

19. Testy dobré shody

V mnoha případech jsme nuceni z daných údajů zjišťovat například typ rozdělení, nezávislost údajů, homogenitu dat. Takovéto možnosti nám nabízí metoda, která se jmenuje testy dobré shody. Ve své podstatě jde o rozšíření testů binomického rozdělení na tzv. multinomické rozdělení. Ukazuje se dále, že jistý speciální poměr vytvářený v průběhu tohoto testu je asymptoticky roven rozdělení χ^2 . Nejdříve uvedeme definici multinomického rozdělení.

Definice 19.1

Nechť $A_1, \dots, A_k$ jsou neslučitelné jevy, z nichž v průběhu náhodného pokusu musí nastat právě jeden. Nechť $P(A_i) = p_i > 0$, pro všechna i . Předpokládejme, že náhodný pokus opakujeme celkem n – krát, označme dále X_i počet výskytů jevu A_i v těchto n opakovaných náhodných pokusech, potom

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} \cdot p_1^{x_1} \dots p_k^{x_k} \quad (19.1)$$

pro nezáporná celá čísla $x_1, \dots, x_k$, jejichž součet je roven n . V ostatních případech je pravděpodobnost v (19.1) rovna nule.

Náhodná veličina daná výše uvedeným rozdělením se nazývá multinomická.

Poznámka 19.2

Z předchozí definice vyplývá, že pokud se budeme zabývat jen případem jednoho náhodného jevu A_i a jedné náhodné veličiny X_i (ostatní náhodné jevy shrneme do náhodného jevu rovnajícího se jejich sjednocení), dostaneme klasický případ binomického rozdělení $Bi(n, p_i)$.

Pro další vyšetřování je tato poznámka velmi důležitá, pomocí ní můžeme totiž sjistit charakteristiky jednotlivých náhodných veličin X_i . Podrobnosti jsou uvedeny v následujícím tvrzení.

Věta 19.3

V multinomickém rozdělení platí

$$E(X_i) = n \cdot p_i, \quad VAR(X_i) = n \cdot p_i \cdot (1 - p_i) \quad \text{pro } i = 1, \dots, k \quad (19.2)$$

a

$$\text{cov}(X_i, X_j) = -n \cdot p_i \cdot p_j \quad \text{pro } i \neq j \quad (19.3)$$

Důkaz:

Uveden například v [1].

Všimněme si, že náhodné hodnoty X_i nejsou podle (19.3) nezávislé (proč?).

Věta 19.4

Nechť platí předchozí označení, potom následující náhodná veličina

$$\chi^2 = \sum_{i=1}^k \frac{(X_i - n \cdot p_i)^2}{n \cdot p_i} \quad (19.4)$$

je asymptoticky χ_{k-1}^2 rozdělení.

Důkaz:

Uveden opět v [1].

Důležitá vlastnost náhodné veličiny uvedené ve vztahu (19.4) je to , že limitně rovna rozdělení χ^2_{k-1} . Veličinám X_i říkáme velmi často empirické četnosti a hodnotám $n \cdot p_i$ teoretické četnosti . Vzorec (19.4) je možno upravit ještě na jiný , pro výpočty užitečný tvar

$$\chi^2 = \frac{1}{n} \cdot \sum_{i=1}^k \frac{X_i^2}{p_i} - n \quad (19.5)$$

V tomto vzorci nezjistíme příspěvky jednotlivých členů do celkového součtu!.

Pomocí výše uvedených vzorců můžeme testovat hypotézu H_0 , že skutečné hodnoty pravděpodobností multinomického rozdělení jsou rovny právě číslům $p_1, \dots, p_k$, hypotéza H_1 je negací nulové hypotézy. Jestliže je součet χ^2 větší nebo roven $\chi^2_{k-1}(\alpha)$, zamítáme na hladině významnosti hypotézu H_0 . Způsob řešení si ukážeme na několika příkladech.

Při řešení je nutno mít na zřeteli to , že tento test je asymptotický , musí tedy být n dostatečně veliké. Obecně je uváděno , že $n \cdot p_i > 5$, pro každé $i = 1, \dots, k$.

Příklad 19.5

V programu Excel jsme pomocí generátoru náhodných čísel vygenerovali 100 náhodných čísel . Ověřte, zda jsou číslice v těchto číslech použity se stejnou pravděpodobností. Dále je uvedena tabulka četností jednotlivých číslic:

0	1	2	3	4	5	6	7	8	9
75	87	90	102	97	86	113	89	95	99

Řešení:

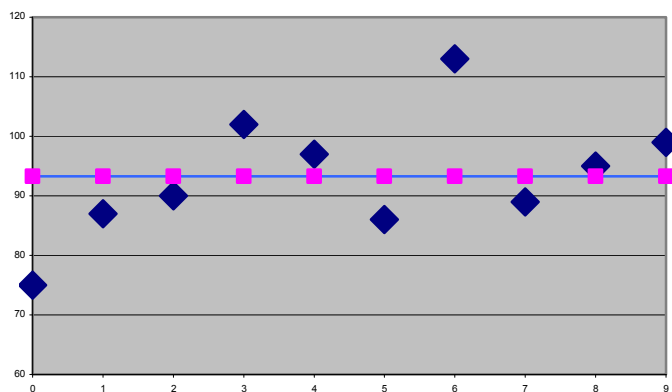
Nejprve si zjistíme celkový počet číslic , ten je roven 933. Tedy teoretická četnost pro jednotlivou číslici je rovna $933 / 10 = 93,3$.

Zjistíme nyní jednotlivé hodnoty sčítanců v (19.4) ,

0	1	2	3	4	5	6	7	8	9
75	87	90	102	97	86	113	89	95	99
3,589389068	0,4254	0,1167	0,8113	0,15	0,5712	4,16	0,1982	0,03098	0,3482

Odtud je vidět , že nejvíce přispívá do součtu cifra 0 , nejméně cifra 8 . Zjistíme nyní hodnotu $\chi^2 = 10,3976$. Kritická hodnota pro $10 - 1 = 9$ stupňů volnosti je rovna 16,91896 . Odtud vyplývá , že nemůžeme zamítnout hypotézu o stejném zastoupení cifer v náhodných číslech.

Dále jsou zobrazeny teoretické a empirické četnosti pro tento případ:



Ověřte pomocí metody χ^2 testu na hladině významnosti 0,05, zda údaje o výšce ve výběru 34 studentů je možno pokládat za výběr z normálního rozdělení $N(\mu, \sigma^2)$, kde $\mu = 178$ cm a $\sigma = 6$ cm.

rozsař třídy		počet
t_{i-1}	t_i	
167,5	172,5	1
172,5	177,5	6
177,5	182,5	11
182,5	187,5	9
187,5	192,5	4
192,5	197,5	3

Řešení:

Nulová hypotéza H_0 je stanovena tak , že výška studentů je typu $N(178;36)$. Nejdříve ztotožníme hodnoty v jednotlivých třídách se středy intervalů tříd (třídy jsou uzavřené !) . Pravděpodobnost , že výška náhodného studenta bude v j – tém intervalu vypočteme pomocí

vztahu $\Phi\left(\frac{t_j - \mu}{\sigma}\right) - \Phi\left(\frac{t_{j-1} - \mu}{\sigma}\right)$, kde Φ je distribuční funkce $N(0;1)$ a μ, σ^2 jsou parametry

vyšetřovaného rozdělení. V prvním intervalu položíme $\Phi\left(\frac{t_{j-1}-\mu}{\sigma}\right)=0$ a v posledním

intervalu bude $\Phi\left(\frac{t_j - \mu}{\sigma}\right) = 1$. Tím získáme hodnoty teoretických pravděpodobností, ty pak

násobíme celkovým počtem studentů a získáme teoretické četnosti. Výsledek je uveden v následující tabulce:

rozsah třídy		počet	normalizace	hodnota distribuční funkce $N(0;1)$		teoretická pravděpodobnost	teoretická četnost	úprava tříd
167,5	172,5	1	-0,916666667	0	0,179659	0,179658642	6,10839383	6,10839383
172,5	177,5	6	-0,083333333	0,179659	0,466793	0,28713454	9,76257435	9,76257435
177,5	182,5	11	0,75	0,466793	0,773373	0,306579539	10,4237043	10,4237043
182,5	187,5	9	1,583333333	0,773373	0,943327	0,169954522	5,77845376	7,70532751
187,5	192,5	4	2,416666667	0,943327	0,992168	0,048841084	1,66059685	
192,5	197,5	3	3,25	0,992168	1	0,007831673	0,2662769	

$\mu =$ **178 cm** počet= **34 studentů**

$$\sigma^2 = 36 \text{ cm}^2$$

Z tabulky je zřejmé, že musíme sloučit poslední tři třídy, protože teoretická četnost těchto tříd je menší než 5 prvků! Nyní již máme k dispozici hodnoty, které nám umožní spočítat hodnotu χ^2 - kvadrát:

rozsah třídy		počet	úprava tříd	chi - kvadrát
t_{j-1}	t_j			
167,5	172,5	1	6,108393825	4,272102982
172,5	177,5	6	9,76257435	1,450126292
177,5	182,5	11	10,42370431	0,031861679
182,5	187,5	9	7,705327511	8,929093748
187,5	192,5	4		
192,5	197,5	3		

součet	34	34	14,6831847
---------------	-----------	-----------	-------------------

Počítáme samozřejmě na hladině významnosti $\alpha = 0,05$, počet stupňů volnosti je tedy roven $n = 4 - 1 = 3$. Kritická hodnota je dána $\chi^2_{0,95;3} = 7,8147$. Je zřejmé, že hodnota testové statistiky vyšla v kritickém oboru, proto zamítáme nulovou hypotézu. Pokud bychom chtěli zjistit p – value tohoto testu získáme číslo 0,002109, které také velmi silně podporuje myšlenku zamítnutí nulové hypotézy.

Příklad 19.7

Předpokládejme, že máme klasickou hrací kostku, s níž provedeme celkem 300 hodů s četnostmi výsledků uvedenými v následující tabulce:

i	1	2	3	4	5	6
p _i	40	55	51	49	46	59

Zjistěte, zda daná kostka je homogenní?

Řešení:

Stanovíme nejprve nulovou hypotézu H_0 : kostka je homogenní a k ní alternativní H_1 : kostka není homogenní. Za předpokladu platnosti H_0 jsou očekávané četnosti jednotlivých hodů stejné a rovné 50. Zjistíme dále hodnotu chi – kvadrát:

i	f _i	očekávané četnosti	(naměřené četnosti - očekávané četnosti) ² / očekávané četnosti
1	40	50	2
2	55	50	0,5
3	51	50	0,02
4	49	50	0,02
5	46	50	0,32
6	59	50	1,62
			$\chi^2=4,48$

Vypočítanou hodnotu nyní porovnáme s kritickou hodnotou rozdělení chi – kvadrát na hladině významnosti $\alpha = 0,05$ s $n=6-1 = 5$ stupni volnosti. Tato hodnota je rovna 11,07. Protože výsledek testovací statistiky je v oboru přijetí, nemůžeme hypotézu H_0 na dané hladině zamítnout. Dokonce krajní mez po kterou bychom přijali hypotézu H_0 je v tomto případě rovna 0,44 (p – value).

Z výše uvedených výpočtů vyplývá, že pokud máme k dispozici výpočetní prostředek, který vypočítává p – value, je možno pomocí nich celkem jednoduše ověřit přijetí či zamítnutí nulové hypotézy.

Velmi častý je případ, kdy pravděpodobnosti $p_1, \dots, p_k$ závisí na určitém množství neznámých parametrů $a_1, \dots, a_l$ (jde samozřejmě o parametry neznámé náhodné veličiny určitého typu). Potom je zapotřebí tyto parametry z dat odhadnout. Někdy se za jednoduché aproximace těchto neznámých parametrů berou jejich bodové odhady, přesným určením postupu, jak získat skutečné hodnoty se zabývá autor v [1]. Pro naše účely bude tento postup vyhovovat.

V případě, že nebudeme základní parametry dané populace znát odhadneme je tedy pro naše účely pomocí bodových odhadů těchto parametrů (s optimálními vlastnostmi), tyto bodové odhady potom použijeme k tomu, abychom získali teoretické četnosti jednotlivých tříd, které v průběhu výpočtu používáme.

Příklad 19.8

Vyšetřovali jsme počet chlapců v rodinách s 5 dětmi . Výsledky jsou uvedené v následující tabulce:

Počet chlapců v rodinách s 5 dětmi	četnost rodin
0	3
1	10
2	22
3	31
4	14
5	4
Celkem rodin	84

Máme rozhodnout , zda tato data podporují myšlenku o tom , že počet chlapců je rozdělen v rodinách pomocí binomického rozdělení.

Řešení:

Protože nemáme stanoven parametr π binomického rozdělení , provedeme nejdříve jeho odhad : $\bar{X} = \frac{223}{84} = 2,65$, $n \cdot p = 2,65 \Rightarrow p = \frac{2,65}{5} = 0,53$. Tím jsme určili druhý nutný parametr binomického rozdělení , nyní můžeme zjistit hodnoty očekávané:

i	pi	očekávané četnosti	po korekci tříd	skutečné četnosti	po korekci	χ^2 - kvadrát
0	0,022935	1,926498059		3		
1	0,129312	10,86216991	12,78867	10	13	0,003492
2	0,291639	24,49765979	24,49766	22	22	0,254649
3	0,328869	27,62502061	27,62502	31	31	0,412325
4	0,185426	15,57580949	19,08865	14	18	0,062087
5	0,04182	3,512842141		4		
		84	84	84	84	0,732553

Hladinu významnosti máme opět stanovenou na 5% , nyní určíme počet stupňů volnosti $n=6 - 1 - 2 = 3$ (počet stupňů volnosti se snížil o počet parametrů , které odhaduji a o třídy, které jsme sloučily) . Budeme tedy hledat kritické hodnoty chí – kvadrát se 3 stupni volnosti . Tato hodnota je rovna 7,8147 . Protože je naše hodnota hluboko pod kritickou hodnotou , hypotézu H_0 nemůžeme zamítnout.

Příklad 19.9

V následující tabulce je uveden rozdělení počtu bodových vad zjištěných při zkoumání 20 videopásků. Rozhodněte , zda je možno považovat počet vad na videopásku za náhodnou veličinu typu Poissonovo rozdělení.

počet vad	0	1	2	3	4	5	6	7
počet pásků	2	3	2	4	2	2	4	1

Řešení:

Nejdříve musíme z daných dat odhadnout parametr λ – střední hodnotu Poissonova rozdělení. Tato střední hodnota je tedy rovna $\lambda = \bar{X} = \frac{68}{20} = 3,4$. Z této hodnoty již můžeme počítat teoretické četnosti.

počet vad	počet pásků	teoretické pravděpodobnosti	teoretické četnosti	po sloučení	praktické po sloučení	chi - kvadrát
0	2	0,03337327	0,667465399			
1	3	0,113469118	2,269382357			
2	2	0,1928975	3,857950007	6,794797764	7	0,006197
3	4	0,218617167	4,372343342			
4	2	0,185824592	3,71649184	8,088835182	6	0,539414
5	2	0,126360723	2,527214452	5,116367054	7	0,693475
6	4	0,071604409	1,432088189			
7	1	0,057853221	1,157064413			

$\lambda =$

3,4

celkem

1,239086

Jak je vidět, vzhledem k teoretickým četnostem je nutno sloučit třídy do celkem 3 nových tříd. Počet stupňů volnosti je tedy roven $n = 3 - 1 - 1 = 1$ (navíc odečítáme 1 za odhadnutý parametr rozdělení). Budeme tedy hledat kritickou hodnotu pro jeden stupeň volnosti a $\alpha = 0,05$. Tato kritická hodnota je rovna 3,8415, p -value = 0,265496. Tato hodnota je velmi vysoká, podporuje tedy nezamítnutí hypotézy H_0 o tom, že daná data pochází z Poissonova rozdělení.

Příklad 19.10

Ověřte na hladině významnosti $\alpha = 5\%$, zda můžeme považovat následující data za normálně rozdělená:

třída	naměřená četnost
pod 9	0
9 - 11	1
12 - 14	3
15 - 17	10
18 - 20	18
21 - 23	27
24 - 26	17
27 - 29	5
nad 29	0

Řešení:

Protože jde o data třídní, musíme nejdříve nahradit otevřené třídy dolními či horními odhady. Tedy dolní odhad bude roven 7,5 a horní odhad bude roven 30.

třída	naměřená četnost	střed tříd	normalizované střed	pravděpodobnosti středů tříd	pravděpodobnost tříd
pod 9	0	7,5	-3,544560185	0,000196675	0,000196675
9 - 11	1	10	-2,893518519	0,001904826	0,00170815
12 - 14	3	13	-2,112268519	0,017331651	0,015426825
15 - 17	10	16	-1,331018519	0,091591521	0,074259871
18 - 20	18	19	-0,549768519	0,291239045	0,199647524
21 - 23	27	22	0,231481481	0,591529572	0,300290527

24 - 26	17	25	1,012731481	0,844405771	0,252876199
27 - 29	5	28	1,793981481	0,96359199	0,119186219
nad 29	0	30	2,314814815	0,989688491	0,026096501

Dále :

třída	naměřená četnost	teoretické četnosti	teoretické četnosti upravené	skutečné četnosti upravené	chi
pod 9	0	0,01593071			
9 - 11	1	0,13836017			
12 - 14	3	1,24957281			
15 - 17	10	6,01504954	7,418913226	14	5,837877031
18 - 20	18	16,1714494	16,17144945	18	0,206759273
21 - 23	27	24,3235327	24,32353266	27	0,294508101
24 - 26	17	20,4829721	20,48297212	17	0,592252664
27 - 29	5	9,65408375	11,76790035	5	3,892323503
nad 29	0	2,1138166			

hodnota test. statistiky 10,82372057

odhad stř. hodnoty 21,1111111

odhad směr. odchylky 3,84

kvantil chi 95% 5,99

4 stupně volnosti

Z výpočtů vedených výše je zřejmé , že na hladině významnosti 5% hypotézu o tom , že data pochází z normálního rozdělení musíme zamítnout, protože ale výsledek není příliš průkazný , použili bychom ještě jiný test např. Kolmogor – Smirnovův test.