Cvičení 13: Porovnání empirického a teoretického rozložení

Úkol 1.: 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:
Testujeme H_0: náhodný výběr $X_1, \ldots, X_{84}$ pochází z Bi(5; 0,5) proti H_1: non H_0.

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.

Úkol 2.: Na jistém nádraží byl sledován počet přijíždějících vlaků za 1 h. Pozorování bylo prováděno celkem 15 dnů (tj. 360 h) a výsledky jsou uvedeny v tabulce:

Počet vlaků za 1 hodinu	0	1	2	3	4	5	6	7 a víc
četnost	27	93	103	58	50	21	6	2

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že počet přijíždějících vlaků za 1 h se řídí Poissonovým rozložením, a to a) testem dobré shody, b) jednoduchým testem Poissonova rozložení.

Řešení:
Testujeme H_0: náhodný výběr $X_1, \ldots, X_{360}$ pochází z Po(λ) proti H_1: non H_0.

Ad a) Nejprve odhadneme parametr λ Poissonova rozložení:

$$\hat{\lambda} = m = \frac{1}{n}\sum_{j=0}^{r} n_j x_{[j]} = \frac{1}{360}\left(27\cdot 0 + 93\cdot 1 + \ldots + 2\cdot 7\right) = 2{,}3$$

Pravděpodobnost, že náhodná veličina s rozložením Po(λ), kde λ = 2,3 bude nabývat hodnot 0, 1, ..., 7 a víc je $p_j = \frac{\lambda^j}{j!}e^{-\lambda} = \frac{2{,}3^j}{j!}e^{-2{,}3}$, $j = 0,1,\ldots,6$, $p_7 = 1-\left(p_0 + p_1 \ldots + p_6\right)$.

Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j
0	27	0,1003	36,0932
1	93	0,2306	83,0143
2	103	0,2652	95,4665
3	58	0,2033	73,1910
4	50	0,1169	43,0848
5	21	0,0538	19,3590
6	6	0,0216	7,4210
7 a víc	2	0,0094	3,3703

Podmínky dobré aproximace nejsou splněny, sloučíme tedy varianty 6 a 7 a víc.

j	n_j	p_j	np_j	$(n_j - np_j)^2 / np_j$
0	27	0,1003	36,0932	2,2909
1	93	0,2306	83,0143	1,2012
2	103	0,2652	95,4665	0,5945
3	58	0,2033	73,1910	3,1529
4	50	0,1169	43,0848	1,4887
5	21	0,0538	19,3590	0,1391
6 a víc	8	0,0300	10,7912	0,7220

K = 2,2909 + 1,2012 + … + 0,7220 = 9,5892, r = 7, p = 1, r – p – 1 = 5, $\chi^2_{0,95}(5) = 11,0705$. Protože 9,5892 < 11,0705, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05. Nepodařilo se tedy prokázat, že počty přijíždějících vlaků za 1 h se neřídí Poissonovým rozložením.

Ad b)
Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$$m = \frac{1}{360}(27 \cdot 0 + 93 \cdot 1 + \ldots + 2 \cdot 7) = 2,3$$

$$s^2 = \frac{1}{359}\left[27 \cdot (0 - 2,3)^2 + 93 \cdot (1 - 2,3)^2 + \ldots + 2 \cdot (7 - 2,3)^2\right] = 2,121448$$

Testová statistika: $K = \frac{(n-1)S^2}{M} = \frac{359 \cdot 2,121448}{2,3} = 331,1304$.

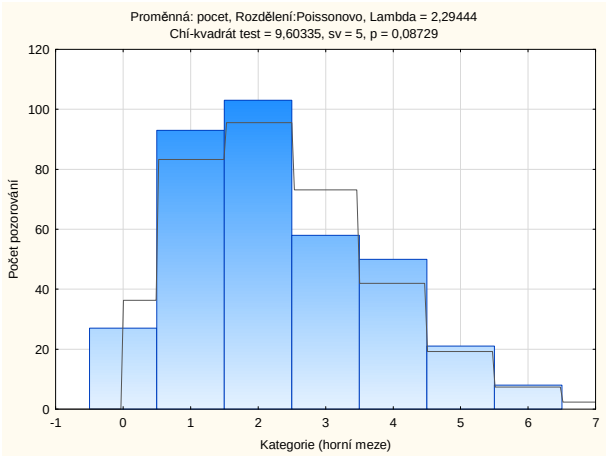
Kritický obor: $W = \langle 0, \chi^2_{0,025}(359)\rangle \cup \langle \chi^2_{0,975}(359), \infty) = \langle 0,308,4\rangle \cup \langle 413,4,\infty)$.

H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:
Načteme datový soubor vlaky.sta.
Ad a) Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – Výpočet.

	Proměnná: pocet, Rozdělení:Poissonovo, Lambda = 2,29444 (vlaky.sta) Chí-kvadrát = 9,60335, sv = 5, p = 0,08729							
Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 0,00000	27	27	7,50000	7,5000	36,29426	36,2943	10,08174	10,0817
1,00000	93	120	25,83333	33,3333	83,27517	119,5694	23,13199	33,2137
2,00000	103	223	28,61111	61,9444	95,53512	215,1045	26,53753	59,7513
3,00000	58	281	16,11111	78,0556	73,06667	288,1712	20,29630	80,0476
4,00000	50	331	13,88889	91,9444	41,91185	330,0831	11,64218	91,6897
5,00000	21	352	5,83333	97,7778	19,23288	349,3160	5,34247	97,0322
< Nekonečno	8	360	2,22222	100,0000	10,68405	360,0000	2,96779	100,0000

V tomto případě je parametr λ Poissonova rozložení neznámý, je odhadnut pomocí výběrového průměru a odhad činí 2,29444. Dále je v záhlaví výstupní tabulky uvedena hodnota testové statistiky (Chí kvadrát = 9,60335), počet stupňů volnosti r – p – 1 = 7 – 1 – 1 = 5 a p-hodnota (0,0879). 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.



Ad b) Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnná počet – OK – zapneme proměnnou vah cetnost – OK. Na záložce Detailní výsledky zaškrtneme Počet platn., Průměr, Rozptyl – Výpočet.
K výstupní tabulce přidáme za proměnnou Rozptyl tři nové proměnné, a to Test. stat., Kvantil1, Kvantil 2.
Do Dlouhého jména proměnné Test. stat. napíšeme:
=(v1-1)*v3/v2
Do Dlouhého jména proměnné Kvantil 1 napíšeme:
=VChi2(0,025;359)
Do Dlouhého jména proměnné Kvantil 2 napíšeme:
=VChi2(0,975;359)
Dostaneme výslednou tabulku:

	Popisné statistiky (vlaky.sta)					
Proměnná	N platných	Průměr	Rozptyl	Test.	Kvantil 1	Kvantil 2
pocet	360	2,294444	2,074621	324,6053	308,4009	413,3862

Vidíme, že testová statistika se nerealizuje v kritickém oboru $W = \langle 0;308,4 \rangle \cup \langle 413,4;\infty)$, tedy H_0 nezamítáme na asymptotické hladině významnosti 0,05.
(Malé rozdíly mezi ručním výpočtem a výpočtem ve STATISTICE plynou ze zaokrouhlovacích chyb.)

Úkol 3.: Jsou známy počty občanů města Brna podle měsíce narození (stav k 31.12.2001).

měsíc narození	počet osob
leden	32309
únor	30126
březen	35010
duben	34761
květen	34955
červen	32883
červenec	33255
srpen	31604
září	31173
říjen	30536
listopad	28571
prosinec	29467
celkem	384650

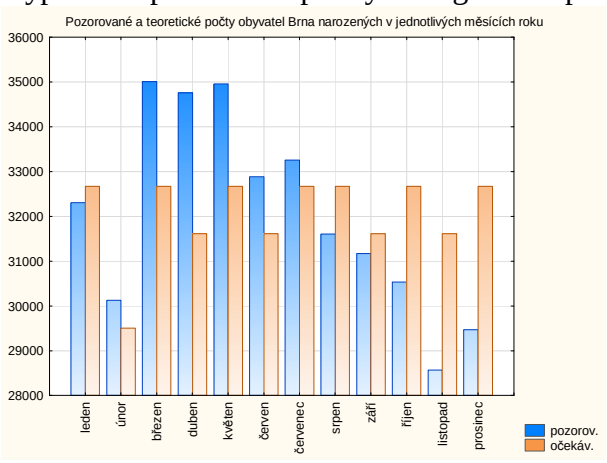
Na asymptotické hladině významnosti 0,05 ověřte hypotézu, že počty narozených jsou pro všechny měsíce stejné. Počty narozených lidí v jednotlivých měsících roku rovněž znázorněte graficky.

Výpočet pomocí systému STATISTICA:
Načteme datový soubor obyvatele_brna.sta. Tento soubor má tři proměnné (počet, délka měsíce a teor. počet) a 12 případů. Proměnná počet obsahuje absolutní četnosti z předchozí tabulky. Proměnná délka měsíce obsahuje počty dnů v jednotlivých měsících roku. Proměnná teor. počet obsahuje teoretické četnosti, tj. její hodnoty získáme tak, že do jejího Dlouhého jména napíšeme:
=384650/(365/v2)
Statistiky – Neparametrická statistika – Pozorované versus očekávané χ^2 – OK - Pozorované četnosti počet, Očekávané četnosti teor. počet - OK – Výpočet. Dostaneme tabulku:

	Pozorované vs. očekávané četnosti (obyvatele_brna.sta Chi-Kvadr. = 1506,153 sv = 11 p = 0,000000 POZN.: Nestejné součty pozor. a oček. četností			
Případ	pozorov. pocet	očekáv. teor. pocet	P - O	(P-O)^2 /O
C: 1	32309,0	32668,9	-359,90	3,965
C: 2	30126,0	29507,4	618,60	12,969
C: 3	35010,0	32668,9	2341,10	167,766
C: 4	34761,0	31615,1	3145,93	313,043
C: 5	34955,0	32668,9	2286,10	159,976
C: 6	32883,0	31615,1	1267,93	50,851
C: 7	33255,0	32668,9	586,10	10,515
C: 8	31604,0	32668,9	-1064,90	34,713
C: 9	31173,0	31615,1	-442,07	6,181
C: 10	30536,0	32668,9	-2132,90	139,254
C: 11	28571,0	31615,1	-3044,07	293,099
C: 12	29467,0	32668,9	-3201,90	313,821
Sčt	384650,0	384650,0	-0,00	1506,153

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}(r-1-p)$. V našem případě je r = 12, p = 0 a $\chi^2_{0,95}(11) = 19{,}675$. Protože $K = 1506{,}153 \geq 19{,}675$, zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5 % jsme prokázali, že obyvatelé Brna se rodí v průběhu roku nerovnoměrně.

Výpočet doplníme sloupkovým diagramem pozorovaných četností a očekávaných četností.



Komentář: Největší rozdíly mezi pozorovanými a očekávanými relativními četnostmi jsou v prosinci, dubnu a listopadu, naopak nejmenší v lednu a září.

Úkol 4.: Firma, která vlastní několik supermarketů, se zajímá, zda zákazníci dávají přednost některému dnu v týdnu pro nákup. Náhodně bylo vybráno 300 zákazníků, kteří měli říci, který den v týdnu nejčastěji nakupují v supermarketu.
Výsledky:

Den	pondělí	úterý	středa	čtvrtek	pátek	sobota	neděle
Počet	10	20	40	40	80	60	50

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že žádný den v týdnu nemá při nakupování v supermarketu přednost před jinými dny.

Návod:
Načteme datový soubor nakupy.sta. Proměnná X obsahuje pozorované absolutní četnosti a Y vypočítané teoretické četnosti (v našem případě 300/7).
Statistiky – Neparametrické statistiky – Pozorované vs. očekávané χ^2 – Proměnné Pozorované X, Očekávané Y, OK – Výpočet. Dostaneme tabulku:

	Pozorované vs. očekávané četnosti (nakupy.sta) Chi-Kvadr. = 78,00000 sv = 6 p = ,000000			
Případ	pozorov. X	očekáv. Y	P - O	(P-O)^2 /O
C: 1	10,0000	42,8571	-32,8571	25,19048
C: 2	20,0000	42,8571	-22,8571	12,19048
C: 3	40,0000	42,8571	-2,8571	0,19048
C: 4	40,0000	42,8571	-2,8571	0,19048
C: 5	80,0000	42,8571	37,1429	32,19048
C: 6	60,0000	42,8571	17,1429	6,85714
C: 7	50,0000	42,8571	7,1429	1,19048
Sčt	300,0000	300,0000	0,0000	78,00000

Komentář: Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Square = 78) a odpovídající p-hodnotu, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota velmi malá, takřka nulová, takže nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že zákazníci nakupují během týdne nerovnoměrně.

Příklad k samostatnému řešení: Do rybníka bylo umístěno 5 pastí, přičemž každá past svítila jiným světlem (bílým, žlutým, modrým, zeleným, červeným). Do těchto pastí se chytilo 56, 72, 41, 53 a 38 jedinců. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že barva světla v pasti nemá vliv na počet chycených jedinců.
Výsledek: Testová statistika nabývá hodnoty 14,1154, kritický obor je $W = \langle 9{,}488; \infty)$, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme. S rizikem omylu nejvýše 0,05 jsme prokázali, že barva světla v pasti má vliv na počet chycených jedinců.