeme postup uvedený například v Jurečková [8] a zkonstruujeme jinou statistiku založenou na pořadí. Pro zjednodušení zápisu označíme $N = m + n$. Všech N veličin $X_1, \ldots, X_m,\ Y_1, \ldots, Y_n$ uspořádáme vzestupně podle velikosti a

pořadí jednotlivých náhodných veličin v tomto sdruženém výběru označíme $R_1, \ldots, R_m, R_{m+1}, \ldots, R_N$.

Novou statistiku získáme tak, že do výrazu pro t-statistiku (1.1) dosadíme na místa hodnot $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ jejich pořadí. Dostáváme tak

$$t_R = \frac{(\frac{1}{n}\sum_{j=m+1}^{N} R_j - \frac{1}{m}\sum_{i=1}^{m} R_i)\sqrt{\frac{1}{m}+\frac{1}{n}}}{\left[\frac{\sum_{i=1}^{N}(R_i-\overline{R})^2}{N-2}\right]^{\frac{1}{2}}} , \tag{2.1}$$

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

Víme, že pořadí $R_1, \ldots, R_N$ jsou permutací čísel $1, \ldots, N$ a každá permutace může nastat s pravděpodobností $\frac{1}{N!}$. Z toho můžeme odvodit několik vztahů, které použijeme ke zjednodušení (2.1):

$$\begin{aligned}
&\overline{R} = \frac{N+1}{2} \\
&\sum_{i=1}^{m} R_i = \frac{N(N+1)}{2} - \sum_{j=m+1}^{N} R_j \\
&\sum_{i=1}^{N}(R_i-\overline{R})^2 = \sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2 = \frac{N(N^2-1)}{12} .
\end{aligned}$$

(**Poznámka:** Použili jsme obecně známé vztahy $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$ a $\sum_{i=1}^{N} i^2 = \frac{N(2N+1)(N+1)}{6}$).

Nyní můžeme výraz (2.1) přepsat:

$$t_R = \frac{(\frac{1}{n}\sum_{j=m+1}^{N} R_j - \frac{1}{m}\left(\frac{N(N+1)}{2} - \sum_{j=m+1}^{N} R_j\right)\sqrt{\frac{1}{m}+\frac{1}{n}}}{\left[\frac{N(N^2-1)}{12(N-2)}\right]^{\frac{1}{2}}} .$$

Je zřejmé, že tato statistika je až na lineární transformaci ekvivalentní součtu pořadí $Y_1, \ldots, Y_n$ ve sdruženém výběru, tedy statistice

$$W = \sum_{i=m+1}^{N} R_i . \tag{2.2}$$

Test založený na statistice W se nazývá Wilcoxonův test. Je lokálně nej- silnějším pořadovým testem proti alternativě posunutí H_1, jestliže F je distribuční funkce logistického rozdělení $F(x) = \frac{1}{1+e^{-x}}$.

Pro menší m a n jsou kritické hodnoty Wilcoxonova testu tabelovány, při větších hodnotách (už při $m > 10$, $n > 10$) můžeme použít aproximaci rozdělení statistiky W pomocí normálního rozdělení. Za platnosti H_0 má totiž $\frac{W-\mathrm{E}W}{\sqrt{\mathrm{var}\, W}}$ při $m \to \infty$ a $n \to \infty$ asymptoticky normální rozdělení $N(0,1)$.

Nyní se podíváme na základní charateristiky Wilcoxonova testu. Při jejich vyšetřování budeme uvažovat obecnější skupinu pořadových testů, které je Wilcoxonův test zástupcem. Nechť $a(i)$ je nějaká funkce definovaná pro $i = 1, \ldots, N$, říkáme jí skórová funkce. Budeme se zabývat testy založenými na statistice tvaru

$$S = \sum_{i=1}^{N} c_i \; a(R_i) \; .$$

Čísla $c_1, \ldots, c_N$ nazýváme regresní koeficienty a statistiku S jednoduchá lineární pořadová statistika. Statistiku W získáme z S tak, že položíme $a(i) = i$ a

$$c_i = \begin{cases} 0 & \text{pro} \quad i = 1, \ldots, m \\ 1 & \text{pro} \quad i = m+1, \ldots, N \; . \end{cases}$$

Věta 3 *Označme*

$$\overline{a} = \frac{1}{N} \sum_{i=1}^{N} a(i) \; , \quad \overline{c} = \frac{1}{N} \sum_{i=1}^{N} c_i \; .$$

Jestliže platí H_0, pak

$$\mathrm{E}S = N\, \overline{a}\, \overline{c} \; , \quad \mathrm{var}\, S = \frac{1}{N-1} \; \sum_{i=1}^{N} (a(i) - \overline{a})^2 \; \sum_{j=1}^{N} (c_j - \overline{c})^2 \; .$$

Důkaz: Jestliže platí nulová hypotéza H_0, nabývá náhodná veličina R_i každé z hodnot $1, \ldots, N$ s pravděpodobností $\frac{1}{N}$. Pak

$$\mathrm{E}a(R_i) = \sum_{i=1}^{N} a(i) \frac{1}{N} = \overline{a} \quad \text{a} \quad \mathrm{E}S = \sum_{i=1}^{N} c_i \, \mathrm{E}a(R_i) = \sum_{i=1}^{N} c_i \overline{a} = N\, \overline{a}\, \overline{c} \; .$$

Důkaz vzorce pro rozptyl lze najít například v knize Anděl [1]. □

Speciálně pro Wilcoxonovu statistiku W dostáváme

$$\mathrm{E}W = \frac{n(N+1)}{2} \quad \text{a} \quad \mathrm{var}\, W = \frac{mn(N+1)}{12} \; .$$

Za platnosti H_0 je rozdělení statistiky W symetrické kolem své střední hodnoty. Tuto vlastnost dokážeme opět s použitím obecnější jednoduché lineární pořadové statistiky:

Věta 4 *Nechť platí H_0 a pro $a(1), \ldots, a(N)$ a $c_1, \ldots, c_N$ platí buď*

$$\begin{array}{ll} (i) & a(i) + a(N-i+1) = konstanta\ , \quad i = 1, \ldots, N \qquad nebo \\ (ii) & c_i + c_{N-i+1} = konstanta\ , \quad i = 1, \ldots, N\ . \end{array}$$

Pak je rozdělení statistiky $S = \sum_{i=1}^{N} c_i\ a(R_i)$ symetrické kolem své střední hodnoty ES.

Důkaz: Nechť platí nejprve možnost (i). Pak

$$2N\overline{a} = \sum_{i=1}^{N} a(i) + \sum_{i=1}^{N} a(N-i+1) = N \cdot konstanta\ .$$

Z toho vyplývá $a(i) + a(N-i+1) = 2\overline{a}$ pro $i = 1, \ldots, N$. Označme $S_1 = \sum_{i=1}^{N} c_i\ a(N-R_i+1)$. Použijeme-li vztah $\mathrm{E}S = N\,\overline{a}\,\overline{c}$ z Věty 3, můžeme psát

$$S_1 = \sum_{i=1}^{N} c_i\ a(N-R_i+1) + \sum_{i=1}^{N} c_i\ a(R_i) - \sum_{i=1}^{N} c_i\ a(R_i) = 2\overline{a} \sum_{i=1}^{N} c_i - S = 2\mathrm{E}S - S\ .$$

Náhodný vektor $(N-R_1+1, \ldots, N-R_N+1)$ má stejné rozdělení jako $(R_1, \ldots, R_N)$, S_1 má tedy stejné rozdělení jako S a platí

$$S_1 - \mathrm{E}S_1 = \mathrm{E}S - S \implies P(S - \mathrm{E}S = s) = P(S_1 - \mathrm{E}S_1 = s) = P(\mathrm{E}S - S = s)$$

pro libovolné s. Tím je symetrie S dokázána.

Nechť nyní platí vztah (ii). Podobně jako v prvním případě odvodíme $c_i + c_{N-i+1} = 2\overline{c}$ pro $i = 1, \ldots, N$. Označíme

$$S_2 = \sum_{i=1}^{N} c_{N-i+1}\ a(R_i) = \sum_{i=1}^{N} c_i\ a(R_{N-i+1}).$$

Náhodný vektor $(R_N, \ldots, R_1)$ má stejné rozdělení jako $(R_1, \ldots, R_N)$, S_2 má tedy stejné rozdělení jako S a podobně jako v předchozím případě platí

$$S_2 = 2\overline{c} \sum_{i=1}^{N}\ a(R_i) - S = 2\mathrm{E}S - S \implies S_2 - \mathrm{E}S_2 = \mathrm{E}S - S\ .$$

Důkaz dokončíme stejně jako v případě (i). □

V praxi se místo statistiky W často používá statistika

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} I[Y_j > X_i] \ , \tag{2.3}$$

kde I je indikátorová funkce. Test založený na U se nazývá Mann-Whitneyův test. Pro veličiny W a U platí

$$W = U + \frac{n(n+1)}{2} \ ; \tag{2.4}$$

v literatuře se často uvádí také vztah

$$U = nm + \frac{m(m+1)}{2} - W' \ ,$$

kde W' je součet pořadí $X_1, \ldots, X_m$ ve sdruženém výběru.

I pro statistiku U existují tabulky kritických hodnot, tady je však nutné poznamenat, že maximální možná hodnota U je mn a pokud při pokusu získáme $U > \frac{mn}{2}$, je třeba použít k vyhledání v tabulce hodnotu $mn - U$. Nulovou hypotézu zamítáme, jestliže je výsledná hodnota menší nebo rovna tabelované kritické hodnotě.

Poznámka: Pokud se mezi pozorováními $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ vyskytují stejné hodnoty, mělo by jim být přiřazeno stejné pořadí. Používá se modifikace testu, kdy každému ze shodných pozorování přiřadíme průměr z jejich pořadí. Získáme-li při pokusu například hodnoty 1, 2, 2, 3, budou mít obě hodnoty 2 pořadí $2,5$. Upravenou statistiku W uvažující průměrná pořadí označíme W_p.

Rozdělení této statistiky za nulové hypotézy závisí na počtech shodných pozorování, není tedy možné pro každou situaci vytvářet tabulku kritických hodnot a musíme použít normální aproximaci. K tomu potřebujeme znát střední hodnotu a rozptyl statistiky W_p. Jejich odvození lze najít například v knize Lehmann [10].

Svoji modifikaci při výskytu shodných pozorování má i statistika U:

$$U_p = \sum_{i=1}^{m} \sum_{j=1}^{n} I[Y_j > X_i] + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I[Y_j = X_i] \ . \tag{2.5}$$

Testy založené na statistikách U_p a W_p jsou ekvivalentní a statistiky jsou ve vztahu analogickém k (2.4), tedy

$$W_p = U_p + \frac{n(n+1)}{2} \ .$$

2.1.2 Jednovýběrový Wilcoxonův test

Tento test se někdy nazývá také Wilcoxonův test symetrie. Nechť $X_1, \ldots, X_N$ je náhodný výběr s hustotou f, která je symetrická kolem bodu η. Chceme testovat hypotézu $H_0' : \eta = \eta_0$ proti alternativě $H_1' : \eta > \eta_0$. Položíme $Z_i = X_i - \eta_0$, přitom předpokládáme, že žádná z veličin X_i se nerovná η_0 (v praxi se tyto hodnoty zpravidla vynechávají). Získáme tak nezávislé náhodné veličiny $Z_1, \ldots, Z_N$, pro které platí

$$P(Z_i \leq u) = F(u - \theta) \quad \text{pro } i = 1, \ldots, N \ ,$$

kde distribuční funkce F je spojitá a kde pro F platí $F(x) + F(-x) = 1 \ \forall x$. Místo H_0' a H_1' pak můžeme ekvivalentně psát $H_0 : \theta = 0$ a $H_1 : \theta > 0$.

Veličiny $Z_1, \ldots, Z_N$ seřadíme podle jejich absolutní hodnoty do neklesající posloupnosti a označíme R_i^+ pořadí $|Z_i|$ mezi $|Z_1|, \ldots, |Z_N|$. Jednovýběrový Wilcoxonův test je založen na statistice

$$W^+ = \sum_{Z_i \geq 0} R_i^+ \ ,$$

tedy na součtu pořadí kladných Z_i mezi $|Z_1|, \ldots, |Z_N|$. Dále označme

$$W^- = \sum_{Z_i \leq 0} R_i^+ \quad \text{a} \quad W^* = \sum_{i=1}^{N} R_i^+ \ \operatorname{sign} Z_i \ .$$

Je zřejmé, že $W^+ + W^- = \frac{N(N+1)}{2}$. Zároveň platí vztah $W^+ - W^- = W^*$ a statistiku W^+ lze tedy vyjádřit také ve tvaru

$$W^+ = \frac{1}{2} W^* + \frac{N(N+1)}{4} \ .$$

Kritické hodnoty pro Wilcoxonův jednovýběrový test jsou tabelovány. I tady však platí, že pokud je hodnota statistiky W^+ větší než polovina maxima, tedy než $\frac{N(N+1)}{4}$, použijeme k porovnání s kritickou hodnotou $\frac{N(N+1)}{2} - W^+$. Nulovou hypotézu tudíž zamítáme, je-li hodnota $\min(W^+, W^-)$ menší nebo rovna tabelované kritické hodnotě. Pro velká N můžeme stejně jako v dvourozměrném případě použít normální aproximaci. Veličina $\frac{W^+ - \mathrm{E}W^+}{\sqrt{\operatorname{var} W^+}}$ má totiž asymptoticky rozdělení $N(0,1)$.

Ke zjištění střední hodnoty a rozptylu statistiky W^+ budeme potřebovat následující lemma:

Lemma 5 *Označme $|Z|_{(i)}$ i-tou pořádkovou statistiku z náhodného výběru $|Z_1|, \ldots, |Z_N|$. Pak platí-li H_0, jsou vektory $(\operatorname{sign} Z_1, \ldots, \operatorname{sign} Z_N)$ a $(|Z|_{(1)}, \ldots, |Z|_{(N)})$ nezávislé.*

Důkaz: Důkaz lze nalézt například v knize Anděl [1]. □

Věta 6 *Platí-li H_0, pak*

$$\mathrm{E}W^+ = \frac{N(N+1)}{4} \quad a \quad \mathrm{var}\, W^+ = \frac{N(N+1)(2N+1)}{24} \ .$$

Důkaz: Za platnosti H_0 je rozdělení $Z_1, \ldots, Z_N$ symetrické kolem 0. Platí tedy $\mathrm{E}\,\mathrm{sign}\, Z_i = 0$. Z tvrzení Lemmatu 5 vyplývá $\mathrm{E}(R_i^+ \,\mathrm{sign}\, Z_i) = (\mathrm{E}R_i^+)(\mathrm{E}\,\mathrm{sign}\, Z_i)$, a proto $\mathrm{E}(R_i^+ \,\mathrm{sign}\, Z_i) = 0$. Z toho dostáváme

$$\mathrm{E}W^* = 0 \implies \mathrm{E}W^+ = \frac{1}{2}\,\mathrm{E}W^* + \frac{N(N+1)}{4} = \frac{N(N+1)}{4} \ .$$

Dále máme

$$\begin{aligned} \mathrm{var}\,(R_i^+ \,\mathrm{sign}\, Z_i) &= \mathrm{E}(R_i^+ \,\mathrm{sign}\, Z_i)^2 = \mathrm{E}(R_i^+)^2 \mathrm{E}(\mathrm{sign}\, Z_i)^2 \\ &= \mathrm{E}(R_i^+)^2 = \frac{1}{N}(1^2 + 2^2 + \ldots + N^2) = \frac{(N+1)(2N+1)}{6} \ , \end{aligned}$$

přičemž jsme využili skutečnost, že za platnosti nulové hypotézy má veličina R_i^+ rovnoměrné rozdělení na množině $\{1, \ldots, N\}$.
Analogicky postupujeme při odvození kovariance:

$$\begin{aligned} \mathrm{cov}\,(R_i^+ \,\mathrm{sign}\, Z_i, R_j^+ \,\mathrm{sign}\, Z_j) &= \mathrm{E}(R_i^+ \,\mathrm{sign}\, Z_i \cdot R_j^+ \,\mathrm{sign}\, Z_j) \\ &= \mathrm{E}(R_i^+ \,\mathrm{sign}\, Z_i)\ \mathrm{E}(R_j^+ \,\mathrm{sign}\, Z_j) = 0 \quad \text{pro } i \neq j \ . \end{aligned}$$

Ze vzorce pro rozptyl součtu náhodných veličin dostáváme

$$\mathrm{var}\, W^* = \sum_{i=1}^{N} \mathrm{var}\, R_i^+ \,\mathrm{sign}\, Z_i = \frac{N(N+1)(2N+1)}{6}$$

a odtud

$$\mathrm{var}\, W^+ = \frac{1}{4}\,\mathrm{var}\, W^* = \frac{N(N+1)(2N+1)}{24} \ . \qquad \square$$

Stejně jako u statistiky W dvouvýběrového Wilcoxonova testu je i rozdělení statistiky W^+ symetrické kolem své střední hodnoty. Vyplývá to z tvrzení Věty 2 (ii). Je-li $h(Z) = W^+$, pak $H(-Z) = W^-$. Již víme, že $W^+ + W^- = \frac{N(N+1)}{2}$, a rozdělení statistiky W^+ je tedy symetrické kolem $\frac{N(N+1)}{4} = \mathrm{E}W^+$.

Statistiku W^+ lze ekvivalentně zapsat i v následujícím tvaru, který odvodil Tukey:

$$W^+ = [\text{počet dvojic } (i,j) \ \ 1 \leq i \leq j \leq N \text{ takových, že } Z_i + Z_j > 0] \ . \quad (2.6)$$

Poznámka: Při výskytu pozorování se shodnými absolutními hodnotami mezi $Z_1, \ldots, Z_N$ se analogicky jako v případě dvouvýběrového Wilcoxonova testu používá upravená statistika W_p^+, která počítá s průměrnými pořadími. Při testování se používá normální aproximace, totiž že za platnosti H_0 má $\frac{W_p^+ - \mathrm{E}W_p^+}{\sqrt{\mathrm{var}\, W_p^+}}$ při $N \to \infty$ asymptoticky normální rozdělení $N(0,1)$. Odvození střední hodnoty a rozptylu lze opět nalézt v knize Lehmann [10].

Stejně jako statistiku W^+ můžeme i W_p^+ zapsat v ekvivalentním tvaru

$$\begin{aligned} W_p^+ = \; & [\text{počet dvojic } (i,j) \;\; 1 \le i \le j \le N \text{ takových, že } Z_i + Z_j > 0] \\ & + \tfrac{1}{2}[\text{počet dvojic } (i,j) \;\; 1 \le i \le j \le N \text{ takových, že } Z_i + Z_j = 0] \; . \end{aligned} \tag{2.7}$$

2.2 Galtonův test

2.2.1 Dvouvýběrový Galtonův test

Galtonův test je příkladem jednoho z prvních použití pořadí ve statistice. Mějme náhodná pozorování $X_1, \ldots, X_n$ a na nich nezávislá pozorování $Y_1, \ldots, Y_n$. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení se spojitou distribuční funkcí F a $Y_1, \ldots, Y_n$ je náhodný výběr z rozdělení se spojitou distribuční funkcí G. Testujeme hypotézu

$$\begin{aligned} H_0 : F(u) &= G(u), \; u \in \mathbf{R} \quad \text{proti alternativě} \\ H_1 : G(u) &\le F(u), \; u \in \mathbf{R}. \end{aligned}$$

Platí-li alternativa, říkáme, že náhodná veličina s distribuční funkcí G je stochasticky větší než náhodná veličina s distribuční funkcí F. Tato alternativa je zobecněním alternativy posunutí uvažované u Wilcoxonova testu.

Všech $2n$ pozorování uspořádáme do společné neklesající posloupnosti. Označme $X_{(i)}$ v pořadí i-tou nejmenší veličinu mezi $X_1, \ldots, X_n$ a $Y_{(i)}$ v pořadí i-tou nejmenší veličinu mezi $Y_1, \ldots, Y_n$. Galtonův test je založen na statistice

$$V = [\text{počet případů, kdy } Y_{(i)} \text{ je větší než } X_{(i)}] = \sum_{i=1}^{n} I[Y_{(i)} > X_{(i)}] \; , \tag{2.8}$$

kde I je indikátorová funkce. Je zřejmé, že ve prospěch alternativy mluví vysoké hodnoty V.

Pro Galtonův test neexistují tabulky kritických hodnot, využívá se skutečnosti, že známe rozdělení veličiny V. Hodges dokázal ve svém článku Hodges [6], že za platnosti nulové hypotézy má statistika V diskrétní rovnoměrné rozdělení na hodnotách $0, \ldots, n$. Víme tedy, že $P(V \ge x) = \frac{n-x+1}{n+1}$, a pro získanou hodnotu V můžeme lehce určit hladinu významnosti.

Věta 7 *Platí-li H_0, má náhodná veličina $V = \sum_{i=1}^{n} I[Y_{(i)} > X_{(i)}]$ rovnoměrné rozdělení na hodnotách $0, \ldots, n$.*

Důkaz: Pro zjednodušení důkazu budeme místo neklesajících posloupností vzniklých uspořádáním $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_n$ uvažovat množinu všech $\binom{2n}{n}$ možných uspořádání n nul a n jedniček. Každé takové posloupnosti přiřadíme hodnotu $v =$ (počet případů, kdy i-tá 1 předchází i-tou 0). Tedy například posloupnost 0,1,1,1,0,0,0,1 bude mít hodnotu $v = 2$, protože druhá 1 předchází druhé 0 a třetí 1 předchází třetí 0. Identifikujeme-li 0 (respektive 1) na i-té pozici v uspořádání nul a jedniček s jevem, že v uspořádané posloupnosti z hodnot $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_n$ je na i-té pozici reprezentant prvního (respektive druhého) výběru, pak je zřejmé, že hodnoty v odpovídají hodnotám statistiky V.

Pro posloupnost délky $2n$ může v nabývat hodnot $0, \ldots, n$. Zavedeme si označení $[a_1, \ldots, a_{2n}]$ pro hodnotu v posloupnosti $a_1, \ldots, a_{2n}$. Podaří-li se nám najít vzájemně jednoznačné zobrazení, které zobrazuje množinu uspořádání s hodnotou $v = x$ na množinu uspořádání s hodnotou $v = x - 1$ pro $x = 1, \ldots, n$, pak počet uspořádání s $v = x$ nezávisí na x, všechny hodnoty v jsou stejně pravděpodobné a z toho vyplývá, že v, a tedy i V, má rovnoměrné rozdělení na $0, \ldots, n$.

Nechť T_n je zobrazení s následujícími vlastnostmi:

(i) Definiční obor T_n tvoří množina všech uspořádání délky $2n$, pro které $v \geq 0$.

(ii) Obor hodnot T_n tvoří množina všech uspořádání délky $2n$, pro které $v < n$.

(iii) T_n je vzájemně jednoznačné zobrazení.

(iv) $[T_n(a_1, \ldots, a_{2n})] = [a_1, \ldots, a_{2n}] - 1$.

T_n splňuje požadavky, které máme na hledané zobrazení.

K důkazu existence takového zobrazení bude třeba uvažovat první místo v posloupnosti, kterému předchází stejný počet 0 a 1. Pro každé uspořádání $a_1, \ldots, a_{2n}$ nechť tedy k je nejmenší přirozené číslo takové, že $a_1 + \cdots + a_{2k} = k$ (k se může rovnat i n). Pak platí

$$[a_1, \ldots, a_{2n}] = [a_1, \ldots, a_{2k}] + [a_{2k+1}, \ldots, a_{2n}] \ . \tag{2.9}$$

Navíc protože a_{2k} je první člen posloupnosti, kterým nastává rovnost počtu 0 a 1, musí být $[a_1, \ldots, a_{2k}]$ rovno buď 0 nebo k. Jestliže $[a_1, \ldots, a_{2k}] = k$, pak určitě $a_1 = 1$, $a_{2k} = 0$ a $[a_2, \ldots, a_{2k-1}] = k - 1$.

Nyní si induktivně definujeme T_n.

1. Nechť $T_1(1,0) = (0,1)$.
   $[0,1] = 0$, $[1,0] = 1$ a T_1 zřejmě splňuje podmínky (i)-(iv).

2. Předpokládejme, že pro $m < n$ máme definováno T_m, které má všechny požadované vlastnosti. Pro libovolnou posloupnost $a_1, \ldots, a_{2n}$ s kladnou hodnotou v nechť $T_n(a_1, \ldots, a_{2n})$ rovná se

   (a) $(a_1, \ldots, a_{2k}, T_{n-k}(a_{2k+1}, \ldots, a_{2n}))$, jestliže $[a_{2k+1}, \ldots, a_{2n}] > 0$,

   (b) $(0, a_{2k+1}, \ldots, a_{2n}, 1, a_2, \ldots, a_{2k-1})$, jestliže $[a_{2k+1}, \ldots, a_{2n}] = 0$.

Zbývá nám ukázat, že zobrazení T_n splňuje podmínky (i)-(iv). Vlastnost (i) má T_n z definice. Podmínka (iv) platí pro (a) díky (2.9) a indukčnímu předpokladu. Pro (b), tedy pro $[a_{2k+1}, \ldots, a_{2n}] = 0$, máme

$$0 < [a_1, \ldots, a_{2k}] = k = [a_1, \ldots, a_{2n}] \quad \text{a} \quad [0, a_{2k+1}, \ldots, a_{2n}, 1] = 0$$

$$\Longrightarrow \ [T_n(a_1, \ldots, a_{2n})] = [0, a_{2k+1}, \ldots, a_{2n}, 1] + [a_2, \ldots, a_{2k-1}] = k - 1 \ .$$

Zobrazení T_n tedy splňuje podmínku (iv). Z toho vyplývá, že obor hodnot T_n je obsažen v množině posloupností s hodnotou $v < n$.

Každá taková posloupnost má svůj vzor vzhledem k T_n. Jestliže totiž $[a_{2k+1}, \ldots, a_{2n}] < n - k$, pak existenci vzoru zaručí možnost (a) díky předpokladu, že T_{n-k} má všechny požadované vlastnosti. Jestliže $[a_{2k+1}, \ldots, a_{2n}] = n - k$, posloupnost $a_{2k+1}, \ldots, a_{2n}$ obsahuje stejný počet 0 a 1 a každá jednička předchází nulu odpovídajícího pořadí. V posloupnosti $1, a_{2k+1}, \ldots, a_{2n}, 0, a_2, \ldots, a_{2k-1}$ nastává tudíž rovnost mezi počtem 0 a 1 až díky poslednímu členu. Dále dostáváme

$$k > [a_1, \ldots, a_{2k}] = 0 \quad \Longrightarrow \quad a_1 = 0,\ a_{2k} = 1,\ [a_2, \ldots, a_{2k-1}] = 0$$

$$\Longrightarrow [1, a_{2k+1}, \ldots, a_{2n}, 0, a_2, \ldots, a_{2k-1}] = n - k + 1 \ .$$

Vzor pro $a_1, \ldots, a_{2n}$ pak získáme díky možnosti (b), neboť

$$T_n(1, a_{2k+1}, \ldots, a_{2n}, 0, a_2, \ldots, a_{2k-1}) = a_1, \ldots, a_{2n} \ .$$

Ukázali jsme, že T_n splňuje podmínku (ii). Navíc je zřejmé, že každá posloupnost s kladnou hodnotou v má svůj jedinečný obraz vzhledem k T_n, a podmínka (iii) je tedy také splněna.
Tím je věta dokázána. □

Z rovnoměrného rozdělení V vyplývá, že $\mathrm{E}V = \frac{n}{2}$, $\operatorname{var} V = \frac{n^2+2n}{12}$ a rozdělení V je symetrické kolem své střední hodnoty.

V praktických úlohách se samozřejmě často setkáváme i s náhodnými výběry, které nemají stejnou velikost. V takových případech můžeme použít přirozené rozšíření Galtonova testu. Nechť $X_1, \ldots, X_{n_1}$ je první náhodný výběr a $Y_1, \ldots, Y_{n_2}$ je druhý náhodný výběr. Jestliže $n_2 = n_1 + k(n_1 + 1)$, pozorování $Y_{(k+1)}, Y_{(2k+2)}, \ldots, Y_{(n_1 k + n_1)}$ dělí druhý náhodný výběr na (n_1+1) částí, z nichž každá obsahuje k pozorování. Pozorování $Y_{(k+1)}, Y_{(2k+2)}, \ldots, Y_{(n_1 k + n_1)}$ budeme považovat za reprezentanty většího náhodného výběru a na jejich základě spočítáme hodnotu V stejným postupem, jako kdybychom ji počítali pro dva náhodné výběry o velikosti n_1. Například pro $n_1 = 3$, $n_2 = 11$ reprezentujeme větší výběr v pořadí třetím, šestým a devátým pozorováním. Náhodná veličina V může nabývat hodnot $0, \ldots, n_1$ a má stále rovnoměrné rozdělení, neboť množině uspořádání, která mají hodnotu $V = x$, odpovídá pro $\forall x = 0, \ldots, n_1$ shodný počet uspořádání v původním problému.

Jestliže $n_2 - n_1$ není dělitelné číslem $(n_1 + 1)$, nemůžeme použít přímo výše popsanou metodu, ale základní myšlenka zůstane stejná. Reprezentanty většího výběru vybíráme tak, aby rozdělili výběr na části, v nichž se počet pozorování liší nejvýše o 1. Lze dokázat, že pokud náhodně vybereme jedno ze všech takových dělení, výsledné V má opět rovnoměrné rozdělení. V praxi však spíše zjistíme rozdělení hodnot V mezi děleními a za hodnotu V pro dané uspořádání bereme přirozené číslo nejbližší střední hodnotě tohoto rozdělení.

Poznámka: Při výskytu většího počtu shodných pozorování mezi $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_n$, můžeme použít modifikaci testu ve tvaru

$$V_p = \sum_{i=1}^{n} I[Y_{(i)} > X_{(i)}] + \frac{1}{2} \sum_{i=1}^{n} I[Y_{(i)} = X_{(i)}] \ .$$

2.2.2 Jednovýběrový Galtonův test

Nechť $Z_1, \ldots, Z_N$ jsou náhodné veličiny s distribuční funkcí

$$P(Z_i \leq u) = F(u - \theta) \quad \text{pro } i = 1, \ldots, N \ ,$$

která je spojitá a pro kterou platí $F(x) + F(-x) = 1 \ \forall x$. Budeme testovat hypotézu $H_0 : \theta = 0$ proti alternativě $H_1 : \theta > 0$.

Základní úvaha je podobná jako u dvouvýběrového problému. Uvažujme náhodné výběry $Z_1, \ldots, Z_N$ a $-Z_1, \ldots, -Z_N$ a nechť $N = 2m$ nebo $N = 2m-1$. Kdybychom použili myšlenku dvouvýběrového Galtonova testu a porovnali odpovídající pořádkové statistiky, získali bychom statistiku $\sum_{i=1}^{N} I[Z_{(i)} > (-Z)_{(i)}]$. Víme však, že $(-Z)_{(i)} = -Z_{(N-i+1)}$, a jednovýběrový

Galtonův test proto můžeme založit na podobné, ale jednodušší statistice

$$V' = \sum_{i=1}^{m} I[Z_{(i)} + Z_{(N-i+1)} > 0] \ . \tag{2.10}$$

Testujeme hypotézu, že je rozdělení náhodných veličin $Z_1, \ldots, Z_N$ symetrické kolem 0. Hypotézu tedy zamítneme, je-li hodnota V' příliš velká.

Stejně jako jsme v případě dvouvýběrového Galtonova testu znali rozdělení statistiky V, známe i rozdělení náhodné veličiny V'. K jeho odvození budeme potřebovat několik pomocných tvrzení.

Lemma 8 *Nechť $U_1, U_2, \ldots$ je posloupnost nezávislých náhodných veličin, pro které platí $P(U_i = 1) = P(U_i = -1) = \frac{1}{2}$ pro $\forall i$. Označme $S_n = \sum_{i=1}^{n} U_i$. Pak pro $L_n = \sum_{i=1}^{n}(I[S_i > 0] + I[S_i = 0, S_{i-1} > 0])$ platí*

$$\begin{array}{lll} P(L_{2n} = 2k) & = \binom{2k}{k}\binom{2n-2k}{n-k}2^{-2n} & k = 0, \ldots, n \\ & = 0 & \textit{jinak} \ . \end{array}$$

Důkaz: Důkaz lze najít například v knize Feller [4]. □

Nechť $a = 2b$ je sudé číslo a mějme $z_1 < \ldots < z_a$ reálná čísla s různými abso- lutními hodnotami. Definujme A jako počet kladných součtů $z_i + z_{a-i+1}$, $i = 1, \ldots, b$. Nyní seřadíme $z_1, \ldots, z_a$ do klesající posloupnosti podle absolutní hodnoty, dostaneme tak posloupnost $t_1, \ldots, t_a$. Později se nám bude hodit poznámka, že pokud se některé z čísel $z_1, \ldots, z_a$ rovná 0, musí nutně být $t_a = 0$. Definujeme si funkce $U(j) = \frac{1}{2}(\operatorname{sign} t_j + 1)$, $S^+(q) = \sum_{j=1}^{q-1} U(j)$ a $S^-(q) = (q-1) - S^+(q)$, kde $q = 1, \ldots, a$. Položíme

$$L = \sum_{q=1}^{a}(I[S^+(q) > S^-(q)] + I[S^+(q) = S^-(q), U(q) = 1]) \ .$$

Situace tedy vypadá následovně: máme posloupnost $U(1), \ldots, U(a)$, tedy posloupnost nul, jedniček a $\frac{1}{2}$, přičemž $\frac{1}{2}$ se může v posloupnosti vyskytovat nejvýše jednou a navíc jen na posledním místě. Pro $q = 1, \ldots, a$ máme

$$\begin{array}{rcl} S^+(q) & = & [\text{počet 1 v prvních } q-1 \text{ prvcích posloupnosti}] \ , \\ S^-(q) & = & [\text{počet 0 v prvních } q-1 \text{ prvcích posloupnosti}] \ . \end{array}$$

Jev $(S^+(q) > S^-(q))$ tedy značí skutečnost, že v prvních $q-1$ prvcích posloupnosti je více 1 než 0. Jev $(S^+(q) = S^-(q), U(q) = 1)$ nastane tehdy, je-li v prvních $q-1$ prvcích posloupnosti stejný počet 0 a 1 a na q-tém místě je 1. Poznamenejme ještě, že a je sudé číslo, nemůže proto nastat jev $(S^+(a) = S^-(a))$.

Lemma 9 *Za výše uvedených podmínek platí* $2A = L$.

Důkaz: Mějme nějaké pevné i z $i = 1, \ldots, b$. Je-li $z_i + z_{a-i+1} > 0$, pak $z_{a-i+1} > 0$ a $|z_{a-i+1}| > |z_i|$.

Nechť nejprve $z_i < 0$. V posloupnosti $U(1), \ldots, U(a)$ je pak i-tá jednička před i-tou nulou. Označme q_1 pořadí i-té jedničky a q_0 pořadí i-té nuly v posloupnosti $U(1), \ldots, U(a)$. V prvních $q_1 - 1$ prvcích posloupnosti musí být počet 1 vyšší nebo rovný počtu 0 a z toho vyplývá, že buď $I[S^+(q_1) > S^-(q_1)] = 1$, nebo $I[S^+(q_1) = S^-(q_1), U(q_1) = 1] = 1$. Dále v prvních $q_0 - 1$ prvcích posloupnosti je nutně více 1 než 0, a proto $I[S^+(q_0) > S^-(q_0)] = 1$.

Nechť nyní $z_i > 0$. V tom případě nahradí nulu uvažovanou v předchozí možnosti jednička a je zřejmé, že i pak bude platit

$$(I[S^+(q_1) > S^-(q_1)] + I[S^+(q_1) = S^-(q_1), U(q_1) = 1]) = 1 \text{ a}$$
$$(I[S^+(q_0) > S^-(q_0)] + I[S^+(q_0) = S^-(q_0), U(q_0) = 1]) = 1 \ ,$$

kde q_0 v tomto případě značí pořadí 1, která v posloupnosti $U(1), \ldots, U(a)$ odpovídá $\frac{1}{2}(\operatorname{sign} z_i + 1)$.

Zbývá nám poslední možnost $z_i = 0$. Pak je v posloupnosti více 1 než 0. Jak už jsme si řekli, 0 je posledním prvkem posloupnosti $t_1, \ldots, t_a$ a nemůže nastat $(S^+(a) = S^-(a))$. Z toho vyplývá, že v prvních $a - 1$ prvcích posloupnosti $U(1), \ldots, U(a)$ je více 1 než 0 a $I[S^+(a) > S^-(a)] = 1$. Navíc samozřejmě stále platí $(I[S^+(q_1) > S^-(q_1)] + I[S^+(q_1) = S^-(q_1), U(q_1)]) = 1$.

Dospěli jsme tedy k závěru, že alespoň $2A$-krát z $q = 1, \ldots, a$ platí $(I[S^+(q) > S^-(q)] + I[S^+(q) = S^-(q), U(q) = 1]) = 1$. Analogickou úvahou pro $b - A$ součtů, pro které $z_i + z_{a-i+1} \leq 0$ zjistíme, že alespoň $2(b - A)$-krát platí $(I[S^+(q_1) > S^-(q_1)] + I[S^+(q_1) = S^-(q_1), U(q_1)]) = 0$. Celkově tedy máme $2A = L$. □

Vraťme se zpátky k jednovýběrovému Galtonovu testu a statistice V'. Nechť $N = 2m$. Z Lemmat 8 a 9 okamžitě vyplývá, že za platnosti nulové hypotézy, tedy jestliže mají náhodné veličiny $Z_1, \ldots, Z_N$ spojité rozdělení symetrické kolem 0, platí

$$\begin{aligned} P(V' = x) = P(L_n = 2x) \ &= \tbinom{2x}{x}\tbinom{2m-2x}{m-x}2^{-n} \quad x = 0, \ldots, m \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{jinak} \ . \end{aligned}$$

Je zřejmé, že rozdělení náhodné veličiny V' je symetrické kolem hodnoty $\frac{m}{2}$, máme tedy $\mathrm{E}V' = \frac{m}{2}$.

Hladinu významnosti můžeme opět lehce určit pomocí

$$P(V' \geq x) = 2^{-n} \sum_{k=x}^{m} \binom{2k}{k}\binom{2m-2k}{m-k} \ .$$

Už pro středně velké výběry můžeme použít aproximaci. Nechť $N = 2m$ nebo $N = 2m - 1$, pak

$$P(V' \geq x) \sim 1 - \frac{2}{\pi} \arcsin\left(\frac{x}{m}\right)^{\frac{1}{2}} . \tag{2.11}$$

Poznámka: Vyskytuje-li se mezi $Z_1, \ldots, Z_N$ větší počet pozorování se stejnou absolutní hodnotou, můžeme použít modifikaci statistiky V' tvaru

$$V'_p = \sum_{i=1}^{m} I[Z_{(i)} + Z_{(N-i+1)} > 0] + \frac{1}{2} \sum_{i=1}^{m} I[Z_{(i)} + Z_{(N-i+1)} = 0] .$$

2.3 Znaménkový test

Znaménkový test je jednovýběrový test. Nechť $Z_1, \ldots, Z_N$ je náhodný výběr z rozdělení se spojitou distribuční funkcí

$$P(Z_i \leq u) = F(u - \theta) \quad \text{pro } i = 1, \ldots, N ,$$

pro kterou platí $F(x) + F(-x) = 1 \ \forall x$. Testujeme hypotézu $H_0 : \theta = 0$ proti alternativě $H_1 : \theta > 0$. Pokud je některé pozorování ze $Z_1, \ldots, Z_N$ nulové, zpravidla ho vynecháme. Znaménkový test testuje uvedenou hypotézu pomocí jednoduché statistiky

$$S_N = [\text{počet } Z_i > 0,\ i = 1, \ldots, N] = \sum_{i=1}^{N} I[Z_i > 0] .$$

Platí-li H_0, je Z_i pro $i = 1, \ldots, N$ kladné s pravděpodobností $\frac{1}{2}$ a S_N má zřejmě binomické rozdělení $Bi(N, \frac{1}{2})$. Odtud vyplývá, že pro N malé můžeme k testu využít kritické hodnoty binomického rozdělení. Hypotézu zamítáme, je-li S_N příliš velké, tedy příliš blízké číslu N.

Pro velká N použijeme normální aproximaci. Z binomického rozdělení S_N víme, že za nulové hypotézy platí

$$\mathrm{E}S_N = \frac{N}{2} \quad \text{a} \quad \operatorname{var} S_N = \frac{N}{4} .$$

Pro $N \geq 20$ tak můžeme využít skutečnosti, že $\frac{S_N - \frac{N}{2}}{\frac{1}{2}\sqrt{N}}$ má pro $N \to \infty$ asymptoticky normální rozdělení $N(0, 1)$.

Poznámka: Případná shodná pozorování nemají na znaménkový test vliv.

Kapitola 3

R-odhady

Nebude-li uvedeno jinak, budeme ve všech úvahách o R-odhadech dále předpokládat, že všechna pozorování, na nichž je odhad založen, jsou různá; respektive, že mají různé absolutní hodnoty u některých odhadů parametru polohy jednoho výběru. Postupům při výskytu shodných pozorování se budeme věnovat v poznámkách.

3.1 Odvození a definice

3.1.1 Dvouvýběrový problém

Nechť $X_1, \ldots, X_m$ a $Y_1, \ldots, Y_n$ jsou nezávislé náhodné výběry, pro které platí

$$\begin{array}{lll} P(X_i \leq u) & = F(u) & \text{pro } i = 1, \ldots, m \quad \text{a} \\ P(Y_j \leq u) & = F(u - \Delta) & \text{pro } j = 1, \ldots, n \ , \end{array} \tag{3.1}$$

kde F je spojitá distribuční funkce. Budeme se snažit nalézt odhad parametru Δ, který bude dobrý i v případě, že náhodné veličiny nemají normální rozdělení. Posunutím $Y_1, \ldots, Y_n$ doleva o vzdálenost Δ získáme náhodný výběr $Y_1 - \Delta, \ldots, Y_n - \Delta$. Náhodné veličiny $X_1, \ldots, X_m$ a $Y_1 - \Delta, \ldots, Y_n - \Delta$ jsou pak nezávislé a stejně rozdělené. Přirozeným odhadem pro Δ pak je velikost posunutí, které potřebujeme, aby se množiny $\{X_1, \ldots, X_m\}$ a $\{Y_1 - \Delta, \ldots, Y_n - \Delta\}$ co nejvíce přiblížily. Definice takového přiblížení mohou být různé. Jednou z možností je považovat tyto množiny za blízké, jestliže polovina nenulových rozdílů $(Y_j - \Delta) - X_i$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ je kladná a polovina záporná. Existuje buď jediná taková hodnota Δ, ta nám pak může sloužit jako odhad, nebo celý interval hodnot - v tom případě je přirozeným odhadem střed tohoto intervalu.

Podíváme-li se na statistiku U (viz (2.3)), vidíme, že naše definice přiblížení souvisí s dvouvýběrovým Wilcoxonovým testem. Obecně máme-li test hypotézy $H_0 : \Delta = 0$, jehož statistika je symetrická kolem nějaké hodnoty μ, můžeme dvě množiny považovat za blízké, jestliže po dosazení jejich prvků dává tato statistika hodnotu μ. Uvažujeme-li totiž test hypotézy $H_0 : \Delta = 0$ proti oboustranné alternativě $H_1 : \Delta \neq 0$, který zamítá hypotézu, jestliže je hodnota jeho statistiky příliš velká nebo příliš malá, je nulové hypotéze nejpříznivější právě střední hodnota rozdělení této statistiky.

Naše úvahy nyní zformalizujeme. Budeme značit $x = (x_1, \ldots, x_m)$, $y = (y_1, \ldots, y_n)$, $X = (X_1, \ldots, X_m)$ a $Y = (Y_1, \ldots, Y_n)$. Mějme statistiku $h(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ pro test hypotézy $H_0 : \Delta = 0$ proti alternativě $H_1 : \Delta > 0$. Dále budeme předpokládat

(A) $h(x_1, \ldots, x_m, y_1 + a, \ldots, y_n + a)$ je neklesající funkcí a pro všechna x a y,

(B) jestliže $\Delta = 0$, rozdělení $h(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ je symetrické kolem pevně dané hodnoty μ, která nezávisí na F, a to pro (i) všechny spojité distribuční funkce F, (ii) všechny spojité distribuční funkce F, pro které platí $F(x) + F(-x) = 1 \ \forall x$.

Položme

$$\Delta^* = \sup\left(\Delta : h(x, y - \Delta) > \mu\right) \quad \text{a} \quad \Delta^{**} = \inf\left(\Delta : h(x, y - \Delta) < \mu\right) . \tag{3.2}$$

Pro h splňující dané podmínky navrhli Hodges a Lehmann používat

$$\hat{\Delta} = \frac{\Delta^* + \Delta^{**}}{2} \tag{3.3}$$

jako odhad pro parametr posunutí Δ.

Příklad 3.1: Nechť $m = n$. Medián hodnot $x_1, \ldots, x_m$ budeme označovat $\tilde{x}$ nebo $\mathrm{med}\,(x)$. Tedy

$$\begin{aligned} \tilde{x} = \mathrm{med}\,(x) &= x_{(k+1)} && \text{jestliže } m = 2k + 1 \\ &= \tfrac{x_{(k)} + x_{(k+1)}}{2} && \text{jestliže } m = 2k \ , \end{aligned}$$

kde $x_{(k)}$ a $x_{(k+1)}$ jsou pořádkové statistiky z $x_1, \ldots, x_m$. Budeme uvažovat statistiku $h(x, y) = \tilde{y} - \tilde{x}$. Statistika zřejmě splňuje podmínku (A). Dále platí $\tilde{y} - \tilde{x} + \tilde{x} - \tilde{y} = 0$, statistika tudíž splňuje podmínku (ii) z Věty 1 pro $\mu = 0$ a tedy i podmínku (B).

$$\begin{aligned} \Delta^* &= \sup\left(\Delta : h(x, y - \Delta) > 0\right) = \sup\left(\Delta : y \widetilde{-} \Delta - \tilde{x} > 0\right) \\ &= \sup\left(\Delta : \tilde{y} - \tilde{x} > \Delta\right) = \tilde{y} - \tilde{x} \ . \end{aligned}$$

Analogicky $\Delta^{**} = \tilde{y} - \tilde{x}$ a $\hat{\Delta} = \tilde{y} - \tilde{x} = h(x, y)$. $\diamond$

V příkladu jsme pro jednoduchost použili statistiku $h(x, y) = \tilde{y} - \tilde{x}$, dále se však budeme zabývat odhady $\hat{\Delta}$ založenými na pořadových testech. Takové odhady nazýváme R-odhady.

Dnes se R-odhady často ekvivalentně definují i jiným způsobem. Vyjdeme ze stejné myšlenky, použijeme ale statistiku speciálního tvaru, konkrétně jednoduchou lineární pořadovou statistiku, pro kterou

$$c_i = \begin{cases} 0 & \text{pro} \quad i = 1, \ldots, m \\ 1 & \text{pro} \quad i = m+1, \ldots, N \end{cases} .$$

Náhodný výběr $X_1, \ldots, X_m, Y_1 - \Delta, \ldots, Y_n - \Delta$ uspořádáme vzestupně podle velikosti. Označme $N = m + n$ a $R_i(\Delta)$ pořadí i-té veličiny z $X_1, \ldots, X_m, Y_1 - \Delta, \ldots, Y_n - \Delta$ v tomto uspořádaném náhodném výběru. Test hypotézy $H_0 : \Delta = 0$ můžeme založit na statistice

$$S(\Delta) = \sum_{i=1}^{N} c_i \; a(R_i(\Delta)) \; , \tag{3.4}$$

kde $a(1), \ldots, a(N)$ jsou dané skóry generované funkcí $J : [0, 1) \to \mathbf{R}$ jako $a(i) = J(\frac{i}{N+1})$. Obvykle se navíc předpokládá, že $J(1-t) = -J(t)$. Například pro dvouvýběrový Wilcoxonův test volíme $J(t) = t - \frac{1}{2}$, tedy $a(i) = \frac{i}{N+1} - \frac{1}{2}$.

Z Vět 3 a 4 víme, že za platnosti H_0 je rozdělení $S(0)$ symetrické kolem $N\,\overline{a}\,\overline{c}$. Jelikož $J(1 - t) = -J(t)$, platí $a(i) = -a(N - i + 1)$, $i = 1, \ldots, N$ a $J(\frac{1}{2}) = 0$. Z toho vyplývá $\sum_{i=1}^{N} a(i) = 0$, $\overline{a} = 0$ a rozdělení $S(0)$ je symetrické kolem 0. Jako odhad parametru posunutí Δ vezmeme hodnotu $\hat{\Delta}$, pro kterou $S(\hat{\Delta}) = 0$. Statistika $S(\Delta)$ však není spojitá, takové $\hat{\Delta}$ tedy nemusí existovat. Podobně jako v předchozí definici proto definujeme

$$\hat{\Delta} = \frac{\Delta^- + \Delta^+}{2} \; ,$$

kde $\Delta^- = \sup\,(\Delta : S(\Delta) > 0)$ a $\Delta^+ = \inf\,(\Delta : S(\Delta) < 0)$.

3.1.2 Jednovýběrový problém

Nechť $Z_1, \ldots, Z_N$ je náhodný výběr z rozdělení s distribuční funkcí

$$P(Z_i \le u) = F(u - \theta) \quad \text{pro } i = 1, \ldots, N \; , \tag{3.5}$$

která je spojitá a pro kterou platí $F(x) + F(-x) = 1 \; \forall x$. Chceme odhadnout parametr polohy θ. Provedeme podobnou úvahu jako v případě dvou

výběrů. Posunutím $Z_1, \ldots, Z_N$ doleva o vzdálenost θ získáme posloupnost nezávislých náhodných veličin $Z_1 - \theta, \ldots, Z_N - \theta$. S přihlédnutím ke statistice jednovýběrového Wilcoxonova testu (viz (2.6)) můžeme θ odhadnout například jako velikost posunutí potřebného k tomu, aby z nenulových součtů $(Z_i - \theta) + (Z_j - \theta)$, $1 \leq i \leq j \leq N$ byla právě polovina kladných a polovina záporných.

Obecně založíme odhad na statistice $h(Z_1, \ldots, Z_N)$ pro test hypotézy $H_0 : \theta = 0$ proti alternativě $H_1 : \theta > 0$. Opět budeme používat obvyklé značení $z = (z_1, \ldots, z_N)$, $Z = (Z_1, \ldots, Z_N)$. O h budeme předpokládat:

(C) $h(z_1 + a, \ldots, z_N + a)$ je neklesající funkcí a pro všechna z,

(D) jestliže $\theta = 0$, rozdělení $h(Z_1, \ldots, Z_N)$ je symetrické kolem pevně dané hodnoty μ, která nezávisí na F, a to pro všechny spojité distribuční funkce F, pro které platí $F(x) + F(-x) = 1 \ \forall x$.

Jako odhad θ budeme používat hodnotu

$$\hat{\theta} = \frac{\theta^* + \theta^{**}}{2} \ , \tag{3.6}$$

kde

$$\theta^* = \sup\left(\theta : h(z - \theta) > \mu\right) \quad \text{a} \quad \theta^{**} = \inf\left(\theta : h(z - \theta) < \mu\right). \tag{3.7}$$

Pro ekvivalentní definici R-odhadů pro jednovýběrový problém použijeme test hypotézy $H_0 : \theta = 0$ založený na statistice

$$S'(\theta) = \sum_{i=1}^{N} c_i \ a(R_i^+(\theta)) \ ,$$

kde $R_i^+(\theta)$ je pořadí $|Z_i - \theta|$ mezi $|Z_1 - \theta|, \ldots, |Z_N - \theta|$ a $c_i = \text{sign}(Z_i - \theta)$.

Skóry $a(1), \ldots, a(N)$ jsou generovány neklesající skórovou funkcí $J : [0, 1) \to \mathbf{R}^+$, pro kterou platí $J(0) = 0$, a to jako $a(i) = J(\frac{i}{N+1})$. Za platnosti H_0 splňuje statistika $S'(0)$ podmínku (ii) z Věty 2 pro $\mu = 0$ a její rozdělení je tedy symetrické kolem 0. Jako odhad θ vezmeme takovou hodnotu $\hat{\theta}$, pro kterou $S'(\hat{\theta}) = 0$, případně

$$\hat{\theta} = \frac{\theta^- + \theta^+}{2} \ ,$$

kde $\theta^- = \sup\left(\theta : S'(\theta) > 0\right)$ a $\theta^+ = \inf\left(\theta : S'(\theta) < 0\right)$.

K odhadu parametru polohy θ jednoho náhodného výběru $Z_1, \ldots, Z_N$ lze použít i dvouvýběrové pořadové testy. Budeme uvažovat statistiku analogickou statistice (3.4). Za první náhodný výběr budeme pro jakékoliv θ považovat $-Z_1+\theta, \ldots, -Z_N+\theta$, za druhý náhodný výběr $Z_1-\theta, \ldots, Z_N-\theta$. Mějme tedy statistiku

$$S'(\theta) = \sum_{i=1}^{2N} c_i \; a(R_i(\theta)) \; ,$$

kde $R_i(\theta)$ je pořadí i-té hodnoty z $-Z_1+\theta, \ldots, -Z_N+\theta, Z_1-\theta, \ldots, Z_N-\theta$ v posloupnosti uspořádané vzestupně z těchto hodnot a

$$c_i = \begin{cases} 0 & \text{pro} \quad i = 1, \ldots, N \\ 1 & \text{pro} \quad i = N+1, \ldots, 2N \; . \end{cases}$$

Skóry $a(1), \ldots, a(2N)$ jsou opět generovány funkcí $J : [0,1) \to \mathbf{R}$, pro kterou platí $J(1-t) = -J(t)$, a to jako $a(i) = J(\frac{i}{2N+1})$. Analogickou úvahou jako v případě statistiky (3.4) dojdeme k definici odhadu parametru θ jako hodnoty $\hat{\theta}$, pro kterou $S'(\hat{\theta}) = 0$. Pokud takové $\hat{\theta}$ neexistuje, definujeme stejně jako v předchozích případech

$$\hat{\theta} = \frac{\theta^- + \theta^+}{2} \; ,$$

kde $\theta^- = \sup(\theta : S'(\theta) > 0)$ a $\theta^+ = \inf(\theta : S'(\theta) < 0)$.

3.2 Příklady odhadů

Naprostou většinu R-odhadů nelze vyjádřit explicitně a musíme je počítat iteračně pomocí numerických metod. Uvedeme si několik odhadů, jejichž tvar umíme odvodit.

3.2.1 Odhad založený na dvouvýběrovém Wilcoxonově testu

Odhad založený na dvouvýběrovém Wilcoxonově testu je nejznámější R-odhad, někdy se nazývá také Hodges-Lehmannův odhad. Budeme pracovat s testovou statistikou v Mann-Whitneyově tvaru. Nechť tedy $h(X,Y)$ je počet dvojic (X_i, Y_j), $i = 1, \ldots, m$, $j = 1, \ldots, n$, pro které platí $X_i < Y_j$. $h(X,Y)$ může nabývat hodnot $0, 1, \ldots, mn$.

V kapitole 2.1.1 o dvouvýběrovém Wilcoxonově testu jsme ukázali, že $h(X,Y) = W - \frac{n(n+1)}{2}$, viz (2.4). Víme, že pro $\Delta = 0$ je rozdělení statistiky

W symetrické kolem $\frac{n(N+1)}{2}$, tudíž rozdělení $h(X,Y)$ je symetrické kolem $\frac{n(N+1)}{2} - \frac{n(n+1)}{2} = \frac{n(N-n)}{2} = \frac{nm}{2}$. Statistika h tedy splňuje podmínku (B), splnění podmínky (A) je zřejmé a odhad můžeme odvodit podle vzorce (3.3).

Rozdíly $Y_j - X_i$ pro $i = 1, \ldots, m$, $j = 1, \ldots, n$ seřadíme do neklesající posloupnosti, jejíž prvky označíme $D_{(1)} < \ldots < D_{(mn)}$. Pro explicitní vyjádření odhadu $\hat{\Delta}$ budeme rozlišovat dva případy. Nechť nejprve $mn = 2k + 1$, tedy mn je liché číslo. Pak je rozdělení $h(X,Y)$ symetrické kolem hodnoty $\mu = k + \frac{1}{2}$, což je hodnota, kterou $h(X,Y)$ nemůže nabývat. Vyjádříme

$$\begin{aligned}
\Delta^* &= \sup\,(\Delta : h(X, Y - \Delta) > k + \frac{1}{2}) \\
&= \sup\,(\Delta : \text{více než } (k + \frac{1}{2})\text{-krát platí } X_i < Y_j - \Delta) \\
&= \sup\,(\Delta : \text{více než } k + \frac{1}{2} \text{ rozdílů } Y_j - X_i \text{ je větších než } \Delta) \\
&= \sup\,(\Delta : D_{(k+1)} > \Delta) = D_{(k+1)}\ .
\end{aligned}$$

Podobně

$$\begin{aligned}
\Delta^{**} &= \inf\,(\Delta : h(X, Y - \Delta) < k + \frac{1}{2}) \\
&= \inf\,(\Delta : \text{méně než } k + \frac{1}{2} \text{ rozdílů } Y_j - X_i \text{ je větších než } \Delta) \\
&= \inf\,(\Delta : D_{(k+1)} \leq \Delta) = D_{(k+1)}
\end{aligned}$$

a celkově dostaneme $\hat{\Delta} = D_{(k+1)}$.

Uvažujme druhý případ, kdy $mn = 2k$, tedy mn je sudé číslo. Potom $\mu = k$ a

$$\begin{aligned}
\Delta^* &= \sup\,(\Delta : h(X, Y - \Delta) > k) \\
&= \sup\,(\Delta : \text{více než } k \text{ rozdílů } Y_j - X_i \text{ je větších než } \Delta) \\
&= \sup\,(\Delta : D_{(k)} > \Delta) = D_{(k)}
\end{aligned}$$

a

$$\begin{aligned}
\Delta^{**} &= \inf\,(\Delta : h(X, Y - \Delta) < k) \\
&= \inf\,(\Delta : \text{méně než } k \text{ rozdílů } Y_j - X_i \text{ je větších než } \Delta) \\
&= \inf\,(\Delta : D_{(k+1)} \leq \Delta) = D_{(k+1)}\ .
\end{aligned}$$

Tentokrát máme $\hat{\Delta} = \frac{D_{(k)} + D_{(k+1)}}{2}$.

V obou případech jsme dospěli k odhadu rovnému mediánu hodnot $D_{(1)}, \ldots, D_{(mn)}$. Odhad posunutí Δ založený na dvouvýběrovém Wilcoxonově testu má tedy tvar

$$\hat{\Delta} = \text{med}\,(Y_j - X_i) \quad i = 1, \ldots, m,\ j = 1, \ldots, n\ . \tag{3.8}$$

Poznámka: Při odvozování odhadu jsme předpokládali, že pozorování $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ jsou různá a také že všechny rozdíly $Y_j - X_i$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ jsou různé. Je ale zřejmé, že v praktických úlohách se často setkáme s tím, že některá pozorování nebo rozdíly mají stejnou hodnotu. S odpovědí na otázku, zda je i v takovém případě Hodges-Lehmannův odhad vhodný, nám pomůže následující úvaha. Předpokládejme, že stejné hodnoty jsou způsobeny zaokrouhlováním, a označme X_i' původní hodnotu pozorování X_i před zaokrouhlením pro $i = 1, \ldots, m$ a Y_j' původní hodnotu Y_j pro $j = 1, \ldots, n$. Je-li zaokrouhlování prováděno na násobky jednotky ϵ, platí pro všechna pozorování $|X_i' - X_i| \leq \frac{\epsilon}{2}$ a $|Y_j' - Y_j| \leq \frac{\epsilon}{2}$ a z toho vyplývá $|(Y_j' - X_i') - (Y_j - X_i)| \leq \epsilon$. Snadno vidíme, že pak i $\text{med}\,(Y_j' - X_i')$ se od $\text{med}\,(Y_j - X_i)$ liší nanejvýš o ϵ. Jestliže je ϵ dostatečně malé vzhledem k Δ, bude odhad $\hat{\Delta}$ založený na pozorováních $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ vždy blízký odhadu, který by byl založen na původních pozorováních bez zaokrouhlování, tedy na různých pozorováních $X_1', \ldots, X_m', Y_1', \ldots, Y_n'$ a následně na rozdílech $Y_j' - X_i'$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, které nenabývají stejných hodnot.

Použití Hodges-Lehmannova odhadu nyní předvedeme na příkladu z knihy Lehmann [10].

Příklad 3.2: Muž se právě přestěhoval do nového domu a má na výběr dvě možné cesty do práce. Obě cesty několikrát vyzkoušel a pokaždé zaznamenal, jak dlouho mu cesta trvala. Časy v minutách jsou uvedeny v následující tabulce:

Tabulka 3.1: Doby trvání cesty v minutách

Cesta A	6,0	5,8	6,5	5,8	6,3	6,0	6,3	6,4	5,9	6,5	6,0
Cesta B	7,3	7,1	6,5	10,2	6,8						

Můžeme přirozeně předpokládat, že rozdělení dob trvání cesty A a cesty B se liší pouze konstantně. Chceme odhadnout o kolik, tedy chceme odhadnout parametr posunutí Δ. Doby trvání cesty A budeme uvažovat jako první

náhodný výběr $X_1, \ldots, X_{11}$, doby trvání cesty B jako druhý náhodný výběr $Y_1, \ldots, Y_5$. V našem značení pak $m = 11$ a $n = 5$.

Protože $mn = 55$, máme 55 rozdílů $Y_j - X_i$, $i = 1, \ldots, 11$, $j = 1, \ldots, 5$ (viz Tabulka 3.2) a Hodges-Lehmannův odhad $\hat{\Delta}$ je roven 28. nejmenšímu z těchto rozdílů. Z uvedené tabulky už snadno zjistíme, že 28. nejmenší rozdíl má hodnotu 0,9, a tudíž dostáváme odhad $\hat{\Delta} = 0,9$.

Tabulka 3.2: Rozdíly $Y_j - X_i$

		X_i										
		5,8	5,8	5,9	6,0	6,0	6,0	6,3	6,3	6,4	6,5	6,5
	6,5	0,7	0,7	0,6	0,5	0,5	0,5	0,2	0,2	0,1	0,0	0,0
	6,8	1,0	1,0	0,9	0,8	0,8	0,8	0,5	0,5	0,4	0,3	0,3
Y_j	7,1	1,3	1,3	1,2	1,1	1,1	1,1	0,8	0,8	0,7	0,6	0,6
	7,3	1,5	1,5	1,4	1,3	1,3	1,3	1,0	1,0	0,9	0,8	0,8
	10,2	4,4	4,4	4,3	4,2	4,2	4,2	3,9	3,9	3,8	3,7	3,7

◇

3.2.2 Odhad založený na dvouvýběrovém Galtonově testu

Dalším R-odhadem, který umíme explicitně vyjádřit, je odhad založený na dvouvýběrovém Galtonově testu. Budeme předpokládat, že máme k dispozici náhodná pozorování $X_1, \ldots, X_n$ a stejný počet na nich nezávislých náhodných pozorování $Y_1, \ldots, Y_n$. Pokud bychom pracovali se dvěma náhodnými výběry s různou velikostí, vypořádáme se s tím stejným způsobem jako při testování pomocí Galtonova testu (viz kapitola 2.2.1).

Nechť $h(X, Y)$ je statistika tvaru (2.8). Tedy $h(X, Y)$ se rovná počtu případů, kdy i-tá pořádková statistika $Y_{(i)}$ náhodného výběru $Y_1, \ldots, Y_n$ je větší než i-tá pořádková statistika $X_{(i)}$ náhodného výběru $X_1, \ldots, X_n$. Statistika $h(X, Y)$ může nabývat hodnot $0, \ldots, n$.

Chceme opět odvodit odhad na základě vzorce (3.3), musíme proto ověřit splnění podmínek (A) a (B). $h(x, y + a)$ je zřejmě neklesající funkcí a pro všechna x a y a podmínka (A) je splněna. Ve Větě 7 jsme dokázali, že statistika $h(X, Y)$ má za platnosti nulové hypotézy rovnoměrné rozdělení na hodnotách $0, \ldots, n$. To nám zajišťuje splnění podmínky (B), neboť rozdělení statistiky $h(X, Y)$ je pak symetrické kolem své střední hodnoty $\frac{n}{2}$.

Rozdíly pořádkových statistik $Y_{(i)} - X_{(i)}$, $i = 1, \ldots, n$ seřadíme vzestupně podle velikosti a označíme $D_{(1)} < \ldots < D_{(n)}$. Stejně jako v případě

Hodges-Lehmannova odhadu se budeme nejprve zabývat možností, kdy n je liché. Nechť $n = 2k + 1$, pak je rozdělení $h(X, Y)$ symetrické kolem hodnoty $\mu = k + \frac{1}{2}$ a můžeme odvodit

$$\begin{aligned}
\Delta^* &= \sup(\Delta : h(X, Y - \Delta) > k + \frac{1}{2}) \\
&= \sup(\Delta : \text{více než } (k + \frac{1}{2})\text{-krát platí } X_{(i)} < (Y - \Delta)_{(i)}) \\
&= \sup(\Delta : \text{více než } (k + \frac{1}{2})\text{-krát platí } X_{(i)} < Y_{(i)} - \Delta) \\
&= \sup(\Delta : \text{více než } k + \frac{1}{2} \text{ rozdílů } Y_{(i)} - X_{(i)} \text{ je větších než } \Delta) \\
&= \sup(\Delta : D_{(k+1)} > \Delta) = D_{(k+1)} \ .
\end{aligned}$$

Analogicky vyjádříme

$$\begin{aligned}
\Delta^{**} &= \inf(\Delta : h(X, Y - \Delta) < k + \frac{1}{2}) \\
&= \inf(\Delta : \text{méně než } k + \frac{1}{2} \text{ rozdílů } Y_{(i)} - X_{(i)} \text{ je větších než } \Delta) \\
&= \inf(\Delta : D_{(k+1)} \leq \Delta) = D_{(k+1)}
\end{aligned}$$

a pro liché n jsme odvodili $\hat{\Delta} = D_{(k+1)}$.

Předpokládejme nyní, že n je sudé, tedy $n = 2k$. Potom $\mu = k$ a

$$\begin{aligned}
\Delta^* &= \sup(\Delta : h(X, Y - \Delta) > k) \\
&= \sup(\Delta : \text{více než } k \text{ rozdílů } Y_{(i)} - X_{(i)} \text{ je větších než } \Delta) \\
&= \sup(\Delta : D_{(k)} > \Delta) = D_{(k)}
\end{aligned}$$

a

$$\begin{aligned}
\Delta^{**} &= \inf(\Delta : h(X, Y - \Delta) < k) \\
&= \inf(\Delta : \text{méně než } k \text{ rozdílů } Y_{(i)} - X_{(i)} \text{ je větších než } \Delta) \\
&= \inf(\Delta : D_{(k+1)} \leq \Delta) = D_{(k+1)} \ .
\end{aligned}$$

V tomto případě dostáváme $\hat{\Delta} = \frac{D_{(k)} + D_{(k+1)}}{2}$.

Opět jsme v obou případech získali odhad rovný mediánu hodnot $D_{(1)}, \ldots, D_{(n)}$. Odhad posunutí Δ založený na dvouvýběrovém Galtonově testu má proto tvar

$$\hat{\Delta} = \text{med}\,(Y_{(i)} - X_{(i)}) \quad i = 1, \ldots, n \ . \tag{3.9}$$

Poznámka: Co se týče výskytu shodných pozorování, situace je stejná jako v případě odhadu založeného na dvouvýběrovém Wilcoxonově testu - v praxi se setkáme se situací, kdy nejsou splněny naše předpoklady o různosti pozorování nebo rozdílů pořádkových statistik. A analogická je i úvaha, která odůvodní vhodnost odhadu (3.9) i při výskytu shodných pozorování.

K ilustraci použití tohoto odhadu využijeme zadání Příkladu 3.2.

Příklad 3.2 (pokračování): Máme k dispozici dva náhodné výběry o velikosti 11 a 5. Velikost náhodných výběrů se liší, ale platí $11 = 5 + 1 * (5 + 1)$ a náhodný výběr $X_1, \ldots, X_{11}$ můžeme reprezentovat pozorováními $X_{(2)}, X_{(4)}, X_{(6)}, X_{(8)}, X_{(10)}$. Přeznačíme

$$X^*_{(1)} = X_{(2)}, \ldots, X^*_{(5)} = X_{(10)}$$

a odhad parametru posunutí Δ založíme na pozorováních $X^*_{(1)}, \ldots, X^*_{(5)}$ a $Y_1, \ldots, Y_5$.

Hledáme medián rozdílů $Y_{(i)} - X^*_{(i)}$, $i = 1, \ldots, 5$. V Tabulce 3.3 vidíme, že 3. nejmenší z těchto rozdílů má hodnotu 1,0, a na základě (3.9) odhadneme $\hat{\Delta} = 1$.

Tabulka 3.3: $Y_{(i)}$, $X^*_{(i)}$ a rozdíly $Y_{(i)} - X^*_{(i)}$

$Y_{(i)}$	6,5	6,8	7,1	7,3	10,2
$X^*_{(i)}$	5,8	6,0	6,0	6,3	6,5
$Y_{(i)} - X^*_{(i)}$	0,7	0,8	1,1	1,0	3,7

◇

3.2.3 Odhad založený na jednovýběrovém Wilcoxonově testu

Odhad parametru polohy θ v jednovýběrovém problému můžeme založit na jednovýběrovém Wilcoxonově testu. Použijeme k tomu tvar testové statistiky (2.6) odvozený Tukeym. Nechť $h(Z)$ označuje počet dvojic (Z_i, Z_j) $1 \leq i \leq j \leq N$ takových, že $Z_i + Z_j > 0$. Statistika $h(Z)$ může nabývat hodnot $0, 1, \ldots, \frac{N(N+1)}{2}$.

Abychom mohli k získání odhadu použít vzorec (3.7) navržený Hodgesem a Lehmannem, potřebujeme, aby $h(Z)$ splňovala podmínky (C) a (D). Je zřejmé, že pro h platí podmínka (C), a splnění (D) jsme ukázali v kapitole 2.1.2 o jednovýběrovém Wilcoxonově testu, kde jsme odvodili, že rozdělení $h(Z)$ je při $\theta = 0$ symetrické kolem hodnoty $\frac{N(N+1)}{4}$.

Označme $K = \frac{N(N+1)}{2}$ a $D_{(1)} < \ldots < D_{(K)}$ podle velikosti uspořádané průměry $\frac{Z_i+Z_j}{2}$, $1 \leq i \leq j \leq N$. Analogicky jako v případě dvou výběrů uvažujeme dvě možnosti. Nechť nejprve $K = 2k+1$. Pak $\mu = \frac{N(N+1)}{4} = k+\frac{1}{2}$ a

$$\begin{aligned}
\theta^* &= \sup\,(\theta : h(Z-\theta) > k+\frac{1}{2}) \\
&= \sup\,(\theta : \text{více než } k+\frac{1}{2} \text{ součtů } (Z_i-\theta)+(Z_j-\theta) \text{ je větších než } 0) \\
&= \sup\,(\theta : \text{více než } k+\frac{1}{2} \text{ průměrů } \frac{(Z_i-\theta)+(Z_j-\theta)}{2} \text{ je větších než } 0) \\
&= \sup\,(\theta : \text{více než } k+\frac{1}{2} \text{ průměrů } \frac{Z_i+Z_j}{2} \text{ je větších než } \theta) \\
&= \sup\,(\theta : D_{(k+1)} > \theta) = D_{(k+1)}\ .
\end{aligned}$$

Dále

$$\begin{aligned}
\theta^{**} &= \inf\,(\theta : h(Z-\theta) < k+\frac{1}{2}) \\
&= \inf\,(\theta : \text{méně než } k+\frac{1}{2} \text{ průměrů } \frac{Z_i+Z_j}{2} \text{ je větších než } \theta) \\
&= \inf\,(\theta : D_{(k+1)} \leq \theta) = D_{(k+1)}
\end{aligned}$$

a celkově jsme získali $\hat{\theta} = D_{(k+1)}$.

Nyní se budeme věnovat druhé možnosti, totiž že $K = 2k$. Potom $\mu = \frac{N(N+1)}{4} = k$ a

$$\begin{aligned}
\theta^* &= \sup\,(\theta : h(Z-\theta) > k) \\
&= \sup\,(\theta : \text{více než } k \text{ průměrů } \frac{Z_i+Z_j}{2} \text{ je větších než } \theta) \\
&= \sup\,(\theta : D_{(k)} > \theta) = D_{(k)}
\end{aligned}$$

a

$$\begin{aligned}
\theta^{**} &= \inf\,(\theta : h(Z-\theta) < k) \\
&= \inf\,(\theta : \text{méně než } k \text{ průměrů } \frac{Z_i+Z_j}{2} \text{ je větších než } \theta) \\
&= \inf\,(\theta : D_{(k+1)} \leq \theta) = D_{(k+1)}\ .
\end{aligned}$$

Tentokrát celkově dostáváme $\hat{\theta} = \frac{D_{(k)}+D_{(k+1)}}{2}$.

V případě lichého i sudého K jsme dospěli k závěru, že odhad parametru polohy θ je rovný mediánu hodnot $D_{(1)}, \ldots, D_{(K)}$. Odhad založený na jednovýběrovém Wilcoxonově testu má tudíž tvar

$$\hat{\theta} = \text{med}\,(\frac{Z_i+Z_j}{2}) \quad 1 \leq i \leq j \leq N\ . \tag{3.10}$$

Poznámka: Předpoklad, že v náhodném výběru $Z_1, \ldots, Z_N$ mají všechna pozorování různou absolutní hodnotu a všechny průměry $\frac{Z_i+Z_j}{2}$, $1 \leq i \leq j \leq N$ jsou různé, nemusí být v praxi vždy splněn. Odhad $\hat{\theta}$ tvaru (3.10) bude však i nadále dostatečně dobrý, což si můžeme ověřit analogickou úvahou jako v případě odhadu parametru posunutí založeném na dvouvýběrovém Wilcoxonově testu.

Použití R-odhadu založeného na jednovýběrovém Wilcoxonově testu ukážeme na příkladu z knihy Conover [3].

Příklad 3.3: Dvanáct párů identických dvojčat podstoupilo psychologický test, který v jistém smyslu měřil míru agresivity osobnosti. Zajímáme se o porovnání prvorozených a druhorozených dvojčat. Vyšší výsledek testu ukazuje na větší míru agresivity, výsledky jsou uvedeny v následující tabulce:

Tabulka 3.4: Výsledky dvojčat v testu agresivity

Prvorozené dvojče	86	71	77	68	91	72	77	91	70	71	88	87
Druhorozené dvojče	88	77	76	64	96	72	65	90	65	80	81	72

Výsledky testu budeme uvažovat jako dvojice (X_i, Y_i), $i = 1, \ldots, 12$, kde X_i je výsledek druhorozeného dvojčete v i-tém páru a Y_i je výsledek prvorozeného dvojčete v i-tém páru. Můžeme předpokládat, že rozdělení výsledků testů prvorozených a druhorozených se liší konstantně. Položíme-li $Z_i = Y_i - X_i$, $i = 1, \ldots, 12$, bude nás zajímat odhad parametru polohy θ rozdělení, ze kterého pochází náhodný výběr $Z_1, \ldots, Z_{12}$. Budeme tedy pracovat s hodnotami

$$-2, -6, 1, 4, -5, 0, 12, 1, 5, -9, 7, 15 \quad .$$

V našem značení máme $N = 12$. Abychom odhadli θ odhadem (3.10), hledáme medián $\frac{N(N+1)}{2} = 78$ průměrů $\frac{Z_i+Z_j}{2}$, $1 \leq i \leq j \leq N$. Z Tabulky 3.5 snadno určíme, že 39. nejmenší průměr má hodnotu 1,5 a 40. nejmenší průměr má také hodnotu 1,5. Parametr polohy θ odhadneme hodnotou $\hat{\theta} = \frac{1,5+1,5}{2} = 1,5$.

Tabulka 3.5: Průměry $\frac{Z_i+Z_j}{2}$

	-9	-6	-5	-2	0	1	1	4	5	7	12	15
-9	-9	-7,5	-7	-5,5	-4,5	-4	-4	-2,5	-2	-1	1,5	3
-6		-6	-5,5	-4	-3	-2,5	-2,5	-1	-0,5	0,5	3	4,5
-5			-5	-3,5	-2,5	-2	-2	-0,5	0	1	3,5	5
-2				-2	-1	-0,5	-0,5	1	1,5	2,5	5	6,5
0					0	0,5	0,5	2	2,5	3,5	6	7,5
1						1	1	2,5	3	4	6,5	8
1							1	2,5	3	4	6,5	8
4								4	4,5	5,5	8	9,5
5									5	6	8,5	10
7										7	9,5	11
12											12	13,5
15												15

◇

3.2.4 Odhad založený na jednovýběrovém Galtonově testu

R-odhad založený na jednovýběrovém Galtonově testu se někdy nazývá také Bickel-Hodgesův odhad. Mějme k dispozici N nezávislých pozorování $Z_1, \ldots, Z_N$, kde $N = 2m$ nebo $N = 2m - 1$. K odvození odhadu použijeme statistiku (2.10), nechť tedy $h(Z)$ určuje počet součtů pořádkových statistik tvaru $Z_{(i)} + Z_{(N-i+1)}$, $i = 1, \ldots, m$, které jsou kladné. $h(Z)$ může nabývat hodnot $0, \ldots, m$.

Potřebujeme ověřit, zda statistika $h(Z)$ splňuje podmínky (C) a (D). $h(z + a)$ je zřejmě neklesající funkcí a pro všechna z a podmínka (C) je splněna. Pro $N = 2m$ jsme v kapitole 2.2.2 odvodili rozdělení statistiky $h(Z)$ za platnosti nulové hypotézy a víme, že je symetrické kolem své střední hodnoty $\frac{m}{2}$. Pro $N = 2m-1$ přesné rozdělení statistiky neznáme, můžeme ho však také pokládat za symetrické kolem $\frac{m}{2}$ vzhledem k tomu, že už pro středně velké výběry lze rozdělení statistiky $h(Z)$ při sudém i lichém N aproximovat stejnou funkcí (viz (2.11)). Máme tedy splněnu i podmínku (D) a můžeme odhad odvodit na základě (3.7).

Vypočteme průměry $\frac{Z_{(i)}+Z_{(N-i+1)}}{2}$, $i = 1, \ldots, m$, uspořádáme je vzestupně podle velikosti a označíme $D_{(1)} < \ldots < D_{(m)}$. Stejně jako u předchozích odhadů budeme rozlišovat dva případy. Nechť $m = 2k+1$, pak střed symetrie $\mu = k + \frac{1}{2}$ a

$$
\begin{aligned}
\theta^* &= \sup\,(\theta : h(Z-\theta) > k + \frac{1}{2}) \\
&= \sup\,(\theta : \text{více než } k + \frac{1}{2} \text{ součtů } (Z-\theta)_{(i)} + (Z-\theta)_{(N-i+1)} \\
&\qquad \text{je větších než } 0) \\
&= \sup\,(\theta : \text{více než } k + \frac{1}{2} \text{ součtů } (Z_{(i)}-\theta) + (Z_{(N-i+1)}-\theta) \\
&\qquad \text{je větších než } 0) \\
&= \sup\,(\theta : \text{více než } k + \frac{1}{2} \text{ průměrů } \frac{(Z_{(i)}-\theta) + (Z_{(N-i+1)}-\theta)}{2} \\
&\qquad \text{je větších než } 0) \\
&= \sup\,(\theta : \text{více než } k + \frac{1}{2} \text{ průměrů } \frac{Z_{(i)} + Z_{(N-i+1)}}{2} \text{ je větších než } \theta) \\
&= \sup\,(\theta : D_{(k+1)} > \theta) = D_{(k+1)}\ .
\end{aligned}
$$

Podobně

$$
\begin{aligned}
\theta^{**} &= \inf\,(\theta : h(Z-\theta) < k + \frac{1}{2}) \\
&= \inf\,(\theta : \text{méně než } k + \frac{1}{2} \text{ průměrů } \frac{Z_{(i)} + Z_{(N-i+1)}}{2} \text{ je větších než } \theta) \\
&= \inf\,(\theta : D_{(k+1)} \le \theta) = D_{(k+1)}
\end{aligned}
$$

a celkově jsme pro m liché dospěli k $\hat{\theta} = D_{(k+1)}$.

Uvažujme druhou možnost $m = 2k$. m je sudé a $\mu = k$. Vyjádříme

$$
\begin{aligned}
\theta^* &= \sup\,(\theta : h(Z-\theta) > k) \\
&= \sup\,(\theta : \text{více než } k \text{ průměrů } \frac{Z_{(i)} + Z_{(N-i+1)}}{2} \text{ je větších než } \theta) \\
&= \sup\,(\theta : D_{(k)} > \theta) = D_{(k)}
\end{aligned}
$$

a

$$
\begin{aligned}
\theta^{**} &= \inf\,(\theta : h(Z-\theta) < k) \\
&= \inf\,(\theta : \text{méně než } k \text{ průměrů } \frac{Z_{(i)} + Z_{(N-i+1)}}{2} \text{ je větších než } \theta) \\
&= \inf\,(\theta : D_{(k+1)} \le \theta) = D_{(k+1)}\ .
\end{aligned}
$$

V případě sudého m jsme odvodili $\hat{\theta} = \frac{D_{(k)}+D_{(k+1)}}{2}$.

V obou případech jsme získali odhad parametru polohy θ rovný mediánu hodnot $D_{(1)}, \ldots, D_{(m)}$. Odhad založený na jednovýběrovém Galtonově testu má tudíž tvar

$$\hat{\theta} = \text{med}\,(\frac{Z_{(i)} + Z_{(N-i+1)}}{2}) \quad i = 1, \ldots, m\ . \tag{3.11}$$

Poznámka: Se situací, kdy se mezi pozorováními $Z_1, \ldots, Z_N$ vyskytují shodné hodnoty nebo všechny průměry $\frac{Z_{(i)}+Z_{(N-i+1)}}{2}$ nejsou různé, se vypořádáme stejnou úvahou jako v případě odhadu založeného na Wilcoxonově testu.

K ilustraci použití Bickel-Hodgesova odhadu využijeme zadání Příkladu 3.3.

Příklad 3.3 (pokračování): Budeme pracovat s náhodným výběrem $Z_1, \ldots, Z_{12}$. Máme $N = 12 = 2*6$ a tedy $m = 6$. Hledáme medián 6 průměrů $\frac{Z_{(i)}+Z_{(N-i+1)}}{2}$. Z Tabulky 3.6 zjistíme, že 3. nejmenší průměr má hodnotu 1,5 a 4. nejmenší průměr má hodnotu 2. Z toho vyplývá, že parametr polohy θ odhadneme hodnotou $\hat{\theta} = \frac{1,5+2}{2} = 1,75$.

Tabulka 3.6: Průměry $\frac{Z_{(i)}+Z_{(N-i+1)}}{2}$

i	1	2	3	4	5	6
$Z_{(i)}$	-9	-6	-5	-2	0	1
$Z_{(N-i+1)}$	15	12	7	5	4	1
$\frac{Z_{(i)}+Z_{(N-i+1)}}{2}$	3	3	1	1.5	2	1

◇

3.2.5 Odhad založený na znaménkovém testu

Dalším testem, na kterém můžeme založit odhad parametru polohy v případě jednoho výběru, je znaménkový test. Jak však uvidíme, získáme v tomto případě R-odhad rovný mediánu pozorování $Z_1, \ldots, Z_N$, což je již známý a běžně používaný odhad.

Nechť se hodnota statistiky $h(Z)$ rovná počtu pozorování $Z_i, i = 1, \ldots, N$, která jsou kladná. Statistika $h(Z)$ může nabývat hodnot $0, \ldots, N$. Opět musíme ověřit, zda statistika $h(Z)$ vyhovuje podmínkám (C) a (D). Splnění podmínky (C) je zřejmé. Z kapitoly 2.3 o znaménkovém testu víme, že $h(Z)$ má za platnosti nulové hypotézy binomické rozdělení $Bi(N, \frac{1}{2})$. Z toho vyplý-

vá, že rozdělení $h(Z)$ je symetrické kolem své střední hodnoty $\frac{N}{2}$ a podmínka (D) je také splněna.

Náhodný výběr $Z_1, \ldots, Z_N$ uspořádáme podle velikosti. Máme $Z_{(1)} < \ldots < Z_{(N)}$. Nechť nejprve N je liché, $N = 2k + 1$. Pak je rozdělení $h(Z)$ symetrické kolem hodnoty $\mu = k + \frac{1}{2}$ a dostáváme

$$\begin{aligned}
\theta^* &= \sup\,(\theta : h(Z - \theta) > k + \frac{1}{2}) \\
&= \sup\,(\theta : \text{více než } k + \frac{1}{2} \text{ rozdílů } (Z_i - \theta) \text{ je větších než } 0) \\
&= \sup\,(\theta : \text{více než } k + \frac{1}{2} \text{ pozorování } Z_i \text{ je větších než } \theta) \\
&= \sup\,(\theta : Z_{(k+1)} > \theta) = Z_{(k+1)} \ .
\end{aligned}$$

Analogicky odvodíme

$$\begin{aligned}
\theta^{**} &= \inf\,(\theta : h(Z - \theta) < k + \frac{1}{2}) \\
&= \inf\,(\theta : \text{méně než } k + \frac{1}{2} \text{ pozorování } Z_i \text{ je větších než } \theta) \\
&= \inf\,(\theta : Z_{(k+1)} \leq \theta) = Z_{(k+1)}
\end{aligned}$$

a celkově jsme získali $\hat{\theta} = Z_{(k+1)}$.

V druhém případě, kdy N je sudé, $N = 2k$, máme $\mu = k$ a

$$\begin{aligned}
\theta^* &= \sup\,(\theta : h(Z - \theta) > k) \\
&= \sup\,(\theta : \text{více než } k \text{ pozorování } Z_i \text{ je větších než } \theta) \\
&= \sup\,(\theta : Z_{(k)} > \theta) = Z_{(k)}
\end{aligned}$$

a

$$\begin{aligned}
\theta^{**} &= \inf\,(\theta : h(Z - \theta) < k) \\
&= \inf\,(\theta : \text{méně než } k \text{ pozorování } Z_i \text{ je větších než } \theta) \\
&= \inf\,(\theta : Z_{(k+1)} \leq \theta) = Z_{(k+1)} \ .
\end{aligned}$$

V případě sudého N jsme odvodili $\hat{\theta} = \frac{Z_{(k)} + Z_{(k+1)}}{2}$.

Jak jsme uvedli již na začátku kapitoly, v případě lichého i sudého N jsme dospěli k odhadu parametru polohy θ rovnému mediánu hodnot $Z_{(1)}, \ldots, Z_{(N)}$. Odhad založený na znaménkovém testu má tudíž tvar

$$\hat{\theta} = \text{med}\,(Z_i) \quad i = 1, \ldots, N \ . \tag{3.12}$$

Poznámka: Také tento odhad můžeme použít i v případě výskytu shodných pozorování. Zdůvodnění by bylo podobné jako u předchozích odhadů.

Použití tohoto odhadu předvedeme opět na zadání Příkladu 3.3.

Příklad 3.3 (pokračování): Znovu budeme pracovat s náhodným výběrem $Z_1, \ldots, Z_{12}$. Jednoduše zjistíme, že mediánem hodnot

$$-2, -6, 1, 4, -5, 0, 12, 1, 5, -9, 7, 15$$

je číslo $\frac{1+1}{2} = 1$. $\diamond$

3.3 Intervaly spolehlivosti

Vedle bodových odhadů parametrů posunutí a polohy můžeme z předchozích úvah odvodit také intervaly spolehlivosti pro tyto parametry. Chceme tedy nalézt takový interval, do kterého patří skutečná hodnota Δ nebo θ s předepsanou pravděpodobností. Uvažujme obecně všechny v kapitole 3.2 odvozené příklady R-odhadů. Nechť $D_{(1)} < \ldots < D_{(K)}$ tvoří vzestupně uspořádanou posloupnost, jejíž medián je roven některému z odvozených odhadů. Například pro Hodges-Lehmannův odhad je $K = mn$ a pro $l = 1, \ldots, K$ je $D_{(l)} = Y_j - X_i$ pro nějaké j a i. Pro všechny uvedené R-odhady platí následující lemma.

Lemma 10 *Pro libovolné $i = 1, \ldots, K$ a libovolné reálné a platí*

$$D_{(i)} \leq a \quad \Longleftrightarrow \quad h(X, Y - a) \leq K - i$$

v případě odhadů parametru posunutí a

$$D_{(i)} \leq a \quad \Longleftrightarrow \quad h(Z - a) \leq K - i$$

v případě odhadů parametru polohy. h je statistika, na které je daný odhad založen.

Důkaz: $D_{(i)} \leq a$ platí právě tehdy, když nejvýše $K - i$ hodnot z $D_{(l)}$, $l = 1, \ldots, K$ je větších než a. Z odvození odhadů vyplývá, že to je právě tehdy, když i $h(X, Y - a)$ nabývá nejvýše hodnoty $K - i$.

Důkaz v případě odhadů parametru polohy je zcela analogický. $\square$

Zaměřme se nyní na odhady parametru posunutí Δ v případě dvou výběrů. Následující úvahy však můžeme analogicky uplatnit i na parametr polohy θ jednoho výběru a jeho R-odhady.

Z Lemmatu 10 automaticky vyplývá, že platí také

$$D_{(i)} > a \quad \Longleftrightarrow \quad h(X, Y - a) \geq K - i + 1 \ .$$

Označíme-li $D_{(0)} = -\infty$ a $D_{(K+1)} = \infty$, snadno dostaneme pro $i = 1, \ldots, K$ ekvivalenci

$$D_{(i)} \leq \Delta < D_{(i+1)} \quad \Longleftrightarrow \quad h(X, Y - \Delta) = K - i \ ,$$

kde Δ je odhadovaný parametr posunutí. Za předpokladu, že Δ je skutečná hodnota parametru, nezávisí rozdělení statistik všech pořadových testů, které jsme použili k odvození odhadů, na této hodnotě. Můžeme proto vyjádřit

$$P_\Delta(D_{(i)} \leq \Delta < D_{(i+1)}) = P_\Delta(h(X, Y - \Delta) = K - i) = P_0(h(X, Y) = K - i) \tag{3.13}$$

pro $i = 1, \ldots, K$.

Hodnoty $D_{(1)}, \ldots, D_{(K)}$ tedy rozdělují reálnou osu na $K + 1$ intervalů, které obsahují Δ se známou pravděpodobností. Vzhledem ke spojitému rozdělení náhodných výběrů, na nichž odhady zakládáme, je pro všechny uvedené R-odhady rozdělení $D_{(i)}$, $i = 1, \ldots, K$ také spojité. Můžeme tedy ekvivalentně uvažovat jak otevřené, tak uzavřené intervaly. Abychom získali interval spolehlivosti $(\overline{\Delta}; \underline{\Delta})$, který obsahuje Δ s předepsanou pravděpodobností α, stačí spojit tolik intervalů $[D_{(i)}; D_{(i+1)})$, aby součet pravděpodobností, že obsahují Δ, byl co nejblíže α. Jestliže

$$[D_{(i)}; D_{(i+1)}), [D_{(i+1)}; D_{(i+2)}), \ldots, [D_{(j-1)}; D_{(j)})$$

jsou vybrané intervaly, je nejpřirozenější volit je tak, aby platilo

$$P(\Delta \leq D_{(i)}) = P(D_{(j)} \leq \Delta) \doteq \frac{1}{2}(1 - \alpha) \ .$$

V tom nám pomůže vztah, který vyplývá z (3.13) a toho, že rozdělení $D_{(i)}$ je spojité:

$$P_\Delta(\Delta \leq D_{(i)}) = P_0(h(X, Y) \geq K - i + 1) = P_0(h(X, Y) \leq i - 1) \ , \tag{3.14}$$

neboť rozdělení statistiky $h(X, Y)$ je symetrické kolem své střední hodnoty. Ze vztahu (3.14) můžeme dále odvodit

$$P_\Delta(D_{(i)} \leq \Delta) = P_0(h(X, Y) \geq i) \ . \tag{3.15}$$

Na Příkladu 3.2 si ukážeme výpočet intervalu spolehlivosti pro Δ na základě dvouvýběrového Wilcoxonova testu.

Příklad 3.2 (pokračování): Máme k dispozici údaje o časech trvání cesty A ($X_1, \ldots, X_{11}$) a cesty B ($Y_1, \ldots, Y_5$). Chceme najít interval spolehlivosti pro parametr Δ s koeficientem spolehlivosti 90% a to na základě dvouvýběrového Wilcoxonova testu. Proto budeme pracovat s $K = mn = 55$ hodnotami $D_{(i)}$, které najdeme v Tabulce 3.2. Statistika h má v tomto případě Mann-Whitneyho tvar (2.3).

Hledáme takové i z $i = 1, \ldots, 55$, pro které platí

$$P_0(h(X,Y) \leq i) \doteq \frac{1}{2}(1 - 0,9) = 0,05 \ .$$

Z tabulky kritických hodnot dvouvýběrového Wilcoxonova testu, kterou lze najít například v článku Verdooren [12], zjistíme, že pro $m = 11$ a $n = 5$ je $P_0(W \leq 27) \doteq 0,05$, kde W je klasická statistika Wilcoxonova testu tvaru (2.2). Víme, že $h(X,Y) = W - \frac{n(n+1)}{2}$, a proto $P_0(h(X,Y) \leq 12) \doteq 0,05$. Rozdělení statistiky h je symetrické kolem $\frac{55}{2}$ a zároveň tedy platí $P_0(h(X,Y) \geq 55 - 12) = P_0(h(X,Y) \geq 43) \doteq 0,05$.

Ze vztahu (3.14) a (3.15) tak víme, že

$$P_\Delta(\Delta \leq D_{(13)}) \doteq 0,05 \quad \text{a} \quad P_\Delta(\Delta \geq D_{(43)}) \doteq 0,05 \ .$$

Označíme-li $\underline{\Delta} = D_{(13)}$ a $\overline{\Delta} = D_{(43)}$, náleží Δ do intervalu $(\underline{\Delta}; \overline{\Delta}) = (0,5; 1,5)$ s požadovanou pravděpodobností přibližně 90%. $\diamond$

Stejným postupem, jakým jsme dospěli ke vztahům (3.13), (3.14) a (3.15), můžeme pro parametr polohy θ jednoho výběru odvodit

$$P_\theta(D_{(i)} \leq \theta < D_{(i+1)}) = P_0(h(Z) = K - i) \ , \tag{3.16}$$

$$P_\theta(\theta \leq D_{(i)}) = P_0(h(Z) \leq i - 1) \tag{3.17}$$

a

$$P_\theta(D_{(i)} \leq \theta) = P_0(h(Z) \geq i) \tag{3.18}$$

pro $i = 1, \ldots, K$.

Nalezení intervalu spolehlivosti pro parametr θ ukážeme na Příkladu 3.3, založíme ho na Wilcoxonově jednovýběrovém testu.

Příklad 3.3 (pokračování): Máme k dispozici náhodné veličiny $Z_1, \ldots, Z_{12}$. Opět hledáme interval spolehlivosti s koeficientem spolehlivosti 90%. V tomto případě $K = 78$ a hodnoty $D_{(i)}$, $i = 1, \ldots 78$ najdeme v Tabulce 3.5. Například v knize Conover [3] nalezneme tabulku kritických hodnot jednovýběrového Wilcoxonova testu. Zjistíme z ní, že $P(h(Z) \leq 18) \doteq 0,05$,

a ze symetrie rozdělení $h(Z)$ plyne $P(h(Z) \geq 60) \doteq 0,05$. Z (3.18) a (3.17) dostaneme

$$P_\theta(\theta \leq D_{(19)}) \doteq 0,05 \quad \text{a} \quad P_\theta(\theta \geq D_{(60)}) \doteq 0,05 \ .$$

$D_{(19)} = -2$, $D_{(60)} = 5,5$ a interval $(-2; 5,5)$ obsahuje θ s požadovanou pravděpodobností přibližně 90%. $\diamond$

Místo intervalu spolehlivosti můžeme někdy chtít určit spíše pouze horní (respektive dolní) mez, to znamená hodnotu, která je s požadovanou pravděpodobností větší (respektive menší) než odhadovaný parametr. Takové meze snadno získáme z (3.14), (3.15) a (3.18), (3.17). Výpočet mezí ukážeme na Příkladu 3.2 a založíme ho na dvouvýběrovém Galtonově testu.

Příklad 3.2 (pokračování): Chceme nalézt horní mez a dolní mez pro parametr Δ. $h(X, Y)$ je v tomto případě statistika dvouvýběrového Galtonova testu a $n = 5$. Pak

$$P_0(h(X, Y) \leq 4) = \frac{5}{6} \doteq 0,83$$

a máme $P_0(\Delta \leq D_{(5)}) \doteq 0,83$. $D_{(5)} = 3,7$ je větší než parametr Δ s pravděpodobností přibližně 83 %.

Dolní mez pro Δ získáme z

$$P_0(h(X, Y) \geq 1) = \frac{5}{6} \doteq 0,83 \ .$$

Platí pak $P_0(D_{(1)} \leq \Delta) \doteq 0,83$ a $D_{(1)} = 0,7$ je s pravděpodobností přibližně 83 % menší než parametr Δ. $\diamond$

Poznámka: S případným výskytem shodných pozorování si poradíme podobně jako u bodových odhadů. Předpokládejme, že stejné hodnoty jsou způsobeny zaokrouhlováním a že platí

$$D'_{(i)} - \epsilon \leq D_{(i)} \leq D'_{(i)} + \epsilon$$

pro $i = 1, \ldots, K$, kde $D'_{(i)}$ jsou hodnoty získané z původních nezaokrouhlených pozorování. Jestliže $D'_{(i)}$ je horní mez pro Δ nebo θ s koeficientem spolehlivosti α, pak $D_{(i)} + \epsilon$ je horní mezí s koeficientem spolehlivosti $\geq \alpha$. Podobně, je-li $D'_{(i)}$ dolní mez s koeficientem spolehlivosti α, je $D_{(i)} - \epsilon$ dolní mezí s koeficientem spolehlivosti $\geq \alpha$. A konečně pro interval spolehlivosti $(D'_{(i)}; D'_{(j)})$ s koeficientem spolehlivosti α platí, že $(D_{(i)} - \epsilon; D_{(j)} + \epsilon)$ je intervalem spolehlivosti s koeficientem $\geq \alpha$.

Kapitola 4

Vlastnosti R-odhadů

4.1 Ekvivariance vzhledem k posunutí

Jednoduchou ale užitečnou vlastností R-odhadů $\hat{\Delta}$ parametru posunutí (definovaných vztahem (3.3)) i R-odhadů $\hat{\theta}$ parametru polohy (definovaných vztahem (3.7)) je to, že jsou ekvivariantní vzhledem k posunutí. Budeme-li totiž v definici (3.2) uvažovat Δ^* a Δ^{**} jako funkce náhodných výběrů, dostáváme pro všechna reálná a

$$\begin{aligned} \Delta^*(x, y+a) &= \sup\left(\Delta : h(x, y+a-\Delta) > \mu\right) \\ &= \sup\left(\Delta + a : h(x, y-\Delta) > \mu\right) = \Delta^*(x,y) + a \end{aligned}$$

a analogicky

$$\Delta^{**}(x, y+a) = \Delta^{**}(x,y) + a \ .$$

Podobně, pokud budeme v definici (3.6) uvažovat θ^* a θ^{**} také jako funkce náhodných výběrů, můžeme pro všechna reálná a psát

$$\begin{aligned} \theta^*(z+a) &= \sup\left(\theta : h(z+a-\theta) > \mu\right) \\ &= \sup\left(\theta + a : h(z-\theta) > \mu\right) = \theta^*(z) + a \end{aligned}$$

a analogicky

$$\theta^{**}(z+a) = \theta^{**}(z) + a \ .$$

Celkově tak pro oba odhady platí

$$\hat{\Delta}(x, y+a) = \hat{\Delta}(x,y) + a \quad \text{pro všechna reálná } a \tag{4.1}$$

$$\hat{\theta}(z+a) = \hat{\theta}(z) + a \quad \text{pro všechna reálná } a \ . \tag{4.2}$$

Z (4.1) a (4.2) přímo vyplývá

$$P_\Delta(\hat{\Delta} - \Delta \leq u) = P_0(\hat{\Delta} \leq u) \tag{4.3}$$

a

$$P_\theta(\hat{\theta} - \theta \leq u) = P_0(\hat{\theta} \leq u) \ , \tag{4.4}$$

kde P_Δ a P_θ značí, že pravděpodobnosti jsou počítány za předpokladu, že Δ a θ jsou skutečné hodnoty odhadovaných parametrů. Těchto vztahů využijeme při zkoumání vlastností rozdělení R-odhadů. Bez újmy na obecnosti budeme moci předpokládat, že $\Delta = 0$ nebo $\theta = 0$, neboť rozdělení v obecném případě lze získat jednoduše posunutím.

Dvouvýběrové pořadové testy splňují pro všechna reálná a vztah

$$h(x + a, y + a) = h(x, y)$$

a tato vlastnost se zřejmě přenáší i na R-odhady parametru posunutí. Platí tedy

$$\hat{\Delta}(x + a, y + a) = \hat{\Delta}(x, y) \quad \text{pro všechna reálná } a \ . \tag{4.5}$$

4.2 Spojitost rozdělení

Zaměřme se nejprve na odhady $\hat{\Delta}$ parametru posunutí v případě dvou výběrů. Zajímavou vlastností R-odhadů je, že má-li vektor $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ absolutně spojité sdružené rozdělení, je i rozdělení odhadu $\hat{\Delta}$ absolutně spojité a to bez ohledu na to, zda statistika, ze které byl odvozen, má absolutně spojité rozdělení nebo nikoliv.

Věta 11 *Nechť h je reálná funkce na $(m+n)$-rozměrném prostoru taková, že $h(x_1, \ldots, x_m, y_1 + a, \ldots, y_n + a)$ je neklesající funkcí a pro všechna x a y. Nechť Δ^* a Δ^{**} jsou definovány pomocí (3.2) a předpokládejme, že $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ je náhodný vektor se sdruženým rozdělením H. Pak je rozdělení Δ^* a Δ^{**} absolutně spojité, jestliže je H absolutně spojité.*

Důkaz: Mějme pevně dána libovolná čísla $t_2, \ldots, t_n$. Pak je funkce $h(x_1, \ldots, x_m, y_1, y_1 + t_2, \ldots, y_1 + t_n)$ neklesající v y_1. Nechť $u(x_1, \ldots, x_m, t_2, \ldots, t_n)$ je taková funkce, že

$h(x_1, \ldots, x_m, y_1, y_1 + t_2, \ldots, y_1 + t_n) < \mu$ jestliže $y_1 < u(x_1, \ldots, x_m, t_2, \ldots, t_n)$
$h(x_1, \ldots, x_m, y_1, y_1 + t_2, \ldots, y_1 + t_n) \geq \mu$ jestliže $y_1 > u(x_1, \ldots, x_m, t_2, \ldots, t_n)$.

Označme na okamžik pro zjednodušení zápisu $u = u(x_1, \ldots, x_m, t_2, \ldots, t_n)$ a $x = (x_1, \ldots, x_m)$. Můžeme vyjádřit

$$h(x, u, u + t_2, \ldots, u + t_n) = \\ h(x, y_1 - (y_1 - u), y_1 + t_2 - (y_1 - u), \ldots, y_1 + t_n - (y_1 - u))$$

a platí

$$\Delta^{**}(x_1, \ldots, x_m, y_1, y_1 + t_2, \ldots, y_1 + t_n) = y_1 - u(x_1, \ldots, x_m, t_2, \ldots, t_n) \ .$$

Nechť H je absolutně spojité. Nechť $\mathcal{A}$ je libovolná množina na reálné ose, jejíž Lebesgueova míra je 0. Definujme

$$S = \{(x_1, \ldots, x_m, y_1, \ldots, y_n) : \Delta^{**}(x_1, \ldots, x_m, y_1, \ldots, y_n) \in \mathcal{A}\} \ .$$

a přímky

$$L(x_1^0, \ldots, x_m^0, t_2^0, \ldots, t_n^0) : \quad x_1 = x_1^0, \ldots, x_m = x_m^0, \\ y_2 = y_1 + t_2^0, \ldots, y_n = y_1 + t_n^0 \ . \qquad (4.6)$$

Snažíme se dokázat, že $P(S) = 0$. Platí, že

$$\Delta^{**}(x_1, \ldots, x_m, y_1, \ldots, y_n) \in \mathcal{A} \Leftrightarrow y_1 - u(x_1, \ldots, x_m, y_2 - y_1, \ldots, y_n - y_1) \in \mathcal{A},$$

a množina bodů, v kterých S protíná libovolnou přímku L z (4.6), má Lebesgueovu míru 0. Z toho vyplývá, že množina S má Lebesgueovu míru 0, a díky předpokladu, že H je absolutně spojité rozdělení, platí tudíž i $P(S) = 0$.

Důkaz pro Δ^* je zcela analogický. □

Z Věty 11 okamžitě vyplývá, že rozdělení odhadu $\hat{\Delta}$ je absolutně spojité, jestliže rozdělení s distribuční funkcí F z (3.1) je absolutně spojité.

Odpovídající tvrzení pro odhad $\hat{\theta}$ parametru polohy v případě jednoho výběru získáme položením $m = 0$ a $n = N$. Platí tedy, že rozdělení $\hat{\theta}$ je absolutně spojité, jestliže rozdělení s distribuční funkcí F z (3.5) je absolutně spojité.

Poznámka: V článku Hodges a Lehmann [7] je navíc uveden důkaz, že rozdělení Δ^* a Δ^{**} je spojité, jestliže rozdělení H je spojité. Tento důkaz je však chybný a uvedené tvrzení neplatí. Protipříklad ukázal ve svém článku Torgersen [11].

4.3 Nestrannost

Mají-li být $\hat{\Delta}$ a $\hat{\theta}$ dobrými odhady parametrů Δ a θ, jejich rozdělení by mělo v nějakém smyslu mít střed rovný skutečné hodnotě parametrů. Uvedeme si podmínky, za kterých je rozdělení $\hat{\Delta}$ a $\hat{\theta}$ symetrické kolem těchto skutečných hodnot. Pak platí $\mathrm{E}\hat{\Delta} = \Delta$, $\mathrm{E}\hat{\theta} = \theta$ a odhady jsou nestranné.

Věta 12 *Rozdělení odhadu $\hat{\Delta}$ definovaného vztahem (3.3) je symetrické kolem Δ, jestliže platí jedna z následujících podmínek:*
(i) rozdělení s distribuční funkcí F z (3.1) je symetrické a $h(x, y)$ splňuje

$$h(x, y) + h(-x, -y) = 2\mu \quad a \quad h(x + a, y + a) = h(x, y)$$

(ii) velikosti náhodných vektorů $X = (X_1, \ldots, X_m)$ a $Y = (Y_1, \ldots, Y_n)$ jsou stejné, tedy $m = n$, a $h(x, y)$ splňuje

$$h(x, y) + h(y, x) = 2\mu \quad a \quad h(x + a, y + a) = h(x, y) \ .$$

Důkaz: (i) Díky (4.3) můžeme bez újmy na obecnosti předpokládat, že $\Delta = 0$. Náhodné vektory X a Y jsou pak stejně rozdělené. Dále vzhledem k $h(x + a, y + a) = h(x, y)$ víme, že platí (4.5), a můžeme předpokládat, že rozdělení s distribuční funkcí F je symetrické kolem 0. Pak mají náhodné vektory (X, Y) a $(-X, -Y)$ stejné rozdělení a tudíž i $\hat{\Delta}(X, Y)$ a $\hat{\Delta}(-X, -Y)$ mají stejné rozdělení.

Naším cílem je dokázat, že $\hat{\Delta}(X, Y)$ a $-\hat{\Delta}(X, Y)$ jsou stejně rozdělené. Nyní nám tedy stačí ukázat, že $\hat{\Delta}(-X, -Y) = -\hat{\Delta}(X, Y)$. Platí totiž

$$\Delta^{**}(-x, -y) = \inf\left(\Delta : h(-x, -y - \Delta) < \mu\right) = \inf\left(\Delta : h(x, y + \Delta) > \mu\right) \ ,$$

kde jsme použili předpoklad $h(x, y) + h(-x, -y) = 2\mu$. Zároveň však

$$-\Delta^{*}(x, y) = \inf\left(-\Delta : h(x, y - \Delta) > \mu\right) = \inf\left(\Delta : h(x, y + \Delta) > \mu\right) \ .$$

Ukázali jsme, že platí

$$\Delta^{**}(-x, -y) = -\Delta^{*}(x, y) \ .$$

Stejným postupem lze odvodit, že

$$\Delta^{*}(-x, -y) = -\Delta^{**}(x, y) \ ,$$

a celkově dostaneme $\hat{\Delta}(-X, -Y) = -\hat{\Delta}(X, Y)$, což jsme chtěli dokázat.

(ii) Opět můžeme předpokládat, že $\Delta = 0$. Vzhledem k tomu, že $m = n$, jsou pak vektory (X, Y) a (Y, X) stejně rozdělené. Tedy i $\hat{\Delta}(X, Y)$ a

$\hat{\Delta}(Y,X)$ mají stejné rozdělení. Budeme chtít ukázat, že $\hat{\Delta}(Y,X) = -\hat{\Delta}(X,Y)$. Vyjádříme

$$\Delta^{**}(y,x) = \inf\left(\Delta : h(y, x-\Delta) < \mu\right) = \inf\left(\Delta : h(x-\Delta, y) > \mu\right) ,$$

kde jsme použili podmínku $h(x,y) + h(y,x) = 2\mu$. Díky podmínce $h(x+a, y+a) = h(x,y)$ můžeme dále psát

$$-\Delta^{*}(x,y) = \inf\left(\Delta : h(x, y+\Delta) > \mu\right) = \inf\left(\Delta : h(x-\Delta, y) > \mu\right) .$$

Podobně jako v případě (i) platí

$$\Delta^{**}(y,x) = -\Delta^{*}(x,y) \quad \text{a} \quad \Delta^{*}(y,x) = -\Delta^{**}(x,y)$$

a celkově máme $\hat{\Delta}(y,x) = -\hat{\Delta}(x,y)$. Z toho vyplývá, že $\hat{\Delta}(X,Y)$ a $-\hat{\Delta}(X,Y)$ mají stejné rozdělení, což jsme chtěli dokázat. □

Věta 13 *Rozdělení odhadu $\hat{\theta}$ definovaného vztahem (3.7) je symetrické kolem θ, jestliže rozdělení s distribční funkcí F z (3.5) je symetrické kolem 0 a $h(z)$ splňuje*

$$h(z) + h(-z) = 2\mu \ .$$

Důkaz: Důkaz je analogický jako důkaz části (i) Věty 12. □

Věta 12 (i) i Věta 13 podmiňují symetrii rozdělení odhadu kolem skutečné hodnoty symetrií rozdělení náhodných výběrů, na nichž je odhad založen. Toto rozdělení však neznáme a můžeme jen těžko říci, zda je symetrické či nikoliv. V mnoha případech ale není předpoklad symetrie neoprávněný, například porovnáváme-li účinnost dvou ošetření na subjektech rozdělených do dvojic $(X_1, Y_1), \ldots, (X_N, Y_N)$, je symetrické rozdělení $Z_i = Y_i - X_i$, $i = 1, \ldots, N$ zaručeno v případě, že přiřazení ošetření ve dvojicích probíhá náhodně.

4.3.1 Mediánová nestrannost

Pokud však rozdělení náhodných výběrů není symetrické, $\hat{\Delta}$ ani $\hat{\theta}$ nemusí být symetricky rozdělené a ani nestranné. I v takovém případě jsou však R-odhady většinou přesně nebo alespoň přibližně mediánově nestranné ve smyslu, že medián jejich rozdělení je roven skutečné hodnotě parametrů. Vyplývá to z následujícího lemmatu a věty.

Lemma 14 *Nechť (X, Y) a Z jsou náhodné vektory se sdruženým absolutně spojitým rozdělením. Pak pro libovolné reálné a splňují odhady $\hat{\Delta}$ a $\hat{\theta}$ nerovnosti*

$$P(h(X, Y - a) < \mu) \leq P(\hat{\Delta} < a) \leq P(h(X, Y - a) \leq \mu) \tag{4.7}$$

a

$$P(h(Z - a) < \mu) \leq P(\hat{\theta} < a) \leq P(h(Z - a) \leq \mu) \ . \tag{4.8}$$

Důkaz: Z definice Δ^* a Δ^{**} vyplývá

$$\Delta^{**} < a \Longrightarrow h(x, y - a) < \mu \Longrightarrow \Delta^{**} \leq a$$

a

$$\Delta^* > a \Longrightarrow h(x, y - a) > \mu \Longrightarrow \Delta^* \geq a \ .$$

V kapitole 4.2 jsme dokázali, že Δ^* a Δ^{**} mají za daných předpokladů absolutně spojité rozdělení, a proto $P(\Delta^{**} < a) = P(\Delta^{**} \leq a)$ a $P(\Delta^* > a) = P(\Delta^* \geq a)$. Platí tak

$$P(\Delta^{**} < a) = P(h(X, Y - a) < \mu)$$

a

$$P(\Delta^* < a) = P(h(X, Y - a) \leq \mu) \ .$$

Protože $P(h(X, Y - a) < \mu) \leq P(h(X, Y - a) \leq \mu)$, vyplývá z těchto vztahů přímo (4.7).

Důkaz (4.8) je zcela analogický. □

Věta 15 *Nechť (X, Y) a Z jsou náhodné vektory se sdruženým absolutně spojitým rozdělením. Nechť*

$$P_0(h(X, Y) = \mu) = \delta \quad a \quad P_0(h(Z) = \mu) = \epsilon \ .$$

Pak platí

$$\frac{1}{2} - \frac{\delta}{2} \leq P_\Delta(\hat{\Delta} \leq \Delta) \leq \frac{1}{2} + \frac{\delta}{2} \tag{4.9}$$

a

$$\frac{1}{2} - \frac{\epsilon}{2} \leq P_\theta(\hat{\theta} \leq \theta) \leq \frac{1}{2} + \frac{\epsilon}{2} \ . \tag{4.10}$$

Indexování P jako obvykle značí, že pravděpodobnost je počítána za předpokladu, že skutečná hodnota parametru se rovná indexu.

Důkaz: Z (4.3) víme, že $P_\Delta(\hat{\Delta} \leq \Delta) = P_0(\hat{\Delta} \leq 0)$, a (4.7) pak dává

$$P_0(h(X,Y) < \mu) \leq P_0(\hat{\Delta} < 0) \leq P(h(X,Y) \leq \mu) \ .$$

Z toho už přímo vyplývá vztah (4.9), protože rozdělení $h(X,Y)$ je symetrické kolem hodnoty μ.

Vztah (4.10) vyplývá stejným způsobem z (4.4) a (4.8). □

Z Věty 15 je okamžitě zřejmé, že odhady $\hat{\Delta}$ a $\hat{\theta}$ jsou mediánově nestranné, jestliže $P_0(h(X,Y) = \mu) = 0$ respektive $P_0(h(Z) = \mu) = 0$. Jako příklad si uveďme Hodges-Lehmannův odhad - jsou-li velikosti obou náhodných výběrů, na kterých odhad zakládáme, liché, je $P_0(h(X,Y) = \mu)$ rovna 0, neboť μ je hodnota, kterou $h(X,Y)$ nemůže nabývat. A i v případech, kdy jsou uvedené pravděpodobnosti nenulové, nabývají většinou relativně malých hodnot a odhady jsou mediánově nestranné alespoň přibližně.

4.4 Asymptotické vlastnosti

V této kapitole se podíváme na vlastnosti R-odhadů při velkých N a ukážeme, že souvisí s asymptotickými vlastnostmi testu, na kterém je daný odhad založen.

4.4.1 Asymptotické rozdělení

Pro dvouvýběrový problém uvažujme posloupnost rozsahů dvou výběrů $m(N)$, $n(N)$ pro $N = 1, 2, \ldots$ a Δ_N posloupnost hodnot parametru posunutí Δ. Předpokládáme, že platí $\lim_{n\to\infty} \frac{m(N)}{N} = \lambda$, a tudíž $\lim_{n\to\infty} \frac{n(N)}{N} = 1 - \lambda$. Pro jednovýběrový problém uvažujme $N = 1, 2, \ldots$ posloupnost rozsahů výběru a θ_N posloupnost hodnot parametru polohy θ. Hodnota statistiky h a střed symetrie jejího rozdělení μ pak závisí na N, budeme proto značit h_N a μ_N. h_N tedy označuje $h(X_1, \ldots, X_{m(N)}, Y_1, \ldots, Y_{n(N)})$ nebo $h(Z_1, \ldots, Z_N)$.

Věta 16 *Nechť $(X_1, \ldots, X_{m(N)}, Y_1, \ldots, Y_{n(N)})$ a $(Z_1, \ldots, Z_N)$ jsou náhodné vektory se sdruženým absolutně spojitým rozdělením. Mějme reálné konstanty $a, c_1, c_2, \ldots$. Nechť*

$$\Delta_N = \frac{-a}{c_N} \quad \textit{nebo} \quad \theta_N = \frac{-a}{c_N} \quad N = 1, 2, \ldots \ .$$

Nechť H je spojitá distribuční funkce náhodné veličiny s nulovou střední hodnotou a jednotkovým rozptylem. Nechť statistika h splňuje

$$\lim_{N\to\infty} P_N(c_N(h_N - \mu_N) \leq u) = H\left(\frac{u + aB}{A}\right) \ , \tag{4.11}$$

kde index N u P_N značí, že pravděpodobnost je počítána za předpokladu, že skutečná hodnota parametru je Δ_N nebo θ_N.

Pak platí pro libovolnou pevnou hodnotu parametru Δ

$$\lim_{N\to\infty} P_\Delta(c_N(\hat{\Delta}_N - \Delta) \le a) = H\left(\frac{aB}{A}\right) \tag{4.12}$$

nebo pro libovolnou pevnou hodnotu parametru θ

$$\lim_{N\to\infty} P_\theta(c_N(\hat{\theta}_N - \theta) \le a) = H\left(\frac{aB}{A}\right) \ . \tag{4.13}$$

Důkaz: Větu dokážeme pro odhad $\hat{\Delta}_N$, důkaz pro $\hat{\theta}_N$ je analogický. Z (4.3) vyplývá, že můžeme uvažovat $\Delta = 0$. Použijeme-li vztahy (4.7) a znovu (4.3), dostáváme

$$\begin{aligned}
&\lim_{N\to\infty} P_0(c_N\hat{\Delta}_N \le a) \\
&\qquad = \lim_{N\to\infty} P_0(h(X_1,\dots,X_{m(N)}, Y_1 - \frac{a}{c_N},\dots,Y_{n(N)} - \frac{a}{c_N}) \le \mu_N) \\
&\qquad = \lim_{N\to\infty} P_N(h(X_1,\dots,X_{m(N)}, Y_1,\dots,Y_{n(N)}) \le \mu_N) \\
&\qquad = \lim_{N\to\infty} P_N(h_N - \mu_N \le 0) = H\left(\frac{aB}{A}\right)
\end{aligned}$$

a věta je dokázána. □

Za daných podmínek bude mít tedy $c_N(\hat{\Delta}_N - \Delta)$ nebo $c_N(\hat{\theta}_N - \theta)$ asymptoticky rozdělení H s nulovou střední hodnotou a rozptylem $\frac{A^2}{B^2}$.

Pro příklad ukážeme asymptotické rozdělení odhadů založených na dvouvýběrovém a jednovýběrovém Wilcoxonově testu. Například z knihy Jurečková [8] zjistíme, že dvouvýběrový Wilcoxonův test splňuje (4.11) pro H distribuční funkci standardního normálního rozdělení, $c_N = \sqrt{N}$ a

$$A = \sqrt{\frac{1}{12}\lambda(1-\lambda)} \quad \text{a} \quad B = \lambda(1-\lambda)\int_{-\infty}^{\infty} f^2(x)dx \ ,$$

kde f je hustota rozdělení s distribuční funkcí F z (3.1). Dostáváme tak, že $\sqrt{N}(\hat{\Delta}_N - \Delta)$ má asymptoticky normální rozdělení s nulovou střední hodnotou a rozptylem

$$\frac{A^2}{B^2} = \frac{1}{12\lambda(1-\lambda)[\int_{-\infty}^{\infty} f^2(x)dx]^2} \ .$$

Pro jednovýběrový Wilcoxonův test platí, že $\sqrt{N}(h_N - \mu_N)$ splňuje (4.11) pro H distribuční funkci standardního normálního rozdělení a

$$A = \sqrt{\frac{1}{3}} \quad \text{a} \quad B = 2\int_{-\infty}^{\infty} f^2(x)dx \ , \tag{4.14}$$

kde f je hustota rozdělení s distribuční funkcí F z (3.5). Můžeme to opět zjistit z knihy Jurečková [8]. $\sqrt{N}(\hat{\theta}_N - \theta)$ má tedy asymptoticky normální rozdělení s nulovou střední hodnotou a rozptylem

$$\frac{A^2}{B^2} = \frac{1}{12[\int_{-\infty}^{\infty} f^2(x)dx]^2} \ . \tag{4.15}$$

4.4.2 Asymptotická relativní vydatnost

Definice: *Mějme $h = \{h_N\}$ a $h' = \{h'_N\}$ dvě posloupnosti testových statistik pro test hypotézy H_0 proti alternativě H_1. Nechť N a N' jsou takové rozsahy výběrů, na kterých je třeba testy založit, aby oba měly stejnou sílu β proti stejné posloupnosti alternativ, pokud hladiny významnosti obou testů konvergují ke stejné limitě α. $\lim_{N\to\infty} \frac{N'}{N}$ při pevném β a α se nazývá asymptotická relativní vydatnost (někdy také Pitmanova vydatnost) testu založeného na posloupnosti statistik h vzhledem k testu založenému na h', pokud tato limita nezávisí na β a α.*

Definice: *Mějme $\hat{\Delta} = \{\hat{\Delta}_N\}$ a $\hat{\Delta}' = \{\hat{\Delta}'_N\}$ posloupnosti dvou asymptoticky nestranných odhadů nějakého parametru. Pokud existuje limita $\lim_{N\to\infty} \frac{\mathrm{var}\,\hat{\Delta}'_N}{\mathrm{var}\,\hat{\Delta}_N}$, nazýváme ji asymptotická relativní vydatnost odhadu $\hat{\Delta}$ vzhledem k $\hat{\Delta}'$.*

Asymptotickou relativní vydatnost značíme ARE. Je zřejmé, že při porovnávání odhadu $\hat{\Delta}$ vzhledem k $\hat{\Delta}'$ jsou pro $\hat{\Delta}$ příznivější vyšší hodnoty ARE - je-li $ARE(\hat{\Delta}, \hat{\Delta}') > 1$, je asymptotický rozptyl $\hat{\Delta}$ menší než asymptotický rozptyl $\hat{\Delta}'$. Výhodnou vlastností některých R-odhadů je skutečnost, že se asymptotická relativní vydatnost dvou odhadů rovná odpovídající Pitmanově vydatnosti testů, na kterých jsou založeny.

Věta 17 *Nechť $\hat{\Delta}_N$ a $\hat{\Delta}'_N$ (respektive $\hat{\theta}_N$ a $\hat{\theta}'_N$) jsou odhady parametru Δ (respektive θ) založené na posloupnostech testových statistik h_N a h'_N. Nechť jsou splněny předpoklady Věty 16 se stejným asymptotickým rozdělením H pro h_N i h'_N. Nechť navíc $c_N = c'_N = \sqrt{N}$. Pak je asymptotická relativní vydatnost odhadu $\hat{\Delta}$ vzhledem k $\hat{\Delta}'$ (respektive $\hat{\theta}$ vzhledem k $\hat{\theta}'$) rovna odpovídající Pitmanově vydatnosti testu založeného na posloupnosti statistik h_N vzhledem k testu založenému na h'_N, pokud tato Pitmanova vydatnost existuje.*

Důkaz: Věta 16 předpokládá, že H je distribuční funkce rozdělení s jednotkovým rozptylem, a proto z (4.12) vyplývá, že $\sqrt{N}\hat{\Delta}_N$ má asymptotický

rozptyl $\frac{A^2}{B^2}$ a analogicky že $\sqrt{N}\hat{\Delta}'_N$ má asymptotický rozptyl $\frac{A'^2}{B'^2}$. Ihned tak dostáváme, že asymptotická relativní vydatnost odhadu $\hat{\Delta}$ vzhledem k $\hat{\Delta}'$ je

$$ARE(\hat{\Delta}, \hat{\Delta}') = \frac{\frac{A'^2}{B'^2}}{\frac{A^2}{B^2}} \ .$$

Podívejme se nyní na Pitmanovu vydatnost testu založeného na posloupnosti statistik h_N vzhledem k testu založenému na h'_N. Uvažujme posloupnost alternativ $\Delta_N = \frac{-a}{\sqrt{N}}$ z Věty 16 a posloupnost testů hypotézy $H_0 : \Delta = 0$ proti alternativě $H_1 : \Delta < 0$ s kritickým oborem $h_N < \mu_N$. Podle (4.11) platí

$$\lim_{N\to\infty} P_N(\sqrt{N}(h_N - \mu_N) \leq 0) = H\left(\frac{aB}{A}\right)$$

a síla testu založeného na h_N se limitně rovná $H\left(\frac{aB}{A}\right)$. Podobně síla testu založeného na posloupnosti statistik $h'_{N'}$ proti alternativám $\Delta'_{N'} = \frac{-a'}{\sqrt{N'}}$ jde k $H\left(\frac{a'B'}{A'}\right)$. K vyjádření Pitmanovy vydatnosti potřebujeme u obou testů stejnou sílu proti stejným alternativám, a proto musí platit

$$\frac{-a}{\sqrt{N}} = \frac{-a'}{\sqrt{N'}} \quad \text{a} \quad \frac{aB}{A} = \frac{a'B'}{A'} \ .$$

Nyní můžeme vyjádřit

$$\frac{N'}{N} = \frac{a'^2}{a^2} = \frac{\frac{A'^2}{B'^2}}{\frac{A^2}{B^2}} \ .$$

Tím jsme větu dokázali pro asymptotickou relativní vydatnost odhadu $\hat{\Delta}$ vzhledem k $\hat{\Delta}'$.

Důkaz pro $\hat{\theta}$ a $\hat{\theta}'$ je analogický. □

Pokud bychom ve vzorcích (3.2) a (3.3) použili k odvození odhadu dvouvýběrový t-test, získali bychom klasický odhad $\overline{Y} - \overline{X}$. Statistika dvouvýběrového t-testu je totiž spojitá vzhledem k Δ a za hypotézy $H_0 : \Delta = 0$ symetrická kolem 0. Obdobně bychom na základě (3.6) a (3.7) a jednovýběrového t-testu získali klasický odhad $\overline{Z}$ v jednovýběrovém problému. Dvouvýběrový i jednovýběrový t-test mají pro všechna rozdělení náhodných výběrů z (3.1) nebo (3.5), která mají konečný rozptyl, asymptoticky normální rozdělení. K porovnání některých R-odhadů i klasických odhadů tak můžeme díky Větě 17 využít Pitmanovy vydatnosti testů. V Tabulkách 4.1-4.4 jsou uvedeny asymptotické relativní vydatnosti několika odhadů pro různá rozdělení s distribuční funkcí F z (3.1) nebo (3.5).

Tabulka 4.1: ARE med $(Y_j - X_i)$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ vzhledem k $\overline{Y} - \overline{X}$

F			
Normální	Rovnoměrné	Dvojitě exponenciální	Logistické
0,955	1	1,5	1,097

Tabulka 4.2: ARE med $(\frac{Z_i+Z_j}{2})$, $1 \leq i \leq j \leq N$ vzhledem k $\overline{Z}$

F			
Normální	Rovnoměrné	Dvojitě exponenciální	Logistické
0,955	1	1,5	1,097

Tabulka 4.3: ARE med (Z_i), $i = 1, \ldots, N$ vzhledem k $\overline{Z}$

F			
Normální	Rovnoměrné	Dvojitě exponenciální	Logistické
0,637	0,333	2	0,823

Tabulka 4.4: ARE med (Z_i), $i = 1, \ldots, N$ vzhledem k med $(\frac{Z_i+Z_j}{2})$, $1 \leq i \leq j \leq N$

F			
Normální	Rovnoměrné	Dvojitě exponenciální	Logistické
0,667	0,333	1,333	0,750

Je dokázáno, že Pitmanova vydatnost Wilcoxonova testu vzhledem k t-testu je $\geq 0,864$ pro libovolné rozdělení s konečným rozptylem. Platí to tedy i pro ARE odhadů založených na Wilcoxonově testu vzhledem k odpovídajícímu klasickému odhadu. Neznáme-li rozdělení náhodného výběru nebo výběrů, na kterých odhad zakládáme, můžeme tudíž poměrně bezpečně použít tyto R-odhady namísto klasických odhadů.

Bickel [2] ukázal, že Galtonův test, a tedy ani odhady na něm založené, nemají asymptoticky normální rozdělení. Nemůžeme je proto porovnávat s ostatními odhady na základě Pitmanovy vydatnosti odpovídajících testů. Některé charakteristiky odhadů založených na Galtonově testu porovnáme s charakteristikami ostatních odhadů na základě simulovaných dat.

4.5 Robustnost

Již v Úvodu jsme za výhodu R-odhadů v porovnání s klasickými odhady označili jejich robustnost. Podívejme se teď tedy na charakteristiky robustnosti některých R-odhadů.

4.5.1 Míra chvostů odhadu

Pro odhad parametru polohy jednoho výběru se jako charakteristika robustnosti používá tzv. míra chvostů odhadu (viz například Jurečková [9]).

Definice: *Mějme $Z_1, \ldots, Z_N$ náhodný výběr z rozdělení se spojitou distribuční funkcí $F(u-\theta)$, $\theta \in \mathbf{R}$ a pro F platí, že $F(x)+F(-x) = 1\ \forall x$. Nechť $\hat{\theta}_N$ je odhad θ ekvivariantní vzhledem k posunutí a založený na N pozorováních. Pravděpodobnosti $P_\theta(\hat{\theta}_N - \theta > a)$, respektive $P_\theta(\hat{\theta}_N - \theta < -a)$, při velkých a nazýváme pravým, respektive levým, chvostem rozdělení odhadu $\hat{\theta}_N$.*

Robustnost odhadu pak můžeme charakterizovat následující mírou chování chvostů při pevném N:

$$B(\hat{\theta}_N; a) = \frac{-\ln P_\theta(|\hat{\theta}_N - \theta| > a)}{-\ln(1-F(a))} = \frac{-\ln P_0(|\hat{\theta}_N| > a)}{-\ln(1-F(a))},\ a > 0\ .$$

Zajímavý je odhad $\hat{\theta}_N$ s co největšími hodnotami $B(\hat{\theta}_N; a)$ pro velké hodnoty $a > 0$.

My budeme dále o rozdělení s distribuční funkcí F předpokládat pouze, že jeho medián je roven 0, a použijeme mírně pozměněnou míru chování chvostů ve tvaru navrženém v článku Zuo [13]:

$$B(\hat{\theta}_N; u, a) = \frac{-\ln P_\theta(u(\hat{\theta}_N - \theta) > a)}{-\ln P_\theta(u(Z-\theta) > a)} = \frac{-\ln P_0(u\hat{\theta}_N > a)}{-\ln P_0(uZ > a)},\ a > 0$$

pro pevné N a u takové, že $|u| = 1$. Díky ekvivarianci odhadu vzhledem k posunutí můžeme předpokládat, že $\theta = 0$. Pro zjednodušení zápisu budeme dále psát P místo P_0. Označme

$$\underline{B}(\hat{\theta}_N; a) = \min_{|u|=1}(B(\hat{\theta}_N; u, a)) \quad \text{a} \quad \overline{B}(\hat{\theta}_N; a) = \max_{|u|=1}(B(\hat{\theta}_N; u, a))\ .$$

Chování chvostů rozdělení R-odhadů závisí na statistice, na které je daný odhad založen.

Věta 18 *Nechť $\hat{\theta}_N$ je R-odhad definovaný vzorci (3.6) a (3.7) na základě statistiky $h(Z)$, která splňuje podmínku (D) pro hodnotu μ. Nechť existují taková čísla K_N a L_N, pro která platí*

$$\begin{aligned}
h(Z_1,\ldots,Z_N) \geq \mu &\implies \sum_{i=1}^{N} I(Z_i \geq 0) \geq K_N \\
h(Z_1,\ldots,Z_N) \leq \mu &\implies \sum_{i=1}^{N} I(Z_i \leq 0) \geq K_N \\
&\overset{a}{} \\
\sum_{i=1}^{N} I(Z_i > 0) \geq L_N &\implies h(Z_1,\ldots,Z_N) > \mu \\
\sum_{i=1}^{N} I(Z_i < 0) \geq L_N &\implies h(Z_1,\ldots,Z_N) < \mu \ .
\end{aligned}$$

Potom platí

$$K_N \leq \liminf_{a\to\infty} \underline{B}(\hat{\theta}_N; a) \leq \limsup_{a\to\infty} \overline{B}(\hat{\theta}_N; a) \leq L_N \ . \tag{4.16}$$

Důkaz: Prostřední nerovnost je zřejmá. Jako první dokážeme levou nerovnost. Nechť nejprve $u\hat{\theta}_N = \hat{\theta}_N$. Pak máme

$$\begin{aligned}
P(u\hat{\theta}_N > a) &= P(\hat{\theta}_N > a) \\
&\leq P(\theta^{**} > a) \\
&\leq P(h(Z_1 - a,\ldots,Z_N - a) \geq \mu) \\
&\leq P(Z_{(N-K_N+1)} - a \geq 0) \\
&= \sum_{s=K_N}^{N} \binom{N}{s} (P(uZ > a))^s (P(uZ \leq a))^{N-s} \\
&= (P(uZ > a))^{K_N} \sum_{s=K_N}^{N} \binom{N}{s} (P(uZ > a))^{s-K_N} (P(uZ \leq a))^{N-s}.
\end{aligned}$$

Podobně pro $u\hat{\theta}_N = -\hat{\theta}_N$ dostáváme

$$\begin{aligned}
P(u\hat{\theta}_N > a) &= P(-\hat{\theta}_N > a) \\
&\leq P(\theta^{*} < -a) \\
&\leq P(h(Z_1 + a,\ldots,Z_N + a) \leq \mu) \\
&\leq P(Z_{(K_N)} + a \leq 0) \\
&= \sum_{s=K_N}^{N} \binom{N}{s} (P(Z < -a))^s (P(Z \geq -a))^{N-s}
\end{aligned}$$

$$= \sum_{s=K_N}^{N} \binom{N}{s} (P(uZ > a))^s (P(uZ \le a))^{N-s}$$

$$= (P(uZ > a))^{K_N} \sum_{s=K_N}^{N} \binom{N}{s} (P(uZ > a))^{s-K_N} (P(uZ \le a))^{N-s}.$$

Označíme-li

$$f(u;a) = \sum_{s=K_N}^{N} \binom{N}{s} (P(uZ > a))^{s-K_N} (P(uZ \le a))^{N-s} \ ,$$

pak můžeme vyjádřit

$$\frac{-\ln P(u\hat{\theta}_N > a)}{-\ln P(uZ > a)} \ge K_N + \frac{-\ln f(u;a)}{-\ln P(uZ > a)} \ .$$

Ukážeme, že druhý člen pravé strany se blíží 0 pro $a \to \infty$ stejnoměrně vzhledem k u, a tím bude první část věty dokázána. Jestliže v $f(u;a)$ nahradíme pravděpodobnosti $P(uZ > a)$ a $P(uZ \le a)$ hodnotou 1, získáme horní mez

$$f(u;a) \le \sum_{s=K_N}^{N} \binom{N}{s} \equiv U \ .$$

Jestliže v $f(u;a)$ vynecháme všechny sčítance kromě prvního, dostaneme

$$f(u;a) \ge \binom{N}{K_N} (P(uZ \le a))^{N-K_N} \ge \binom{N}{K_N} (P(|Z| \le a))^{N-K_N}$$

a pro dostatečně velká a, pro která $P(|Z| \le a) \ge \frac{1}{2}$, získáme dolní mez

$$f(u;a) \ge \binom{N}{K_N} (\frac{1}{2})^{N-K_N} \equiv L \ .$$

Víme tedy, že $\ln L \le \ln f(u;a) \le \ln U$ a $\ln f(u;a)$ je stejnoměrně omezená vzhledem k u. Nyní si už stačí jen uvědomit, že $P(uZ > a) \le P(|Z| > a)$ jde k 0 stejnoměrně vzhledem k u pro $a \to \infty$, a proto $\ln P(uZ > a)$ se blíží k $-\infty$ stejnoměrně vzhledem k u pro $a \to \infty$. Z toho vyplývá, že

$$\frac{-\ln f(u;a)}{-\ln P(uZ > a)} \text{ se blíží k 0 stejnoměrně vzhledem k } u \text{ pro } a \to \infty \ .$$

Nyní dokážeme poslední nerovnost v (4.16). Uvažujme takové kladné a, pro které je statistika $h(Z_1 - a, \ldots, Z_N - a)$ spojitá vzhledem k a. Takové a

existuje díky předpokladu, že $h(Z+a)$ je neklesající funkcí a pro všechna Z, a $h(Z_1-a,\ldots,Z_N-a)$ má tedy nejvýše spočetně bodů nespojitosti vzhledem k a. Nechť opět nejprve $u\hat{\theta}_N = \hat{\theta}_N$. Odvodíme

$$\begin{aligned} P(u\hat{\theta}_N > a) &= P(\hat{\theta}_N > a) \\ &\geq P(h(Z_1-a,\ldots,Z_N-a) > \mu) \\ &\geq P(Z_{(N-L_N+1)} - a > 0) \\ &= \sum_{s=L_N}^{N} \binom{N}{s} (P(uZ>a))^s (P(uZ\leq a))^{N-s} \\ &= (P(uZ>a))^{L_N} \sum_{s=L_N}^{N} \binom{N}{s} (P(uZ>a))^{s-L_N} (P(uZ\leq a))^{N-s}. \end{aligned}$$

Obdobně pro $u\hat{\theta}_N = -\hat{\theta}_N$

$$P(u\hat{\theta}_N > a) \geq (P(uZ>a))^{L_N} \sum_{s=L_N}^{N} \binom{N}{s} (P(uZ>a))^{s-L_N} (P(uZ\leq a))^{N-s}.$$

Analogicky jako v předchozí části důkazu označíme

$$g(u;a) = \sum_{s=L_N}^{N} \binom{N}{s} (P(uZ>a))^{s-L_N} (P(uZ\leq a))^{N-s}$$

a platí

$$\frac{-\ln P(u\hat{\theta}_N > a)}{-\ln P(uZ>a)} \leq L_N + \frac{-\ln g(u;a)}{-\ln P(uZ>a)} \ .$$

Analogicky také můžeme ukázat, že

$$\frac{-\ln g(u;a)}{-\ln P(uZ>a)} \text{ se blíží k 0 stejnoměrně vzhledem k } u \text{ pro } a \to \infty \ ,$$

a požadovaná nerovnost je dokázána. □

Větu 18 nyní uplatníme na některé R-odhady. Snažíme se najít hodnoty K_N a L_N.

Odhad med $(\frac{Z_i+Z_j}{2})$ $1 \leq i \leq j \leq N$ je založený na statistice jednovýběrového Wilcoxonova testu $h(Z_1,\ldots,Z_N) = W^+ = \sum_{Z_i\geq 0} R_i^+$, jejíž rozdělení je symetrické kolem $\mu = \frac{N(N+1)}{4}$. Použijeme-li značení z kapitoly 2.1.2 o jednovýběrovém Wilcoxonově testu, víme, že pro $W^* = W^+ - W^-$ platí:

$$\begin{aligned} h(Z_1,\ldots,Z_N) \leq \mu &\iff W^*(Z_1,\ldots,Z_N) \leq 0 \\ h(Z_1,\ldots,Z_N) \geq \mu &\iff W^*(Z_1,\ldots,Z_N) \geq 0 \ . \end{aligned}$$

Chceme najít takové číslo K_N, že z $W^*(Z_1,\dots,Z_N) \geq 0$ (respektive ≤ 0) vyplývá, že mezi $Z_1,\dots,Z_N$ je nejméně K_N nezáporných (respektive nekladných) hodnot. Uvažujme tedy situaci, kdy všechny kladné hodnoty Z_i mají větší absolutní hodnotu než všechny záporné hodnoty Z_i. Hledáme nejmenší takové K_N, že

$$\frac{(2N-K_N+1)K_N}{2} - \frac{(N-K_N)(N-K_N+1)}{2} \geq 0 \ .$$

Vyřešením nerovnice získáme $K_N \geq \frac{2N+1-\sqrt{2N^2+2N+1}}{2}$.

Podobnou úvahou dospějeme k

$$\frac{L_N(1+L_N)}{2} + \frac{(N+L_N+1)(N-L_N)}{2} > 0$$

a $L_N > \frac{\sqrt{2N^2+2N+1}-1}{2}$. Celkově tedy máme

$$K_N = \left\lceil \frac{2N+1-\sqrt{2N^2+2N+1}}{2} \right\rceil \quad \text{a} \quad L_N = \left\lfloor \frac{\sqrt{2N^2+2N+1}-1}{2} \right\rfloor + 1,$$

kde $\lfloor x \rfloor$ značí největší celé číslo $\leq x$ a $\lceil x \rceil$ nejmenší celé číslo $\geq x$.

Nechť $N = 2m$ nebo $N = 2m-1$. Uvažujme odhad $\operatorname{med}\left(\frac{Z_{(i)}+Z_{(N-i+1)}}{2}\right)$ $i = 1,\dots,m$ založený na statistice jednovýběrového Galtonova testu ve tvaru $h(Z_1,\dots,Z_N) = \sum_{i=1}^{m} I[Z_{(i)}+Z_{(N-i+1)} > 0]$. Rozdělení této statistiky je za nulové hypotézy symetrické kolem hodnoty $\mu = \frac{m}{2}$.

Jestliže $h(Z) \geq \frac{m}{2}$ (respektive $\leq \frac{m}{2}$), je zřejmé, že počet kladných (respektive záporných) veličin Z_i, $i = 1,\dots,N$ je $\geq \frac{m}{2}$. Z toho vyplývá

$$K_N = \left\lfloor \frac{m+1}{2} \right\rfloor .$$

Pro určení L_N budeme rozlišovat dvě situace. Nechť nejprve $N = 2m$. Je-li více než $\frac{m}{2}$ náhodných veličin z $Z_{(i)}$, $i = 1,\dots,m$ kladných, máme jistotu, že $h(Z) > \frac{m}{2}$, neboť z uspořádanosti $Z_{(i)}$ pak plyne, že $Z_{(i)} > 0$ pro $i = m+1,\dots,2m$. Naopak, je-li více než $\frac{m}{2}$ veličin z $Z_{(i)}$, $i = m+1,\dots,2m$ záporných, musí být $h(Z) < \frac{m}{2}$, protože zároveň musí platit $Z_{(i)} < 0$ pro $i = 1,\dots,m$. Celkově tedy dostáváme

$$L_N = m + \left\lceil \frac{m+1}{2} \right\rceil .$$

Analogickou úvahou pro $N = 2m-1$ získáme

$$L_N = m - 1 + \left\lceil \frac{m+1}{2} \right\rceil .$$

Pro odhad med (Z_i) $i = 1, \ldots, N$ snadno odvodíme, že

$$K_N = \left\lfloor \frac{N+1}{2} \right\rfloor \quad \text{a} \quad L_N = \left\lfloor \frac{N+2}{2} \right\rfloor .$$

Pro srovnání uveďme, že pro klasický odhad $\overline{Z}$ je $K_N = 1$ a $L_N = N$ (viz Zuo [13]).

4.5.2 Bod selhání

Často používanou charakteristikou robustnosti odhadu je jeho bod selhání.

Definice: *Uvažujme odhad $\hat{\theta}_N$ založený na náhodném výběru $Z_1, \ldots, Z_N$. V tomto výběru nahradíme několik pozorování libovolnými hodnotami. Bod selhání $\epsilon_N^*(\hat{\theta}_N)$ odhadu $\hat{\theta}_N$ definujeme jako nejmenší počet pozorování, jejichž nahrazení může způsobit, že odhad půjde za všechny meze.*

Bod selhání souvisí s mírou chvostů odhadu, jak ukáže následující věta.

Věta 19 *Nechť $\hat{\theta}_N(Z)$ je R-odhad definovaný vzorci (3.6) a (3.7) na základě pozorování $Z = (Z_1, \ldots, Z_N)$ a statistiky $h(Z)$ a K_N a L_N jsou odpovídající hodnoty z Věty 18. Pak platí, že*

$$K_N \leq \epsilon_N^*(\hat{\theta}_N) \leq L_N \ .$$

Důkaz: Nejprve sporem ukážeme, že $K_N \leq \epsilon_N^*(\hat{\theta}_N)$. Nechť existuje $K_N - 1$ hodnot $Z_1^*, \ldots, Z_{K_N-1}^*$ takových, že pokud jimi nahradíme $K_N - 1$ pozorování z $Z_1, \ldots, Z_N$, odhad $\hat{\theta}_N(Z_1^*, \ldots, Z_{K_N-1}^*, Z_{K_N}, \ldots, Z_N)$ může jít za všechny meze. Bez újmy na obecnosti můžeme předpokládat, že

$$\hat{\theta}_N(Z_1^*, \ldots, Z_{K_N-1}^*, Z_{K_N}, \ldots, Z_N) \geq M = \max_{i=1,\ldots,N}(|Z_i|) + 1 \ .$$

Víme, že $\theta_N^{**} \geq \theta_N^*$, a proto $\theta_N^{**}(Z_1^*, \ldots, Z_{K_N-1}^*, Z_{K_N}, \ldots, Z_N) \geq M$. Z definice odhadu vyplývá, že pak $h(Z_1^* - \theta, \ldots, Z_{K_N-1}^* - \theta, Z_{K_N} - \theta, \ldots, Z_N - \theta) \geq \mu$ pro libovolné $\theta < M$. Označme

$$U = (U_1, \ldots, U_N) = (Z_1^* - \theta, \ldots, Z_{K_N-1}^* - \theta, Z_{K_N} - \theta, \ldots, Z_N - \theta) \ .$$

Z předpokladu Věty 18 potom víme, že platí $\sum_{i=1}^N I(U_i \geq 0) \geq K_N$ pro libovolné $\theta < M$. Položíme-li $\theta = \max(Z_{K_N}, \ldots, Z_N) + \frac{1}{2}$, pak θ je menší než M, ale zřejmě $\sum_{i=1}^N I(U_i \geq 0) < K_N$, což je spor.

Důkaz $\epsilon_N^*(\hat{\theta}_N) \leq L_N$ je jednoduchý. Položme $Z_i^* = z$ pro $i = 1, \ldots . L_N$, kde $z \to \infty$. Označme

$$U = (U_1, \ldots, U_N) = (Z_1^* - \theta, \ldots, Z_{L_N}^* - \theta, Z_{L_N+1} - \theta, \ldots, Z_N - \theta) \ .$$

Pak $\sum_{i=1}^N I(U_i > 0) \geq L_N$ pro libovolné θ a díky předpokladu Věty 18 také $h(U) > \mu$ pro libovolné θ. Tedy $\theta_N^*(U) \to \infty$ a věta je dokázána. □

Kapitola 5

Simulace

5.1 Výsledky výpočtů

V této kapitole prozkoumáme vlastnosti R-odhadů na základě numerických výpočtů na simulovaných datech. Všechny výpočty byly provedeny v statistickém programu R (viz `http://www.r-project.org`). Zdrojové kódy vybraných funkcí jsou uvedeny v kapitole 5.2.

V případě dvou výběrů byly pro jednoduchost uvažovány dva výběry stejného rozsahu. Mějme tedy náhodný výběr $X_1, \ldots, X_n$ a náhodný výběr $Y_1 \ldots, Y_n$, jejichž rozdělení se liší pouze posunutím. V případě jednoho výběru mějme náhodný výběr $Z_1, \ldots, Z_N$ ze symetrického rozdělení. Pomocí programu R byly simulovány náhodné výběry z normálního ($N(\mu, \sigma^2)$), rovnoměrného ($R(a, b)$), Cauchyho ($C(a, b)$), logistického ($L(a, b)$) a dvojitě exponenciálního ($DE(a, b)$) rozdělení. Výpočtené hodnoty jsou založeny na 1000 opakováních simulace pro každou situaci.

Výpočty byly provedeny pro R-odhady odvozené v kapitole 3.2 a pro srovnání také pro klasické odhady $\overline{Y} - \overline{X}$ a $\overline{Z}$. Kvůli jednoduchosti zápisu zavedeme následující značení odhadů:

Dvouvýběrový problém

P2	...	$\overline{Y} - \overline{X}$	
W2	...	$\operatorname{med}(Y_j - X_i)$	$i = 1, \ldots, n,\ j = 1, \ldots, n$
G2	...	$\operatorname{med}(Y_{(i)} - X_{(i)})$	$i = 1, \ldots, n$

Jednovýběrový problém

P1	...	$\overline{Z}$	
W1	...	$\operatorname{med}(\frac{Z_i + Z_j}{2})$	$1 \leq i \leq j \leq N$
G1	...	$\operatorname{med}(\frac{Z_{(i)} + Z_{(N-i+1)}}{2})$	$i = 1, \ldots, m$
Z1	...	$\operatorname{med}(Z_i)$	$i = 1, \ldots, N$

Pro každý z těchto odhadů byly pro různá rozdělení a rozsahy výběrů 20 a 100 spočítány základní výběrové charakteristiky: průměr, rozptyl, 25%-kvantil a 75%-kvantil. Výsledky zaokrouhlené na čtyři desetinná místa jsou obsaženy v Tabulkách 5.1-5.6.

Simulovány byly následující konkrétní situace:

Dvouvýběrový problém

$$\begin{array}{lll} N2 & \dots & X \sim N(0,1), \quad Y \sim N(2,1) \\ R2 & \dots & X \sim R(0,1), \quad Y \sim R(2,3) \\ C2 & \dots & X \sim C(0,1), \quad Y \sim C(2,1) \\ DE2 & \dots & X \sim DE(0,1), \quad Y \sim DE(2,1) \\ L2 & \dots & X \sim L(0,1), \quad Y \sim L(2,1) \end{array}$$

Jednovýběrový problém

$$\begin{array}{lll} N1 & \dots & Z \sim N(0,1) \\ R1 & \dots & Z \sim R(-1,1) \\ C1 & \dots & Z \sim C(0,1) \\ DE1 & \dots & Z \sim DE(0,1) \\ L1 & \dots & Z \sim L(0,1) \end{array}$$

Tabulka 5.1: Průměr vybraných odhadů parametru posunutí Δ vypočtený na základě simulovaných dat

n=20	$N2$	$R2$	$C2$	$DE2$	$L2$
P2	1,9956	2,0002	2,8489	1,9918	2,0119
W2	2,0000	2,0003	2,0105	1,9860	2,0082
G2	1,9992	2,0009	2,0070	1,9861	2,0104
n=100					
P2	2,0017	2,0011	4,0522	2,0115	2,0094
W2	2,0020	2,0012	2,0084	2,0103	2,0088
G2	2,0034	2,0011	2,0102	2,0102	2,0076

Tabulka 5.2: Výběrový rozptyl vybraných odhadů parametru posunutí Δ vypočtený na základě simulovaných dat

n=20	$N2$	$R2$	$C2$	$DE2$	$L2$
P2	0,1107	0,0080	1546,247	0,1954	0,3334
W2	0,1151	0,0091	0,3572	0,1396	0,2986
G2	0,1165	0,0086	0,3966	0,1473	0,3071
n=100					
P2	0,0190	0,0017	4472,727	0,0415	0,0633
W2	0,0196	0,0018	0,0677	0,0279	0,0555
G2	0,0198	0,0017	0,0725	0,0287	0,0570

Tabulka 5.3: Výběrové α-kvantily vybraných odhadů parametru posunutí Δ vypočtené na základě simulovaných dat

n=20	α	$N2$	$R2$	$C2$	$DE2$	$L2$
P2	25%	1,7774	1,9405	0,0977	1,7056	1,6228
	75%	2,2125	2,0571	4,2525	2,2805	2,4031
W2	25%	1,7794	1,9399	1,6343	1,7428	1,6354
	75%	2,2246	2,0613	2,3871	2,2183	2,3719
G2	25%	1,7768	1,9460	1,6085	1,7388	1,6276
	75%	2,2286	2,0585	2,3983	2,2277	2,3883
n=100						
P2	25%	1,9096	1,9719	0,1879	1,8715	1,8381
	75%	2,0950	2,0273	4,0164	2,1541	2,1840
W2	25%	1,9074	1,9715	1,8338	1,8939	1,8557
	75%	2,0996	2,0277	2,1820	2,1229	2,1603
G2	25%	1,9078	1,9753	1,8363	1,8967	1,8580
	75%	2,1018	2,0264	2,1807	2,1225	2,1662

Tabulka 5.4: Průměr vybraných odhadů parametru polohy θ vypočtený na základě simulovaných dat

N=20	$N1$	$R1$	$C1$	$DE1$	$L1$
P1	0,0022	0,0065	-4,6368	-0,0095	0,0188
W1	0,0017	0,0075	0,0049	-0,0085	0,0158
G1	0,0036	0,0062	0,0084	-0,0081	0,0200
Z1	-0,0033	0,0094	-0,0063	-0,0138	0,0126
N=100					
P1	-0,0017	-0,0011	2,5762	0,0009	0,0088
W1	-0,0013	-0,0013	-0,0081	0,0015	0,0081
G1	-0,0017	-0,0009	-0,0079	0,0004	0,0084
Z1	-0,0011	-0,0021	-0,0082	0,0031	0,0106

Tabulka 5.5: Výběrový rozptyl vybraných odhadů parametru polohy θ vypočtený na základě simulovaných dat

N=20	$N1$	$R1$	$C1$	$DE1$	$L1$
P1	0,0486	0,0187	13055,46	0,0959	0,1665
W1	0,0516	0,0227	0,2254	0,0691	0,1587
G1	0,0504	0,0205	0,3645	0,0789	0,1601
Z1	0,0709	0,0460	0,1375	0,0623	0,1901
N=100					
P1	0,0106	0,0033	18879,63	0,0202	0,0321
W1	0,0110	0,0035	0,0361	0,0141	0,0289
G1	0,0110	0,0035	0,0467	0,0158	0,0292
Z1	0,0155	0,0094	0,0260	0,0123	0,0382

Tabulka 5.6: Výběrové α-kvantily vybraných odhadů parametru polohy θ vypočtené na základě simulovaných dat

N=20	α	$N1$	$R1$	$C1$	$DE1$	$L1$
P1	25%	-0,1441	-0,0870	-1,1199	-0,2067	-0,2530
	75%	0,1564	0,1024	0,9679	0,1878	0,2894
W1	25%	-0,1428	0,0075	-0,2686	-0,1732	-0,2444
	75%	0,1548	0,0227	0,2734	0,1529	0,2775
G1	25%	-0,1462	-0,0793	-0,3005	-0,1797	-0,2471
	75%	0,1557	0,0999	0,3114	0,1722	0,2722
Z1	25%	-0,1855	-0,1536	-0,2309	-0,1635	-0,2796
	75%	0,1664	0,1619	0,2255	0,1393	0,3148
N=100						
P1	25%	-0,0694	-0,0403	-1,0280	-0,0970	-0,1115
	75%	0,0697	0,0380	1,0303	0,0910	0,1307
W1	25%	-0,0692	-0,0404	-0,1359	-0,0744	-0,1108
	75%	0,0687	0,0383	0,1187	0,0782	0,1221
G1	25%	-0,0712	-0,0395	-0,1467	-0,0822	-0,1120
	75%	0,0676	0,0339	0,1309	0,0846	0,1228
Z1	25%	-0,0851	-0,0667	-0,1134	-0,0670	-0,1167
	75%	0,0824	0,0611	0,1020	0,0744	0,1412

Vidíme, že kromě situace, kdy vybíráme z Cauchyho rozdělení, mají zkoumané odhady při všech ostatních rozděleních srovnatelné charakteristiky. A to jak v případě porovnáváme-li různé R-odhady mezi sebou, tak i při porovnávání R-odhadů s klasickými odhady. Ani při výběru z normálního rozdělení nic nenaznačuje, že by klasické odhady měly výrazně lepší vlastnosti. Naopak, pokud chceme odhadnout parametr na základě náhodných výběrů z Cauchyho rozdělení, je z výsledků simulací zřejmé, že klasický odhad se může poměrně hodně odchýlit od skutečné hodnoty parametru. Pro $X_1, \ldots, X_N$ náhodný výběr z Cauchyho rozdělení totiž platí, že $\overline{X}$ má pro libovolně velké N opět Cauchyho rozdělení a klasické odhady mají v tomto případě v porovnání s jinými odhady vydatnost 0. Pro náhodné výběry z Cauchyho rozdělení dáme tedy přednost nějakému R-odhadu.

V Úvodu jsme uvedli, že nevýhodou klasických odhadů je jejich citlivost k odlehlým pozorováním. Porovnáme je v tomto ohledu s R-odhady i s pomocí simulovaných dat. Simulovali jsme náhodné výběry s rozsahem 20 z normálního rozdělení s jedním a pěti odlehlými pozorováními. Používáme označení

$$
\begin{array}{lll}
N2-1 & \ldots & X \sim N(0,1), \quad Y: 19 \text{ pozorování } \sim N(2,1) \\
 & & \qquad\qquad\quad\; 1 \text{ pozorování } \sim N(2,10) \\
N2-5 & \ldots & X \sim N(0,1), \quad Y: 15 \text{ pozorování } \sim N(2,1) \\
 & & \qquad\qquad\quad\; 5 \text{ pozorování } \sim N(2,10)
\end{array}
$$

$$
\begin{array}{lll}
N1-1 & \ldots & X: 19 \text{ pozorování } \sim N(0,1) \\
 & & \qquad\quad 1 \text{ pozorování } \sim N(0,10) \\
N1-5 & \ldots & X: 15 \text{ pozorování } \sim N(0,1) \\
 & & \qquad\quad 5 \text{ pozorování } \sim N(0,10)
\end{array}
$$

Výsledky zaokrouhlené na čtyři desetinná místa jsou uvedeny v Tabulkách 5.7-5.9.

Tabulka 5.7: Průměr vybraných odhadů v případě výskytu odlehlých pozorování vypočtený na základě simulovaných dat

	$N2-1$	$N2-5$
P2	1,9962	2,0024
W2	2,0128	1,9922
G2	2,0119	1,9939
	$N1-1$	$N1-5$
P1	0,0213	-0,0358
W1	0,0104	-0,0134
G1	0,0100	-0,0146
Z1	0,0097	-0,0061

Tabulka 5.8: Výběrový rozptyl vybraných odhadů v případě výskytu odlehlých pozorování vypočtený na základě simulovaných dat

	$N2-1$	$N2-5$
P2	0,3339	1,3370
W2	0,1090	0,1817
G2	0,1117	0,1811
	$N1-1$	$N1-5$
P1	0,2995	1,3176
W1	0,0597	0,1519
G1	0,0613	0,1964
Z1	0,0755	0,1285

Tabulka 5.9: Výběrové α-kvantily vybraných odhadů v případě výskytu odlehlých pozorování vypočtené na základě simulovaných dat

	α	$N2-1$	$N2-5$
P2	25%	1,5867	1,2412
	75%	2,3995	2,7573
W2	25%	1,7796	1,6836
	75%	2,2447	2,2813
G2	25%	1,7839	1,6987
	75%	2,2453	2,2871
	α	$N1-1$	$N1-5$
P1	25%	-0,3512	-0,8606
	75%	0,3766	0,7474
W1	25%	-0,1450	-0,2801
	75%	0,1773	0,2506
G1	25%	-0,1574	-0,3045
	75%	0,1773	0,2675
Z1	25%	-0,1723	-0,2543
	75%	0,2042	0,2353

Již v případě výskytu jen jednoho odlehlého pozorování se na výběrovém rozptylu a kvantilech odhadů projevuje větší nepřesnost klasických odhadů v porovnání s R-odhady. Při výskytu pěti odlehlých pozorování je tento rozdíl ještě více zřejmý.

Dalším nezanedbatelným hlediskem, z kterého posuzujeme odhady, je jejich výpočetní složitost. Simulovali jsem situace N1 a N2 pro rozsahy výběrů 20 a 100. Pro každou situaci jsme na základě 1000 simulací 1000krát spočetli odhad a pro jednotlivé odhady změřili celkovou dobu trvání výpočtu v sekundách. Naměřené hodnoty jsou uvedeny v Tabulce 5.10.

Tabulka 5.10: Doba trvání 1000 výpočtů pro vybrané odhady na základě simulovaných dat

	N2			
	$P2$	$W2$	$G2$	
n=20	$<$ 1s	14s	3s	
n=100	1s	380s	6s	
	N1			
	$P1$	$W1$	$G1$	$Z1$
N=20	$<$ 1s	11s	3s	2s
N=100	$<$ 1s	178s	5s	

Je zřejmé, že i v době výkonných počítačů může hrát výpočetní složitost svoji roli. Porovnáváme-li odhady založené na Wilcoxonově testu s odhady založenými na Galtonově testu, vychází nám v tomto ohledu výhodnější druhé R-odhady. Vzhledem k tomu, že jinak mají tyto odhady podobné charakteristiky, jeví se odhady založené na Galtonově testu jako vhodné alternativy ke známějším odhadům založeným na Wilcoxonově testu.

5.2 Zdrojové kódy

Ukážeme zde zdrojové kódy základních procedur pro simulace z normálního rozdělení. Zdrojové kódy pro ostatní simulované situace se liší pouze příkazem pro generování náhodných výběrů z daného rozdělení. K výpočtu odhadů založených na jednovýběrovém i dvouvýběrovém Wilcoxonově testu byla použita funkce `wilcox.test`, která provede Wilcoxonův test na základě zadaných dat a při zadání parametru `conf.int=TRUE` vypočte odhad odpovídajícího parametru podle vzorců (3.8) a (3.10).

Poznámka: V programu R obsahuje funkci pro nasimulování náhodného výběru z dvojitě exponenciálního rozdělení balík `rmutil`, který lze stáhnout na adrese `http://popgen.unimaas.nl/~jlindsey/rcode.html`. Ostatní rozdělení lze simulovat pomocí funkcí z balíku `stats` základně obsaženého v R.

Dvouvýběrový problém:

```
normal2<-function(n,mu1,sigma1,mu2,sigma2){
P2<-c(1:1000)
W2<-c(1:1000)
G2<-c(1:1000)
D<-c(1:n)
for (i in 1:1000) {
  x<-rnorm(n,mu1,sigma1)
  y<-rnorm(n,mu2,sigma2)
  P2[i]<-mean(y)-mean(x)
  W2[i]<-wilcox.test(y,x,conf.int=TRUE)$estimate
  sx<-sort(x)
  sy<-sort(y)
  for (k in 1:n){
    D[k]<-(sy[k]-sx[k])
  }
  G2[i]<-median(D) }
```

```
print("P2")
print(mean(P2))
print(var(P2))
print(quantile(P2,c(0.25,0.75)))
print("W2")
print(mean(W2))
print(var(W2))
print(quantile(W2,c(0.25,0.75)))
print("G2")
print(mean(G2))
print(var(G2))
print(quantile(G2,c(0.25,0.75)))
}
```

Jednovýběrový problém:

```
normal1<-function(n,mu,sigma){
P1<-c(1:1000)
W1<-c(1:1000)
G1<-c(1:1000)
Z1<-c(1:1000)
m<-(n+1)%/%2
D<-c(1:m)
for (i in 1:1000) {
  x<-rnorm(n,mu,sigma)
  P1[i]<-mean(x)
  W1[i]<-wilcox.test(x,conf.int=TRUE)$estimate
  sx<-sort(x)
  for (k in 1:m){
    D[k]<-(sx[k]+sx[n-k+1])/2
  }
  G1[i]<-median(D)
  Z1[i]<-median(x)
}
print("P1")
print(mean(P1))
print(var(P1))
print(quantile(P1,c(0.25,0.75)))
print("W1")
print(mean(W1))
print(var(W1))
print(quantile(W1,c(0.25,0.75)))
```

```
print("G1")
print(mean(G1))
print(var(G1))
print(quantile(G1,c(0.25,0.75)))
print("Z1")
print(mean(Z1))
print(var(Z1))
print(quantile(Z1,c(0.25,0.75)))
}
```

Při porovnávání výpočetní složitosti jednotlivých odhadů byla pro odhady založené na jednovýběrovém i dvouvýběrovém Wilcoxonově testu použita funkce počítající odhady přímo na základě vzorců (3.8) a (3.10), abychom zamezili ovlivnění výsledku prováděním testu.

```
w2normal<-function(n,mu1,sigma1,mu2,sigma2){
W2<-c(1:1000)
D<-c(1:(n*n))
for (i in 1:1000) {
  x<-rnorm(n,mu1,sigma1)
  y<-rnorm(n,mu2,sigma2)
  p<-0
  for (j in 1:n){
    for (k in 1:n){
      p<-p+1
      D[p]<-(y[j]-x[k])
    }
  }
  W2[i]<-median(D)
}
print("Konec")
}
```

```
w1normal<-function(n,mu,sigma){
W1<-c(1:1000)
m<-(n*(n+1))/2
D<-c(1:m)
for (i in 1:1000) {
  x<-rnorm(n,mu,sigma)
  p<-0
  for (j in 1:n){
    for (k in j:n){
      p<-p+1
      D[p]<-(x[j]+x[k])/2
    }
  }
  W1[i]<-median(D)
}
print("Konec")
}
```

Kapitola 6

Příklady

Použití R-odhadů ještě ilustrujeme na několika příkladech s reálnými daty. Data byla získána z knihy Hand a kol. [5]. Všechny příklady byly opět spočítány v programu R.

Příklad 6.1: Skupina 12 samic králíka byla v době od 6. do 18. dne březosti podrobena speciální léčbě. 30. den březosti byly samice obětovány a spočítán počet živých zárodků pro každou z nich. Pro účely porovnání máme k dispozici také údaje kontrolní skupiny 12 samic, které léčbu nepodstoupily. Výsledky pokusu pro obě skupiny jsou uvedeny v Tabulce 6.1. Cílem je určit, jaký vliv má léčba na počet živých zárodků.

Tabulka 6.1: Počet živých zárodků pro obě skupiny samic

S léčbou	11	7	7	6	7	9	7	7	1	6	11	6
Bez léčby	3	8	12	4	9	7	6	3	7	9	10	8

Zřejmě můžeme předpokládat, že počet živých zárodků u léčených i neléčených samic má až na posunutí stejné rozdělení. K odhadu tohoto posunutí můžeme tedy použít R-odhady pro dvouvýběrový problém. Počty živých zárodků u samic, které podstoupily léčbu, budeme uvažovat jako náhodný výběr $Y_1, \ldots, Y_{12}$; počty živých zárodků u kontrolní skupiny jako náhodný výběr $X_1, \ldots, X_{12}$. Pak

$$\begin{aligned} \hat{\Delta}_1 &= \mathrm{med}_{i=1,\ldots,12, j=1,\ldots,12} \, (Y_j - X_i) = 0 \\ \hat{\Delta}_2 &= \mathrm{med}_{i=1,\ldots,12} \, (Y_{(i)} - X_{(i)}) = 0 \end{aligned}$$

Výsledky ukazují, že léčba nemá na počet živých zárodků žádný vliv. $\diamond$

Příklad 6.2: Uvažujeme rozdělení lidí na dva základní typy podle způsobu chování: typ A je charakterizován ctižádostivým a agresivním chováním, typ B označuje uvolněné a nesoutěživé lidi. Byla zkoumána hladina cholesterolu u obézních mužů středního věku a to jak u mužů typu A, tak u mužů typu B. Naměřené hodnoty hladiny cholesterolu v mg na 100 ml krve jsou uvedeny v Tabulce 6.2. Zajímá nás, jak typ chování ovlivňuje hladinu cholesterolu.

Tabulka 6.2: Hladiny cholesterolu v mg na 100 ml

Typ A	233	291	312	250	246	197	268
	224	239	239	254	276	234	181
	248	252	202	218	212	325	
Typ B	344	185	263	246	224	212	188
	250	148	169	226	175	242	252
	153	183	137	202	194	213	

Opět můžeme předpokládat, že hodnoty hladiny cholesterolu mužů typu A a B jsou až na posunutí stejně rozdělené. Toto posunutí odhadneme pomocí dvouvýběrových R-odhadů. Za náhodný výběr $Y_1, \ldots, Y_{20}$ budeme považovat naměřené hladiny cholesterolu u mužů typu A, za náhodný výběr $X_1, \ldots, X_{20}$ hodnoty naměřené u mužů typu B. Potom

$$\begin{aligned} \hat{\Delta}_1 &= \mathrm{med}_{i=1,\ldots,20, j=1,\ldots,20} (Y_j - X_i) = 37 \\ \hat{\Delta}_2 &= \mathrm{med}_{i=1,\ldots,20} (Y_{(i)} - X_{(i)}) = 40 \end{aligned}$$

Ctižádostivý člověk se sklony k agresivnímu chování bude mít tedy zřejmě vyšší hladinu cholesterolu než pohodový člověk. $\diamond$

Příklad 6.3: Chceme otestovat účinky léku Captopril na krevní tlak pacientů s vysokým tlakem. Máme k dispozici data 15 pacientů, u každého z nich byl změřen systolický krevní tlak těsně před podáním léku a dvě hodiny po podání. Naměřené hodnoty jsou uvedeny v Tabulce 6.3.

Tabulka 6.3: Systolický krevní tlak

Pacient	Před podáním léku	Po podání léku	Rozdíl
1	210	201	9
2	169	165	4
3	187	166	21
4	160	157	3
5	167	147	20
6	176	145	31
7	185	168	17
8	206	180	26
9	173	147	26
10	146	136	10
11	174	151	23
12	201	168	33
13	198	179	19
14	148	129	19
15	154	131	23

Výsledky pokusu budeme uvažovat jako dvojice (X_i, Y_i), $i = 1, \ldots, 15$, kde X_i je naměřený krevní tlak po podání léku a Y_i před podáním léku pro i-tého pacienta. Můžeme předpokládat, že rozdělení hodnot naměřených před a po podání léku se liší konstantně. Položíme $Z_i = Y_i - X_i$ $i = 1, \ldots, 15$ a pomocí R-odhadů odhadneme parametr polohy θ rozdělení, ze kterého pochází náhodný výběr $Z_1, \ldots, Z_{15}$. Máme

$$\begin{aligned}
\hat{\theta}_1 &= \mathrm{med}_{1\le i\le j\le 15}\,(\frac{Z_i + Z_j}{2}) = 19,75 \\
\hat{\theta}_2 &= \mathrm{med}_{i=1,\ldots,8}\,(\frac{Z_{(i)} + Z_{(N-i+1)}}{2}) = 19 \\
\hat{\theta}_3 &= \mathrm{med}_{i=1,\ldots,15}\,(Z_i) = 20
\end{aligned}$$

Můžeme tedy prohlásit, že lék Captopril má pozitivní účinky při snižování systolického krevního tlaku. $\diamond$

Příklad 6.4: Máme k dispozici záznamy o počtu vražd na 100 000 obyvatel z roků 1960 a 1970 pro 30 jihoamerických měst. Data jsou uvedena v Tabulce 6.4. Chceme posoudit, jak se změnil počet vražd za dané desetiletí.

Tabulka 6.4: Počet vražd na 100 000 obyvatel v jihoamerických městech

Město	1960	1970	Rozdíl	Město	1960	1970	Rozdíl
1	10,1	20,4	10,3	16	7,9	8,2	0,3
2	10,6	22,1	11,5	17	4,5	12,6	8,1
3	8,2	10,2	2	18	8,1	17,8	9,7
4	4,9	9,8	4,9	19	17,7	13,1	-4,6
5	11,5	13,7	2,2	20	11	15,6	4,6
6	17,3	24,7	7,4	21	10,8	14,7	3,9
7	12,4	15,4	3	22	12,5	12,6	0,1
8	11,1	12,7	1,6	23	8,9	7,9	-1
9	8,6	13,3	4,7	24	4,4	11,2	6,8
10	10	18,4	8,4	25	6,4	14,9	8,5
11	4,4	3,9	-0,5	26	3,8	10,5	6,7
12	13	14	1	27	14,2	15,3	1,1
13	9,3	11,1	1,8	28	6,6	11,4	4,8
14	11,7	16,9	5,2	29	6,2	5,5	-0,7
15	9,1	16,2	7,1	30	3,3	6,6	3,3

Záznamy za rok 1970 budeme uvažovat jako pozorování Y_i, $i = 1, \ldots, 30$, záznamy za rok 1960 jako pozorování X_i, $i = 1, \ldots, 30$. Položíme $Z_i = Y_i - X_i$, $i = 1, \ldots, 30$ a pomocí R-odhadů odhadneme parametru polohy θ rozdělení, ze kterého pochází náhodný výběr $Z_1, \ldots, Z_{30}$. Pak platí

$$\begin{aligned}
\hat{\theta}_1 &= \mathrm{med}_{1 \le i \le j \le 30} \left(\frac{Z_i + Z_j}{2}\right) = 4,1 \\
\hat{\theta}_2 &= \mathrm{med}_{i=1,\ldots,15} \left(\frac{Z_{(i)} + Z_{(N-i+1)}}{2}\right) = 4,2 \\
\hat{\theta}_3 &= \mathrm{med}_{i=1,\ldots,30} \left(Z_i\right) = 4,25
\end{aligned}$$

Asi není překvapením, že počet vražd na 100 000 obyvatel byl v roce 1970 v porovnání s rokem 1960 vyšší. $\diamond$

Literatura

[1] J. Anděl. *Základy matematické statistiky.* Preprint. Matematicko-fyzikální fakulta Univerzity Karlovy, Praha, 2002.

[2] P. J. Bickel and J. L. Hodges, Jr. The asymptotic theory of Galton's test and a related simple estimate of location. *Ann. Math. Statist.*, 38:73–89, 1967.

[3] W.J. Conover. *Practical Nonparametric Statistics.* John Wiley, New York, 1971.

[4] W. Feller. *An introduction to probability theory and its applications. 2nd ed.* A Wiley Publication in Mathematical Statistics. New York: John Wiley & Sons, Inc., 461 p. , 1957.

[5] David J. Hand, F. Daly, K. McConway, D. Lunn, and E. Ostrowski. *Handbook of small data sets.* Boca Raton, FL: Chapman & Hall/ CRC Press. 464 p., 1994.

[6] J. L. Hodges, Jr. Galton's rank-order test. *Biometrika*, 42:261–262, 1955.

[7] J. L. Hodges, Jr. and E. L. Lehmann. Estimates of location based on rank tests. *Ann. Math. Statist.*, 34:598–611, 1963.

[8] J. Jurečková. *Pořadové testy.* Státní pedagogické nakladatelství Praha, 1981.

[9] J. Jurečková. *Robustní statistické metody.* Nakladatelství Karolinum, Praha, 2001.

[10] E. L. Lehmann. *Nonparametrics: statistical methods based on ranks.* Holden-Day Inc., San Francisco, Calif., 1975.

[11] E. N. Torgersen. A counterexample on translation invariant estimators. *Ann. Math. Statist.*, 42(4):1450–1451, 1971.

[12] L. R. Verdooren. Extended tables of critical values for Wilcoxon's test statistic. *Biometrika*, 50:177–186, 1963.

[13] Yijun Zuo. Finite sample tail behavior of Hodges-Lehmann type estimators. *Statistics*, 35(4):557–568, 2001.