Neparametrické testy o mediánech

Osnova:

- jednovýběrové a párové testy
- dvouvýběrové testy
- neparametrické obdoby jednofaktorové analýzy rozptylu

Motivace: Při aplikaci t-testů či analýzy rozptylu by měly být splněny určité předpoklady:

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

Pokud nejsou tyto předpoklady splněny, použijeme tzv. neparametrické testy, které nevyžadují předpoklad o konkrétním typu rozložení (např. normálním), stačí např. předpokládat, že distribuční funkce rozložení, z něhož náhodný výběr pochází, je spojitá.

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

V této kapitole se omezíme na ty neparametrické testy, které se týkají mediánů.

Jednovýběrové testy (Jde o neparametrické obdoby jednovýběrového t-testu a párového t-testu.)

Znaménkový test a jeho asymptotická varianta

Nechť $X_1,\dots,X_n$ je náhodný výběr ze spojitého rozložení. Nechť $x_{0,50}$ je mediánem tohoto rozložení a c je reálná konstanta. Testujeme hypotézu $H_0 : x_{0,50} = c$ proti oboustranné alternativě $H_1 : x_{0,50} \neq c$ (resp. proti levostranné alternativě $H_1 : x_{0,50} < c$ resp. proti pravostranné alternativě $H_1 : x_{0,50} > c$).

Znaménkový test se nejčastěji používá jako párový test, kdy máme náhodný výběr ze spojitého dvourozměrného rozložení $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix},\dots,\begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ a testujeme hypotézu o rozdílu mediánů, tj. $H_0 : x_{0,50} - y_{0,50} = c$ proti $H_1 : x_{0,50} - y_{0,50} \neq c$ (resp. proti jednostranným alternativám). Přejdeme k rozdílům $Z_1 = X_1 - Y_1,\dots,Z_n = X_n - Y_n$ a testujeme hypotézu o mediánu těchto rozdílů, tj. $H_0 : z_{0,50} = c$.

a) Utvoříme rozdíly $D_i = X_i - c$ pro jednovýběrový test resp. $D_i = Z_i - c$ pro párový test, $i = 1,\dots,n$. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)

b) Zavedeme statistiku S_Z^+, která udává počet těch rozdílů D_i, které jsou kladné. S_Z^+ je součtem náhodných veličin s alternativním rozložením (i-tá veličina nabývá hodnoty 1, když i-tý rozdíl je kladný a hodnoty 0, když je záporný). Platí-li H_0, pak pravděpodobnost kladného i záporného rozdílu je stejná, tedy $S_Z^+ \sim Bi\left(n,\frac{1}{2}\right)$. Z vlastností binomického rozložení plyne, že $E\left(S_Z^+\right) = \frac{n}{2}$, $D\left(S_Z^+\right) = \frac{n}{4}$.

c) Stanovíme kritický obor $W = \langle 0,k_1\rangle \cup \langle k_2,n\rangle$

(Nezáporná celá čísla k_1, k_2 lze najít v tabulkové příloze.)

d) H_0 zamítáme na hladině významnosti α, když $S_Z^+ \in W$.

Asymptotická varianta testu

Pro velká n (prakticky $n > 20$) lze využít asymptotické normality statistiky S_Z^+.

Testová statistika $U_0 = \frac{S_Z^+ - E(S_Z^+)}{\sqrt{D(S_Z^+)}} = \frac{S_Z^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$ má za platnosti H_0 asymptoticky rozložení $N(0,1)$.

Kritický obor pro oboustranný test: $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$.

Kritický obor pro levostranný test: $W = (-\infty, -u_{1-\alpha}\rangle$.

Kritický obor pro pravostranný test: $W = \langle u_{1-\alpha}, \infty)$.

Příklad na jednovýběrový znaménkový test:
U 10 náhodně vybraných vzorků benzínu byly zjištěny následující hodnoty oktanového čísla:
98,2 96,8 96,3 99,8 96,9 98,6 95,6 97,1 97,7 98,0.
Na hladině významnosti 0,05 testujte hypotézu, že medián oktanového čísla je 98 proti oboustranné alternativě.

Řešení:
rozdíly $x_i - 98$: 0,2 -1,2 -1,7 1,8 -1,1 0,6 -2,4 -0,9 -0,3 0,0
$S_Z^+ = 3$, nenulových rozdílů je 9. Ve statistických tabulkách najdeme pro $n = 9$ a $\alpha = 0,05$ kritické hodnoty $k_1 = 1$, $k_2 = 8$. Protože kritický obor $W = \langle 0,1 \rangle \cup \langle 8,9 \rangle$ neobsahuje hodnotu 3, nemůžeme H_0 zamítnout na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými a 10 případy. Do proměnné X napíšeme hodnoty oktanového čísla a do proměnné konst uložíme číslo 98.
Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných X, 2. seznam proměnných konst – OK – Znaménkový test.

	Znaménkový test (oktanove cislo) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet různých	procent v < V	Z	Úroveň p
X & konst	9	66,66667	0,666667	0,504985

Vidíme, že nenulových hodnot n = 9. Z nich záporných je 66,7%, tj. 6. Hodnota testové statistiky $S_Z^+ = 9 - 6 = 3$. Asymptotická testová statistika U_0 (zde označená jako Z) se realizuje hodnotou 0,6667. Odpovídající asymptotická p-hodnota je 0,505, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu, že medián oktanového čísla je 98.

Upozornění: V tomto případě není splněna podmínka pro využití asymptotické normality statistiky S_Z^+, tj. n > 20. Je tedy vhodnější najít v tabulkách kritické hodnoty pro znaménkový test. Pro n = 9 a $\alpha = 0,05$ jsou kritické hodnoty $k_1 = 1$, $k_2 = 8$. Protože kritický obor $W = \langle 0,1 \rangle \cup \langle 8,9 \rangle$ neobsahuje hodnotu 3, nezamítáme H_0 na hladině významnosti 0,05.
Dostáváme týž výsledek jako při použití asymptotického testu.

Příklad na párový znaménkový test

U 9 náhodně vybraných manželských párů byl zjištěn průměrný roční příjem (v tisících Kč).

číslo páru	1	2	3	4	5	6	7	8	9
příjem manžela	216	336	384	432	456	528	552	600	1872
příjem manželky	336	240	192	336	384	288	960	312	576

Na hladině významnosti 0,05 testujte hypotézu, že mediány příjmů manželů a manželek jsou stejné.

Řešení:

Jedná se o párový test. Vypočteme rozdíly mezi příjmy manželů a manželek, čímž úlohu převedeme na jednovýběrový test.

Testujeme $H_0 : z_{0,50} = 0$ proti oboustranné alternativě $H_1 : z_{0,50} \neq 0$, kde $z_{0,50}$ je medián rozložení, z něhož pochází rozdílový náhodný výběr $Z_1 = X_1 - Y_1, \ldots, Z_9 = X_9 - Y_9$.

Vypočtené rozdíly $x_i - y_i$: -120 96 192 96 72 240 -408 288 1296

Testová statistika $S_z^+ = 7$.

Ve statistických tabulkách najdeme pro $n = 9$ a $\alpha = 0{,}05$ kritické hodnoty $k_1 = 1,\ k_2 = 8$.

Protože kritický obor $W = \langle 0,1 \rangle \cup \langle 8,9 \rangle$ neobsahuje hodnotu 7, nemůžeme H_0 zamítnout na hladině významnosti 0,05.

Neprokázaly se tedy významné rozdíly v mediánech příjmů manželů a manželek.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými a 9 případy. Do proměnné X napíšeme příjmy manželů, do proměnné Y příjmy manželek.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných X, 2. seznam proměnných Y – OK – Znaménkový test.

Dvojice proměnných	Počet různých	procent v < V	Z	Úroveň p
X & Y	9	22,22222	1,333333	0,182422

Vidíme, že nenulových hodnot n = 9. Z nich záporných je $22,\overline{2}\%$, tj. 2. Hodnota testové statistiky $S_Z^+ = 9 - 2 = 7$. Asymptotická testová statistika U_0 (zde označená jako Z) se realizuje hodnotou $1,\overline{3}$. Odpovídající asymptotická p-hodnota je 0,1824, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu, že mediány příjmů manželů a manželek jsou stejné.

Jednovýběrový Wilcoxonův test a jeho asymptotická varianta

Frank Wilcoxon (1892 – 1965): Americký statistik a chemik

Nechť $X_1, ..., X_n$ je náhodný výběr ze spojitého rozložení s hustotou $\varphi(x)$, která je symetrická kolem mediánu $x_{0,50}$, tj.
$\varphi(x_{0,50} + x) = \varphi(x_{0,50} - x)$. Nechť c je reálná konstanta.
Testujeme hypotézu H_0: $x_{0,50} = c$
proti oboustranné alternativě H_1: $x_{0,50} \neq c$ nebo
proti levostranné alternativě H_1: $x_{0,50} < c$ nebo
proti pravostranné alternativě H_1: $x_{0,50} > c$.

Postup provedení testu:

a) Utvoříme rozdíly $D_i = X_i - c$, $i = 1, ..., n$. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)

b) Absolutní hodnoty $|D_i|$ uspořádáme vzestupně podle velikosti a spočteme pořadí R_i.

c) Zavedeme statistiky

$S_W^+ = \sum_{D_i>0} R_i^+$, což je součet pořadí přes kladné hodnoty D_i,

$S_W^- = \sum_{D_i<0} R_i^-$, což je součet pořadí přes záporné hodnoty D_i.

Přitom platí, že součet $S_W^+ + S_W^- = n(n+1)/2$.

Je-li H_0 pravdivá, pak $E(S_W^+) = n(n+1)/4$ a $D(S_W^+) = n(n+1)(2n+1)/24$.

d) Testová statistika = $\min(S_W^+, S_W^-)$ pro oboustrannou alternativu,
= S_W^+ pro levostrannou alternativu,
= S_W^- pro pravostrannou alternativu.

e) H_0 zamítáme na hladině významnosti α, když testová statistika je menší nebo rovna tabelované kritické hodnotě.

Asymptotická varianta jednovýběrového Wilcoxonova testu:

Pro $n \geq 30$ lze využít asymptotické normality statistiky S_W^+.

Platí-li H_0, pak $U_0 = \frac{S_W^+ - E(S_W^+)}{\sqrt{D(S_W^+)}} = \frac{S_W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \approx N(0,1)$.

Kritický obor:

pro oboustrannou alternativu $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$,

pro levostrannou alternativu $W = (-\infty, -u_{1-\alpha}\rangle$,

pro pravostrannou alternativu $W = \langle u_{1-\alpha}, \infty)$

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Předpoklady použití jednovýběrového Wilcoxonova testu:

- rozložení, z něhož daný náhodný výběr pochází, je spojité
- hustota tohoto rozložení je symetrická kolem mediánu
- sledovaná veličina X má aspoň ordinální charakter

(Není-li splněn předpoklad o symetrii hustoty kolem mediánu, lze použít např. znaménkový test.)

Příklad: U 12 náhodně vybraných zemí bylo zjištěno procento populace starší 60 let:

4,9 6,0 6,9 17,6 4,5 12,3 5,7 5,3 9,6 13,5 15,7 7,7.

Na hladině významnosti 0,05 testujte hypotézu, že medián procenta populace starší 60 let je 12 proti oboustranné alternativě.

Řešení:

Testujeme hypotézu H_0: $x_{0,50} = 12$ proti oboustranné alternativě H_1: $x_{0,50} \neq 12$.

Vypočteme rozdíly pozorovaných hodnot od čísla 12: -7,1 -6,0 -5,1 5,6 -7,5 0,3 -6,3 -6,7 -2,4 1,5 3,7 -4,3.

Absolutní hodnoty těchto rozdílů uspořádáme vzestupně podle velikosti. Kladné rozdíly přitom označíme červeně:

usp. $\lvert x_i - 12 \rvert$	0,3	1,5	2,4	3,7	4,3	5,1	5,6	6	6,3	6,7	7,1	7,5
pořadí	1	2	3	4	5	6	7	8	9	10	11	12

$S_W^+ = 1 + 2 + 4 + 7 = 14$,
$S_W^- = 3 + 5 + 6 + 8 + 9 + 10 + 11 + 12 = 64$,
$n = 12$, $\alpha = 0,05$, tabelovaná kritická hodnota pro $n = 12$ a $\alpha = 0,05$ je 13,
testová statistika = $\min(S_W^+, S_W^-) = \min(14,64) = 14$.
Protože $14 > 13$, H_0 nezamítáme na hladině významnosti 0,05. Znamená to, že na hladině významnosti 0,05 se nepodařilo prokázat, že aspoň v polovině zemí by se podíl populace nad 60 let odlišoval od 12 %.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 12 případy. Do proměnné procento napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 12.
Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných rozdil, Druhý seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (populace_nad_60) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
procento & konst	12	14,00000	1,961161	0,049861

Výstupní tabulka poskytne hodnotu testové statistiky SW^+ (zde označena T), hodnotu asymptotické testové statistiky U_0 a p-hodnotu pro U_0. V tomto případě je p-hodnota 0,049861, tedy nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. Tento výsledek je v rozporu s výsledkem, ke kterému jsme dospěli při přesném výpočtu. Je to způsobeno tím, že není splněna podmínka pro využití asymptotické normality statistiky SW^+, tj. $n \geq 30$.

Párový Wilcoxonův test

Nechť $(X_1, Y_1), ..., (X_n, Y_n)$ je náhodný výběr ze spojitého dvourozměrného rozložení.

Testujeme $H_0: x_{0,50} - y_{0,50} = c$ proti $H_1: x_{0,50} - y_{0,50} \neq c$ (resp. proti jednostranným alternativám).

Utvoříme rozdíly $Z_i = X_i - Y_i$, $i = 1, ..., n$ a testujeme hypotézu o mediánu $z_{0,50}$, tj. $H_0: z_{0,50} = c$ proti $H_1: z_{0,50} \neq c$.

Příklad: K zjištění cenových rozdílů mezi určitými dvěma druhy zboží bylo náhodně vybráno 15 prodejen a byly zjištěny ceny zboží A a ceny zboží B: (11,10), (14,11), (11,9), (13,9), (11,9), (10,9), (12,10), (10,8), (12,11), (11,9), (13,10), (14,10), (14,12), (19,15), (14,12). Na hladině významnosti 0,05 je třeba testovat hypotézu, že medián cenových rozdílů činí 3 Kč.

Řešení:Testujeme H_0: $z_{0,50} = 3$ proti oboustranné alternativě H_1: $z_{0,50} \neq 3$, kde $z_{0,50}$ je medián rozložení, z něhož pochází rozdílový náhodný výběr $Z_1 = X_1 - Y_1$, … $Z_{15} = X_{15} - Y_{15}$.Vypočteme rozdíly mezi cenou zboží A a cenou zboží B, čímž úlohu převedeme na jednovýběrový test. Výpočty uspořádáme do tabulky:

č. prodejny	cena zboží A	cena zboží B	rozdíl	|rozdíl-medián|	pořadí
1	11	10	1	2	12
2	14	11	3	0	-
3	11	9	2	1	5,5
4	13	9	4	1	**5,5**
5	11	9	2	1	5,5
6	10	9	1	2	12
7	12	10	2	1	5,5
8	10	8	2	1	5,5
9	12	11	1	2	12
10	11	9	2	1	5,5
11	13	10	3	0	-
12	14	10	4	1	**5,5**
13	14	12	2	1	5,5
14	19	15	4	1	**5,5**
15	14	12	2	1	5,5

(Tučně jsou vytištěna pořadí pro kladné hodnoty rozdíl - medián.)
$S_W^+ = 5,5 + 5,5 + 5,5 = 16,5$,
$S_W^- = 12 + 5,5 + 5,5 + 12 + 5,5 + 5,5 + 12 + 5,5 + 5,5 + 5,5 = 74,5$,
$n = 13$, $\alpha = 0,05$, tabelovaná kritická hodnota = 17, testová statistika = $\min(S_W^+, S_W^-) = \min(16,5; 74,5) = 16,5$. Protože $16,5 \leq 17$, H_0 zamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor se čtyřmi proměnnými A, B, rozdíl, konst a 15 případy. Do proměnných A, B napíšeme ceny zboží A a B, do proměnné rozdíl uložíme rozdíl cen A a B a do proměnné konst uložíme číslo 3.
Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných rozdil, 2. seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (ceny zbozi) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
rozdil & konst	15	16,50000	2,026684	0,042696

Testová statistika (zde označená jako T) nabývá hodnoty 16,5, asymptotická testová statistika (označená jako Z) nabývá hodnoty 2,026684, odpovídající asymptotická p-hodnota je 0,042696, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme.

Příklad (na asymptotickou variantu Wilcoxonova testu):

30 náhodně vybraných osob mělo nezávisle na sobě bez předchozího nácviku odhadnout, kdy od daného signálu uplyne právě 1 minuta. Byly získány následující výsledky (v sekundách):

53 48 45 55 63 51 66 56 50 58 61 51 64 63 59 47 46 58 52 56 61 57 48 62 54 49 51 46 53 58.

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že medián rozložení, z něhož daný náhodný výběr pochází, je 60 sekund proti oboustranné alternativě (nulová hypotéza vlastně tvrdí, že polovina osob délku jedné minuty podhodnotí a druhá nadhodnotí).

Řešení:

Testujeme H_0: $x_{0,50} = 60$ proti oboustranné alternativě H_1: $x_{0,50} \neq 60$.

Obvyklým způsobem stanovíme statistiku $S_W^+ = 55$.

Asymptotická testová statistika:

$$U_0 = \frac{S_W^+ - E\left(S_W^+\right)}{\sqrt{D\left(S_W^+\right)}} = \frac{S_W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{55 - \frac{30(30+1)}{4}}{\sqrt{\frac{30(30+1)(2.30+1)}{24}}} = -3{,}65$$

Kritický obor:

$$W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right) = \left(-\infty, -u_{0,975}\right\rangle \cup \left\langle u_{0,975}, \infty\right) = \left(-\infty, -1{,}96\right\rangle \cup \left\langle 1{,}96, \infty\right).$$

Testová statistika se realizuje v kritickém oboru, tedy H_0 zamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že pravděpodobnost nadhodnocení jedné minuty není stejná jako pravděpodobnost podhodnocení.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 30 případy. Do proměnné odhad napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 60.
Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných odhad, 2. seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (odhad minuty) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
odhad & konst	30	55,00000	3,650880	0,000261

Testová statistika (zde označená jako T) nabývá hodnoty 55, asymptotická testová statistika (označená jako Z) nabývá hodnoty 3,65088, odpovídající asymptotická p-hodnota je 0,000261, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme.

Dvouvýběrové testy (Jedná se o neparametrickou obdobu dvouvýběrového t-testu)

Dvouvýběrový Wilcoxonův test a jeho asymptotická varianta

Nechť $X_1, ..., X_n$ a $Y_1, ..., Y_m$ jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit pouze posunutím. Označme $x_{0,50}$ medián prvního rozložení a $y_{0,50}$ medián druhého rozložení. Na hladině významnosti 0,05 testujeme hypotézu, že distribuční funkce těchto rozložení jsou shodné neboli mediány jsou shodné proti alternativě, že jsou rozdílné, tj.

$H_0: x_{0,50} - y_{0,50} = 0$ proti $H_1: x_{0,50} - y_{0,50} \neq 0$.

Postup provedení testu:

a) Všech $n + m$ hodnot $X_1, ..., X_n$ a $Y_1, ..., Y_m$ uspořádáme vzestupně podle velikosti.

b) Zjistíme součet pořadí hodnot $X_1, ..., X_n$ a označíme ho T_1.
Součet pořadí hodnot $Y_1, ..., Y_m$ označíme T_2.

c) Vypočteme statistiky $U_1 = mn + n(n+1)/2 - T_1$, $U_2 = mn + m(m+1)/2 - T_2$.
Přitom platí $U_1 + U_2 = mn$.

d) Pokud $\min(U_1,U_2) \leq$ tabelovaná kritická hodnota (pro dané rozsahy výběrů m, n a dané α), pak nulovou hypotézu o totožnosti obou distribučních funkcí zamítáme na hladině významnosti α. V tabulkách: $n = \min\{m,n\}$ a $m = \max\{m,n\}$.

Asymptotická varianta dvouvýběrového Wilcoxonova testu:

Pro velká n, m (n, m > 30) lze využít asymptotické normality statistiky U_1.

Platí-li H_0, pak $U_0 = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \approx N(0,1)$, kde $U_1 = \min(U_1, U_2)$.

Kritický obor:

pro oboustrannou alternativu $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$,

pro levostrannou alternativu $W = (-\infty, -u_{1-\alpha}\rangle$,

pro pravostrannou alternativu $W = \langle u_{1-\alpha}, \infty)$

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Předpoklady použití dvouvýběrového Wilcoxonova testu:

- dané dva náhodné výběry jsou nezávislé
- rozložení, z nichž dané dva náhodné výběry pocházejí, jsou spojitá
- distribuční funkce těchto rozložení se mohou lišit pouze posunutím
- sledovaná veličina má aspoň ordinální charakter

(Není-li splněn předpoklad, že distribuční funkce se mohou lišit pouze posunutím, lze použít např. dvouvýběrový Kolmogorovův – Smirnovův test.)

Příklad:
Bylo vybráno 10 polí stejné kvality. Na čtyřech z nich se zkoušel nový způsob hnojení, zbylých šest bylo ošetřeno starým způsobem. Pole byla oseta pšenicí a sledoval se její hektarový výnos. Je třeba zjistit, zda nový způsob hnojení má týž vliv na průměrné hektarové výnosy pšenice jako starý způsob hnojení.
hektarové výnosy při novém způsobu: 51 52 49 55
hektarové výnosy při starém způsobu: 45 54 48 44 53 50
Test proveďte na hladině významnosti 0,05.

Řešení:
Na hladině významnosti 0,05 testujeme H_0: $x_{0,50} - y_{0,50} = 0$ proti oboustranné alternativě H_1: $x_{0,50} - y_{0,50} \neq 0$.

usp. hodnoty	44	45	48	**49**	50	**51**	**52**	53	54	**55**		
pořadí x-ových hodnot						4		6	7			10
pořadí y-ových hodnot		1		2	3		5			8	9	

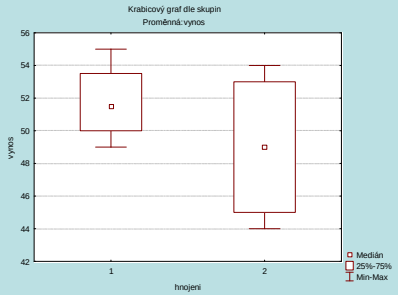
$T_1 = 4 + 6 + 7 + 10 = 27$, $T_2 = 1 + 2 + 3 + 5 + 8 + 9 = 28$
$U_1 = 4.6 + 4.5/2 - 27 = 7$, $U_2 = 4.6 + 6.7/2 - 28 = 17$
Kritická hodnota pro $\alpha = 0,05$, $\min(4,6) = 4$, $\max(4,6) = 6$ je 2. Protože $\min(7,17) = 7 > 2$, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že nový způsob hnojení má na hektarové výnosy pšenice stejný vliv jako starý způsob.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 10 případy. Do proměnné vynos napíšeme zjištěné hodnoty a do proměnné hnojeni napíšeme 4x číslo 1 pro nový způsob hnojení a 6x číslo 2 pro starý způsob hnojení.
Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných vynos, Nezáv. (grupov.) proměnná hnojeni – OK – M-W U test.
Upozornění: Ve STATISTICE je dvouvýběrový Wilcoxonův test uveden pod názvem Mannův – Whitneyův test.

	Mann-Whitneyův U test (vynos) Dle proměn. hnojeni Označené testy jsou významné na hladině p <,05000									
Proměnná	Sčt poř. skup. 1	Sčt poř. skup. 2	U	Z	Úroveň p	Z upravené	Úroveň p	N platn. skup. 1	N platn. skup. 2	2*1str. přesné p
vynos	27,00000	28,00000	7,000000	1,066004	0,286423	1,066004	0,286423	4	6	0,352381

Ve výstupní tabulce jsou součty pořadí T_1, T_2, hodnota testové statistiky
$\min(U_1, U_2)$ označená U, hodnota asymptotické testové statistiky U_0 (označená Z), asymptotická p-hodnota pro U_0 a přesná p-hodnota (ozn. 2*1str. přesné p – ta se používá pro rozsahy výběrů pod 30). V našem případě přesná p-hodnota = 0,352381, tedy H_0 nezamítáme na hladině významnosti 0,05.
Výpočet je vhodné doplnit krabicovým diagramem.



Je zřejmé, že výnosy při novém způsobu hnojení jsou vesměs nižší než při starém způsobu a také vykazují mnohem větší variabilitu.

Dvouvýběrový Kolmogorovův - Smirnovův test

Nechť $X_1,\dots,X_n$ a $Y_1,\dots,Y_m$ jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit nejenom posunutím, ale také tvarem.

Testujeme hypotézu H_0: distribuční funkce těchto rozložení jsou shodné (tj. všech n + m veličin pochází z téhož rozložení) proti alternativě H_1: distribuční funkce jsou rozdílné.

Nechť $F_1(x)$ je výběrová distribuční funkce 1. výběru a $F_2(y)$ je výběrová distribuční funkce 2. výběru.

Testová statistika $D = \max_{-\infty<x<\infty} |F_1(x) - F_2(x)|$.

H_0 zamítáme na hladině významnosti α, když $D \geq D_{n,m}(\alpha)$, kde $D_{n,m}(\alpha)$ je tabelovaná kritická hodnota.

Pro větší rozsahy n,m lze kritickou hodnotu aproximovat vzorcem $\sqrt{\frac{n+m}{2nm}\ln\frac{2}{\alpha}}$.

Příklad: Výrobce určitého výrobku se má rozhodnout mezi dvěma dodavateli polotovarů vyrábějících je různými technologiemi. Rozhodující je procentní obsah určité látky.
1. technologie: 1,52 1,57 1,71 1,34 1,68
2. technologie: 1,75 1,67 1,56 1,66 1,72 1,79 1,64 1,55
Na hladině významnosti 0,05 posuďte pomocí dvouvýběrového K-S testu, zda je oprávněný předpoklad, že obě technologie poskytují stejné procento účinné látky.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 13 případy. Do proměnné X napíšeme zjištěné hodnoty a do proměnné ID napíšeme 5x číslo 1 pro první technologii a 8x číslo 2 pro starý druhou technologii.
Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných X, Nezáv. (grupov.) proměnná ID – OK – Kolmogorov-Smirnovův 2-výběrový test.

Proměnná	Max záp rozdíl	Max klad rozdíl	Úroveň p	Průměr skup. 1	Průměr skup. 2	Sm.odch. skup. 1	Sm.odch. skup. 2	N platn. skup. 1	N platn. skup. 2
obsah	-0,400000	0,025000	p > .10	1,564000	1,667500	0,147411	0,085147	5	8

Ve výstupní tabulce pro dvouvýběrový K-S test dostaneme maximální záporný a maximální kladný rozdíl mezi hodnotami obou výběrových distribučních funkcí, dolní omezení pro p-hodnotu (p > 0,1), průměry, směrodatné odchylky a rozsahy obou výběrů. Jelikož p-hodnota převyšuje hladinu významnosti 0,05, na této hladině nelze nulovou hypotézu zamítnout.

Kruskalův - Wallisův test

William Kruskal (1919 – 2005): Americký matematik

Wilson Allen Wallis (1912 – 1988): Americký matematik

Nechť je dáno $r \geq 3$ nezávislých náhodných výběrů o rozsazích $n_1, \ldots, n_r$. Předpokládáme, že tyto výběry pocházejí ze spojitých rozložení. Označme $n = n_1 + \ldots + n_r$. Na asymptotické hladině významnosti α chceme testovat hypotézu, že všechny tyto výběry pocházejí z téhož rozložení.

Postup testu:

a) Všech n hodnot seřadíme do rostoucí posloupnosti.
b) Určíme pořadí každé hodnoty v tomto sdruženém výběru.
c) Označme T_j součet pořadí těch hodnot, které patří do j-tého výběru, j = 1, ..., r (kontrola: musí platit $T_1 + ... + T_r = n(n+1)/2$).
d) Testová statistika má tvar: $Q = \frac{12}{n(n+1)} \sum_{j=1}^{r} \frac{T_j^2}{n_j} - 3(n+1)$. Platí-li H_0, má statistika Q asymptoticky rozložení $\chi^2(r-1)$.
e) Kritický obor: $W = \langle \chi^2_{1-\alpha}(r-1), \infty)$.
f) H_0 zamítneme na asymptotické hladině významnosti α, když $Q \geq \chi_{1-\alpha}{}^2(r-1)$.

Příklad: V roce 1980 byly získány tři nezávislé výběry obsahující údaje o průměrných ročních příjmech (v tisících dolarů) čtyř sociálních skupin ve třech různých oblastech USA.
jižní oblast: 6 10 15 29
pacifická oblast: 11 13 17 131
severovýchodní oblast: 7 14 28 25
Na hladině významnosti 0,05 testujte hypotézu, že příjmy v těchto oblastech se neliší.

Řešení:
Výpočty uspořádáme do tabulky

Usp. hodnoty	6	7	10	11	13	14	15	17	25	28	29	131
Pořadí 1.výběru	1		3				7				11	
Pořadí 2.výběru				4	5			8				12
Pořadí 3.výběru		2				6			9	10		

$T_1 = 1 + 3 + 7 + 11 = 22,$
$T_2 = 4 + 5 + 8 + 12 = 29,$
$T_3 = 2 + 6 + 9 + 10 = 27\ ,$

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^{r} \frac{T_j^2}{n_j} - 3(n+1) = \frac{12}{12 \cdot 13} \left(\frac{22^2}{4} + \frac{29^2}{4} + \frac{27^2}{4} \right) - 3 \cdot 13 = 0{,}5,$$

$$W = \langle \chi^2_{1-\alpha}(r-1), \infty \rangle = \langle \chi^2_{0,95}(2), \infty \rangle = \langle 5{,}991, \infty \rangle$$

Protože $Q < 5{,}991$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.
Rozdíly mezi průměrnými ročními příjmy v uvedených třech oblastech se neprokázaly.

Mediánový test

Výchozí situace je stejná jako u K-W testu

Postup testu:

a) Všech n hodnot uspořádáme do rostoucí posloupnosti.

b) Najdeme medián $x_{0,50}$ těchto n hodnot.

c) Označme P_j počet hodnot v j-tém výběru, které jsou větší nebo rovny mediánu $x_{0,50}$.

d) Testová statistika má tvar $Q_M = 4\sum_{j=1}^{r} \frac{P_j^2}{n_j} - n$. Platí-li H_0, má statistika Q_M asymptoticky rozložení $\chi^2(r-1)$.

d) Kritický obor: $W = \langle \chi^2_{1-\alpha}(r-1), \infty)$.

e) H_0 zamítneme na asymptotické hladině významnosti α, když $Q_M \geq \chi_{1-\alpha}{}^2(r-1)$.

Příklad:
Pro data o průměrných ročních příjmech proveďte mediánový test. Hladinu významnosti volte 0,05.

Řešení:
Usp. hodnoty 6 7 10 11 13 14 15 17 25 28 29 131

Medián je průměr 6. a 7. uspořádané hodnoty: $x_{0,50} = \frac{14+15}{2} = 14{,}5.$

V prvním výběru existují 2 hodnoty, které jsou větší nebo rovny 14,5, stejně tak i ve druhém a třetím výběru,
tedy $P_1 = P_2 = P_3 = 2$.

Testová statistika: $Q_M = 4\sum_{j=1}^{r}\frac{P_j^2}{n_j} - n = 4\left[\frac{1}{4}\left(2^2+2^2+2^2\right)\right] - 12 = 0$

Kritický obor: $W = \left\langle \chi^2_{1-\alpha}(r-1), \infty \right) = \left\langle \chi^2_{0,95}(2), \infty \right) = \left\langle 5{,}991, \infty \right)$

Protože $Q_M < 5{,}991$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Metody mnohonásobného porovnávání

Zamítneme-li hypotézu, že všechny náhodné výběry pocházejí z téhož rozložení, zajímá nás, které dvojice náhodných výběrů se liší na zvolené hladině významnosti. Testujeme H_0: k-tý a l-tý náhodný výběr pocházejí z téhož rozložení, k, l = 1, .., r, k ≠ l proti H_1: aspoň jedna dvojice výběrů pochází z různých rozložení.

a) Neményiho metoda (Peter Neményi 1927 – 2002: Americký matematik maďarského původu)

- Všechny výběry mají týž rozsah p (třídění je vyvážené).
- Vypočteme $|T_l - T_k|$.
- V tabulkách najdeme kritickou hodnotu (pro dané p, r, α).
- Pokud $|T_l - T_k| \geq$ tabelovaná kritická hodnota, pak na hladině významnosti α zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

b) Obecná metoda mnohonásobného porovnávání

- Vypočteme $\left|\frac{T_l}{n_l} - \frac{T_k}{n_k}\right|$.
- Ve speciálních statistických tabulkách najdeme kritickou hodnotu $h_{KW}(\alpha)$. Při větších rozsazích výběrů je možno ji nahradit kvantilem $\chi^2_{1-\alpha}(r-1)$.

- Jestliže $\left|\frac{T_l}{n_l} - \frac{T_k}{n_k}\right| \geq \sqrt{\frac{1}{12}\left(\frac{1}{n_l} + \frac{1}{n_k}\right) n(n+1) h_{KW}(\alpha)}$, pak na hladině významnosti α zamítáme hypotézu,

- že l-tý a k-tý výběr pocházejí z téhož rozložení.

Příklad:
Čtyři laboranti provedli analytické stanovení procenta niklu v oceli. Každý hodnotil pět vzorků.
Laborant A: 4,15 4,26 4,10 4,30 4,25
Laborant B: 4,38 4,40 4,29 4,39 4,45
Laborant C: 4,23 4,16 4,20 4,24 4,27
Laborant D: 4,41 4,31 4,42 4,37 4,43
Na asymptotické hladině významnosti 0,05 testujte hypotézu, že všechny čtyři náhodné výběry pocházejí ze stejného rozložení. Pokud nulovou hypotézu zamítnete, zjistěte, které dvojice výběrů se liší.

Výpočet pomocí systému STATISTICA:
Vytvoříme nový datový soubor o dvou proměnných a 20 případech. Do proměnné nikl napíšeme změřené hodnoty, do proměnné laborant napíšeme 5x1 pro 1. laboranta atd. až 5x4 pro 4. laboranta.
Statistiky – Neparametrická statistika – Porovnání více nezávislých vzorků - OK – Seznam závislých proměnných nikl, Nezáv. (grupovací) proměnná laborant – OK – Summary: Kruskal-Wallis ANOVA & Median test. Ve dvou výstupních tabulkách se objeví výsledky K-W testu a mediánového testu.

Závislá: nikl	Kruskal-Wallisova ANOVA založ. na poř.; nikl (nikl v oceli) Nezávislá (grupovací) proměnná laborant Kruskal-Wallisův test: H (3, N= 20) =13,77714 p =,0032		
	Kód	Počet platných	Součet pořadí
1	1	5	29,00000
2	2	5	75,00000
3	3	5	27,00000
4	4	5	79,00000

Závislá: nikl	Mediánový test, celk. medián = 4,29500; nikl (nikl v oceli Nezávislá (grupovací) proměnná : laborant Chi-Kvadr. = 13,60000 sv = 3 p = ,0035				
	1	2	3	4	Celkem
<= Medián: pozorov.	4,00000	1,00000	5,00000	0,00000	10,00000
očekáv.	2,50000	2,50000	2,50000	2,50000	
poz.-oč.	1,50000	-1,50000	2,50000	-2,50000	
> Medián: pozorov.	1,00000	4,00000	0,00000	5,00000	10,00000
očekáv.	2,50000	2,50000	2,50000	2,50000	
poz.-oč.	-1,50000	1,50000	-2,50000	2,50000	
Celkem: oček.	5,00000	5,00000	5,00000	5,00000	20,00000

Oba testy zamítají hypotézu o shodě mediánů v daných čtyřech skupinách na asymptotické hladině významnosti 0,05.

Nyní provedeme mnohonásobné porovnávání, abychom zjistili, které dvojice laborantů se liší. Zvolíme Vícenás. porovnání průměrného pořadí pro vš. skupiny.

Závislá: nikl	Vícenásobné porovnání p hodnot (oboustr.); nikl (nikl v oceli Nezávislá (grupovací) proměnná: laborant Kruskal-Wallisův test: H (3, N= 20) =13,77714 p =,0032			
	1 R:5,8000	2 R:15,000	3 R:5,4000	4 R:15,800
1		0,083641	1,000000	0,045158
2	0,083641		0,061779	1,000000
3	1,000000	0,061779		0,032664
4	0,045158	1,000000	0,032664	

Tabulka obsahuje p-hodnoty pro porovnání dvojic skupin. Vidíme, že na hladině významnosti 0,05 se liší laboranti A, D a laboranti C, D.

Grafické znázornění výsledků