

18. Neparametrické hypotézy

V této části se budeme zabývat speciální částí teorie statistických hypotéz tzv. neparametrickými hypotézami nebo jinak řečeno neparametrickými statistickými testy. Neparametrické se nazývají proto, že při konstrukci těchto testů nejsou využity znalosti parametrů rozdělení náhodných veličin, ale jiné způsoby práce s daty.

18.1 Znaménkový test

Jde o jeden z nejjednodušších testů. Necht' $X_1, \dots, X_n$ je výběr ze **spojitého** rozdělení s *mediánem* $\tilde{x}$. Testujeme hypotézu $H_0 : \tilde{x} = x_0$, kde x_0 je dané číslo. Test má jak oboustrannou alternativu, tak i jednostrannou.

$H_1 : \tilde{x} \neq x_0$. Jde tedy o oboustrannou alternativu. Nejdříve vytvoříme rozdíly $X_1 - x_0, \dots, X_n - x_0$. Pokud je některý z těchto rozdílů nulový vynecháme ho. Tím získáme množinu čísel. Z této množiny nás bude zajímat jen **počet** rozdílů s kladným znaménkem, tento počet označme **B**. Pokud by platila hypotéza H_0 , pak by náhodná veličina **B** byla typu **Bi(n ; 0,5)** (předpokládáme, hodnoty jsou vesměs různé od x_0). Hypotézu H_0 zamítneme, jestliže bude hodnota **B** velmi malá nebo naopak bude téměř rovna n . Kritické hodnoty tohoto testu jsou tabelovány v tabulkách, ale můžete si je zjistit sami pomocí distribuční funkce binomického rozdělení **Bi(n;0,5)**.

Platí tedy

$$P(B \leq k_1) \leq \frac{\alpha}{2}, \quad P(B \geq k_2) \leq \frac{\alpha}{2}. \quad (18.1)$$

Pro hodnoty n malé ($n < 20$) se v tabulkách naleznou čísla k_1 a k_2 . Hypotézu zamítneme, jestliže $B \leq k_1$ nebo $B \geq k_2$.

V případě velkých hodnot ($n > 19$) se provádí normalizace náhodné veličiny **B**. Náhodná veličina

$$U = \frac{2 \cdot B - n}{\sqrt{n}} \quad (18.2)$$

se podle centrální limitní věty asymptoticky blíží k **N(0,1)**. Hypotézu H_0 tedy zamítneme v tomto případě, když $|U| \geq u\left(\frac{\alpha}{2}\right)$.

Příklad 18.1

Necht' jsou naměřeny hodnoty 15,6 ; 14,8; 18,2; 14,7; 16,3; 20,7; 19,4; 18,2; 17,4; 16,4; 14,9; 15,1; 16,5. Pomocí znaménkového testu rozhodněte, zda 18 může být mediánem dat.

Řešení:

Od hodnot x_i odečteme očekávanou hodnotu mediánu a dostaneme:

$$\begin{array}{cccccccccccccccc} 15,6 & 14,8 & 18,2 & 14,7 & 16,3 & 20,7 & 19,4 & 18,2 & 17,4 & 16,4 & 14,9 & 15,1 & 16,5 \\ -2,4 & -3,2 & 0,2 & -3,3 & -1,7 & 2,7 & 1,4 & 0,2 & -0,6 & -1,6 & -3,1 & -2,9 & -1,5 \end{array}$$

Počet čísel kladných je roven 4. Pro hodnotu $n = 13$ jsou kritické hodnoty na hladině významnosti 0,05 rovny 2 a 9. Protože číslo 4 leží mezi těmito hodnotami, nemůžeme vyloučit možnost, že je mediánem našich dat.

Příklad 18.2

Při měření hmotnosti výlovku ryb (v tunách) byly zjištěny následující hodnoty:

1	2	3	4	5	6	7	8	9	10	11	12	13
2,64	1,537	1,020	1,162	1,263	3,215	0,31	0,14	1,774	1,823	0,899	2,743	6,397
14	15	16	17	18	19	20	21	22	23	24	25	26
2,021	0,623	0,107	0,262	2,704	2,781	1,004	1,344	6,954	1,479	0,123	7,19	3,62

Máme ověřit hypotézu , že medián výlovku ryb je roven 2 t.

Řešení:

Opět zjistíme hodnotu počtu kladných rozdílů mezi x_i a hodnotou 2 . Těchto čísel je 10 . Hodnota $B = 10$. Ověření provedeme pomocí vztahu (18.2) , tento vztah je aproximativně

rozdělení $N(0,1)$. Vypočteme hodnotu U , $|U| = \left| \frac{2 \cdot 10 - 26}{\sqrt{26}} \right| = |-1,1767| = 1,1767$, porovnáme

– li tuto hodnotu s kritickou hodnotou $u_{0,975} = 1,96$, zjišťujeme , že hodnota U neleží v kritickém oboru. Nemůžeme tedy vyvrátit možnost , že číslo 2 je mediánem výlovku ryb. Pokud bychom prováděli výpočet klasickou metodou , měli bychom pro $n = 26$ k dispozici kritické hodnoty 7 a 18 , a ani pomocí této metody nemůžeme vyvrátit možnost , že 2 je mediánem hmotnosti výlovku.

18.2 Jednovýběrový Wilcoxonův test

Tento test je opět založen na pořadových statistikách. Předpokládáme , že daná jednorozměrná data byla vybrána z populace popsané **spojitou náhodnou** veličinou . Nulová hypotéza je postavená tak , že zjišťujeme , zda distribuční funkce $F(x)$ je symetrická kolem počátku (resp. kolem obecného bodu a) tedy zda platí :

$$\forall x \in (-\infty, \infty) \quad F(x) = 1 - F(-x),$$

oproti ní stojí alternativní hypotéza , která je negací tohoto výroku. Z požadavků na rozdělení vyplývá , že mediánem rozdělení je **0** (resp. bod **a**) , pokud existuje střední hodnota rozdělení je také rovna **0** (resp. **a**) .

Nechť $X_1, \dots, X_n$ je náhodný výběr z výše uvedeného rozdělení. Položíme $Y_i = X_i - 0$ (resp. $X_i - a$) . Náhodné veličiny uspořádáme vzestupně podle jejich absolutní hodnoty , tedy $|Y|_{(1)} \leq |Y|_{(2)} \leq \dots \leq |Y|_{(n)}$.

Symbolem R_i^+ označíme pořadí veličiny $|Y_i|$. Zavedme dále označení

$$R^+ = \sum_{Y_i \geq 0} R_i^+ \quad , \quad R^- = \sum_{Y_i < 0} R_i^+ \quad (18.3)$$

Podle zavedení těchto hodnot jistě platí $R^+ + R^- = \frac{n(n+1)}{2}$. Hypotézu H_0 na hladině α zamítáme, jestliže číslo $\min(R^+, R^-) \leq w_n(\alpha)$, kde $w_n(\alpha)$ je tabelovaná kritická hodnota. Tuto hodnotu získáte v tabulce kritických hodnot jednovýběrového Wilcoxonova testu.

Věta 18.3

Platí – li hypotéza H_0 , potom

$$E(R^+) = \frac{n(n+1)}{4}, \quad VAR(R^+) = \frac{n(n+1)(2n+1)}{24} \quad (18.4).$$

Důkaz:

Viz [1].

Dá se dokázat, že náhodná veličina R^+ má asymptoticky rozdělení normální. Můžeme tedy test hypotézy H_0 založit na náhodné veličině

$$U = \frac{R^+ - \frac{n(n+1)}{4}}{\frac{n(n+1)(2n+1)}{24}} \quad (18.5)$$

Hypotézu H_0 potom zamítneme, jestliže $|U| \geq u\left(\frac{\alpha}{2}\right)$, na hladině, která se s rostoucím n přibližuje hodnotě α .

Příklad 18.4

V průběhu deseti za sebou jdoucích dnů si pacient měřil 10x tep. Můžeme na základě těchto měření prohlásit, že medián naměřených hodnot je roven 75 tepům?

měř. č.1	měř. č.2	měř. č.3	měř. č.4	měř. č.5	měř. č.6	měř. č.7	měř. č.8	měř. č.9	měř. č.10
76 tepů	76 tepů	74 tepů	77 tepů	79 tepů	81 tepů	83 tepů	67 tepů	65 tepů	90 tepů

Řešení:

Pokud by medián hodnoty tepů byl 75, pak dostaneme hodnoty Y_i takovéto

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
1	1	-1	2	4	6	8	-8	-10	15

V této tabulce máme dokonce hodnoty $|Y_i|$ seřazené podle velikosti. Vypočteme dále hodnoty $R^+ = 35$, $R^- = 20$. Hodnota $R = 20$, tuto hodnotu porovnáme s kritickou hodnotou v tabulkách, $w_{10}(0,05) = 8$, protože hodnota R je větší než 8, nemůžeme na dané hladině hypotézu H_0 zamítnout.

Pokud bychom řešili tuto úlohu pomocí normální aproximace, získali bychom hodnotu $U = \frac{35 - 27,5}{\sqrt{96,25}} = 0,764471$, tato hodnota je menší než 1,96, tedy i touto metodou nemůžeme hypotézu H_0 zamítnout.

V případě jednovýběrového Wilcoxonova testu je důležitým předpokladem i symetrie hustoty f kolem mediánu. K zamítnutí hypotézy H_0 může tedy dojít také tehdy, když je sice medián roven nule (resp. a), ale hustota je výrazně nesymetrická.

Můžeme tedy tento test použít i na jednoduché testování, zda vybraná data mohou pocházet z rozdělení normálního!

18.3 Dvouvýběrový Wilcoxonův test

Nechť je $X_1, \dots, X_m$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F_1 , nechť dále je $Y_1, \dots, Y_n$ náhodný výběr ze spojitého rozdělení s distribuční funkcí F_2 , který je na výběru $X_1, \dots, X_m$ nezávislý. Hypotézu H_0 sestavujeme tak, že $F_1 = F_2$, alternativní hypotéza je oboustranná tedy $F_1 \neq F_2$.

Při konstrukci testu nejdříve vytvoříme z výše uvedených výběrů **sdružený výběr** $X_1, \dots, X_m, Y_1, \dots, Y_n$. Hodnoty v tomto výběru uspořádáme vzestupně podle velikosti. Označíme S_1 součet všech pořadí hodnot $X_1, \dots, X_m$ a S_2 pořadí hodnot $Y_1, \dots, Y_n$ ve sdruženém výběru. Zřejmě podobně jako v jednovýběrovém Wilcoxonově testu platí

$$S_1 + S_2 = \frac{(n+m) \cdot (n+m+1)}{2}.$$
 V případě náhodných veličin S_1 a S_2 jde o tzv. lineární pořadové statistiky. Jejich parametry jsou uvedeny v následující větě.

Věta 18.5

Platí – li hypotéza H_0 , potom

$$\begin{aligned} E(S_1) &= \frac{m \cdot (m+n+1)}{2}, & VAR(S_1) &= \frac{m \cdot n \cdot (m+n+1)}{12}, \\ E(S_2) &= \frac{n \cdot (m+n+1)}{2}, & VAR(S_2) &= \frac{m \cdot n \cdot (m+n+1)}{12} \end{aligned} \quad (18.6)$$

Důkaz:

Uveden v [1].

Vzhledem k vazbě mezi náhodnými veličinami S_1 a S_2 , stačí vyšetřovat jen jednu z nich, zpravidla je to S_1 . Stále častěji se místo této veličiny vyšetřuje

$$U_1 = m \cdot n + \frac{m \cdot (m+1)}{2} - S_1,$$

Test založený na náhodné veličině U_1 se nazývá **Mann – Whitneyův** test. Zavedeme – li ještě označení

$$U_2 = m \cdot n + \frac{m \cdot (m+1)}{2} - S_2,$$

platí $U_1 + U_2 = m \cdot n$. Náhodná veličina U_1 udává počet případů, kdy $X_i < Y_j$, náhodná veličina U_2 udává počet případů, kdy $X_i > Y_j$. Opět podobně jako v případě jednovýběrového testu stanovíme $S = \min(U_1, U_2)$. Pokud hodnota S je menší nebo rovna kritické hodnotě uvedené v tabulce kritických hodnot dvouvýběrového Wilcoxonova testu, zamítá se na dané hladině α hypotéza H_0 . Označení výběrů přitom volíme tak, aby $m \geq n$. Na základě výsledků věty 18.5 zřejmě platí

$$E(U_1) = \frac{m \cdot n}{2}, \quad VAR(U_1) = \frac{m \cdot n \cdot (m+n+1)}{12} \quad (18.7)$$

,protože $U_2 = m \cdot n - U_1$. Je $E(U_1) = E(U_2)$, $VAR(U_1) = VAR(U_2)$. Dá se dokázat, že veličiny S_1, U_1 asymptoticky normální rozdělení, tedy

$$U = \frac{U_1 - E(U_1)}{\sqrt{VAR(U_1)}} \quad (18.8)$$

Pokud $|U| \geq u\left(\frac{\alpha}{2}\right)$, zamítneme hypotézu H_0 na hladině blíží se hodnotě α . Tentotest

se dá použít v případě, že $m > 10$ a $n > 10$.

Dvouvýběrový Wilcoxonův test je velmi citlivý zejména na případ posunutí, tedy na případ, kdy $F_1(x) = F_2(x - \Delta)$, kde $\Delta \neq 0$.

Příklad 18.6

Dvěma způsoby jsme prováděli analýzu procentního obsahu zlata ve vzorcích rudy. Výsledky analýz jsou uvedeny v následující tabulce:

hodnoty X	3,3	3,4	3,4	3,2	3,5	3,3	3,2			
hodnoty Y	3,2	3,5	3,4	3,4	3,4	3,3	3,5	3,5	3,4	3,5

Řešení:

Seřadíme uvedené hodnoty do jedné tabulky, zjistíme pořadí hodnot X:

3,2	3,2	3,2	3,3	3,3	3,3	3,4	3,4	3,4	3,4	3,4	3,4	3,5	3,5	3,5	3,5	3,5
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Odlišným podkladem jsou označeny hodnoty X. Zjistíme a spočteme hodnoty $U_1 = 58$ a $U_2 = 12$. Hodnoty $S_1 = 40$ a $S_2 = 30$. V tabulkách dvouvýběrového Wilcoxonova testu zjistíme pro hodnoty $m = 10$ a $n = 7$ (první hodnota je podle úmluvy vždy větší nebo rovna druhé hodnotě) kritickou hodnotu 14. Tedy zamítáme na hladině významnosti 0,05 hypotézu H_0 o stejném způsobu měření.

Použijeme – li přechod na rozdělení $N(0,1)$ získáme

$$U = \frac{58 - 35}{\sqrt{105}} = 2,24457, \text{ protože je tato hodnota větší než } 1,96 \text{ zamítáme i tímto}$$

způsobem hypotézu H_0 .

Příklad 18.7

Ověřme, zda existuje významný rozdíl v hektarových výnosech brambor při použití dvou metod hnojení. V následující tabulce jsou uvedeny hektarové výnosy:

hodnoty X	58	55	57	59	54	52	50	53	56
hodnoty Y	67	60	71	68	56	53	62		

Řešení:

Nejdříve seřadíme hodnoty X a Y dohromady a setřídíme je podle velikosti:

50	52	53	53	54	55	56	56	57	58	59	60	62	67	68	71
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Opět jsme odlišným způsobem označili hodnoty X. Podobně jako v předchozím příkladu zjistíme hodnoty $S_1 = 54$ a $S_2 = 9$, stejně tak zjistíme, že hodnoty $U_1 = 54$ a $U_2 = 9$, protože kritická hodnota pro případ $m = 9$ a $n = 7$ je rovna 12, zamítneme hypotézu H_0 na hladině významnosti 0,05.

Pokud bychom počítali aproximaci pomocí rozdělení $N(0,1)$ získáme hodnotu:

$$U = \frac{54 - 31,5}{\sqrt{42}} = 2,31455. \text{ Tato hodnota je větší než } 1,96. \text{ Zamítáme tedy i touto metodou}$$

hypotézu H_0 o stejných výnosech pomocí obou metod hnojení.

18.4 Jednovýběrový Kolmogor – Smirnovův test

Nechť $X_1, \dots, X_n$ je náhodný výběr. Mějme dále určité spojitě rozdělení s distribuční funkcí F . Stanovme hypotézu H_0 : Distribuční funkce rozdělení, z něhož pochází náhodný výběr je F .

Nejdříve vytvoříme empirickou distribuční funkci F_n , vytvořenou z hodnot výběru $X_1, \dots, X_n$. Položíme

$$D_n = \sup_x |F_n(x) - F(x)| \quad (18.9)$$

zřejmě velké hodnoty D_n budou podporovat hypotézu H_1 . Pro malé hodnoty n nacházíme kritické hodnoty v tabulkách, pro velké hodnoty n se provádí aproximace

$$D_n \doteq \sqrt{\frac{1}{2n} \cdot \ln\left(\frac{2}{\alpha}\right)} \quad (18.10)$$

Hypotézu H_0 zamítáme, když $D_n \geq D_n(\alpha)$.

Tento test můžeme používat bez jakýchkoli úprav na testování, zda daná data pochází z rozdělení, které známe včetně všech možných parametrů, které toto rozdělení definují. Tedy bez úprav se nehodí přímo na testování, zda daná data pochází z normálního rozdělení. V případě, že by dané parametry byly odhadnuty z výběru, změnil se velmi výrazně hodnoty kritických hodnot D_n . Pro takovéto případy jsou kritické hodnoty odhadnuty pomocí simulačních metod viz Lilliefors, Iman.

Příklad 18.8

V následující tabulce je uvedeno 20 náhodných čísel vygenerovaných generátorem náhodných čísel v programu Excel. Protože tato čísla mají pocházet z rozdělení rovnoměrného na intervalu (0,1), budeme testovat hypotézu H_0 : Dané údaje jsou popsány pomocí rozdělení $Ro(0,1)$.

1	2	3	4	5	6	7	8	9	10
0,040391	0,061493	0,077673	0,086749	0,123528	0,136589	0,168583	0,327113	0,376007	0,376199

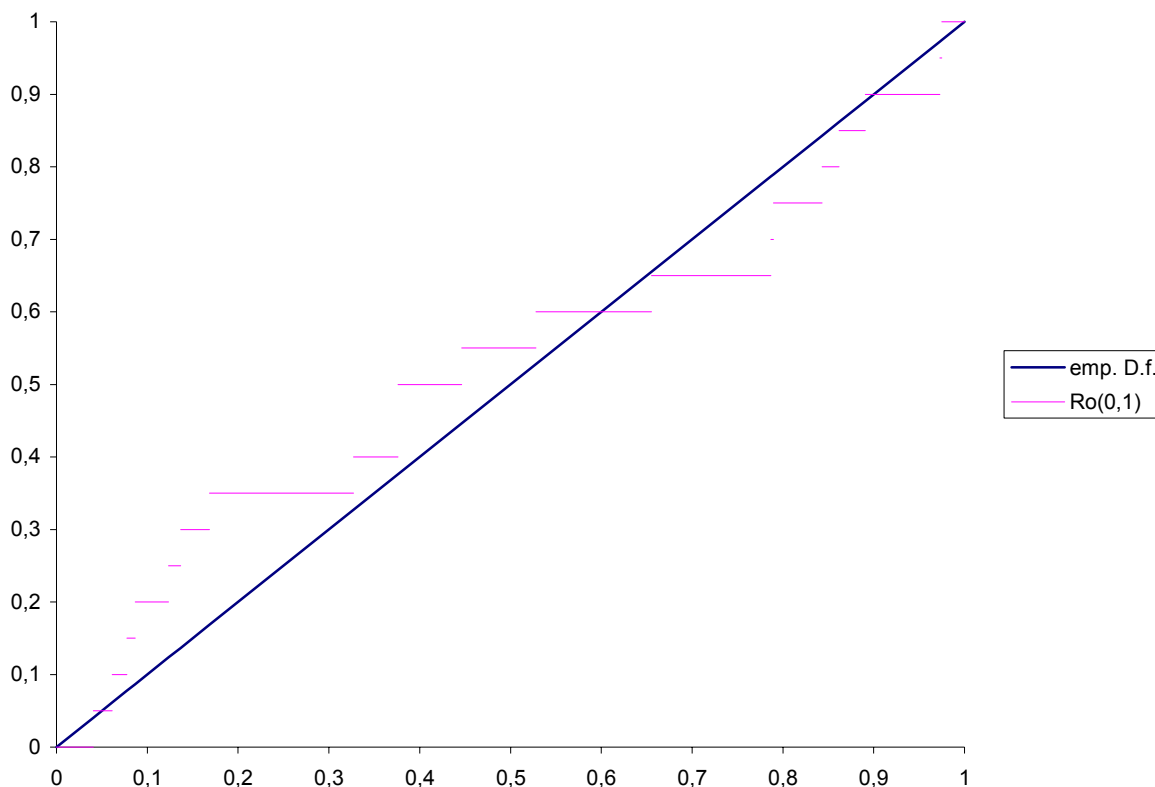
11	12	13	14	15	16	17	18	19	20
0,446321	0,528015	0,655216	0,786571	0,789621	0,842974	0,861987	0,890659	0,972825	0,974874

Řešení:

Nejdříve sestavíme empirickou distribuční funkci. Protože vygenerované hodnoty jsou jedinečné (jsou od sebe různé) bude konstrukce empirické distribuční funkce takováto:

Pro hodnoty $x \leq 0,040391$ bude $F_{20}(x)=0$, v každé hodnotě x_i bude mít empirická distribuční funkce skok stejný a rovný $\frac{1}{20}=0,05$ (jestliže by se některá z hodnot x_i

vyskytovala např. p – krát, pak by měla funkce F_n skok $\frac{p}{n}$ v bodě x_i). Dále je zobrazena empirická distribuční funkce F_{20} a distribuční funkce rovnoměrného rozdělení na intervalu (0,1):



Výpočtem zjistíme, že v našem případě je hodnota $D_{20} = 0,18147$. V tabulkách zjistíme, že pro hladinu významnosti je kritická hodnota pro $n = 20$ rovna $0,29408$. Protože D_{20} je menší než kritická hodnota, nelze zamítnout hypotézu o výběru z rozdělení $Ro(0,1)$.

18.5 Dvouvýběrový Kolmogor – Smirnovův test

Nechť je $X_1, \dots, X_m$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F_1 , necht' dále je $Y_1, \dots, Y_n$ náhodný výběr ze spojitého rozdělení s distribuční funkcí F_2 , který je na výběru $X_1, \dots, X_m$ nezávislý. Hypotézu H_0 sestavujeme tak, že $F_1 = F_2$, alternativní hypotéza je oboustranná tedy $F_1 \neq F_2$.

Označíme – li F_{1m} a F_{2n} empirické distribuční funkce jednotlivých výběrů, vyplývá z obecných vět, že s rostoucími hodnotami m , n se tyto funkce blíží ke skutečným distribučním funkcím F_1 a F_2 . Označme dále

$$D_{m,n} = \sup_x |F_{1m}(x) - F_{2n}(x)|.$$

Celý test je založen na následující větě:

Věta 18.9

Označme $M = \frac{m \cdot n}{m + n}$. Necht' dále je

$$K(\lambda) = 1 - 2 \cdot \sum_{k=1}^{\infty} (-1)^{k+1} \cdot \exp(-2 \cdot k^2 \cdot \lambda^2) \quad (18.11)$$

Potom pro každé λ platí

$$\lim_{m,n \rightarrow \infty} P(\sqrt{M} \cdot D_{m,n} < \lambda) = K(\lambda).$$

Důkaz : Proveden např. v Hájek a Šidák (1967).

Funkci $K(\lambda)$ často aproximujeme pomocí jejích počátečních členů $1 - 2e^{-2\lambda^2}$. Tato hodnota je rovna $1 - \alpha$, jestliže $\lambda = \sqrt{\frac{1}{2} \cdot \ln\left(\frac{2}{\alpha}\right)}$, aproximativní kritická hodnota pro $D_{m,n}$ je tedy rovna $\sqrt{\frac{1}{2M} \cdot \ln\left(\frac{2}{\alpha}\right)} = D_{m,n}^*(\alpha)$. Jsou – li čísla m , n malá porovnává se hodnota $D_{m,n}$ s tabelovanými hodnotami kritických hodnot na dané hladině významnosti.

Jestliže jsou hodnoty m , n větší ($m + n > 35$), vypočítáme nejdříve hodnotu $\lambda_0 = \sqrt{M} \cdot D_{m,n}$, pomocí ní vypočítáme $K(\lambda_0)$. Jestliže je tato hodnota větší nebo rovna $1 - \alpha$, zamítneme hypotézu H_0 na hladině, která se s rostoucím rozsahem blíží k číslu α . Při velkých hodnotách aproximujeme kritickou hodnotu číslem $D_{m,n}^*(\alpha)$. Hypotézu H_0 pak zamítáme, jestliže je $D_{m,n} \geq D_{m,n}^*(\alpha)$.

Příklad 18.9

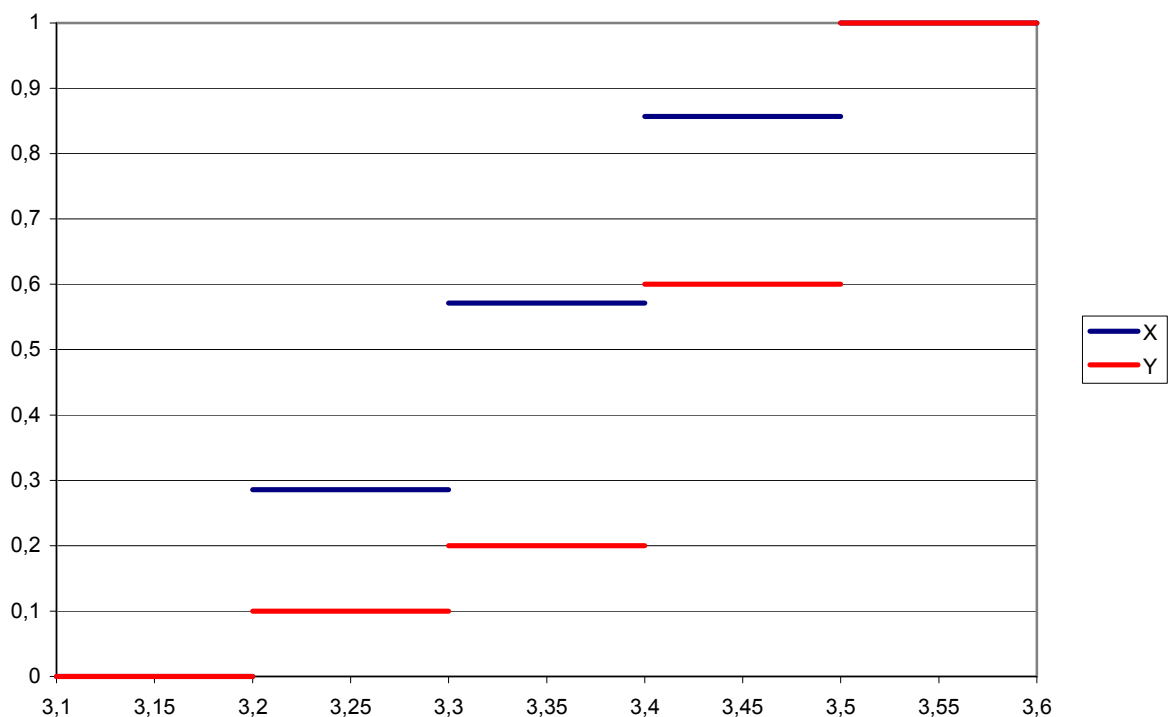
Budeme vyšetřovat hodnoty uvedené v příkladě 18.6. Máme tedy k dispozici data:

hodnoty X	3,3	3,4	3,4	3,2	3,5	3,3	3,2			
hodnoty Y	3,2	3,5	3,4	3,4	3,4	3,3	3,5	3,5	3,4	3,5

Řešení:

Podobně jako v předchozí části si zobrazíme empirické distribuční funkce:

Empirické distribuční funkce



Zřejmě je hodnota $D_{10,7} = 0,371429$, kritická hodnota nalezená v tabulkách dvouvýběrového Kolmogorova – Smirnova testu pro hladinu významnosti 0,05 je rovna $\frac{46}{70} = 0,657142857$. Tedy na základě zjištěných údajů nemůžeme zamítnout hypotézu H_0 , že oba výběry pocházejí ze základních souborů se stejnými distribučními funkcemi. Pokud bychom použili aproximovaných hodnot je v našem případě $D_{7,10}^*(0,05) = 0,669279734$, tedy ani v tomto případě hypotézu H_0 nemůžeme zamítnout.

18.6 Kruskalův – Wallisův test

Jde o test , který je zobecněním dvouvýběrového Wilcoxonova testu . Je používán na data , která se velmi liší od dat typu normálního.

Nechť $X_{i1}, \dots, X_{in_i}$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F_i , $i=1, \dots, N$. Nechť jsou všechny na sobě navzájem nezávislé. Hypotézu H_0 postavíme takto.

$H_0 : F_1(x) = \dots = F_N(x)$ pro všechna x

Alternativní hypotéza H_1 je stanovena v případě neplatnosti hypotézy H_0 . Podobně jako v předchozích případech vytvoříme sdružený náhodný výběr z hodnot Y_{ij} o rozsahu $n = n_1 + \dots + n_N$ (předpokládáme , že $n_i > 5, i=1, \dots, N$). Hodnoty Y_{ij} se uspořádají do neklesající posloupnosti a určí se pořadí R_{ij} v rámci každé veličiny Y_{ij} ze sdruženého náhodného výběru.

V rámci každé skupiny zjistím hodnotu $R_i = \sum_{j=1}^{n_i} R_{ij}$, zřejmě je součet všech hodnot R_i roven $\frac{n \cdot (n+1)}{2}$.

Testovou statistikou je

$$Q = \frac{12}{n \cdot (n+1)} \cdot \sum_{i=1}^N \frac{R_i^2}{n_i} - 3 \cdot (n+1) \quad (18.12)$$

Ukazuje se , že při platnosti hypotézy H_0 je rozdělení Q asymptoticky χ^2 rozdělení , jestliže hodnoty všech n_i rostou nade všechny meze. Výpočtem se zjistí , že $E(Q) = N-1$, tedy půjde o rozdělení χ^2 s $N-1$ stupni volnosti. Hypotézu H_0 zamítáme, jestliže $Q \geq \chi_{N-1}^2(\alpha)$.

Kruskal – Wallisův test je velmi citlivý na případná posunutí v argumentech jednotlivých distribučních funkcí. Jestliže dojde k zamítnutí H_0 , snažíme se dále rozhodnout , které dvojice se ve sdruženém výběru od sebe významně liší.

Pro tento případ označíme symbolem $r_i = \frac{R_i}{n_i}$, kde $i=1, \dots, N$. Jestliže označíme symbolem $kw_{N-1}(\alpha)$ kritickou hodnotu Kruskal – Wallisova testu (v případě malých hodnot N ji najdeme přímo v tabulkách tohoto testu, pro větší rozsahy ji aproximujeme výše uvedeným rozdělením $\chi_{N-1}^2(\alpha)$) , potom prohlásíme , že se distribuční funkce i – tého a j – tého výběru od sebe významně liší , jestliže platí

$$|r_i - r_j| > \sqrt{\frac{1}{12} \cdot \left(\frac{1}{n_i} - \frac{1}{n_j} \right) \cdot n \cdot (n+1) \cdot kw_{N-1}(\alpha)} \quad (18.13)$$

Tento postup opakujeme pro všechny možné dvojice.

Příklad 18.10

U přijímacích zkoušek se sledují počty bodů z testů z matematiky, chceme posoudit, zda se výsledky významně odlišují. Náhodně vybereme z jednotlivých sledovaných typů škol 8 studentů. Hladina významnosti je stanovena na 0,05. Dále následují data:

stud.	gymnasium státní	gymnasium soukromé	SEŠ	prům. škola
1	78	84	32	93
2	95	42	72	74
3	47	54	65	58
4	78	76	41	85
5	85	68	67	60
6	96	61	52	72
7	75	71	53	67
8	83	66	70	59

Řešení:

Jednotlivé hodnoty přepíšeme do jediné tabulky a přidáme sloupec, v němž uvedeme pořadí jednotlivého prvku ve sruženém výběru, v poslední řádce čteme hodnoty R_i .

stud.	gymnasium státní	pořadí	gymnasium soukromé	pořadí	SEŠ	pořadí	prům. škola	pořadí
1	78	24,5	84	27	32	1	93	30
2	95	31	42	3	72	19,5	74	21
3	47	4	54	7	65	12	58	8
4	78	24,5	76	23	41	2	85	28,5
5	85	28,5	68	16	67	14,5	60	10
6	96	32	61	11	52	5	72	19,5
7	75	22	71	18	53	6	67	14,5
8	83	26	66	13	70	17	59	9
		192,5			118	77	140,5	

Dosazením známých hodnot zjistíme velikost testovací statistiky

$$Q = \frac{12}{32 \cdot 33} \cdot \left(\frac{192,5^2}{8} + \frac{118^2}{8} + \frac{77^2}{8} + \frac{140,5^2}{8} \right) - 3,33 = 9,877130682$$

Kritická hodnota χ^2 s 3 stupni volnosti je 7,81. Protože je hodnota testové statistiky větší než kritická hodnota χ^2 , zamítáme hypotézu H_0 o shodě výsledků testů studentů z různých typů škol.

Protože budeme chtít posoudit na základě dat, které dvojice se vzájemně liší, použijeme vztahu (18.13). Vzhledem k tomu, že všechny skupiny mají stejný počet členů, bude také pravá strana ve výrazu (18.13) konstantní a rovna 13,10801. Zjistíme tedy absolutní hodnoty rozdílů stojících ve výrazu (18.13) na levé straně. Uvedeme tyto rozdíly v přehledné tabulce:

	gymnasium soukromé	SEŠ	prům. škola
gymnasium státní	9,3125	14,4375	6,5
gymnasium soukromé		5,125	2,8125
SEŠ			7,9375

Z vypočtených hodnot je jasné , že významný rozdíl mezi výsledky studentů jednotlivých typů škol je v případě státních gymnázií a SEŠ.

18.7 Friedmanův test

Nechť jako v minulém případě jsou Y_{ij} nezávislé náhodné veličiny spojitého typu s distribučními funkcemi F_{ij} . Friedmanovým testem testujeme hypotézu H_0 , že F_{ij} nezávisí na j , na i záviset může. V podstatě jde o rozšíření Wilcoxonova testu pro dva závislé výběry , na více na N závislých výběrů.

Test se používá k ověření shody úrovně sledovaného znaku v souborech vytvořených na základě N závislých výběrů se stejným rozsahem jednotek. Typickou úlohou je případ , kdy jednu skupinu jednotek sledujeme v určitých časových obdobích . Úlohou je potom posoudit, zda sledovaný znak závisí na daných podmínkách.

Pro každou hodnotu Y_{ij} ($i=1,\dots,I$ a $j=1,\dots,J$) zjistíme pořadí R_{ij} v dané skupině i . Tvar testovací statistiky je potom roven

$$Q = \frac{12}{I.J.(J+1)} \cdot \sum_{j=1}^J \left(\sum_{i=1}^I R_{ij} \right)^2 - 3.I.(J+1) \quad (18.14)$$

Dá se opět dokázat, že při platnosti hypotézy H_0 má náhodná veličina Q tvar χ^2 s $J-1$ stupni volnosti. Hypotézu H_0 zamítáme , jestliže hodnota Q překročí kritickou hodnotu na hladině významnosti α . Stejně jako v předchozích testech pro větší hodnoty za tuto hodnotu budeme brát $\chi^2_{J-1}(\alpha)$.

Zamítneme – li H_0 zajímá nás většinou , které dvojice se od sebe významně liší. Označme

$$R_j = \sum_{i=1}^I R_{ij} \quad (18.15)$$

Podobně jako u Kruskal – Wallisova testu budeme zjišťovat zda $|R_u - R_t|$ je větší nebo rovno tabelované kritické hodnotě. Asymptoticky lze kritické hodnoty určit pomocí vztahu

$$q_{J,\infty}(\alpha) \cdot \sqrt{\frac{1}{12} \cdot I.J.(J+1)} \quad (18.16)$$

Kritické hodnoty $q_{J,\infty}(\alpha)$ se opět hledají ve speciálních tabulkách .

Příklad 18.11

Máme k dispozici údaje z průzkumu cen u tří prodejen , u nichž byly sledovány ceny 7 druhů zboží. Chceme posoudit , zda ve sledované době byla významně odlišná úroveň cen v těchto prodejnách.

a	4,5	5	8,2
b	12,5	11,7	14,1
c	5,7	5,5	5,6
d	1,4	1,5	1,8
e	27,8	32,6	28,1
f	72,4	89,4	72,5
g	8,4	6,7	7,8

Řešení:

Z dat je zřejmé , že je $I = 7$ a $J = 3$. Nejdříve upravíme předchozí tabulku tak , aby se v jednotlivých řádcích vyskytovali jen pořadí jednotlivých hodnot. Pro jednotlivé prodejny tyto hodnoty sečteme , umocníme a dosadíme do vzorce (18.14).

zboží	1. Prodejna	2. Prodejna	3. Prodejna
a	1	2	3
b	2	1	3
c	3	1	2
d	1	2	3
e	1	3	2
f	1	3	2
g	3	1	2
R_j	12	13	17
R_j²	144	169	289

Tedy $Q = \frac{12}{7.3.4} (144 + 169 + 289) - 3.7.4 = 86 - 84 = 2$. Z tabulek Friedmanova testu

vyplývá, že kritická hodnota pro $\alpha = 0,05$; $I = 7$ a $J = 3$ je rovna 7,143. Hodnota Q není větší než kritická hodnota, proto nemůže na hladině významnosti 0,05 zamítnout předpoklad o shodě cen jednotlivých prodejnách.

Příklad 18.12

Celkem 10 studentů řešilo stejně obtížné úlohy s přemísťováním předmětů (každý celkem 100 v jedné úloze) za různých teplotních podmínek. Pomocí Friedmanova testu rozhodněte, zda se výkony v těchto podmínkách lišily. Dále jsou uvedeny data jednotlivých studentů:

student	t=0°C	t=30°C	t=50°C
1	70	65	48
2	96	92	72
3	82	91	80
4	78	86	82
5	91	90	79
6	98	95	87
7	92	90	88
8	96	94	78
9	78	90	85
10	86	92	90

Řešení:

Podobně jako v předchozím případě zjistíme, že $I = 10$ a $J = 3$. Dále zjistíme pořadí:

student	t=0°C	t=30°C	t=50°C
1	3	2	1
2	3	2	1
3	2	3	1
4	1	3	2
5	3	2	1
6	3	2	1
7	3	2	1
8	3	2	1
9	1	3	2
10	1	3	2
R_j	23	24	13
R_j²	529	576	169

Z těchto hodnot již můžeme spočítat testovou statistiku .

$$Q = \frac{12}{10 \cdot 3 \cdot 4} \cdot (529 + 576 + 169) - 3 \cdot 10 \cdot 4 = 7,4$$
 . Kritická hodnota pro dané hodnoty I a J je rovna 6,2 . Tedy zamítáme nulovou hypotézu , výsledky studentů se na hladině významnosti 0,05 významně liší. Zjistíme dále hodnoty rozdílů částečných součtů pořadí:

R1-R2=	-1
R2-R3=	11
R1-R3=	10

Z tabulky Friedmanova testu pro mnohonásobná porovnávání je kritická hodnota rovna 10,5. Tuto hranici překračuje R2 a R3 tedy na hladině významnosti 0,05 prokázán významný rozdíl mezi výsledky při teplotě 30°C a 50°C.