9.4 Kruskalův-Wallisův test

Tento test je neparametrickou obdobou analýzy rozptylu jednoduchého třídění a je zobecněním dvouvýběrového Wilcoxonova testu. Bývá používán zejména tehdy, jde-li o výběry z rozdělení značně se lišících od normálního.

$\ldots, Y_{In_I}$ je výběr z nějakého rozdělení se spojitou distribuční funkcí $F_i, i = 1, \ldots, I$. Nechť všechny tyto výběry jsou na sobě nezávislé. Budeme testovat hypotézu

$$H_0 : F_1(x) = \cdots = F_I(x) \qquad \text{pro všechna } x$$

proti alternativě H_1, že H_0 neplatí. Všechny veličiny Y_{ij} dohromady vytvoří *sdružený náhodný výběr* o rozsahu $n = n_1 + \cdots + n_I$. Uspořádají se do rostoucí posloupnosti a určí se pořadí R_{ij} každé veličiny Y_{ij} ve sdruženém výběru. Toto pořadí můžeme zapsat do schématu uvedeného v tab. 9.8.

Tabulka 9.8: Kruskalův-Wallisův test

Výběr	Pořadí veličin ve sdruženém náhodném výběru				Součet pořadí
1	R_{11}	R_{12}	$\ldots$	R_{1n_1}	T_1
2	R_{21}	R_{22}	$\ldots$	R_{2n_2}	T_2
......	......	......	......	......	......
I	R_{I1}	R_{I2}	$\ldots$	R_{In_I}	T_I

Celkový součet všech pořadí je $T_1 + \cdots + T_I = n(n+1)/2$. Jako testová statistika se použije

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

Nejprve vyšetříme obecný tvar statistik tohoto typu. Přitom použijeme označení zavedené již v odst. 8.3.

Věta 9.11 *Nechť $X_1, \ldots, X_N$ je náhodný výběr ze spojitého rozdělení a nechť $g_1, \ldots, g_I$ je rozklad množiny $\{1, \ldots, N\}$ na disjunktní neprázdné podmnožiny. Nechť g_i má n_i prvků, $i = 1, \ldots, I$. Nechť R_i je pořadí X_i, $i = 1, \ldots, N$. Nechť $a(i)$ jsou daná čísla, $i = 1, \ldots, N$. Označme*

$$\bar{a} = \frac{1}{N} \sum a(i), \qquad \sigma_a^2 = \frac{1}{N} \sum [a(i) - \bar{a}]^2.$$

Nechť

$$Q_j = \sum_{i \in g_j} a(R_i), \qquad Q = \frac{N-1}{N\sigma_a^2} \sum_{j=1}^{I} \frac{1}{n_j} (Q_j - \mathsf{E}Q_j)^2.$$

Pak $\mathsf{E}Q = I - 1$.

Důkaz. Máme

$$Q_j = \sum_{i=1}^{N} c_i(j)\, a(R_i),$$

kde $c_i(j) = 1$ pro $i \in g_j$ a $c_i(j) = 0$ pro $i \notin g_j$. Označme

$$\bar{c}(j) = \frac{1}{N}\sum_{i=1}^{N} c_i(j).$$

Podle věty 8.6 platí

$$\operatorname{var} Q_j = \frac{N^2}{N-1}\sigma_a^2\sigma_c^2,$$

kde

$$\sigma_c^2 = \frac{1}{N}\sum_{i=1}^{N}[c_i(j) - \bar{c}(j)]^2 = \frac{1}{N}\sum_{i=1}^{N} c_i^2(j) - [\bar{c}(j)]^2$$
$$= \frac{n_j}{N} - \left(\frac{n_j}{N}\right)^2 = \frac{n_j}{N}\left(1 - \frac{n_j}{N}\right).$$

Proto

$$\mathsf{E}Q = \frac{N-1}{N\sigma_a^2}\sum_{j=1}^{I}\frac{1}{n_j}\operatorname{var} Q_j$$
$$= \frac{N-1}{N\sigma_a^2}\sum_{j=1}^{I}\frac{1}{n_j}\frac{N^2}{N-1}\sigma_a^2\frac{n_j}{N}\left(1 - \frac{n_j}{N}\right)$$
$$= \sum_{j=1}^{I}\left(1 - \frac{n_j}{N}\right) = I - 1. \qquad \square$$

V případě Kruskalova-Wallisova testu je $a(i) = i$, $Q_j = T_j$ a $N = n$. V důkazu věty 8.4 jsme vypočítali, že

$$\sigma_a^2 = \frac{1}{12}(N+1)(N-1).$$

Podle věty 8.6 dále máme

$$\mathsf{E}Q_j = N\bar{a}\bar{c}(j) = N\frac{N+1}{2}\frac{n_j}{N} = \frac{1}{2}(N+1)n_j,$$

takže

$$Q = \frac{12(N-1)}{N(N+1)(N-1)}\sum_{j=1}^{I}\frac{1}{n_j}\left[T_j - \frac{1}{2}n_j(N+1)\right]^2$$
$$= \frac{12}{N(N+1)}\left[\sum_{j=1}^{I}\frac{T_j^2}{n_j} - \frac{1}{4}N(N+1)^2\right]$$
$$= \frac{12}{N(N+1)}\sum_{j=1}^{I}\frac{T_j^2}{n_j} - 3(N+1).$$

... $= n$, dostaneme tím výraz uvedený na začátku. $\square$

Místo Q se někdy užívá označení H a mluví se pak o *H testu*. Je-li v datech víc než 25 % shod, používá se *korigovaná statistika*

$$Q_{korig} = \frac{Q}{1 - (n^3 - n)^{-1}\sum(t_i^3 - t_i)},$$

kde $t_1, t_2, \ldots$ jsou počty shodných pozorování v jednotlivých skupinkách veličin majících tutéž hodnotu.

Dá se dokázat (viz Hájek a Šidák 1967), že za platnosti H_0 má Q asymptoticky χ^2 rozdělení, když všechna n_i rostou nade všechny meze. Jelikož jsme dokázali, že $\mathsf{E}Q = I - 1$, půjde o asymptotické χ^2 rozdělení mající $I - 1$ stupňů volnosti. Proto hypotézu H_0 zamítáme, když $Q \geq \chi^2_{I-1}(\alpha)$.

Kruskalův-Wallisův test je citlivý zejména na ty případy, kdy se jednotlivé distribuční funkce od sebe liší posunutím. Zamítne-li se H_0, je třeba obvykle rozhodnout, které dvojice výběrů se od sebe významně liší. U analýzy rozptylu byla k tomu účelu použita Tukeyova metoda. U Kruskalova-Wallisova testu se postupuje následovně (viz Miller 1966). Označme $t_i = T_i/n_i$, $i = 1, \ldots, I$. Nechť $h_{I-1}(\alpha)$ je kritická hodnota Kruskalova-Wallisova testu na hladině α. Při malých rozsazích výběru se $h_{I-1}(\alpha)$ najde ve speciálních tabulkách a při větších rozsazích se použije výše zmíněné aproximace $h_{I-1}(\alpha) \doteq \chi^2_{I-1}(\alpha)$. Prohlásíme, že se distribuční funkce i-tého a j-tého výběru od sebe signifikantně liší, jakmile platí

$$|t_i - t_j| > \sqrt{\frac{1}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)n(n+1)\,h_{I-1}(\alpha)}. \tag{9.5}$$

Pravděpodobnost, že alespoň u jedné z $I(I-1)/2$ dvojic distribučních funkcí F_i, F_j bude vypočteno, že se F_i a F_j signifikantně liší, ačkoli ve skutečnosti platí hypotéza H_0, přitom nepřekročí α.

Je-li rozsah všech výběrů stejný, řekněme $n_1 = \cdots = n_I = m$ (takže jde o *vyvážené třídění*), lze použít *Neményiovy metody* založené na Tukeyově myšlence uplatněné již při analýze rozptylu (viz Neményi 1963 a Miller 1966). Pro menší hodnoty m a I jsou kritické hodnoty pro $|T_i - T_j|$ uvedeny v tabulce T15. Při větších hodnotách se užije následující postup. Nechť $q_{I,\infty}(\alpha)$ je kritická hodnota rozpětí I nezávislých náhodných veličin s rozdělením $N(0,1)$. Najde se v posledním řádku tabulky T11, resp. T12, a zavádí se takto. Je-li $\xi_1, \ldots, \xi_I$ je výběr z $\mathsf{N}(0,1)$, označíme $R = \xi_{(I)} - \xi_{(1)}$ jeho *rozpětí*. Pak $q_{I,\infty}(\alpha)$ je číslo definované podmínkou

$$\mathsf{P}[R \geq q_{I,\infty}(\alpha)] = \alpha.$$

Prohlásíme, že se F_i a F_j od sebe liší, když

$$|t_i - t_j| > q_{I,\infty}(\alpha)\sqrt{\frac{1}{12}I(Im+1)}. \tag{9.6}$$

Ačkoliv při vyváženém třídění lze užít obou zde uvedených metod, dává se přednost Neményiově metodě, protože je citlivější.

Další testy, které lze užít místo Kruskalova-Wallisova testu, popisují Bhapkar a Deshpande (1968).

Je třeba upozornit na to, že přiřazování pořadí náhodným veličinám je transformace sice monotónní, ale nelineární. Právě tato nelinearita může někdy vést k paradoxním výsledkům, jak ukazuje následující příklad (viz Haunsperger a Saari 1991).

Příklad 9.12 V továrně jsou dvě dílny. V každé z nich pracuje 6 stejných strojů. Ale 2 jsou od výrobce A, 2 od výrobce B a 2 od výrobce C. Výkonnost strojů je uvedena v tabulkách 9.9 – 9.11.

Tabulka 9.9: Dílna č. 1

Výrobce	Výkonnost strojů		Pořadí		Součet pořadí
A	5,89	5,98	3	5	8
B	5,81	5,90	2	4	6
C	5,80	5,99	1	6	7

Tabulka 9.10: Dílna č. 2

Výrobce	Výkonnost strojů		Pořadí		Součet pořadí
A	5,69	5,74	3	5	8
B	5,63	5,71	2	4	6
C	5,62	6,00	1	6	7

Tabulka 9.11: Celá továrna

Výrobce	Výkonnost strojů				Pořadí				Součet pořadí
A	5,89	5,98	5,69	5,74	8	10	3	5	26
B	5,81	5,90	5,63	5,71	7	9	2	4	22
C	5,80	5,99	5,62	6,00	6	11	1	12	30

V dílně 1 je z hlediska součtu pořadí uspořádání výrobců strojů dáno relací $A \succ C \succ B$. Rovněž v dílně 2 platí $A \succ C \succ B$, dokonce součet pořadí je pro každého výrobce stejný v dílně 1 jako v dílně 2. V celé továrně je však naproti tomu uspořádání $C \succ A \succ B$, které je v naprostém rozporu s uspořádáním v dílnách. $\diamond$

o paradoxech vznikajících při porovnávání několika výběrů a o příbuzné tématice pojednávají články Haunsperger (1992) a Saari (1989).

Příklad 9.13 Použijeme dat z příkladu 9.9. Hodnoty uvedené v tab. 9.3 seřadíme podle velikosti a přiřadíme jim jejich pořadí. Dostaneme tab. 9.12.

Tabulka 9.12: Pořadí hektarových výnosů brambor

Odrůda	Pořadí výnosů brambor						
A	8	6	15	17	12	10	22
B	20	27	26	25	24	14	28
C	21	11	9	23	16	18	19
D	4	2	3	5	13	1	7

Máme $n_i = 7$ pro $i = 1, 2, 3, 4$ a postupně vypočteme

$$T_1 = 90, \quad T_2 = 164, \quad T_3 = 117, \quad T_4 = 35, \quad Q = 18{,}369.$$

Protože $Q \geq \chi^2_3(0{,}05) = 7{,}81$, zamítáme hypotézu, že všechny čtyři dílčí výběry pocházejí z rozdělení se stejnou distribuční funkcí. Pro mnohonásobná porovnávání nejdřív vypočteme

$$t_1 = 12{,}857, \quad t_2 = 23{,}429, \quad t_3 = 16{,}714, \quad t_4 = 5.$$

Kritická hodnota podle (9.5) je rovna 12,3. V našem případě jsou ve všech třídách stejné rozsahy, a tak můžeme vypočítat také kritickou hodnotu podle (9.6). Ta vyjde 11,3. Všimněte si, že je také rovna kritické hodnotě z tabulky T15 dělené rozsahem podskupin 7, tj. číslu 79,1/7. Použijeme tuto druhou kritickou hodnotu, jelikož je menší a tudíž výhodnější. Diference hodnot $t_i - t_j$ jsou uvedeny v tab. 9.13. Významnost je označena hvězdičkou. Výsledek shrneme tak, že odrůdy

Tabulka 9.13: Hodnoty $t_i - t_j$

i \ j	2	3	4
1	−10,57	−3,86	7,86
2		6,72	18,43*
3			11,71*

seřadíme podle hodnot t_i a podtržením souvislou čarou vyznačíme ty skupiny, mezi jejichž členy jsou jen nevýznamné rozdíly. Dostaneme tab. 9.14.

Je tedy prokázáno, že se D významně liší od C i od B. Tyto závěry jsou méně určité než při zpracování analýzou rozptylu v příkladu 9.9. $\diamond$

$0{,}712 < F_{2,5}(0{,}05) = 5{,}79$.

Zjistíme, které podklady se od sebe významně liší. Protože $J = 4$, $f_e = 5$, pro $\alpha = 0{,}05$ máme $q_{4,5}(0{,}05) = 5{,}22$. Kritická hodnota pro rozdíly sloupcových průměrů pak je 8,93. Průměry seřadíme do neklesající posloupnosti. Dostaneme tab. 9.25. Souvislou čárou podtrhneme vždy tu skupinu, mezi jejímiž členy jsou rozdíly menší než 8,93. Je tedy statisticky významný rozdíl mezi podkladem H_4 a H_3 a také mezi podklady H_4 a H_2. Ostatní rozdíly jsou statisticky nevýznamné. ◊

9.8 Friedmanův test

Nechť Y_{ij} jsou nezávislé náhodné veličiny se spojitými distribučními funkcemi F_{ij} pro $i = 1, \dots, I$, $j = 1, \dots, J$. *Friedmanovým testem* se testuje hypotéza H_0, že F_{ij} nezávisí na j (zatímco na i záviset může).

Pro každé i zvlášť se určí pořadí R_{ij} veličiny Y_{ij}. Jde tedy jen o určení pořadí mezi veličinami $Y_{i1}, \dots, Y_{iJ}$. Za platnosti H_0 je splněna podmínka

H_0' : pro každé i je vektor $(R_{i1}, \dots, R_{iJ})'$ roven kterékoli permutaci čísel $1, \dots, J$ se stejnou pravděpodobností $1/J!$ a všechny vektory $(R_{i1}, \dots, R_{iJ})'$ pro $i = 1, \dots, I$ jsou na sobě nezávislé.

Protože všechny vlastnosti Friedmanova testu jsou odvozovány pouze za předpokladu platnosti H_0', lze často tento test použít za obecnějších podmínek, než vyplývá z jeho původní formulace.

Teoretický tvar statistiky Friedmanova testu je

$$Q = \frac{12}{IJ(J+1)} \sum_{j=1}^{J} \left[\sum_{i=1}^{I} R_{ij} - \frac{1}{2} I(J+1) \right]^2 . \tag{9.14}$$

Věta 9.17 *Platí-li H_0', pak* $\mathsf{E}Q = J - 1$.

Důkaz. Snadno se zjistí, že $\mathsf{E}R_{ij} = (J+1)/2$ pro všechna i a j. Proto

$$\begin{aligned}\left[\sum_{i=1}^{I} R_{ij} - \frac{1}{2} I(J+1) \right]^2 &= \left[\sum_{i=1}^{I} (R_{ij} - \mathsf{E}R_{ij}) \right]^2 \\ &= \sum_{i=1}^{I} (R_{ij} - \mathsf{E}R_{ij})^2 + \sum_{i \neq l}\sum (R_{ij} - \mathsf{E}R_{ij})(R_{lj} - \mathsf{E}R_{lj})\end{aligned}$$

Vzhledem k nezávislosti máme $\mathrm{cov}(R_{ij}, R_{lj}) = 0$ pro $i \neq l$. Dále platí $\mathrm{var}\, R_{ij} = (J+1)(J-1)/12$. Proto

$$\begin{aligned} \mathrm{E}Q &= \frac{12}{IJ(J+1)} \sum_{j=1}^{J} \left[\sum_{i=1}^{I} \mathrm{var}\, R_{ij} + \sum_{i \neq l} \sum \mathrm{cov}(R_{ij}, R_{lj}) \right] \\ &= \frac{12}{IJ(J+1)} JI \frac{(J+1)(J-1)}{12} = J - 1. \qquad \square \end{aligned}$$

Dá se dokázat, že za platnosti H_0' má Q při $I \to \infty$ asymptoticky χ^2 rozdělení, které má $J-1$ stupňů volnosti.

Věta 9.18 *Platí*

$$Q = \frac{12}{IJ(J+1)} \sum_{j=1}^{J} \left(\sum_{i=1}^{I} R_{ij} \right)^2 - 3I(J+1). \tag{9.15}$$

Důkaz. Tvrzení se získá úpravou vzorce (9.14). $\square$

Výpočty se provádějí podle vzorce (9.15). Hypotézu H_0' (a tedy také hypotézu H_0) zamítáme, když Q překročí kritickou hodnotu na hladině α (viz tab. T16). Při větších hodnotách I se za tuto kritickou hodnotu bere $\chi^2_{J-1}(\alpha)$. Podrobně o kritických hodnotách a jejich aproximacích pojednává Michaelis (1971).

Zamítneme-li H_0, zajímá nás, pro které dvojice j a t se distribuční funkce F_{ij} a F_{it} od sebe významně liší. Označme

$$R_{.j} = \sum_{i=1}^{I} R_{ij}.$$

Jakmile $|R_{.j} - R_{.t}|$ je větší nebo rovno tabelované kritické hodnotě (viz tab. T17), zamítne se rovnost $F_{ij} = F_{it}$. Tato porovnání se dělají pro všechny dvojice $j < t$. Asymptoticky jsou kritické hodnoty pro tato mnohonásobná porovnávání rovny

$$q_{J,\infty}(\alpha) \sqrt{\frac{1}{12} IJ(J+1)}. \tag{9.16}$$

Kritické hodnoty $q_{J,\infty}(\alpha)$ byly definovány v odst. 9.4. Vzorec (9.16) se používá už při $I > 5$.

Příklad 9.19 Celkem 31 dobrovolníků bylo náhodně rozděleno do dvou skupin o 16 a 15 osobách. Každý člen první skupiny dostal citronovou šťávu a 30 ml alkoholu, členové druhé skupiny jen citronovou šťávu. Všichni pak byli podrobeni Danielovu zrychlenému inteligenčnímu testu a zjišťoval se jejich tep před pokusem

Tabulka 9.26: Tep osob

První skupina

i	$t=1$	$t=2$	$t=3$
1	63	66	77
2	60	70	80
3	89	80	88
4	82	80	83
5	70	72	71
6	53	69	66
7	85	93	92
8	86	81	84
9	58	65	68
10	70	72	71
11	80	76	72
12	95	83	82
13	78	90	84
14	87	82	86
15	85	93	92
16	83	98	88

Druhá skupina

i	$t=1$	$t=2$	$t=3$
1	70	64	69
2	66	63	61
3	95	82	80
4	78	71	80
5	64	59	65
6	100	88	84
7	86	85	92
8	65	58	67
9	66	77	73
10	90	70	82
11	76	65	74
12	88	77	86
13	84	81	64
14	83	72	86
15	104	99	86

($t = 1$), na začátku pokusu ($t = 2$) a na konci pokusu ($t = 3$). Lehmacher a Wall (1978) uvádějí (převzaté) výsledky pokusu (tab. 9.26).

Pomocí Friedmanova testu vyšetříme pro každou skupinu zvlášť, zda se tep mění jen náhodně nebo zda se do jeho změn promítá nějaký systematický vliv času. Proto u každé sledované osoby zjištěné výsledky nahradíme jejich pořadím. Vznikne tak tab. 9.27.

Nejprve budeme analyzovat první skupinu. Zde máme $I = 16$, $J = 3$ a podle vzorce (9.15) dostaneme $Q = 2{,}375$. Protože $Q < \chi^2_2(0{,}05) = 5{,}99$, nemůžeme zamítnout hypotézu, že se u první skupiny tepová frekvence mění jen náhodně. U druhé skupiny je $I = 15$, $J = 3$ a $Q = 8{,}533$. Protože $Q \geq \chi^2_2(0{,}05) = 5{,}99$, zamítáme hypotézu, že se u druhé skupiny mění tepová frekvence jen náhodně. Dále tedy vypočteme

$$R_{.1} - R_{.2} = 16, \qquad R_{.1} - R_{.3} = 8, \qquad R_{.2} - R_{.3} = -8.$$

Příslušná kritická hodnota z tabulky T17 je 12,8. Mimochodem, její aproximace pomocí (9.16) činí rovněž 12,8. Kritickou hodnotu překračuje pouze $|R_{.1} - R_{.2}|$. Tím je u druhé skupiny prokázán signifikantní rozdíl mezi tepovou frekvencí před pokusem a na začátku pokusu. $\diamond$

Tabulka 9.27: Tabulka pořadí

První skupina			
i	t = 1	2	3
1	1	2	3
2	1	2	3
3	3	1	2
4	2	1	3
5	1	3	2
6	1	3	2
7	1	3	2
8	3	1	2
9	1	2	3
10	1	3	2
11	3	2	1
12	3	2	1
13	1	3	2
14	3	1	2
15	1	3	2
16	1	3	2
Σ	27	35	34

Druhá skupina			
i	t = 1	2	3
1	3	2	1
2	3	2	1
3	3	2	1
4	2	1	3
5	2	1	3
6	3	2	1
7	2	1	3
8	2	1	3
9	1	3	2
10	3	1	2
11	3	1	2
12	3	1	2
13	3	2	1
14	2	1	3
15	3	2	1
Σ	38	22	30

Někdy se místo Friedmanova testu užívá *Andersonův-Kannemannův test* (viz Kannemann 1976, Küchenhoff a Lehmacher 1985), který nyní popíšeme.

Friedmanův test odpovídá situaci, kdy na každém z I objektů (resp. bloků) je aplikováno J ošetření. Leckdy nejde o ošetření v pravém slova smyslu, ale prostě je nějaká veličina na každém sledovaném objektu zaznamenávána v J časových okamžicích. Opět označme R_{ij} pořadí, které je připsáno j-tému ošetření na i-tém bloku. Nechť D_{jm} je počet bloků, ve kterých ošetření j dostalo pořadí m. Matice

$$D = \begin{pmatrix} D_{11} & \cdots & D_{1J} \\ \ldots & \ldots & \ldots \\ D_{J1} & \cdots & D_{JJ} \end{pmatrix}$$

se nazývá *incidenční*. Označme

$$\chi^2_{AK} = \frac{J-1}{I}\sum_{j=1}^{J}\sum_{m=1}^{J}\left(D_{jm} - \frac{I}{J}\right)^2.$$

Platí-li hypotéza H_0, pak χ^2_{AK} má asymptoticky rozdělení $\chi^2_{(J-1)^2}$. V případě $\chi^2_{AK} \geq \chi^2_{(J-1)^2}(\alpha)$ se H_0 zamítne. Je nutné upozornit na to, že D není kontingenční

tabulka (srv. kap. 11), a proto limitní rozdělení veličiny χ^2_{AK} muselo být v citované literatuře odvozeno jiným způsobem. Andersonův-Kannemannův test je citlivější proti větší třídě alternativ než Friedmanův test.

Příklad 9.20 Použijeme data z příkladu 9.19. Pro první skupinu osob dostaneme

$$D = \begin{pmatrix} 10 & 1 & 5 \\ 4 & 5 & 7 \\ 2 & 10 & 4 \end{pmatrix}, \qquad \chi^2_{AK} = 10.$$

Protože $\chi^2_{AK} \geq \chi^2_4(0{,}05) = 9{,}49$, tímto testem se zamítá hypotéza, že se v první skupině tepová frekvence mění během času jen náhodně. Pro druhou skupinu osob vyjde

$$D = \begin{pmatrix} 1 & 5 & 9 \\ 9 & 5 & 1 \\ 5 & 5 & 5 \end{pmatrix}, \qquad \chi^2_{AK} = 8{,}533.$$

V tomto případě $\chi^2_{AK} < \chi^2_4(0{,}05) = 9{,}49$, takže Andersonův-Kannemannův test ve druhé skupině nezamítá (na rozdíl od Friedmanova testu) hypotézu o náhodném kolísání tepu během pokusu. ◊

o výběr z rozdělení R(0, 1). Tento příklad vzhledem k velmi malému rozsahu výběru je třeba chápat jen jako ilustraci, ne jako skutečný test použitého generátoru. Pro zajímavost ještě uveďme, že aproximace kritické hodnoty $D_{10}(0{,}05)$ vypočtená podle vzorce (10.16) činí 0,429. ◇

10.6 Kolmogorovův-Smirnovův test

V tomto odstavci pojednáme o *jednovýběrovém Kolmogorovově-Smirnovově testu*. Nechť $X_1, \ldots, X_n$ je náhodný výběr z nějakého rozdělení se spojitou distribuční funkcí. Chceme testovat hypotézu H_0, že tato distribuční funkce je F. Nechť F_n je empirická distribuční funkce odpovídající výběru $X_1, \ldots, X_n$. Položme

$$D_n = \sup_x |F_n(x) - F(x)|. \tag{10.15}$$

Vzhledem k větám 8.9 a 8.10 budou velké hodnoty veličiny D_n svědčit proti hypotéze H_0. Je-li n malé, používají se speciální tabulky kritických hodnot (viz tab. T18). Při větších hodnotách n se vychází z limitního vztahu

$$\lim_{n\to\infty} \mathsf{P}(\sqrt{n}\, D_n < \lambda) = K(\lambda),$$

kde funkce $K(\lambda)$ je definována vzorcem (8.6). Při velkých hodnotách n se kritické hodnoty $D_n(\alpha)$ pro náhodnou veličinu D_n aproximují pomocí

$$D_n(\alpha) \doteq \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}} \tag{10.16}$$

(viz Likeš a Laga 1978.) To znamená, že se H_0 zamítá, když $D_n \geq D_n(\alpha)$.

Všimněme si, že se při výpočtu veličiny D_n podle (10.15) můžeme omezit jen na ta x, v nichž F_n má skok. Zato však v těchto bodech je třeba vyšetřit nejen

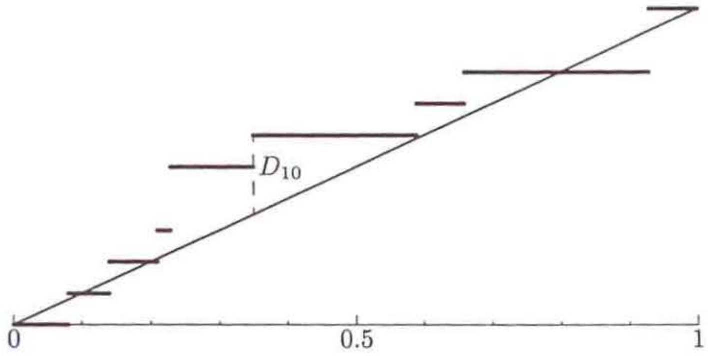
samotný rozdíl $F_n(x) - F(x)$, ale i limitu tohoto výrazu zprava (na to se často zapomíná).

Je třeba zdůraznit, že hypotéza H_0 musí určovat distribuční funkci F zcela jednoznačně, včetně všech jejích případných parametrů. Kolmogorovův-Smirnovův test se tedy může použít třeba k testování hypotézy, že náhodný výběr $X_1, \ldots, X_n$ pochází z rovnoměrného rozdělení R(0, 1). To se hodí např. při testování generátorů náhodných čísel. Nelze však pomocí tohoto testu bez dalších modifikací ověřovat hypotézu, že náhodný výběr pochází z normálního rozdělení. Tato hypotéza totiž neurčuje hodnoty parametrů μ a σ^2. V případě, že bychom parametry odhadli z výběru a za funkci F vzali distribuční funkci normálního rozdělení s těmito odhadnutými parametry, změnilo by se výrazně rozdělení testové statistiky D_n. Protože však uživatelé statistických metod chtěli aplikovat Kolmogorovův-Smirnovův test i na případy, kdy se některé parametry funkce F odhadují z výběru, byly příslušné kritické hodnoty určeny pomocí simulačních studií. Pro normální rozdělení viz Lilliefors (1967) a Iman (1982), pro exponenciální rozdělení viz Lilliefors (1969) a Iman (1982). Zobecnění na výběry z diskrétních rozdělení je uvedeno v pracích Conover (1972) a Mantel (1974). Test hypotézy, že výběr pochází z Poissonova rozdělení, lze nalézt v článku Campbell a Oprian (1979). Kolmogorovův-Smirnovův test byl zobecněn i na cenzorované výběry (viz Barr a Davison 1973).

Příklad 10.7 Prvních deset čísel z generátoru náhodných čísel bylo

0,93, 0,35, 0,66, 0,93, 0,14, 0,23, 0,08, 0,23, 0,21, 0,59.

Je třeba zjistit, zda lze tyto veličiny skutečně pokládat za výběr z R(0, 1). Empirická i teoretická distribuční funkce jsou znázorněny na obr. 10.1.



Obrázek 10.1: Empirická a teoretická distribuční funkce

Výpočtem se zjistí, že $D_{10} = 0{,}27$. V tabulce T18 se najde kritická hodnota $D_{10}(0{,}05) = 0{,}409\,25$. Protože $D_{10} < D_{10}(0{,}05)$, nelze zamítnout hypotézu, že jde

8.4 Dvouvýběrový Kolmogorovův-Smirnovův test

Dříve, než přistoupíme k formulaci tohoto dvouvýběrového testu, uvedeme některá pomocná tvrzení. Nechť $X_1, \ldots, X_m$ je náhodný výběr z rozdělení, které má distribuční funkci F. Budiž x dané reálné číslo. Zaveďme náhodné veličiny

$$\xi_i(x) = \begin{cases} 1, & \text{je-li } X_i < x, \\ 0, & \text{je-li } X_i \geq x \end{cases}$$

pro $i = 1, \ldots, m$. Položme

$$F_m(x) = \frac{1}{m} \sum_{i=1}^{m} \xi_i(x). \tag{8.5}$$

Funkce $F_m(x)$ je *empirická distribuční funkce*. Při konkrétní realizaci výběru je totožná s empirickou distribuční funkcí, která byla zavedena v odst. 1.5. Ukážeme, že se s rostoucím m funkce $F_m(x)$ blíží skutečné distribuční funkci $F(x)$.

Věta 8.9 *Pro každé x platí*

$$F_m(x) \to F(x) \quad \textit{skoro jistě pro} \quad m \to \infty.$$

Důkaz. Pro každé pevně zvolené x jsou veličiny $\xi_i(x)$ nezávislé a mají stejné rozdělení. Platí pro ně

$$\mathsf{P}[\xi_i(x) = 1] = F(x), \qquad \mathsf{E}\xi_i(x) = F(x).$$

Jelikož $F_m(x)$ je dána vzorcem (8.5), z Kolmogorovovy věty 6.4 ihned plyne dokazované tvrzení. □

Dá se však dokázat ještě silnější tvrzení.

Věta 8.10 (Glivenkova) *Označme $D_m = \sup_x |F_m(x) - F(x)|$. Pak platí*

$$\mathsf{P}\left(\lim_{m\to\infty} D_m = 0\right) = 1.$$

Důkaz. Viz Gnedenko (1954). □

Nyní již přejdeme k vlastnímu tématu tohoto odstavce. Nechť $X_1, \ldots, X_m$ je náhodný výběr z rozdělení se spojitou distribuční funkcí F a nechť $Y_1, \ldots, Y_n$ je na něm nezávislý náhodný výběr z rozdělení se spojitou distribuční funkcí G. Budeme se zabývat testem hypotézy $H_0 : F = G$ proti alternativě $H_1 : F \neq G$. Označíme F_m empirickou distribuční funkci prvního výběru a G_n druhého výběru. Z vět 8.9 a 8.10 vyplývá, že se funkce F_m a G_n při rostoucích m a n blíží distribučním funkcím F a G. Označme

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)|.$$

Platí-li H_0, pak podle Glivenkovy věty $D_{m,n} \to 0$ skoro jistě při $m \to \infty$, $n \to \infty$. Přesnější výsledek, na němž se pak dá založit test, je popsán v následující větě.

Věta 8.11 (Smirnovova) *Označme $M = mn/(m+n)$. Nechť*

$$K(\lambda) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2\lambda^2). \tag{8.6}$$

Pak pro každé λ platí

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

Důkaz. Viz Smirnov (1944). Moderní varianta důkazu se najde v knize Hájek a Šidák (1967). □

Rozdělení veličiny $D_{m,n}$ pro konečné hodnoty m, n je uvedeno v knize Hájek a Šidák (1967).

Funkce $K(\lambda)$ se aproximuje pomocí počátečních členů $1 - 2e^{-2\lambda^2}$ (viz Likeš a Laga 1978). Pak

$$\mathsf{P}\left(D_{m,n} < \frac{\lambda}{\sqrt{M}}\right) \doteq 1 - 2e^{-2\lambda^2}.$$

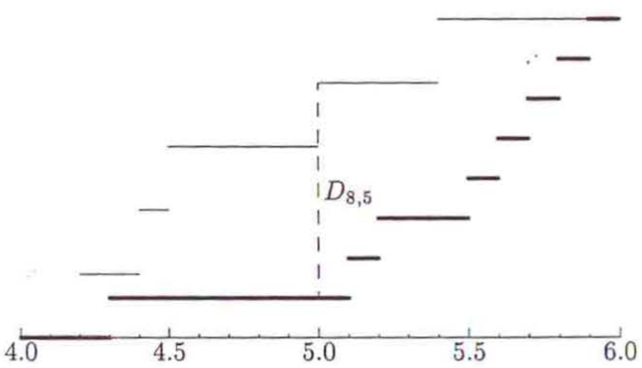
Výraz na pravé straně je roven $1-\alpha$ pro $\lambda = \lambda_\alpha = \sqrt{\frac{1}{2}\ln\frac{2}{\alpha}}$. Aproximativní kritická hodnota je tudíž

$$D^*_{m,n}(\alpha) = \frac{\lambda_\alpha}{\sqrt{M}} = \sqrt{\frac{1}{2M}\ln\frac{2}{\alpha}}.$$

Praktické provedení Kolmogorovova-Smirnovova testu tedy spočívá v tom, že se z výběrů $X_1, \ldots, X_m$ a $Y_1, \ldots, Y_n$ vypočtou empirické distribuční funkce F_m a G_n a veličina $D_{m,n}$. Jsou-li čísla m a n malá, porovná se $D_{m,n}$ s přesnými kritickými hodnotami $D_{m,n}(\alpha)$ (viz tab. T9 a T10). V případě větších hodnot m, n se využije věty 8.11. Položí se $\lambda_0 = \sqrt{M}\, D_{m,n}$ a vypočte se hodnota $K(\lambda_0)$. Pokud vyjde $K(\lambda_0) \geq 1 - \alpha$, zamítne se H_0 na hladině, která se s rostoucími rozsahy výběrů blíží číslu α. Přitom se při větších hodnotách m a n kritická hodnota pro veličinu $D_{m,n}$ obvykle aproximuje číslem $D^*_{m,n}(\alpha)$. Hypotéza H_0 se pak zamítá, když $D_{m,n} \geq D^*_{m,n}(\alpha)$.

Kolmogorovův-Smirnovův test byl zobecněn na případ porovnávání tří a více výběrů v článku Kiefer (1959). Kritické hodnoty tabelovali Wolf a Naus (1973). Viz též Domański (1990).

Příklad 8.12 Budeme analyzovat data z příkladu 8.2. Na obr. 8.1 jsou znázorněny empirické distribuční funkce obou výběrů.



Obrázek 8.1: Empirické distribuční funkce

Snadno se zjistí, že v našem příkladě $D_{8,5} = 0{,}675$. Kritická hodnota podle tab. T9 činí 0,75. (Mimochodem, aproximativní kritická hodnota činí $D^*_{8,5}(0{,}05) = 0{,}774$.) Protože $D_{8,5} < 0{,}75$, Kolmogorovův-Smirnovův test nezamítá hypotézu o tom, že oba výběry pocházejí ze základních souborů se stejnými distribučními funkcemi.

Oba výběry mají malé rozsahy. Užití věty 8.11 se doporučuje až pro $m+n > 35$. Kdybychom přesto limitní postup použili alespoň jako velice hrubou aproximaci, dostali bychom $\lambda_0 = 1{,}184$, $K(\lambda_0) = 0{,}879$. Protože $K(\lambda_0) < 0{,}95$, nemůžeme nulovou hypotézu zamítnout. ◇