Cvičení 10: 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:
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 χ2 – 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.: 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 histogramu posuďte vizuálně předpoklad normality.

Výpočet pomocí systému STATISTICA:

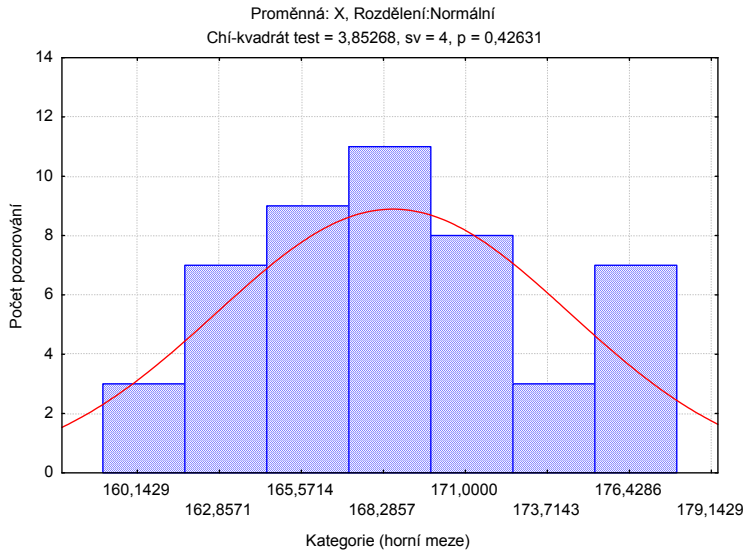
Načteme datový soubor vyska.sta. Statistiky - Prokládání rozdělení – ponecháme implicitní nastavení na normální rozložení – OK – Proměnná X – OK – na záložce Parametry změníme Počet kategorií na 7 (podle Sturgesova pravidla) – Výpočet.

Horní hranice	Proměnná: X, Rozdělení:Normální (vyska.sta) Chí-kvadrát = 1,09280, sv = 1 (uprav.) , p = 0,29585 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.



Samostatný úkol: Tentýž úkol proveďte zvlášť pro studentky oboru informatika a národní hospodářství.

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 pravděpodobnost narození je pro všechny měsíce stejná. (Pravděpodobnost narození pro libovolný měsíc získáte tak, že počet dnů v tomto měsíci podělíte počtem dnů v roce.) Počty narozených lidí v jednotlivých měsících roku rovněž znázorněte graficky.

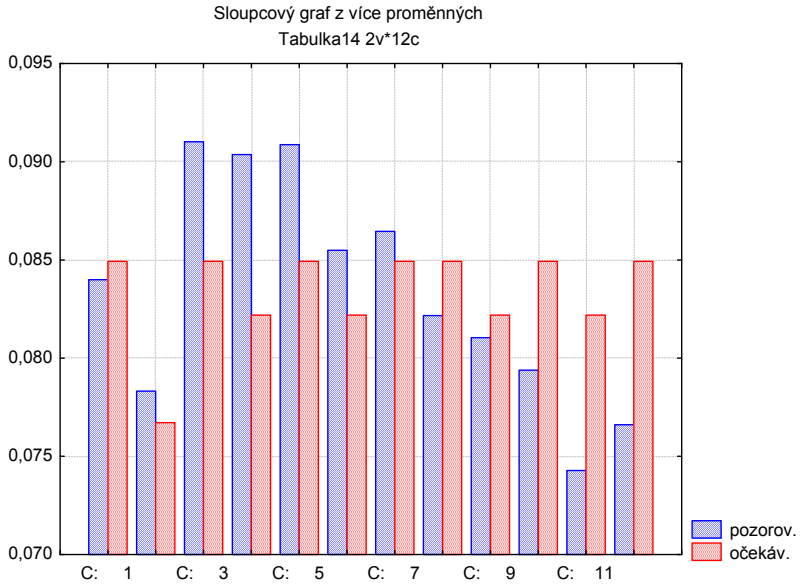
Výpočet pomocí systému STATISTICA:
Načteme datový soubor obyvatele_brna.sta. Tento soubor má tři proměnné (X, X1 a Y) a 12 případů. Proměnná X obsahuje absolutní četnosti z předchozí tabulky. Proměnné X1 obsahu-

je relativní četnosti, tj. v jejím Dlouhém jméně je napsáno = X/384650. Proměnná Y obsahuje očekávané relativní četnosti, tj. její hodnoty jsou vždy počet dní v měsíci/365.
Statistiky – Neparametrická statistika – Pozorované versus očekávané χ^2 – OK - Pozorované četnosti X1, Očekávané četnosti Y - OK – Výpočet. Dostaneme tabulku:

Případ	Pozorované vs. očekávané četnosti (obyvatele_Brna.sta Chi-Kvadr. = ,0039156 sv = 11 p = 1,000000			
	pozorov. X1	očekáv. Y	P - O	(P-O)^2 /O
C: 1	0,083996	0,084932	-0,000936	0,000010
C: 2	0,078321	0,076712	0,001608	0,000034
C: 3	0,091018	0,084932	0,006086	0,000436
C: 4	0,090370	0,082192	0,008179	0,000814
C: 5	0,090875	0,084932	0,005943	0,000416
C: 6	0,085488	0,082192	0,003296	0,000132
C: 7	0,086455	0,084932	0,001524	0,000027
C: 8	0,082163	0,084932	-0,002769	0,000090
C: 9	0,081043	0,082192	-0,001149	0,000016
C: 10	0,079386	0,084932	-0,005545	0,000362
C: 11	0,074278	0,082192	-0,007914	0,000762
C: 12	0,076607	0,084932	-0,008324	0,000816
Sčt	1,000000	1,000000	0,000000	0,003916

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. $\chi^2_{0,95}(11) = 19,675$. Protože K = 0,0039282 < 19,675,nezamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

Výpočet doplníme sloupkovým diagramem pozorovaných relativních četností a očekávaných relativních 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: D 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.