Kapitola 11.: Porovnání empirického a teoretického rozložení

Cíl kapitoly

Po prostudování této kapitoly budete umět
- testovat hypotézu, že daný náhodný výběr pochází z rozložení s danou diskrétní či spojitou distribuční funkcí
- ověřovat podmínky dobré aproximace pro testy dobré shody
- pomocí jednoduchých testů testovat hypotézu, že daný náhodný výběr pochází z exponenciálního či Poissonova rozložení

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 8 hodin studia.

11.1. Motivace

Možnost použití statistických testů je podmíněna nějakými předpoklady o datech. Velmi často je to předpoklad o typu rozložení, z něhož získaná data pocházejí. Mnoho testů je založeno na předpokladu normality. (Testování normality bylo probráno ve 2. kapitole.) Opomíjení předpokladů o typu rozložení může v praxi vést i ke zcela zavádějícím výsledkům, proto je nutné věnovat tomuto problému patřičnou pozornost.

V této kapitole se seznámíme s testem dobré shody, který je (po splnění určitých předpokladů) použitelný k ověření shody empirického rozložení s jakýmkoliv teoretickým rozložením. Tato univerzálnost je ovšem provázena poněkud sníženou silou testu. Proto byly pro některá rozložení vyvinuty speciální testy využívající charakteristických vlastností těchto rozložení. Zde uvedeme tzv. jednoduché testy exponenciálního a Poissonova rozložení.

11.2. Testy dobré shody

11.2.1. Popis testu

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$.

a) Je-li distribuční funkce spojitá, pak data rozdělíme do r třídicích intervalů $\left(u_j, u_{j+1}\right\rangle$, $j = 1, ..., r$. Zjistíme absolutní četnost n_j j-tého třídicího intervalu a vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.

b) Má-li distribuční funkce nejvýše spočetně mnoho bodů nespojitosti, pak místo třídicích intervalů použijeme varianty $x_{[j]}$, $j = 1, …, r$. Pro variantu $x_{[j]}$ zjistíme absolutní četnost n_j a vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat variantou $x_{[j]}$. Platí-li nulová hypotéza, pak $p_j = \Phi\left(x_{[j]}\right) - \lim\limits_{x \to x_{[j]}-} \Phi(x) = P\left(X = x_{[j]}\right)$.

Testová statistika: $K = \sum_{j=1}^{r} \frac{\left(n_j - np_j\right)^2}{np_j}$. Platí-li nulová hypotéza, pak $K \approx \chi^2(r-1-p)$, kde p je počet odhadovaných parametrů daného rozložení. (Např. pro normální rozložení p = 2, protože z dat odhadujeme střední hodnotu a rozptyl.) Pokud žádný parametr nemusíme odhadovat, hovoříme o úplně specifikovaném problému. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}(r-1-p)$. Aproximace se považuje za vyhovující, když $np_j \geq 5$, j = 1, ..., r.

Upozornění: Při nesplnění podmínky $np_j \geq 5$, j = 1, ..., r je třeba některé intervaly resp. varianty slučovat, což vede ke ztrátě informace. Ve spojitém případě je hodnota testové statistiky K silně závislá na volbě třídicích intervalů

11.2.2. Příklad: (Testování shody empirického a teoretického rozložení při úplně specifikovaném problému)
Ze souboru rodin s pěti dětmi bylo náhodně vybráno 84 rodin a byl zjišťován počet chlapců:

Počet chlapců	0	1	2	3	4	5
Počet rodin	3	10	22	31	14	4

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozložení počtu chlapců se řídí binomickým rozložením Bi(5; 0,5).

Řešení:
Pravděpodobnost, že náhodná veličina s rozložením Bi(5; 0,5) bude nabývat hodnot $p_0, ..., p_5$ je $p_j = \binom{5}{j}\frac{1}{32}, j = 0,1,\ldots,5$.
Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j
0	3	0,03125	84.0,03125=2,625
1	10	0,15625	84.0,15625=13,125
2	22	0,3125	84.0,3125=26,25
3	31	0,3125	84.0,3125=26,25
4	14	0,15625	84.0,15625=13,125
5	4	0,03125	84.0,03125=2,625

Podmínky dobré aproximace nejsou splněny, sloučíme tedy první dvě varianty a poslední dvě varianty.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
0 a 1	13	0,1875	84.0,1875=15,75	0,480159
2	22	0,3125	84.0,3125=26,25	0,688095
3	31	0,3125	84.0,3125=26,25	0,859524
4 a 5	18	0,1875	84.0,1875=15,75	0,321429

Vypočteme realizaci testové statistiky: K = 0,48059 + 0,688095 + 0,859524 + 0,321429 = 2,3492, počet tříd r = 4, počet odhadovaných parametrů p = 0, r – p - 1 = 3, kritický obor $W = \langle \chi^2_{1-\alpha}(r-p-1), \infty) = \langle \chi^2_{0,95}(3), \infty) = \langle 7,8147; \infty)$. Protože $K \notin W$, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor se dvěma proměnnými a čtyřmi případy. Proměnná nj obsahuje zjištěné četnosti (po sloučení variant), proměnná npj pak teoretické četnosti.
Statistiky – Neparametrická statistika – Pozorované vs. očekávané $\chi 2$ – OK – Proměnné – Pozorované četnosti nj, očekávané četnosti npj – OK – Výpočet.

Případ	Pozorované vs. očekávané četnosti (Tabulka1 Chi-Kvadr. = 2,349206 sv = 3 p = ,503161			
	pozorov. nj	očekáv. npj	P - O	(P-O)^2 /O
C: 1	13,00000	15,75000	-2,75000	0,480159
C: 2	22,00000	26,25000	-4,25000	0,688095
C: 3	31,00000	26,25000	4,75000	0,859524
C: 4	18,00000	15,75000	2,25000	0,321429
Sčt	84,00000	84,00000	0,00000	2,349206

V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (2,349206), počet stupňů volnosti = 3 a p-hodnota (0,503161). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

11.2.3. Příklad: (Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému – diskrétní případ)
V tabulce jsou roztříděny fotbalové zápasy určité soutěže podle počtu vstřelených branek.

Počet branek	0	1	2	3	4 a víc
Počet zápasů	19	30	17	10	8

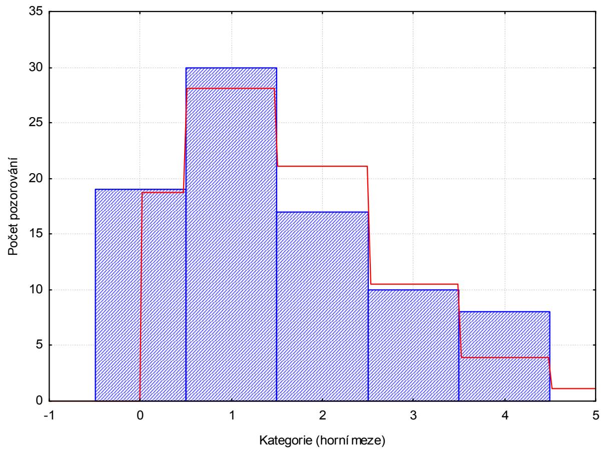
Na asymptotické hladině významnosti 0,05 testujte hypotézu, že jde o výběr z Poissonova rozložení.

Výpočet pomocí systému STATISTICA:
Vytvoříme datový soubor s dvěma proměnnými a 5 případy. Proměnná POCET obsahuje počet vstřelených branek, proměnná CETNOST pak počet zápasů, v nichž bylo dosaženo zjištěného počtu branek.
Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – - Výpočet.

Kategorie	Proměnná: POCET, Rozdělení:Poissonovo, Lambda = 1,500 (branky.sta) Chí-kvadrát = 2,07051, sv = 3, p = 0,55790							
	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 0,00000	19	19	22,61905	22,6190	18,74294	18,74294	22,31302	22,3130
1,00000	30	49	35,71429	58,3333	28,11440	46,85733	33,46952	55,7825
2,00000	17	66	20,23810	78,5714	21,08580	67,94313	25,10214	80,8847
3,00000	10	76	11,90476	90,4762	10,54290	78,48603	12,55107	93,4358
< Nekonečno	8	84	9,52381	100,0000	5,51397	84,00000	6,56424	100,0000

V tomto případě je parametr λ Poissonova rozložení neznámý, je odhadnut pomocí výběrového průměru a odhad činí 1,5. Podmínky dobré aproximace jsou splněny, dokonce všechny teoretické četnosti jsou větší než 5. Dále je v záhlaví výstupní tabulky uvedena hodnota testového kritéria (2,07051), počet stupňů volnosti r – p – 1 = 5 – 1 – 1 = 3 a p-hodnota (0,5578). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Pro vytvoření grafu se vrátíme do Proložení diskrétních rozložení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení.



11.2.4. Příklad: (Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému – spojitý případ)
U 48 studentek VŠE v Praze byla zjišťována výška (v cm):

165	170	170	179	170	168	174	162	167	165	170	173	183	176	165	168
171	178	168	168	169	163	172	184	176	175	176	169	168	170	166	160
167	162	162	166	170	168	155	162	169	166	160	169	165	163	168	163

Pomocí testu dobré shody testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí N-P grafu posuďte vizuálně předpoklad normality.

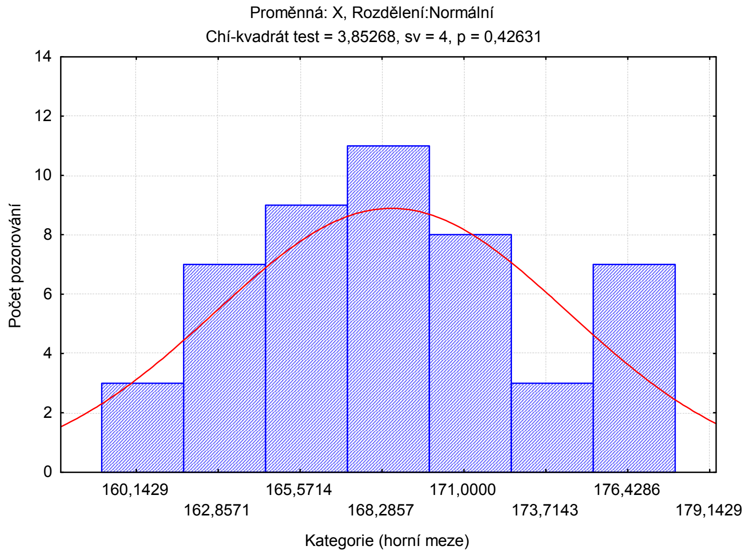
Výpočet pomocí systému STATISTICA:
Statistiky - Prokládání rozdělení – ponecháme implicitní nastavení na normální rozložení – OK – Proměnná X – OK – na záložce Parametry změníme Počet kategorií na 7 (podle Sturgesova pravidla) – Výpočet.

	Proměnná: X, Rozdělení:Normální (vyska.sta) Chí-kvadrát = 1,09280, sv = 1 (uprav.) , p = 0,29585							
Horní hranice	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 157,14286	1	1	2,08333	2,0833	1,19706	1,19706	2,49387	2,4939
162,28571	6	7	12,50000	14,5833	5,51484	6,71189	11,48924	13,9831
167,42857	12	19	25,00000	39,5833	13,46220	20,17409	28,04624	42,0293
172,57143	19	38	39,58333	79,1667	15,89146	36,06555	33,10721	75,1366
177,71429	6	44	12,50000	91,6667	9,07700	45,14255	18,91042	94,0470
182,85714	2	46	4,16667	95,8333	2,50365	47,64620	5,21594	99,2629
< Nekonečno	2	48	4,16667	100,0000	0,35380	48,00000	0,73708	100,0000

Při tomto roztřídění dat do 7 intervalů nejsou splněny podmínky dobré aproximace, ve třech intervalech jsou teoretické četnosti pod 5. Změníme tedy dolní mez na 159 a horní na 178.

Horní hranice	Proměnná: X, Rozdělení:Normální (vyska.sta) Chí-kvadrát = 3,85268, sv = 4, p = 0,42631							
	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 161,71429	3	3	6,25000	6,2500	5,722996	5,72300	11,92291	11,9229
164,42857	7	10	14,58333	20,8333	5,675946	11,39894	11,82489	23,7478
167,14286	9	19	18,75000	39,5833	7,862633	19,26157	16,38048	40,1283
169,85714	11	30	22,91667	62,5000	8,812455	28,07403	18,35928	58,4876
172,57143	8	38	16,66667	79,1667	7,991516	36,06555	16,64899	75,1366
175,28571	3	41	6,25000	85,4167	5,863558	41,92910	12,21575	87,3523
< Nekonečno	7	48	14,58333	100,0000	6,070896	48,00000	12,64770	100,0000

V tomto případě jsou podmínky dobré aproximace splněny. Testová statistika se realizuje hodnotou 3,85268, p-hodnota je 0,42631, tedy na asymptotické hladině významnosti 0,05 hypotézu o normalitě nezamítáme. Podívejme se ještě na histogram s proloženou Gaussovou křivkou: Na záložce Základní výsledky zvolíme Graf pozorovaného a očekávaného rozdělení.



11.3. Jednoduchý test exponenciálního a Poissonova rozložení

11.3.1. Jednoduchý test exponenciálního rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z exponenciálního rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim Ex(\lambda)$ je $E(X) = 1/\lambda$ a rozptyl je $D(X) = 1/\lambda^2$.

Test založíme na statistice $K = \frac{(n-1)S^2}{M^2}$, která se v případě platnosti H_0 asymptoticky řídí rozložením $\chi^2(n-1)$. Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty)$. Jestliže $K \in W$, H_0 zamítáme na asymptotické hladině významnosti α.

11.3.2. Příklad

Byla zkoumána doba životnosti 45 součástek (v hodinách). Průměrná životnost byla m = 99,93 a rozptyl $s^2 = 7328,91$. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z exponenciálního rozložení.

Řešení:

Testovou statistiku K vypočteme podle vzorce $K = \frac{(n-1)S^2}{M^2}$. Kritický obor má tvar: $W = \langle 0; \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1); \infty)$. V našem případě $K = 32{,}2924$, $W = \langle 0; 27{,}575\rangle \cup \langle 64{,}202; \infty)$, H_0 tedy nezamítáme na asymptotické hladině významnosti 0,05.

11.3.3. Jednoduchý test Poissonova rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z Poissonova rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim Po(\lambda)$ je $E(X) = \lambda$ a rozptyl je $D(X) = \lambda$. Test založíme na statistice $K = \frac{(n-1)S^2}{M}$, která se v případě platnosti H_0 asymptoticky řídí rozložením $\chi^2(n-1)$. Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty)$. Jestliže $K \in W$, H_0 zamítáme na asymptotické hladině významnosti α.

11.3.4. Příklad

Studujeme rozložení počtu pacientů, kteří během 75 dnů přijdou na pohotovost. Osmihodinovou pracovní dobu rozdělíme do půlhodinových intervalů a v každém intervalu zjistíme počet příchozích pacientů:

Počet pacientů	Pozorovaná četnost
0	79
1	188
2	282
3	275
4	196
5	114
6	45
7	10
8	7
9	3
10	1

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z Poissonova rozložení.

Řešení:

Celkový počet pacientů je n = 1200. Realizaci výběrového průměru M získáme jako vážený průměr počtu pacientů (m = 2,8033) a realizaci výběrového rozptylu S^2 získáme jako vážený rozptyl počtu pacientů ($s^2 = 2{,}7086$). Testovou statistiku vypočteme podle vzorce $K = \frac{(n-1)S^2}{M}$, tedy K = 1158,5, kritický obor

$$W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty) = \langle 0, \chi^2_{0,025}(1199)\rangle \cup \langle \chi^2_{0,975}(1199), \infty) =$$
$$= \langle 0; 1104{,}93\rangle \cup \langle 1296{,}86; \infty).$$

Protože testová statistika se nerealizuje v kritickém oboru, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Shrnutí

K ověření shody empirického rozložení s teoretickým rozložením se používají různé metody. Zvláštní postavení mezi nimi zaujímají metody zaměřené na ověřování normality dat. S nimi jsme se seznámili ve 2. kapitole.

Obecně je na ověření předpokladu o typu rozložení, z něhož pochází daný náhodný výběr, určen *chí-kvadrát test dobré shody*. Ten je založen na porovnání empirických četností jednotlivých variant či třídicích intervalů s tzv. *teoretickými četnostmi*. Velké odchylky mezi empirickými a teoretickými četnostmi vedou k velkým hodnotám testového kritéria a tudíž k zamítnutí nulové hypotézy. Test dobré shody lze aplikovat pouze při splnění *předpokladů dobré aproximace*.

Pro ověřování shody empirických dat s exponenciálním či Poissonovým rozložením byly vyvinuty *jednoduché testy*, které využívají pouze znalosti rozsahu výběru, výběrového průměru a výběrového rozptylu.

Kontrolní otázky

1. Popište provedení testu dobré shody pro náhodný výběr z diskrétního rozložení a pro náhodný výběr ze spojitého rozložení.
2. Jakým rozložením se asymptoticky řídí testová statistika testu dobré shody v případě platnosti nulové hypotézy?
3. Za jakých podmínek lze použít test dobré shody?
4. Popište jednoduchý test exponenciálního rozložení a Poissonova rozložení.

Autokorekční test

1. Při 600 hodech kostkou byly zjištěny tyto četnosti: 85 x jednička, 99 x dvojka, 91 x trojka, 108 x čtyřka, 119 x pětka, 98 x šestka. Příspěvek šestky do testové statistiky K je
a) 0,04
b) 0
c) 4

2. Uvažme zadání z otázky 1. Pokud je pravdivá hypotéza, že kostka je homogenní, pak testová statistika K se asymptoticky řídí chí-kvadrát rozložením s počtem stupňů volnosti
a) 4
b) 6
c) 5

3. Na základě náhodného výběru rozsahu 537 z diskrétního rozložení je na asymptotické hladině významnosti 0,01 testem dobré shody ověřována hypotéza, že tento výběr pochází z Poissonova rozložení, přičemž parametr λ není znám. V datech se vyskytuje 5 variant náhodné veličiny X. Kritický obor pro test nulové hypotézy má tvar:
a) $W = \langle 0;0,072 \rangle \cup \langle 12,838;\infty)$
b) $W = \langle 12,838;\infty)$
c) $W = \langle 9,348;\infty)$

4. Jednoduchým testem provedeným na hladině významnosti 0,05 chceme ověřit hypotézu, že náhodný výběr rozsahu 43 pochází z exponenciálního rozložení, přičemž výběrový průměr

nabyl hodnoty 20,2558 a výběrová směrodatná odchylka 22,5051. Testová statistika se realizuje hodnotou
a) 51,8457
b) 2,3037
c) 46,664

Správné odpovědi: 1a) 2c) 3b) 4c)

Příklady

1. Ve svých pokusech pozoroval J. G. Mendel 10 rostlin hrachu a na každé z nich počet žlutých a zelených semen. Výsledky pokusu:

č. rostliny	1	2	3	4	5	6	7	8	9	10
počet žlutých	25	32	14	70	24	20	32	44	50	44
počet zelených	11	7	5	27	13	6	13	9	14	18
celkem	36	39	19	97	37	26	45	53	64	62

Z genetických modelů vyplývá, že pravděpodobnost výskytu žlutého semene by měla být 0,75 a zeleného 0,25. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že výsledky Mendelových pokusů se shodují s modelem.
Výsledek:
Testová statistika K = 1,797495, kritický obor $W = \langle \chi^2_{0,95}(9), \infty) = \langle 16,9; \infty)$, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

2. Při 60 hodech kostkou jsme dosáhli těchto výsledků: 9 x jednička, 11 x dvojka, 10 x trojka, 13 x čtyřka, 11 x pětka a 6 x šestka. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že kostka je homogenní.
Výsledek:
Testová statistika K = 2,8, kritický obor $W = \langle \chi^2_{0,95}(5), \infty) = \langle 11,07; \infty)$, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

3. Ze záznamů autosalónu byl ve 100 náhodně vybraných dnech zjištěn počet prodaných aut.

Počet prodaných aut za den	0	1	2	3	4	5 a víc
Počet dnů	9	43	29	11	5	3

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že počet prodaných aut za den se řídí Poissonovým rozložením.
Výsledek:
Odhad parametru λ získaný pomocí výběrového průměru je 1,7.
Testová statistika K = 10,8891, kritický obor $W = \langle \chi^2_{0,95}(4), \infty) = \langle 9,488; \infty)$, nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05.

4. Při parlamentních volbách získaly 4 nejsilnější strany 30 %, 20 %, 15 % a 10 % hlasů, zbytek hlasů byl rozdělen mezi ostatní strany. Při volbách do obecního zastupitelstva v jedné obci získaly zmíněné strany (ve stejném pořadí) 1400, 900, 900 a 600 hlasů z 5000

odevzdaných hlasů. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozložení hlasů při parlamentních a místních volbách (v uvedené obci) je stejné.
Výsledek:
Testová statistika K = 68,67, kritický obor $W = \langle \chi^2_{0,95}(4), \infty) = \langle 9,488; \infty)$, nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme prokázali, že rozložení hlasů při parlamentních volbách a volbách v uvedené obci se liší.