MA012 Statistika II

3. Neparametrické metody

Ondřej Pokora (pokora@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

(podzim 2015)

Motivace k neparametrickým metodám

Obvyklé podmínky **parametrických statistických metod**:

- normalita dat; pro výběry větších rozsahů ($n \geq 30$) nemá mírné porušení normality závažný dopad na výsledky
- homogenita rozptylů náhodných výběrů
- intervalový či poměrový charakter dat

Pokud nejsou tyto předpoklady splněny, používáme tzv. **neparametrické metody a testy**, které nevyžadují předpoklad o konkrétním typu rozložení. Většina zde uvedených testů navíc patří mezi tzv. **pořadové testy**, což jsou neparametrické testy založené na pořadích náhodných veličin v uspořádaném náhodném výběru.

Nevýhodou je skutečnost, že ve srovnání s klasickými parametrickými testy jsou neparametrické testy slabší, tzn. že nepravdivou hypotézu zamítají s menší pravděpodobností než testy parametrické.

Uspořádaný výběr, pořadí a pořádkové statistiky

Nechť $(X_1, X_2, \ldots, X_n)$ je náhodný výběr rozsahu n.

Definice 1 (uspořádaný náhodný výběr)

Uspořádaný náhodný výběr je vektor

$$(X_{(1)}, X_{(2)}, \ldots, X_{(n)}), \qquad \text{kde} \;\; X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},$$

a náhodná veličina $X_{(i)}$ se nazývá i-tá **pořádková statistika**.

Definice 2 (pořadí)

Pořadím R_i veličiny X_i je myšleno pořadí X_i v uspořádaném náhodném výběru $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$. Pokud se hodnoty neopakují, máme

$$R_i = \left|\{k : X_k \leq X_i\}\right|.$$

V praxi se může stát, že se některé hodnoty pozorování opakují. Takovým veličinám se zpravidla přiřazuje tzv. **průměrné pořadí**, které je aritmetickým průměrem pořadí veličin ve skupině veličin se stejnou hodnotou.

$$R_i = \left|\{k : X_k < X_i\}\right| + 1 + \tfrac{1}{2}\left|\{k \neq i : X_k = X_i\}\right|$$

Uspořádaný výběr, pořadí a pořádkové statistiky

Příklad 1

Pro náhodný výběr (2; 1,8; 2,1; 2,4; 1,9; 2,1; 2; 1,8; 2,3; 2,1) sestavte uspořádaný náhodný výběr a (průměrná) pořadí.

řešení

$X_{(i)}$	1,8	1,8	1,9	2	2	2,1	2,1	2,1	2,3	2,4
pořadí (i)	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	7	7	7	9	10

i	1	2	3	4	5	6	7	8	9	10
X_i	2	1,8	2,1	2,4	1,9	2,1	2	1,8	2,3	2,1
pořadí R_i	4	1	6	10	3	7	5	2	9	8
průměrné R_i	4,5	1,5	7	10	3	7	4,5	1,5	9	7

Všimněte si, že součty obou variant pořadí jsou stejné.

Příklad

Příklad 2

Deset testovaných osob mělo nezávisle na sobě a bez předchozího nácviku odhadnout dobu jedné minuty od zaznění zvukového signálu. Byly získány tyto výsledky (v sekundách):

i	1	2	3	4	5	6	7	8	9	10
X_i	53	48	45	55	63	51	66	56	50	58

Testujte hypotézu, že polovina testovaných osob dobu jedné minuty podhodnotila a polovina nadhodnotila.

Znaménkový test (sign test)

Předpokládáme, že $(X_1, \ldots, X_n)$ je náhodný výběr z rozdělení pravděpodonosti spojitého typu s mediánem $\widetilde{x}$. To znamená, že musí platit

$$P(X_i < \widetilde{x}) = P(X_i > \widetilde{x}) = \frac{1}{2}, \qquad i = 1, \ldots, n.$$

Chceme testovat hypotézu, že medián rozdělení pravděpodobnosti náhodného vektoru $(X_1, \ldots, X_n)$ je rovný zvolenému číslu $x_0 \in \mathbb{R}$

$$H_0 : \widetilde{x} = x_0 \qquad\qquad H_1 : \widetilde{x} \neq x_0.$$

Počítáme rozdíly $X_i - x_o$ od testovaného mediánu a označíme počet kladných rozdílů jako S^+,

$$S^+ = \left|\{i : X_i > x_o\}\right|.$$

Zavedeme indikátorové náhodné veličiny $\xi_1, \ldots, \xi_n$ předpisem

$$\xi_i = \begin{cases} 1, & X_i > x_0, \\ 0, & X_i \leq x_0. \end{cases}$$

Znaménkový test (sign test)

Potom můžeme psát $S^+ = \xi_1 + \cdots + \xi_n.$.

Jaké rozdělení pravděpodobnosti má náhodná veličina S^+ za H_0?

Znaménkový test (sign test)

Potom můžeme psát $S^+ = \xi_1 + \cdots + \xi_n.$.

Jaké rozdělení pravděpodobnosti má náhodná veličina S^+ za H_0? $S^+ \sim Bi\left(n,\ \frac{1}{2}\right)$

Věta 3 (Znaménkový test pro malá n)

Pokud

$$S^+ \leq k_\alpha \qquad \textit{nebo} \qquad S^+ \geq n - k_\alpha,$$

zamítneme H_0. Při levostranné, resp. pravostranné, alternativě se použije jen první, resp. jen druhá, podmínka. Hladina významnosti testu je rovna nejvýše α.

Číslo k_α je tabelovaná tzv. kritická hodnota, definovaná jako největší z čísel z množiny $\{0, \ldots, n\}$, pro něž platí

$$P\left(S^+ \leq k_\alpha\right) = \frac{1}{2^n} \sum_{i=0}^{k_\alpha} \binom{n}{i} \leq \frac{\alpha}{2}, \qquad P\left(S^+ \geq n - k_\alpha\right) = \frac{1}{2^n} \sum_{i=n-k_\alpha}^{n} \binom{n}{i} \leq \frac{\alpha}{2}.$$

Při levostranné, resp. pravostranné, alternativě se použije jen první, resp. druhá, podmínka s pravou stranou rovnou α.

Znaménkový test (sign test)

Jaké střední hodnoty a rozptyly mají náhodné veličiny ξ_i a S^+?

Znaménkový test (sign test)

Jaké střední hodnoty a rozptyly mají náhodné veličiny ξ_i a S^+?

$$\mathrm{E}\xi_i = \frac{1}{2}, \quad \mathrm{D}\xi_i = \frac{1}{4}, \quad \mathrm{E}S^+ = \frac{n}{2}, \quad \mathrm{D}S^+ = \frac{n}{4}.$$

Podle Moivreovy-Laplaceovy centrální limitní věty dostáváme

$$n \to \infty \qquad \Longrightarrow \qquad U = \frac{S^+ - \mathrm{E}S^+}{\sqrt{\mathrm{D}S^+}} \overset{as.}{\sim} N(0;1).$$

Věta 4 (Znaménkový test (asymptotická varianta))

Při použití testovací statistiky $\qquad U = \dfrac{2\,S^+ - n}{\sqrt{n}}$

hypotézu H_0 zamítneme, pokud

$|U| \geq u_{1-\alpha/2}$, *resp. pokud* $|U| \geq u_{1-\alpha}$ *při jednostranné alternativě.*

Hladina významnosti testu se s rostoucím n blíží k α.

příklad 2: znaménkový test

$$H_0 : \widetilde{x} = 60, \qquad H_1 : \widetilde{x} \neq 60$$

i	1	2	3	4	5	6	7	8	9	10
X_i	53	48	45	55	63	51	66	56	50	58
$X_i - 60$	-7	-12	-15	-5	3	-9	6	-4	-10	-2

$$n = 10, \qquad S^+ = 2, \qquad U = \frac{4 - 10}{\sqrt{10}} = -1{,}897, \quad k_{0{,}05} = 1, \quad u_{0{,}975} = 1{,}96$$

příklad 2: znaménkový test

```
SIGN.test (X, md=60)

  One-sample Sign-Test
data:  X
s = 2, p-value = 0.1094
alternative hypothesis: true median is not equal to 60
95 percent confidence interval:
 48.64889 61.37778
sample estimates:
median of x
         54
                  Conf.Level  L.E.pt  U.E.pt
Lower Achieved CI     0.8906 50.0000 58.0000
Interpolated CI       0.9500 48.6489 61.3778
Upper Achieved CI     0.9785 48.0000 63.0000
```

Znaménkový test (sign test)

- Používáme jej zejména v případě, kdy rozdělení pravděpodobnosti veličin X_i je výrazně sešikmené. T-test vyžadující normalitu náhodného výběru by v takovém případě dával zkreslené závěry.
- Test má poměrně malou sílu, je žádoucí mít větší rozsah n náhodného výběru.
- Testování pomocí statistiky U a aproximace normálním rozdělením se v praxi používá pro $n \geq 20$. Korekce nespojitosti není povinná, ale jejím použitím urychlujeme konvergenci k normálnímu rozdělení.
- Pokud jsou některé rozdíly $X_i - x_0$ nulové (což má sice teoreticky nulovou pravděpodobnost, ale v praxi se stát může), pak se tyto složky náhodného výběru vynechají a test se provede jen pro zbylé rozdíly s odpovídajícím sníženým n.

Párový znaménkový test

Pro párový náhodný výběr $\big((Y_1, Z_1), \ldots, (Y_n, Z_n)\big)$ z dvourozměrného rozdělení spojitého typu testujeme

$$H_0 : \widetilde{z} - \widetilde{y} = x_0 \qquad\qquad H_1 : \widetilde{z} - \widetilde{y} \neq x_0.$$

Vytvoříme rozdíly $X_i = Z_i - Y_i$ a na nich provedeme znaménkový test.

Jednovýběrový Wilcoxonův test (signed-rank test)

Předpokládáme, že $(X_1, \dots, X_n)$ je náhodný výběr z rozdělení pravděpodobnosti spojitého typu s hustotou $f(x)$, která je symetrická kolem mediánu $\widetilde{x}$, tj. platí

$$P(X_i < \widetilde{x}) = \int_{-\infty}^{\widetilde{x}} f(x)\,\mathrm{d}x = \int_{\widetilde{x}}^{\infty} f(x)\,\mathrm{d}x = P(X_i > \widetilde{x}) = \frac{1}{2}, \qquad i = 1, \dots, n.$$

Chceme testovat hypotézu, že medián rozdělení pravděpodobnosti náhodného vektoru $(X_1, \dots, X_n)$ je rovný zvolenému číslu $x_0 \in \mathbb{R}$

$$H_0 : \widetilde{x} = x_0 \qquad\qquad H_1 : \widetilde{x} \neq x_0$$

Předpokládáme, že žádná ze složek $X_1, \dots, X_n$ není rovna testovanému mediánu x_0, a označíme $Y_i = X_i - x_0$ rozdíly od testovaného mediánu.

Jednovýběrový Wilcoxonův test (signed-rank test)

Veličiny $Y_1, \ldots, Y_n$ seřadíme do neklesající posloupnosti podle jejich absolutní hodnoty:

$$|Y|_{(1)} \leq |Y|_{(2)} \leq \cdots \leq |Y|_{(n)}$$

Pořadí veličiny $|Y_i|$ v takto seřazené posloupnosti označíme jako R_i^+.

Označme S^+ a S^- součty pořadí R_i^+ zvlášť pro kladné a záporné rozdíly Y_i,

$$S^+ = \sum_{Y_i > 0} R_i^+, \qquad S^- = \sum_{Y_i < 0} R_i^+.$$

Čemu je rovný součet $S^+ + S^-$?

Jednovýběrový Wilcoxonův test (signed-rank test)

Veličiny $Y_1, \dots, Y_n$ seřadíme do neklesající posloupnosti podle jejich absolutní hodnoty:

$$|Y|_{(1)} \leq |Y|_{(2)} \leq \cdots \leq |Y|_{(n)}$$

Pořadí veličiny $|Y_i|$ v takto seřazené posloupnosti označíme jako R_i^+.

Označme S^+ a S^- součty pořadí R_i^+ zvlášť pro kladné a záporné rozdíly Y_i,

$$S^+ = \sum_{Y_i > 0} R_i^+, \qquad S^- = \sum_{Y_i < 0} R_i^+.$$

Čemu je rovný součet $S^+ + S^-$?

$$S^+ + S^- = \tfrac{n(n+1)}{2}$$

Jednovýběrový Wilcoxonův test (signed-rank test)

Alternativní forma výpočtu (signed-rank):

Věta 5

$$S^+ = \frac{n(n+1)}{4} + \frac{S}{2}, \quad S^- = S^+ - S, \quad \textit{kde} \quad S = \sum_{i=1}^{n} R_i^+ \operatorname{sgn} Y_i.$$

Věta 6 (Jednovýběrový Wilcoxonův test)

Pokud

$$\min\{S^+,\ S^-\} \leq w_\alpha(n),$$

zamítneme H_0. Při levostranné, resp. pravostranné, alternativě zamítneme H_0,

$$\textit{pokud} \quad S^+ \leq w_\alpha(n), \quad \textit{resp.} \ \ S^- \leq w_\alpha(n).$$

Číslo $w_\alpha(n)$ je tabelovaná kritická hodnota Wilcoxonova testu.

Jednovýběrový Wilcoxonův test (signed-rank test)

Věta 7

Za platnosti H_0 jsou vektory $(\operatorname{sgn} Y_1, \ldots, \operatorname{sgn} Y_1)$ a $(|Y|_{(1)}, \ldots, |Y|_{(n)})$ stochasticky nezávislé, S^+ má asymptoticky normální rozdělení a platí

$$\mathrm{E}S^+ = \frac{n(n+1)}{4}, \qquad \mathrm{D}S^+ = \frac{n(n+1)(2n+1)}{24},$$

$$\mathrm{E}S = 0, \qquad \mathrm{D}S = \frac{n(n+1)(2n+1)}{6}.$$

Standardizací S^+ obdržíme statistiku U,

$$n \to \infty \qquad \Longrightarrow \qquad U = \frac{S^+ - \mathrm{E}S^+}{\sqrt{\mathrm{D}S^+}} \overset{as.}{\sim} N(0;1).$$

Jednovýběrový Wilcoxonův test (signed-rank test)

Věta 8 (Jednovýběrový Wilcoxonův (asymptotická varianta))

Při použití asymptotické statistiky $U = \dfrac{S^+ - \mathrm{E}S^+}{\sqrt{\mathrm{D}S^+}}$ *zamítneme* H_0, *pokud*

$$|U| \geq u_{1-\alpha/2}, \qquad \text{resp. pokud } |U| \geq u_{1-\alpha} \text{ při jednostranné alternativě.}$$

Hladina významnosti testu se s rostoucím n *blíží k* α.

Analogicky lze využít standardizaci statistiky S na U.

Párový Wilcoxonův test

Pro párový náhodný výběr $\big((Y_1, Z_1), \ldots, (Y_n, Z_n)\big)$ z dvourozměrného rozdělení spojitého typu vytvoříme rozdíly $X_i = Z_i - Y_i$ a na nich pomocí jednovýběrového Wilcoxonova testu testujeme hypotézu o náhodné veličině $X = Z - Y$:

$$H_0 : \widetilde{x} = x_0 \qquad\qquad H_1 : \widetilde{x} \neq x_0.$$

příklad 2: Wilcoxonův signed-rank test

$$H_0 : \widetilde{x} = 60, \qquad H_1 : \widetilde{x} \neq 60$$

i	1	2	3	4	5	6	7	8	9	10
X_i	53	48	45	55	63	51	66	56	50	58
$Y = X_i - 60$	-7	-12	-15	-5	3	-9	6	-4	-10	-2
R_i^+	6	9	10	4	2	7	5	3	8	1
$\mathrm{sgn} Y_i$	-1	-1	-1	-1	1	-1	1	-1	-1	-1

$$S = -41, \quad S^+ = 7, \quad S^- = 48, \quad n = 10, \quad w_{0,05}(10) = 8$$

$$\mathrm{E}S^+ = 27{,}5, \quad \mathrm{D}S^+ = 96{,}25, \quad U = -2{,}09, \quad u_{0,975} = 1{,}96$$

příklad 2: Wilcoxonův signed-rank test

$$H_0: \widetilde{x} = 60, \qquad H_1: \widetilde{x} \neq 60$$

i	1	2	3	4	5	6	7	8	9	10
X_i	53	48	45	55	63	51	66	56	50	58
$Y = X_i - 60$	-7	-12	-15	-5	3	-9	6	-4	-10	-2
R_i^+	6	9	10	4	2	7	5	3	8	1
$\mathrm{sgn} Y_i$	-1	-1	-1	-1	1	-1	1	-1	-1	-1

$$S = -41, \quad S^+ = 7, \quad S^- = 48, \quad n = 10, \quad w_{0,05}(10) = 8$$

$$\mathrm{E}S^+ = 27{,}5, \quad \mathrm{D}S^+ = 96{,}25, \quad U = -2{,}09, \quad u_{0,975} = 1{,}96$$

```
wilcox.test (X, mu=60)

  Wilcoxon signed rank test
data:  X
V = 7, p-value = 0.03711
alternative hypothesis: true location is not equal to 60
```

Jednovýběrový Wilcoxonův test (signed-rank test)

- Wilcoxonův signed-rank test používáme pro testování mediánu rozdělení pravděpodobnosti náhodného výběru, pocházejího ze spojitého rozdělení pravděpodobnosti s hustotou symetrickou kolem mediánu. Sledovaná náhodná veličina musí mít alespoň ordinální charakter.
- Wilcoxonův test předpokládá symetrii hustoty pravděpodobnosti sledované veličiny kolem mediánu. Při nesymetrii hustoty pravděpodobnosti sledované veličiny může k zamítnutí H_0 dojít i tehdy, platí-li $\widetilde{x} = x_0$. V případě nesymetrie hustoty kolem mediánu použijeme např. znaménkový test.
- Pokud jsou některé rozdíly $X_i - x_0$ nulové (což má sice teoreticky nulovou pravděpodobnost, ale v praxi se stát může), pak se tyto složky náhodného výběru zpravidla vynechají a pořadí se počítají jen pro zbylé rozdíly.
- Asymptotická varianta testu se obvykle používá pro $n \geq 30$.
- T-test je analogií pro testování střední hodnoty v normálním rozdělení pravděpodobnosti.

Příklad

Příklad 3

Na celkem 13 polích stejné kvality půdy byly testovány 2 způsoby hnojení. Na 8 polích se zkoušel nový způsob A, zbývajících 5 polí bylo ošetřeno způsobem B. Tabulka uvádí výnosy pšenice (v tunách / hektar) na pokusných polích.

hnojení	*výnosy*
A	*(5,7; 5,5; 4,3; 5,9; 5,2; 5,6; 5,8; 5,1)*
B	*(5,0; 4,5; 4,2; 5,4; 4,4)*

Je třeba zjistit, zda způsob hnojení má vliv na výnosy pšenice.

Dvouvýběrový Wilcoxonův test (rank-sum test)

Porovnáváme dva stochasticky nezávislé náhodné výběry

- $(X_1, \ldots, X_m)$ rozsahu m z rozdělení psti. s distribuční funkcí $F(x)$,
- $(Y_1, \ldots, Y_n)$ rozsahu n z rozdělení psti. s distribuční funkcí $G(y)$.

Chceme testovat hypotézu rovnosti obou distribučních funkcí

$$H_0 : F = G \qquad\qquad H_1 : F \neq G$$

Oba výběry umístíme do tzv. sdruženého výběru

$$(Z_1, \ldots, Z_{m+n}) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$$

a ten uspořádáme do neklesající posloupnosti

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

Pořadí veličin $(X_1, \ldots, X_m)$, resp. $(Y_1, \ldots, Y_n)$, v takto seřazeném sdruženém výběru označíme

$$R_1, \ldots, R_m, \qquad \text{resp.} \quad R_{m+1}, \ldots, R_{m+n}.$$

Dvouvýběrový Wilcoxonův test (rank-sum test)

Označme T_1 a T_2 součty pořadí X-ových a Y-ových hodnot,

$$T_1 = \sum_{i=1}^{m} R_i, \qquad T_2 = \sum_{j=m+1}^{m+n} R_j.$$

Dále spočítáme statistiky

$$U_1 = m\,n + \frac{m(m+1)}{2} - T_1, \qquad U_2 = m\,n + \frac{n(n+1)}{2} - T_2,$$

podle nichž se test nazývá také **Mannův-Whitneyův U-test**.

Spočítejte součty $T_1 + T_2$ a $U_1 + U_2$.

Dvouvýběrový Wilcoxonův test (rank-sum test)

Označme T_1 a T_2 součty pořadí X-ových a Y-ových hodnot,

$$T_1 = \sum_{i=1}^{m} R_i, \qquad T_2 = \sum_{j=m+1}^{m+n} R_j.$$

Dále spočítáme statistiky

$$U_1 = m\,n + \frac{m(m+1)}{2} - T_1, \qquad U_2 = m\,n + \frac{n(n+1)}{2} - T_2,$$

podle nichž se test nazývá také **Mannův-Whitneyův U-test**.

Spočítejte součty $T_1 + T_2$ a $U_1 + U_2$.

$$T_1 + T_2 = \frac{(m+n)(m+n+1)}{2}, \qquad U_1 + U_2 = m\,n$$

Dvouvýběrový Wilcoxonův test (rank-sum test)

Věta 9

Za platnosti H_0 je

$$\mathrm{E}T_1 = \frac{m(m+n+1)}{2}, \qquad \mathrm{D}T_1 = \frac{m\,n(m+n+1)}{12},$$

$$\mathrm{E}U_1 = \mathrm{E}U_2 = \frac{m\,n}{2}, \qquad \mathrm{D}U_1 = \mathrm{D}U_2 = \frac{m\,n(m+n+1)}{12}.$$

Standardizací menší ze statistik $U_1,\ U_2$ obdržíme statistiku U_{MW},

$$n \to \infty \qquad \Longrightarrow \qquad U_{MW} = \frac{\min\{U_1,\ U_2\} - \mathrm{E}U_1}{\sqrt{\mathrm{D}U_1}} \overset{as.}{\sim} N(0;1).$$

Dvouvýběrový Wilcoxonův test (rank-sum test)

Věta 10 (Mannův-Whitneyův-Wilcoxonův test)

Nechť $m \geq n$. *Hypotézu* H_0 *zamítneme, pokud*

$$\min\{U_1,\ U_2\} \leq w_\alpha(m,n).$$

Číslo $w_\alpha(m,n)$ je tabelovaná kritická hodnota dvouvýběrového Wilcoxonova testu.

Věta 11 (Mannův-Whitneyův-Wilcoxonův test (asymptotická varianta))

Při použití asymptotické statistiky

$$U_{MW} = \frac{2\min\{U_1,\ U_2\} - m\,n}{\sqrt{m\,n(m+n+1)\,/\,3}}$$

zamítneme H_0, *pokud*

$|U_{MW}| \geq u_{1-\alpha/2}$, *resp. pokud* $|U_{MW}| \geq u_{1-\alpha}$ *při jednostranné alternativě.*

Hladina významnosti testu se s rostoucím n *blíží k* α.

příklad 3: Dvouvýběrový Wilcoxonův (rank-sum) test

$$H_0 : F_A = F_B, \qquad H_1 : F_A \neq F_B$$

	4,2	4,3	4,4	4,5	5,0	5,1	5,2	5,4	5,5	5,6	5,7	5,8	5,9	
R_i pro A		2				6	7		9	10	11	12	13	$T_1 = 70$
R_j pro B	1		3	4	5			8						$T_2 = 21$

$$T_1 = 70, \quad U_1 = 6, \quad T_2 = 21, \quad U_2 = 34, \quad \min\{U_1, U_2\} = 6, \quad w_{0,05}(8;5) = 6$$

$$U_{MW} = \frac{12 - 40}{\sqrt{40 \times 14/3}} = -2{,}049, \quad u_{0,975} = 1{,}96$$

příklad 3: Dvouvýběrový Wilcoxonův (rank-sum) test

$$H_0 : F_A = F_B, \qquad H_1 : F_A \neq F_B$$

	4,2	4,3	4,4	4,5	5,0	5,1	5,2	5,4	5,5	5,6	5,7	5,8	5,9	
R_i pro A		2				6	7		9	10	11	12	13	$T_1 = 70$
R_j pro B	1		3	4	5			8						$T_2 = 21$

$T_1 = 70$, $U_1 = 6$, $T_2 = 21$, $U_2 = 34$, $\min\{U_1, U_2\} = 6$, $w_{0,05}(8;5) = 6$

$$U_{MW} = \frac{12 - 40}{\sqrt{40 \times 14/3}} = -2{,}049, \quad u_{0,975} = 1{,}96$$

```
wilcox.test (X, Y)

  Wilcoxon rank sum test
data:  X and Y
W = 34, p-value = 0.04507
alternative hypothesis: true location shift is not equal to
   0
```

Dvouvýběrový Wilcoxonův test (rank-sum test)

- Dvouvýběrový Wilcoxonův test = Mannův-Whitneyův U-test = Mannův-Whitneyův-Wilcoxonův test = Wilcoxon rank-sum test
- Test předpokládá, že dané dva náhodné výběry jsou stochasticky nezávislé, sledované veličiny mají alespoň ordinální charakter a pochází ze spojitých rozdělení pravděpodobnosti.
- Asymptotická varianta testu se obvykle používá při $m > 10$, $n > 10$.
- Ačkoliv je test originálně zformulován pro obecnou alternativu nerovnosti distribučních funkcí, je dokázáno, že je citlivý zejména při testování hypotézy

$$H_0 : G(x) = F(x) \qquad\qquad H_1 : G(x) = F(x - \Delta),$$

tj. že distribuční funkce, a tedy i mediány, se liší pouze posunutím Δ. Není-li splněn předpoklad, že distribuční funkce se mohou lišit pouze posunutím, používá se např. dvouvýběrový Kolmogorovův-Smirnovův test.

Van der Waerdenův test

Porovnáváme dva stochasticky nezávislé náhodné výběry

- $(X_1, \dots, X_m)$ rozsahu m z rozdělení s hustotou psti. $f(x)$,
- $(Y_1, \dots, Y_n)$ rozsahu n z rozdělení s hustotou psti. $g(x) = f(x - \Delta)$.

Testujeme hypotézu, že posun Δ je nulový, tedy že oba výběry pocháazí ze stejného rozdělení pravděpodobnosti spojitého typu,

$$H_0 : \Delta = 0 \qquad\qquad H_1 : \Delta \neq 0.$$

Postupujeme stejně jako u dvouvýběrového Wilcoxonova testu.

Van der Waerdenův test je založen na statistice využívající pořadí X-ového výběru,

$$S = \sum_{i=1}^{m} \Phi^{-1}\left(\frac{R_i}{m+n+1}\right),$$

kde Φ^{-1} označuje kvantilovou funkci $N(0;1)$ rozdělení.

Van der Waerdenův test

Věta 12

Za platnosti H_0 *je rozdělení pravděpodobnosti statistiky* S *symetrické kolem střední hodnoty*

$$\mathrm{E}S = 0 \ \text{ a platí } \ \mathrm{D}S = \frac{m\,n}{(m+1)(m+n+1)} \sum_{i=1}^{m+n} \left[\Phi^{-1}\left(\frac{i}{m+n+1} \right) \right]^2 .$$

Pro malá m, n lze testovat pomocí tabulek kritických hodnot, pro větší rozsahy náhodných výběrů využíváme standardizaci a aproximaci normálním rozdělením.

Věta 13 (Van der Waerdenův test (asymptotická varianta))

Při použití asymptotické statistiky $\quad U_W = \dfrac{S}{\sqrt{\mathrm{D}S}}$

zamítneme H_0, *pokud*

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

Hladina významnosti testu se s rostoucím n *blíží k* α.

Mediánový test

Mediánový test používáme k testování stejné hypotézy jako u Van der Waerdenova testu.

Testovací statistika

$$S = \frac{m}{2} + \frac{1}{2}\sum_{i=1}^{m} \operatorname{sgn}\left(R_i - \frac{m+n+1}{2}\right)$$

je rovna počtu těch veličin z X-ového náhodného výběru, které jsou větší než medián sdruženého výběru; přitom pokud je $m+n$ liché číslo a medián sdruženého výběru patří do X-ového výběru, je tento počet zvýšen o $\frac{1}{2}$.

Mediánový test je vhodný zejména v případě cenzorovaných výběrů, kdy pro některé extrémně malé či extrémně velké hodnoty víme jen to, že jsou menší či větší než nějaká mez, ale jejich přesné hodnoty přitom neznáme.

Mediánový test

Věta 14

Za platnosti H_0 je rozdělení pravděpodobnosti statistiky S symetrické kolem střední hodnoty

$$\mathrm{E}S = \frac{m}{2} \text{ a platí } \mathrm{D}S = \begin{cases} \dfrac{m\,n}{4(m+n-1)}, & \text{pro } m+n \text{ liché,} \\ \dfrac{m\,n}{4(m+n)}, & \text{pro } m+n \text{ sudé.} \end{cases}$$

Pro větší rozsahy náhodných výběrů využíváme standardizaci a aproximaci normálním rozdělením.

Věta 15 (Mediánový test (asymptotická varianta))

Při použití asymptotické statistiky $\quad U_M = \dfrac{S - \mathrm{E}S}{\sqrt{\mathrm{D}S}}$

zamítneme H_0, pokud

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

Hladina významnosti testu se s rostoucím n blíží k α.

Jednoduché třídění: neparametrický přístup

Kruskalův-Wallisův test je neparametrickou analogií analýzy rozptylu jednoduchého třídění a je zobecněním dvouvýběrového Wilcoxonova testu pro porovnání 3 a více výběrů. Místo standardní analýzy rozptylu jej používáme zejména tehdy, jde-li o výběry z rozdělení pravděpodobnosti značně se lišících od normálního.

Předpoklady:

- uvažujeme jeden faktor A s $a > 2$ úrovněmi,
- pro každou úroveň $i = 1, \ldots, a$ faktoru máme náhodný výběr $\left(Y_{i1}, \ldots, Y_{in_i}\right)$ rozsahu n_i z rozdělení pravděpodobnosti s distribuční funkcí $F_i(x)$,
- tyto náhodné výběry jsou vzájemně stochasticky nezávislé.

Testujeme hypotézu, že faktor A nemá vliv na rozdělení pravděpodobnosti sledované veličiny Y, tzn. testujeme rovnost distribučních funkcí

$$H_0 : F_1 = F_2 = \cdots = F_a, \qquad H_1 : \exists\, i \neq j : F_i \neq F_j$$

Příklad

Příklad 4

U čtyř odrůd brambor (označených symboly A, B, C, D) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky uvádí tabulka:

odrůda	hmotnost (v kg)				
A	0,9	0,8	0,6	0,9	
B	1,3	1,0	1,3		
C	1,3	1,5	1,6	1,1	1,5
D	1,1	1,2	1,0		

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

Jednoduché třídění: neparametrický přístup

Náhodné veličiny zapíšeme ve tvaru tabulky známé z analýzy rozptylu:

faktor A	veličiny
1	$(Y_{1\,1}, \ldots, Y_{1\,n_1})$
$\vdots$	$\ldots$
i	$(Y_{i\,1}, \ldots, Y_{i\,n_i})$
$\vdots$	$\ldots$
a	$(Y_{a\,1}, \ldots, Y_{a\,n_a})$

Dále se však již postup od analýzy rozptylu liší. Všechny náhodné veličiny Y_{ij} dohromady vytvoří tzv. sdružený náhodný výběr $(Y_{1\,1}, \ldots, Y_{a\,n_a})$ o rozsahu

$$n = \sum_{i=1}^{a} n_i.$$

Ze sdruženého náhodného výběru vytvoříme uspořádaný náhodný výběr

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)},$$

v němž (průměrné) pořadí náhodné veličiny Y_{ij} označíme jako R_{ij}.

Jednoduché třídění: neparametrický přístup

Jednotlivá (průměrná) pořadí v uspořádaném sdružené výběru zapíšeme do tabulky spolu s řádkovými rozsahy a řádkovými součty pořadí

$$T_i = \sum_{j=1}^{n_i} R_{ij}, \qquad i = 1, \dots, a.$$

faktor A	pořadí	rozsah	$\sum$ pořadí
1	$(R_{11}, \dots, R_{1n_1})$	n_1	T_1
$\vdots$	$\dots$	$\vdots$	$\vdots$
i	$(R_{i1}, \dots, R_{in_i})$	n_i	T_i
$\vdots$	$\dots$	$\vdots$	$\vdots$
a	$(R_{a1}, \dots, R_{an_a})$	n_a	T_a
celkem		n	$\frac{n(n+1)}{2}$

Přitom platí

$$\sum_{i=1}^{a} T_i = \frac{n(n+1)}{2}$$

Jednoduché třídění: Kruskalův-Wallisův test

Kruskalův-Wallisův test je založen na testovací statistice

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{a} \frac{T_i^2}{n_i} - 3(n+1).$$

Věta 16

Střední hodnota testovací statistiky je rovna $\mathrm{E}Q = a - 1.$

Věta 17 (Kruskalův-Wallisův test)

Hypotézu H_0 zamítneme, pokud

$$Q \geq h_\alpha(a-1).$$

Za platnosti H_0 má statistika Q asymptoticky χ^2-rozdělení pravděpodobnosti,

$$n \to \infty \quad \Longrightarrow \quad Q \overset{as.}{\sim} \chi^2(a-1), \quad \text{a} \quad h_\alpha(a-1) \approx \chi^2_{1-\alpha}(a-1).$$

Číslo $h_\alpha(a-1)$ je tabelovaná kritická hodnota testu,
pro velká n ji aproximujeme kvantily $\chi^2(a-1)$-rozdělení pravděpodobnosti.

Jednoduché třídění: Kruskalův-Wallisův test

Pokud je v souboru více než 25 % shod, obvykle se k testování používá místo statistiky Q její korigovaná varianta

$$Q_k = \frac{Q}{K}, \qquad K = 1 - \frac{\sum\limits_k m_k(m_k^2 - 1)}{n(n^2 - 1)},$$

kde sčítací index k prochází všemi skupinami veličin majících stejnou hodnotu a m_k označuje počet shodných pozorování v k-té skupině.

příklad 4: Kruskalův-Wallisův test

i	hmotnost Y_{ij}					pořadí R_{ij}					n_i	T_i
1	0,9	0,8	0,6	0,9		3,5	2,0	1,0	3,5		4	10
2	1,3	1,0	1,3			11	5,5	11			3	27,5
3	1,3	1,5	1,6	1,1	1,5	11	13,5	15	7,5	13,5	5	60,5
4	1,1	1,2	1,0			7,5	9,0	5,5			3	22
$\sum$											15	120

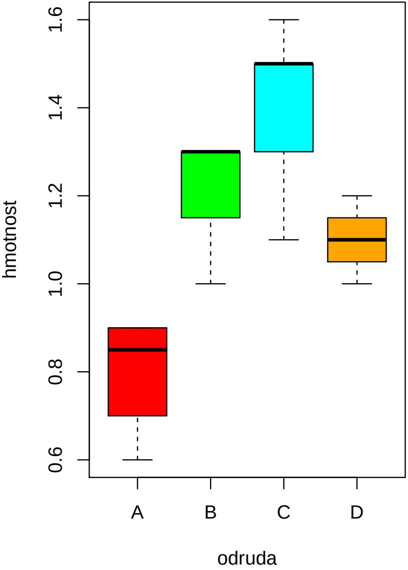
$$Q = 10{,}523,\ Q_k = 10{,}676 \quad > \quad \chi^2_{0{,}95}(3) = 7{,}815,$$

zamítáme tedy hypotézu o rovnosti distribučních funkcí ($\alpha = 0,05$).

```
library (agricolae)
KWtest <- with (tabulka, kruskal (hmotnost, odruda))
KWtest
Mtest <- with (tabulka, Median.test (hmotnost, odruda))
Mtest
```

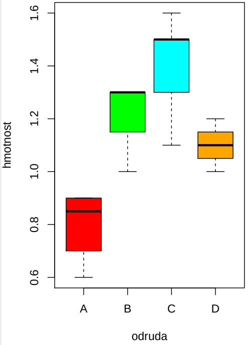
příklad 4: Kruskalův-Wallisův test

```
$statistics
     Chisq     p.chisq
  10.67585 0.01361427
$parameters
  Df ntr  t.value
   3   4 2.200985
$rankMeans
  odruda  hmotnost r
1      A  2.500000 4
2      B  9.166667 3
3      C 12.100000 5
4      D  7.333333 3
$groups
  trt      means  M
1   C 12.100000  a
2   B  9.166667 ab
3   D  7.333333  b
4   A  2.500000  c
```



příklad 4: Mediánový test

```
$statistics
     Chisq    p.chisq Median
  6.428571 0.09252244    1.1
$parameters
  Df ntr
   3   4
$Medians
  trt Median grather lessEqual
1   A   0.85       0         4
2   B   1.30       2         1
3   C   1.50       4         1
4   D   1.10       1         2
$comparison
        Median     Chisq     pvalue sig
A and B   0.90 7.0000000 0.008150972  **
A and C   1.10 5.7600000 0.016395072   *
A and D   0.90 7.0000000 0.008150972  **
B and C   1.30 2.8800000 0.089686022   .
B and D   1.15 0.6666667 0.414216178
C and D   1.25 4.8000000 0.028459737   *
```



Kruskalův-Wallisův test: mnohonásobné porovnávání

Pokud hypotézu H_0 zamítneme, je třeba rozhodnout, které dvojice dvojice výběrů podle úrovně faktoru A, tedy které dvojice distribučních funkcí F_i, F_j, se od sebe významně liší.

Dvojice výběrů pro úrovně faktoru $A = i$ a $A = j$ se významně liší, pokud

$$\left| \frac{T_i}{n_i} - \frac{T_j}{n_j} \right| > \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) h_\alpha(a-1)}\,.$$

Při vyváženém třídění, kdy $n_i = p$ pro $i = 1, \ldots, a$ a $n = a\,p$, se z důvodu větší citlivost dává přednost tzv. **Neményiově metodě** založené na Tukeyově myšlence v analýze rozptylu. Dvojice výběrů se významně liší, pokud $|T_i - T_j|$ překročí příslušnou tabelovanou kritickou hodnotu.

Jednoduché třídění: mediánový test

Mediánový test pro jednoduché třídění je založen na testovací statistice

$$Q_M = 4\sum_{i=1}^{a} \frac{A_i^2}{n_i} - n,$$

kde veličina A_i, $i = 1, \ldots, a$, je rovna počtu veličin i-tého výběru $(Y_{i1}, \ldots, Y_{in_i})$ větších než medián $\widetilde{Y}$ sdruženého výběru.

Navíc, pokud je celkový rozsah n lichý, zvětší se o $\frac{1}{2}$ to A_i, pro nějž medián $\widetilde{Y}$ sdruženého výběru patří do i-tého výběru $(Y_{i1}, \ldots, Y_{in_i})$.

Věta 18 (Mediánový test)

Při $\min\{n_1, \ldots, n_a\} \to \infty$ *hypotézu* H_0 *zamítneme, pokud*

$$Q_M \geq \chi^2_{1-\alpha}(a-1).$$

Mediánový test: Neményiova metoda

V případě vyváženého třídění lze při zamítnutí hypotézy H_0 mediánový test doplnit Neményiovou metodou mnohonásobného porovnávání.

Zavedeme indikátorové náhodné veličiny

$$\xi_{ij} = \begin{cases} 1, & Y_{ij} > \widetilde{Y} \\ 0, & Y_{ij} \leq \widetilde{Y} \end{cases} \quad \text{a označíme} \quad \overline{Z}_{i\bullet} = \frac{1}{p} \sum_{j=1}^{p} Z_{ij}.$$

Dvojice výběrů se významně liší, pokud

$$\sqrt{2p}\, \left| \overline{Z}_{i\bullet} - \overline{Z}_{j\bullet} \right|$$

překročí příslušnou tabelovanou kritickou hodnotu.

Pořadové testy v R

- Znaménkový test
  `SIGN.test (X, md=x0)` (`library (BSDA)`)
- Jednovýběrový Wilcoxonův signed-rank test
  `wilcox.test (X, mu=x0)`
- Dvouvýběrový Wilcoxonův rank-sum test
  `wilcox.test (X, Y)`
- Párový Wilcoxonův test
  `wilcox.test (X, Y, paired=TRUE)`
- Kruskalův-Wallisův test
  `kruskal (Y, group)` (`library (agricolae)`)
- Mediánový test
  `Median.test (Y, group)` (`library (agricolae)`)
- Van der Waerdenův test
  `waerden.test (Y, group)` (`library (agricolae)`)