Kapitola 6. : Neparametrické testy o mediánech

Cíl kapitoly

Po prostudování této kapitoly budete umět
- provádět testy hypotéz o mediánu jednoho spojitého rozložení
- hodnotit shodu dvou nezávislých náhodných výběrů ze spojitých rozložení
- hodnotit shodu aspoň tří nezávislých náhodných výběrů ze spojitých rozložení a identifikovat dvojice významně odlišných náhodných výběrů

Časová zátěž
Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 5 hodin studia.

6.1. Motivace

Při používání t-testů či analýzy rozptylu by měl být splněn předpoklad normality dat. Pro výběry větších rozsahů ($n \geq 30$) nemá mírné porušení normality závažný dopad na výsledky. Někdy se však setkáváme s výběry malých rozsahů, které pocházejí z výrazně nenormálních rozložení. Pro práci s nimi byly vytvořeny tzv. neparametrické testy, které nevyžadují konkrétní typ rozložení (např. normální), stačí např. předpokládat, že distribuční funkce rozložení, z něhož náhodný výběr pochází, je spojitá.

Tyto neparametrické testy se rovněž používají v situacích, kdy zkoumaná data nemají intervalový či poměrový charakter, ale pouze ordinální charakter.

Ve srovnání s klasickými parametrickými testy jsou však 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.

6.2. Jednovýběrové testy

Jde o neparametrické obdoby jednovýběrového t-testu a párového t-testu.

6.2.1. Znaménkový test

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

b) Zavedeme statistiku S_Z^+, která udává počet těch rozdílů, 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.

Pro oboustrannou alternativu: $W = \langle 0, k_1 \rangle \cup \langle k_2, n \rangle$,

pro levostrannou alternativu: $W = \langle 0, k_1 \rangle$,

pro pravostrannou alternativu: $W = \langle k_2, n \rangle$.

(Nezáporná celá čísla k_1, k_2 pro oboustranný test i pro jednostranné testy lze najít ve statistických tabulkách.)

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

Pro velká n (prakticky $n > 20$) lze využít asymptotické normality statistiky S_Z^+. Testová statistika $U_0 = \frac{S_Z^+ - E\left(S_Z^+\right)}{\sqrt{D\left(S_Z^+\right)}} = \frac{S_Z^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$ má za platnosti H_0 asymptoticky rozložení $N(0,1)$.

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

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

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

Aproximace rozložením $N(0,1)$ se zlepší, když použijeme tzv. korekci na nespojitost. Testová statistika pak má tvar $U_0 = \frac{S_Z^+ - \frac{n}{2} \pm \frac{1}{2}}{\sqrt{\frac{n}{4}}}$, přičemž $\frac{1}{2}$ přičteme, když $S_Z^+ < \frac{n}{2}$ a odečteme v opačném případě.

6.2.2. Příklad

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.

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 7, nezamítáme H_0 na hladině významnosti 0,05. Dostáváme týž výsledek jako při použití asymptotického testu.

6.2.3. Jednovýběrový Wilcoxonův test

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: $x_{0,50} \neq c$ (resp. proti levostranné alternativě H_1: $x_{0,50} < c$ resp. proti pravostranné alternativě H_1: $x_{0,50} > c$).
a) Utvoříme rozdíly $Y_i = X_i - c$, i = 1, ..., n. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)
b) Absolutní hodnoty $|Y_i|$ uspořádáme vzestupně podle velikosti a spočteme pořadí R_i.
c) Zavedeme statistiku $S_W^+ = \sum_{Y_i > 0} R_i^+$, což je součet pořadí přes kladné hodnoty Y_i. Analogicky zavedeme statistiku $S_W^- = \sum_{Y_i < 0} R_i^-$, což je součet pořadí přes záporné hodnoty Y_i. Přitom platí, že součet $S_W^+ + S_W^- = n(n+1)/2$. Za platnosti H_0 statistika S_W^+ má střední hodnotu $E(S_W^+) = n(n+1)/4$ a rozptyl $D(S_W^+) = n(n+1)(2n+1)/24$.
d) Stanovíme testovou statistiku:
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.

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 má tvar: $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$. (Analogicky pro jednostranné alternativy.) H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Wilcoxonův test se hodí jen pro výběr ze symetrického rozložení. Není-li tento předpoklad splněn, lze použít např. znaménkový test.

6.2.4. Příklad

U 12 náhodně 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:

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 tučně:

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

$S_W^+ = 14$, $S_W^- = 64$, $n = 12$, $\alpha = 0,05$, tabelovaná kritická hodnota = 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 %.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor se dvěma proměnnou a dvanácti případy. První proměnnou nazveme PROCENTA, druhou KONSTANTA. Do proměnné PROCENTA napíšeme zjištěná procenta populace starší 60 let:a do proměnné KONSTANTA vyplníme čísly 12 (do Dlouhého jména proměnné KONSTANTA napíšeme =12).
Statistika – Neparametrická statistika – Porovnání dvou závislých vzorků (proměnné) – OK. Proměnné – 1. seznam proměnných – PROCENTA, 2. seznam proměnných – KONSTANTA, OK, Wilcoxonův párový test. Dostaneme tabulku

	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

V této tabulce je symbolem T označena testová statistika $\min(S_W^+, S_W^-)$, symbolem Z realizace asymptotické testové statistiky U_0. Uvedená p-hodnota je vypočítána pro realizaci asymptotické testové statistiky U_0. Protože $p \leq 0,05$, hypotézu H_0: $x_{0,50} = 12$ zamítáme na asymptotické hladině významnosti 0,05. Pokud bychom chtěli provést přesný test a nikoliv pouze asymptotický, vyhledali bychom ve statistických tabulkách kritickou hodnotu jednovýběrového Wilcoxonova testu pro $n = 12$, $\alpha = 0,05$ (viz výše). Protože tato hodnota je 13, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

6.2.5. Párový Wilcoxonův test

Nechť $(X_1, Y_1), ..., (X_n\ Y_n)$ je náhodný výběr ze spojitého dvourozměrného rozložení. Testujeme H_0: $x_{0,50} - y_{,50} = c$ 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: $z_{0,50} \neq c$.

6.2.6. 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:

Jedná se o párový test. 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.
$S_W^+ = 16{,}5$, $S_W^- = 74{,}5$, n = 15, $\alpha = 0{,}05$, 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, tedy s rizikem omylu nejvýše 5% jsme prokázali, že medián cenových rozdílů se liší od 3 Kč.

Řešení pomocí systému STATISTICA:
Otevřeme nový datový soubor se čtyřmi proměnnými a 15 případy. První proměnnou nazveme CENA A, druhou CENA B, třetí ROZDÍL a čtvrtou KONSTANTA. Do proměnných CEANA A a CENA B zapíšeme ceny zboží A a B, do Dlouhého jména proměnné ROZDÍL napíšeme = v1-v2 a proměnnou KONSTANTA vyplníme samými trojkami. Nyní provedeme párový Wilcoxonův test:
Statistika – Neparametrická statistika – Porovnání dvou závislých vzorků (proměnné) – OK.

Proměnné – 1. seznam proměnných – ROZDÍL, 2. seznam proměnných – KONSTANTA, OK, Wilcoxonův párový test. Dostaneme tabulku

	Wilcoxonův párový test (priklad734) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
rozdíl & konst	15	16,50000	2,026684	0,042696

Podobně jako v příkladu 6.2.4. je symbolem T označena testová statistika $\min(S_W^+, S_W^-)$, symbolem Z realizace asymptotické testové statistiky U_0. Uvedená p-hodnota je vypočítána pro realizaci asymptotické testové statistiky U_0. Protože $p \leq 0{,}05$, hypotézu H_0: $z_{0,50} = 3$ zamítáme na asymptotické hladině významnosti 0,05. Pokud bychom chtěli provést přesný test a nikoliv pouze asymptotický, vyhledali bychom ve statistických tabulkách kritickou hodnotu jednovýběrového Wilcoxonova testu pro n = 13, $\alpha = 0{,}05$ (viz výše). Protože tato hodnota je 17 a testová statistika 16,5, nulovou hypotézu zamítáme na hladině významnosti 0,05.

6.3. Dvouvýběrové pořadové testy

Jedná se o neparametrickou obdobu dvouvýběrového t-testu.

6.3.1. Dvouvýběrový Wilcoxonův test

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. Testujeme hypotézu, že distribuční funkce těchto rozložení jsou shodné neboli mediány jsou shodné proti alternativě, že jsou rozdílné.

Všech n + m hodnot $X_1, ..., X_n$ a $Y_1, ..., Y_m$ uspořádáme vzestupně podle velikosti. Zjistíme součet pořadí hodnot $X_1, ..., X_n$ a označíme ho T_1. Součet pořadí hodnot $Y_1, ..., Y_m$ označíme T_2. Vypočteme statistiky $U_1 = mn + n(n+1)/2 - T_1$, $U_2 = mn + m(m+1)/2 - T_2$. Přitom platí $U_1 + U_2 = mn$. 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 se používá označení: $n = \min\{m, n\}$ a $m = \max\{m, n\}$.

Pro velká n, m (prakticky n, m > 30) lze využít asymptotické normality statistiky U_1. V případě platnosti H_0 má statistika $U_0 = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$ asymptoticky rozložení N(0,1). Kritický obor pro oboustrannou alternativu má tvar: $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$. (Analogicky pro jednostranné alternativy.) H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Dvouvýběrový Wilcoxonův test se používá v situacích, kdy distribuční funkce rozložení, z nichž dané dva nezávislé náhodné výběry pocházejí, se mohou lišit pouze posunutím.

6.3.2. 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.

x: nový způsob	51	52	49	55		
y: starý způsob	45	54	48	44	53	50

Řešení

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	

$T_1 = 4 + 6 + 7 + 10 = 27$, $T_2 = 1 + 2 + 3 + 5 + 8 + 9 = 28$

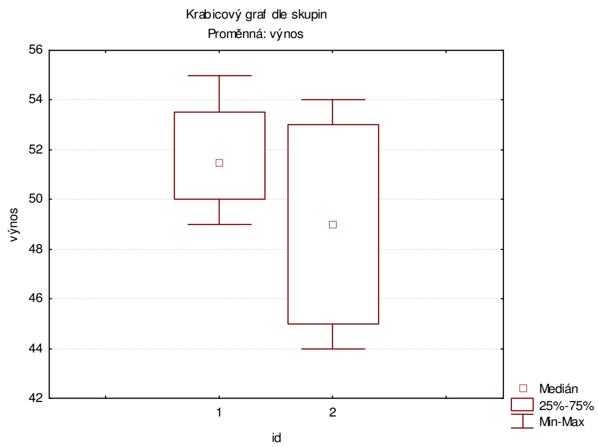
$U_1 = 4.6 + 4.5/2 - 27 = 7$, $U_2 = 4.6 + 6.7/2 - 28 = 17$

Kritická hodnota pro $\alpha = 0{,}05$, min(4,6) = 4, max(4,6) = 6 je 2. Protože min(7,17) > 2, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že nový způsob hnojení má na hektarové výnosy pšenice stejný vliv jako starý způsob.

Řešení pomocí systému STATISTICA:
Otevřeme nový datový soubor se dvěma proměnnými VÝNOS a ID a 10 případy. Do proměnné VÝNOS zapíšeme hektarové výnosy pšenice a do proměnné ID, která slouží k rozlišení nového a starého způsobu hnojení, napíšeme 4 krát jedničku a 6 krát dvojku. Nyní provedeme dvouvýběrový Wilcoxonův test, který je ve STATISTICE uveden pod názvem Mannův – Whitneyův test:
Statistika – Neparametrická statistika – Porovnání dvou nezávislých vzorků (skupiny) – OK. Proměnné – Seznam závislých proměnných – VÝNOS, Nezáv. (grupov.) proměnné - ID – OK, Mann-Whitneyův U test. Dostaneme tabulku

	Mann-Whitneyův U test (Hnojeni.sta) Dle proměn. id Označené testy jsou významné na hladině p <,05000									
Proměnná	Sčt poř. skup. 1	Sčt poř. skup. 2	U	Z	p-hodn.	Z upravené	p-hodn.	N platn. skup. 1	N platn. skup. 2	2*1str. přesné p
x	27,00000	28,00000	7,000000	0,959403	0,337356	0,959403	0,337356	4	6	0,352381

Zde je symbolem U označena testová statistika $\min(U_1,U_2)$. V našem případě U = 7, odpovídající p-hodnotu najdeme v posledním sloupci pod označením 2*1 str. přesné p. Protože 0,352381 > 0,05, nezamítáme na hladině významnosti 0,05 nulovou hypotézu. Výpočet ještě doplníme krabicovým diagramem. Na záložce Zákl. výsledky vybereme Krabicový graf dle skupin, OK, proměnná VÝNOS, OK. Dostaneme graf



Je zřejmé, že medián hektarových výnosů při starém způsobu hnojení je menší než při novém způsobu a také vidíme, že variabilita hektarových výnosů při starém způsobu hnojení je větší než při novém způsobu.

6.3.3. 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, že distribuční funkce těchto rozložení jsou shodné, tj., že všech n+m veličin pochází z téhož rozložení proti alternativě, že distribuční funkce jsou rozdílné. Nechť je empirická distribuční funkce 1. výběru a $F_2(y)$ je empirická distribuční funkce 2. výběru. Jako testová statistika slouží $D = \max\limits_{-\infty<x<\infty} |F_1(x) - F_2(x)|$. 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 $\sqrt{\frac{n+m}{2nm}\ln\frac{2}{\alpha}}$.

6.3.4. Příklad

Na data z příkladu 6.3.2. aplikujte dvouvýběrový K-S test.

Výpočet pomocí systému STATISTICA:
Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných X, Nezáv. (grupov.) proměnná ID – OK – Kolmogorov-Smirnovův 2-výběrový test.

	Kolmogorov-Smirnovův test (Hnojeni.sta) Dle proměn. id Označené testy jsou významné na hladině p <,05000								
Proměnná	Max záp rozdíl	Max klad rozdíl	p-hodn.	Průměr skup. 1	Průměr skup. 2	Sm.odch. skup. 1	Sm.odch. skup. 2	N platn. skup. 1	N platn. skup. 2
x	-0,083333	0,500000	p > .10	51,75000	49,00000	2,500000	4,098780	4	6

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.

6.4. Kruskalův – Wallisův test a mediánový test (neparametrické obdoby analýzy rozptylu jednoduchého třídění)

6.4.1. Formulace problému

Nechť je dáno $r \geq 3$ nezávislých náhodných výběrů o rozsazích $n_1, \dots, n_r$. Předpokládáme, že tyto výběry pocházejí ze spojitých rozložení. Označme $n = n_1 + \dots + n_r$. Chceme testovat hypotézu, že všechny tyto výběry pocházejí z téhož rozložení.

6.4.2. Kruskalův – Wallisův test

Všech n hodnot seřadíme do rostoucí posloupnosti a určíme pořadí každé hodnoty. Označme T_j součet pořadí těch hodnot, které patří do j-tého výběru, j = 1, ..., r (kontrola: musí platit $T_1 + ... + T_r = n(n+1)/2$). 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 $\chi^2(r-1)$. H_0 tedy zamítneme na asymptotické hladině významnosti α, když $Q \geq \chi_{1-\alpha}^2(r-1)$.

6.4.3. Mediánový test

Testová statistika má tvar $Q_M = 4\sum_{j=1}^{r}\frac{P_j^2}{n_j} - n$, kde P_j je počet hodnot v j-tém výběru, které jsou větší nebo rovny mediánu vypočtenému ze všech n hodnot. Platí-li H_0, má statistika Q_M asymptoticky rozložení $\chi^2(r-1)$. H_0 tedy zamítneme na asymptotické hladině významnosti α, když $Q_M \geq \chi_{1-\alpha}^2(r-1)$.

6.4.4. Metody mnohonásobného porovnávání

Zamítneme-li H_0, zajímá nás, které dvojice náhodných výběrů se liší na zvolené hladině významnosti.

a) Neményiho metoda
Používá se v případě, že všechny výběry mají týž rozsah p. Je-li $|T_l - T_k| \geq$ tabelovaná kritická hodnota (pro dané p, r, α), pak na hladině významnosti α zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

b) Obecná metoda mnohonásobného porovnávání
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. Kritickou hodnotu $h_{KW}(\alpha)$ najdeme ve speciálních statistických tabulkách. Při větších rozsazích výběrů je možno ji nahradit kvantilem $\chi_{1-\alpha}^2(r-1)$.

7.4.5. Příklad

U přijímacích zkoušek na vysokou školu sledujeme počet bodů z matematiky. Chceme posoudit, zda výsledky jsou závislé na typu absolvované střední školy. Náhodně vybereme osm písemných zkoušek studentů každého z uvažovaných tří typů škol:

číslo písemky	gymnázium	číslo písemky	SEŠ	číslo písemky	SPŠ
1	78	9	30	17	93
2	95	10	84	18	74
3	84	11	65	19	58
4	78	12	41	20	85
5	85	13	67	21	60
6	96	14	52	22	72
7	90	15	92	23	67
8	83	16	70	24	59

Na hladině významnosti 0,05 testujte hypotézu, že výsledky studentů z gymnázií, SEŠ a SPŠ se neliší. Zamítnete-li nulovou hypotézu, vyšetřete, které dvojice typů škol se od sebe liší na hladině významnosti 0,05.

Řešení

Kruskalův – Wallisův test

usp. hodnoty	pořadí 1. výběru	pořadí 2. výběru	pořadí 3. výběru
30	1		
41	2		
52	3		
58		4	
59		5	
60		6	
65	7		
67	8,5		
67		8,5	
70	10		
72		11	
74		12	
78			13,5
78			13,5
83			15
84			16,5
84	16,5		
85			18,5
85		18,5	
90			20
92	21		
93		22	
95			23
96			24

Součet pořadí pro jednotlivé výběry: $T_1 = 69$, $T_2 = 87$, $T_3 = 144$,

Realizace testové statistiky: $Q = \frac{12}{24 \cdot 25}\left(\frac{69^2}{8} + \frac{87^2}{8} + \frac{144^2}{8}\right) - 3 \cdot 25 = 7{,}665$,

Kritický obor. $W = \langle \chi^2_{0,95}(2), \infty) = \langle 5{,}991, \infty)$.

Protože $Q \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05. Rozdíly mezi počty bodů u přijímací zkoušky z matematiky u studentů ze sledovaných tří typů středních škol se prokázaly s rizikem omylu nejvýše 0,05.

Mediánový test

Medián všech 24 hodnot je 76. V 1. výběru leží nad mediánem 8 hodnot, ve 2. výběru 2 hodnoty, ve 3. výběru 2 hodnoty.

Realizace testové statistiky: $Q_M = 4\left[\frac{1}{8}\left(8^2 + 2^2 + 2^2\right)\right] - 24 = 12$,

Kritický obor. $W = \langle \chi^2_{0,95}(2), \infty) = \langle 5{,}991, \infty)$.

Nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor se dvěma proměnnými X a ID a s 24 případy. Do proměnné X zapíšeme počty bodů, do proměnné ID, která slouží jako identifikátor typu školy, napíšeme 8

krát jedničku, 8 krát dvojku a 8 krát trojku. Nyní provedeme Kruskalův – Wallisův a mediánový test.
Statistika – Neparametrická statistika – Porovnání více nezávislých vzorků (skupiny) – OK. Proměnné – Závisle proměnné – X, Nezáv. (grupov.) proměnná - ID – OK, Shrnutí: Kruskal-Wallis ANOVA a mediánový test, Výpočet. Pro K-W test dostaneme tabulku

Závislá: X	Kruskal-Wallisova ANOVA založ. na poř.; X (body_u_zkousky.sta) Nezávislá (grupovací) proměnná : ID Kruskal-Wallisův test: H (2, N= 24) =7,678354 p =,0215		
	Kód	Počet platných	Součet pořadí
gymnazium	1	8	144,0000
SEŠ	2	8	69,0000
SPŠ	3	8	87,0000

Testová statistika se realizuje hodnotou 7,678, počet stupňů volnosti je 2, odpovídající p-hodnota = 0,0215, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o shodě mediánů.
Pro mediánový test máme tabulku

Závislá: X	Mediánový test, celk. medián = 76,0000; X (body_u_zkousky.sta) Nezávislá (grupovací) proměnná : ID Chi-Kvadr. = 12,00000 sv = 2 p = ,0025			
	gymnazium	SEŠ	SPŠ	Celkem
<= Medián: pozorov.	0,00000	6,00000	6,00000	12,00000
očekáv.	4,00000	4,00000	4,00000	
poz.-oč.	-4,00000	2,00000	2,00000	
> Medián: pozorov.	8,00000	2,00000	2,00000	12,00000
očekáv.	4,00000	4,00000	4,00000	
poz.-oč.	4,00000	-2,00000	-2,00000	
Celkem: oček.	8,00000	8,00000	8,00000	24,00000

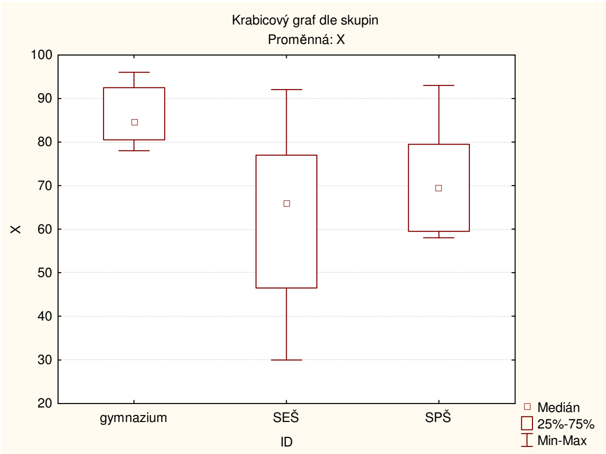
Realizace testové statistiky = 12, počet stupňů volnosti = 2, odpovídající p-hodnota = 0,0025, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o shodě mediánů.

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

Závislá: X	Vícenásobné porovnání p hodnot (oboustr.); X (body_u_zkousky.sta Nezávislá (grupovací) proměnná : ID Kruskal-Wallisův test: H (2, N= 24) =7,678354 p =,0215		
	gymnazium R:18,000	SEŠ R:8,6250	SPŠ R:10,875
gymnazium		0,024030	0,131634
SEŠ	0,024030		1,000000
SPŠ	0,131634	1,000000	

Tabulka obsahuje p-hodnoty pro porovnání dvojic skupin. Vidíme, že na hladině významnosti 0,05 se liší gymnázium a SEŠ.

Výpočet ještě doplníme krabicovým diagramem. Na záložce Zákl. výsledky vybereme Krabicový graf , proměnná X, OK, Typ krabicového grafu Medián/Kvartily/Rozpětí, OK. Dostaneme graf



Vidíme, že mediány se liší velice výrazně, zvláště pro gymnázia a SEŠ. Variabilita počtu bodů je nejmenší pro gymnázia, největší pro SEŠ.

Shrnutí

V některých situacích se setkáváme s náhodnými výběry malých rozsahů, které pocházejí z výrazně nenormálních rozložení. V takových případech nelze použít klasické testy založené na předpokladu normality, které byly popsány ve 4., 5. a 6. kapitole. Místo nich používáme neparametrické testy, které nepotřebují splnění předpokladu no□m□lit³, stačí např. předpokládat spojitost distribuční funkce rozložení, z něhož daný náhodný výběr pochází.

Pro testování hypotéz□ o□mediánu používáme *jednovýběrový* či *párový Wilcoxonův test,* což je neparametri□ká obdoba jednovýběrového či párového t.testu.

Máme-li testovat hypotézu o shodě mediánů dvou rozložení, která se mohou lišit jen posunutím (tj. testujeme hypotézu o shodě těchto dvou rozložení), aplikujeme *dvouvýběrový Wilcoxonův test* – neparametrickou obdobu dvouvýběrového t-testu.

Jako neparametrická obdoba analýzy rozptylu jednoduchého třídění slouží *Kruskalův – Wallisův test* nebo *mediánový test*. Při zamítnutí nulové hypotézy identifikujeme dvojice odlišných výběrů pomocí *metod mnohonásobného porovnávání*, a to buď *obecnou metodu mnohonásobného porovnávání* nebo *Neméniyho metodu*.

Při provádění neparametrických testů potřebujeme speciální tabulky kritických hodnot. Jsou obsaženy v příloze A tohoto učebního textu.

Všechny uvedené testy jsou implementovány v systému STATISTICA.

Kontrolní otázky

1. V jakých situacích používáme neparametrické testy?
2. Jaká je nevýhoda neparametrických testů oproti testům parametrickým?
3. Jak vypočítáme pořadí čísla v dané posloupnosti čísel?
4. Popište rozdíl mezi jednovýběrovým a párovým Wilcoxonovým testem.
5. Jaké podmínky musí být splněny pro dvouvýběrový Wilcoxonův test?
6. K čemu slouží Kruskalův-Wallisův test?
7. Jak provedeme mediánový test?
8. Které metody mnohonásobného porovnávání znáte?

Autokorekční test

1. Máme za úkol zjistit, zda tři nezávislé výběry pocházejí z téhož rozložení. Přitom všechny mají malý rozsah (menší než 30) a vykazují odchylky od normálního rozložení. Jaký test použijeme?
a) Analýzu rozptylu jednoduchého třídění,
b) mediánový test,
c) Kruskalův-Wallisův test.

2. Testujeme hypotézu, že dva nezávislé náhodné výběry pocházejí z téhož rozložení. Oba výběry mají malý rozsah (menší než 30) a diagnostické grafy i testy normality poukazují na závažnější odchylky od normálního rozložení. Jaký test použijeme?
a) Párový Wilcoxonův test,
b) dvouvýběrový t-test,
c) dvouvýběrový Wilcoxonův test.

3. Pomocí K-W testu testujeme na asymptotické hladině významnosti 0,05 hypotézu, že pět nezávislých náhodných výběrů o rozsazích 4, 7, 5, 4, 5 pochází z téhož rozložení. Kritický obor má tvar:
a) $W = \langle 9{,}488; \infty)$,
b) $W = \langle 0{,}711; \infty)$,
c) $W = \langle 0; 9{,}488)$.

4. Máme dvourozměrný náhodný výběr z dvourozměrného rozložení, které se výrazně liší od normálního rozložení. K testování hypotézy, že mediány obou složek tohoto rozložení jsou stejné, použijeme
a) jednovýběrový t-test,
b) dvouvýběrový Wilcoxonův test,
c) párový Wilcoxonův test.

Správné odpovědi: 1b),c)   2c)   3a)   4c)

Příklady

1. 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.
Výsledek:
Použijeme jednovýběrový Wilcoxonův test. Testová statistika se realizuje hodnotou 12, tabelovaná kritická hodnota pro $\alpha = 0{,}05$ a $a = 9$ je 5. Protože $12 > 5$, H_0 nezamítáme na hladině významnosti 0,05.

2. 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 Wilcoxonova testu, zda je oprávněný předpoklad, že obě technologie poskytují stejné procento účinné látky.
Výsledek:

Testová statistika se realizuje hodnotou 12, tabelovaná kritická hodnota pro $\alpha = 0{,}05$, min(5,8) = 5, max(5,8) = 8 je 6. Protože min(28,12) > 6, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že obě technologie poskytují stejné procento účinné látky.

3. Výrobce koláčů v prášku má 4 nové recepty a chce zjistit, zda se jejich kvalita liší. Upekl proto 5 koláčů z každého druhu a dal je porotě k ohodnocení.

recept	počet bodů				
A	72	88	70	87	71
B	85	89	86	82	88
C	94	94	88	87	89
D	91	93	92	95	94

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že recepty se neliší.
Výsledek:
Použijeme Kruskalův – Wallisův test. Všech 20 hodnot uspořádáme vzestupně podle velikosti a stanovíme součet pořadí pro recepty A, B, C, D: $T_1 = 23{,}5$, $T_2 = 37{,}5$, $T_3 = 66$, $T_4 = 83$. Testová statistika:

$$Q = \frac{12}{20 \cdot 21}\left(\frac{23{,}5^2}{5} + \frac{37{,}5^2}{5} + \frac{66^2}{5} + \frac{83^2}{5}\right) - 3 \cdot 21 = 12{,}45,$$ $\chi_{0{,}95}{}^2(3) = 7{,}81$. Protože $Q \geq 7{,}81$, H_0 zamítáme na asymptotické hladině významnosti 0,05. Neményiho metoda prokázala, že na hladině významnosti 0,05 se liší recepty A a D.

4. U osmi osob byl změřen systolický krevní tlak před pokusem a po něm.

č. osoby	1	2	3	4	5	6	7	8
tlak před	130	185	162	136	147	181	128	139
tlak po	139	190	175	135	155	175	158	149

Na hladině významnosti 0,05 testujte hypotézu, že pokus neovlivní systolický krevní tlak
Výsledek:
Párový Wilcoxonův test poskytl p-hodnotu 0,04995, tedy H_0 zamítáme na hladině významnosti 0,05.

5. Majitel obchodu chtěl zjistit, zda velikost nákupů (v dolarech) placených kreditními kartami Master/EuroCard a Visa jsou přibližně stejné. Náhodně vybral 7 nákupů placených Master/EuroCard a 9 placených Visou:

Master/EuroCard	42	77	46	73	78	33	37		
Visa	39	10	119	68	76	126	53	79	102

Lze na hladině významnosti 0,05 tvrdit, že mediány nákupů placených těmito dvěma typy karet se shodují?
Výsledek:
Dvouvýběrový Wilcoxonův test poskytl p-hodnotu 0,2523, H_0 tedy nezamítáme na hladině významnosti 0,05.

6. Z produkce tří podniků vyrábějících televizory bylo vylosováno 10, 8 a 12 kusů. Byly získány následující výsledky zjišťování citlivosti těchto televizorů v mikrovoltech:

podnik	citlivost										
1. podnik	420	560	600	490	550	570	340	480	510	460		
2. podnik	400	420	580	470	470	500	520	530				
3. podnik	450	700	630	590	420	590	610	540	740	690	540	670

Ověřte na hladině významnosti 0,05 hypot1zu o shodě úrovně citlivosti televizorů v jednotlivých podnicích.
Výsledek:
K-W test poskytl testovou statistiku 8,304653, počet stupňů volnosti = 2, odpovídající p-hodnota = 0,0157, tedy H_0 zamítáme na asymptotické hladině významnosti 0,05. Liší se výrobky podniků 2 a 3.