Cvičení 11: Hodnocení kontingenčních tabulek

Úkol 1.: Testování hypotézy o nezávislosti, měření síly závislosti
V roce 1950 zkoumali Yule a Kendall barvu očí a vlasů u 6800 mužů.

Barva očí	Barva vlasů			
	světlá	kaštanová	černá	rezavá
modrá	1768	807	180	47
šedá nebo zelená	946	1387	746	53
hnědá	115	438	288	16

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti barvy očí a barvy vlasů. Vypočtěte Cramérův koeficient. Simultánní četnosti znázorněte graficky.

Návod:
Testujeme hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny. Testová statistika má tvar:

$$K = \sum_{j=1}^{r}\sum_{k=1}^{s}\frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$$

. Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$, kde r, s jsou počty variant jednotlivých proměnných.
Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.
V našem případě zjistíme, že K = 1088,15, r = 3 , s = 4, $\chi^2_{1-\alpha}((r-1)(s-1)) = \chi^2_{0,95}(6) = 12{,}592$ a protože hodnota testové statistiky $K = 1088{,}15 \geq 12{,}592$, zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.
Cramérův koeficient: $V = \sqrt{\dfrac{K}{n(m-1)}}$, kde $m = \min\{r,s\}$. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi X a Y, čím blíže je 0, tím je tato závislost volnější.
Význam hodnot Cramérova koeficientu:
mezi 0 až 0,1 … zanedbatelná závislost,
mezi 0,1 až 0,3 … slabá závislost,
mezi 0,3 až 0,7 … střední závislost,
mezi 0,7 až 1 … silná závislost.

Otevřeme datový soubor oci_vlasy.sta. Před provedením testu je zapotřebí ověřit podmínky dobré aproximace: Statistiky – Základní statistiky/tabulky – Kontingenční tabulky - Specif. tabulky – List 1 OCI, List 2 VLASY, OK, Váhy - CETNOST, Stav zapnuto, OK – na záložce Možnosti zaškrtneme Očekávané četnosti – Výpočet.

Souhrnná tab.: Očekávané četnosti (oci_vlasy.sta)
Četnost označených buněk > 10
Pearsonův chí-kv. : 1088,15, sv=6, p=0,00000

OCI	VLASY světlá	VLASY kaštanová	VLASY černá	VLASY rezavá	Řádk. součty
modrá	1167,259	1085,976	500,902	47,8622	2802,000
šedá nebo zelená	1304,731	1213,875	559,895	53,4990	3132,000
hnědá	357,010	332,149	153,202	14,6388	857,000
Vš.skup.	2829,000	2632,000	1214,000	116,0000	6791,000

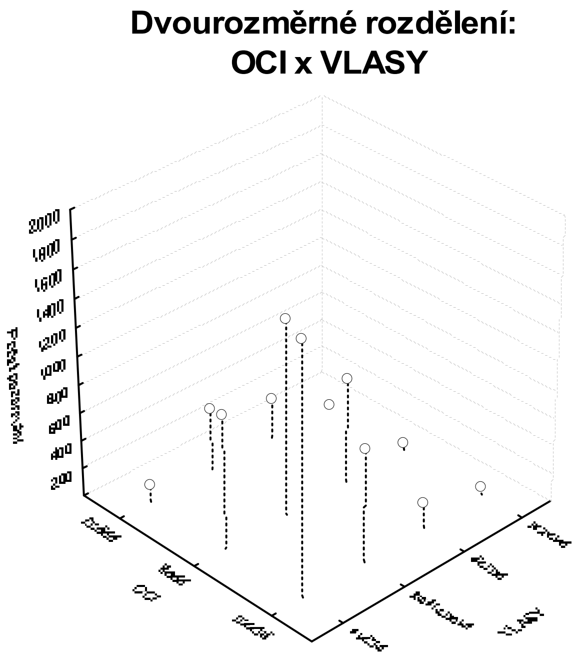
Podmínky dobré aproximace jsou splněny. Všechny teoretické četnosti jsou větší než 5. Nyní budeme testovat hypotézu o nezávislosti proměnných OCI, VLASY.

Návrat do Výsledky; kontingenční tabulky – na záložce Detaily zaškrtneme Pearsonův & M-L Chi - kvadrát, Phi & Cramerovo V – Detailní výsledky – Detailní 2 rozm. tabulky.

Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	1088,149	df=6	p=0,0000
M-V chí-kvadr.	1155,669	df=6	p=0,0000
Fí	,4002923		
Kontingenční koeficient	,3716246		
Cramér. V	,2830494		

Ve výstupní tabulce najdeme mj. hodnotu testové statistiky (Pearsonův chí-kv = 1088,149) s počtem stupňů volnosti (sv = 6) a odpovídající p-hodnotou (p = 0,0000), dále Cramérův koeficient (V = 0,283). Protože p-hodnota je mnohem menší než 0,05, nulovou hypotézu o nezávislosti barvy očí a barvy vlasů zamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient svědčí o slabé závislosti barvy očí a vlasů.

Pro grafické znázornění četností se vrátíme do Výsledky; kontingenční tabulky – Detailní výsledky – 3D histogramy. Po vytvoření grafu 2 krát poklepeme levým tlačítkem myši na pozadí grafu:
Rozvržení grafu – Typ Šipky – OK. Graf lze natáčet pomocí volby Zorný bod.



Úkol 2.: Fisherův faktoriálový test
100 náhodně vybraných osob bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

preferovaný nápoj	pohlaví	
	muž	žena
A	20	30
B	30	20

Na hladině významnosti 0,05 testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Návod: Vytvoříme nový datový soubor o třech proměnných NAPOJ, POHLAVI, CETNOST a čtyřech případech. Do proměnné NAPOJ napíšeme dvakrát pod sebe 1 (nápoj A) a dvakrát pod sebe 2 (nápoj B). Do proměnné POHLAVI napíšeme jedničku (1 – muž) a dvojku (2 – žena) a znovu jedničku a dvojku. D proměnné CETNOST napíšeme uvedené četnosti. Statistiky – Základní statistiky/tabulky – Kontingenční tabulky - Specif. tabulky – List 1 NAPOJ, List 2 POHLAVI, OK, Váhy - CETNOST, Stav zapnuto, OK – na záložce Možnosti zaškrtneme Fisher exakt, Yates, McNemar (2x2) – Detailní výsledky – Detailní 2-rozm. tabulky.

Statist.	Statist. : POHLAVI(2) x NAPOJ(2) (kap11_2)		
	Chí-kvadr.	sv	p
Pearsonův chí-kv.	4,000000	df=1	p=,04550
M-V chí-kvadr.	4,027103	df=1	p=,04478
Yatesův chí-kv.	3,240000	df=1	p=,07186
Fisherův přesný, 1-str.			p=,03567
2-stranný			p=,07134
McNemarův chí-kv. (A/D)	,0250000	df=1	p=,87437
(B/C)	,0166667	df=1	p=,89728

Ve výstupní tabulce je mimo jiné uvedena p-hodnota pro oboustranný a jednostranný test. V našem případě se jedná o oboustranný test (nevíme, zda muži více preferují nápoj A či nápoj B než ženy), zajímáme se tedy o Fisherův přesný, 2-str. Ta je 0,07134. Protože p-hodnota je větší než 0,05, nezamítáme na hladině významnosti 0,05 hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Úkol 3.: Podíl šancí

Pro údaje z úkolu 2 vypočtěte podíl šancí a sestrojte 95% asymptotický interval spolehlivosti pro podíl šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti 0,05 hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Návod: Nejprve zopakujme teorii:

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá podíl šancí (odds ratio). Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		$n_{j.}$
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
$n_{.k}$	a+c	b+d	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za 1. okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$. Podíl šancí je $OR = \frac{ad}{bc}$. Považujeme ho za odhad skutečného podílu šancí ορ. Pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro logaritmus skutečného podílu šancí ln ορ lze na asymptotické hladině významnosti α testovat hypotézu o nezávislosti nominálních veličin X a Y. Asymptotický 100(1-α)% interval spolehlivosti pro přirozený logaritmus skutečného podílu šancí má meze:

$\ln OR \pm \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2}$. Jestliže interval spolehlivosti nezahrne 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

V našem případě podíl šancí vypočteme ručně. $OR = \frac{ac}{bd} = \frac{20 \cdot 20}{30 \cdot 30} = \frac{4}{9} = 0,\overline{4}$. Dolní a horní mez intervalu spolehlivosti pro OR zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a dvou případech. Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:

=log(4/9)-sqrt(1/20+1/30+1/30+1/20)*VNormal(0,975;0;1)

a analogicky do Do Dlouhého jména proměnné HM napíšeme vzorec pro horní mez:

=log(4/9)+sqrt(1/20+1/30+1/30+1/20)*VNormal(0,975;0;1)

	1 DM	2 HM
1	-1,61108	-0,01078

Výsledek: -1,61108 < ln ορ < -0,01078 s pravděpodobností přibližně 0,95. Protože tento interval spolehlivosti neobsahuje 0**,** na asymptotické hladině významnosti 0,05 zamítáme hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.
Tento výsledek je v rozporu s výsledkem, ke kterému dospěl Fisherův přesný test. Je to způsobeno tím, že test pomocí asymptotického intervalu spolehlivosti je pouze přibližný.

Příklady k samostatnému řešení

Příklad 1.: Zajímá nás, zda má lokalita v ČR vliv na objem exportu do sousedních zemí. Sledujeme lokality: Ostrava, Brno, Plzeň, Praha a země: Slovensko, Rakousko, Německo, Polsko, USA). Máme k dispozici tato data:

	Kam:				
Odkud:	Slovensko	Rakousko	Německo	Polsko	USA
Ostrava	350	216	189	626	46
Brno	387	489	274	126	115
Plzeň	52	83	264	132	51
Praha	484	594	737	447	141

Řešení:
Načteme datový soubor export.sta. Proměnná EXPORT obsahuje objem exportu pro zvolenou kombinaci LOKALITA, ZEMĚ.

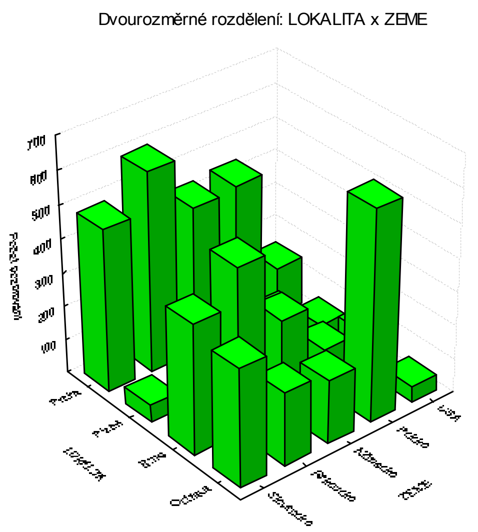
Testová statistika K ze vzorce (11.1) nabývá hodnoty 821,59, odpovídající p-hodnota je velmi blízká nule, tedy na asymptotické hladině významnosti 0,05 považujeme za prokázanou závislost objemu exportu na lokalitě v České republice. Podmínky dobré aproximace jsou splněny, jak vidíme z následující tabulky:

Souhrnná tab.: Očekávané četnosti (export.sta)
Četnost označených buněk > 10
Pearsonův chí-kv. : 821,587, sv=12, p=0,00000

LOKALITA	ZEME Slovensko	ZEME Rakousko	ZEME Německo	ZEME Polsko	ZEME USA	Řádk. součty
Ostrava	330,106	358,371	301,840	345,146	91,5375	1427,000
Brno	321,778	349,330	294,226	336,438	89,2282	1391,000
Plzeň	134,633	146,161	123,105	140,767	37,3335	582,000
Praha	486,484	528,138	444,829	508,649	134,9008	2103,000
Vš.skup.	1273,000	1382,000	1164,000	1331,000	353,0000	5503,000

Cramérův koeficient nabývá hodnoty 0,223, tedy mezi sledovanými proměnnými existuje slabá závislost.

Zjištěná data ještě znázorníme graficky:



Příklad 2.: 200 respondentů, z nichž bylo 73 žen, hodnotilo úroveň jistého časopisu. 34 žen ji hodnotilo kladně, stejně jako 47 mužů. Ostatní respondenti se o úrovni časopisu vyjádřili záporně. Vypočtěte a interpretujte podíl šancí časopisu na kladné hodnocení a na asymptotické hladině významnosti 0,05 testujte pomocí asymptotického intervalu spolehlivosti pro podíl šancí hypotézu, že hodnocení úrovně časopisu nezávisí na pohlaví respondenta. Proveďte též Fisherův přesný test a vypočtěte Cramérův koeficient.

Řešení:

Sestavíme čtyřpolní kontingenční tabulku simultánních absolutních četností:

hodnocení časopisu	pohlaví respondenta		$n_{j.}$
	muž	žena	
kladné	47	34	81
záporné	80	39	119
$n_{.k}$	127	73	200

Kladné hodnocení časopisu pozorujeme u 37% mužů a u 46,6 % žen.
Vypočítáme podíl šancí časopisu na kladné hodnocení.

$OR = \frac{ad}{bc} = \frac{47 \cdot 39}{34 \cdot 80} = 0,673897,$ což znamená, že u mužů je 0,674 x menší šance na kladné hodnocení časopisu než u žen.

Dále provedeme výpočty pro stanovení intervalu spolehlivosti.

$$\ln OR = -0,39468, \quad \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{47}+\frac{1}{34}+\frac{1}{80}+\frac{1}{39}} = 0,298, \; u_{0,975} = 1,96$$

$$\ln d = -0,39468 - 0,298 \cdot 1,96 = -0,97876, \ln h = -0,39468 + 0,298 \cdot 1,96 = 1,89476$$

Protože interval (-0,97876; 1,89476) obsahuje číslo 0, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti hodnocení úrovně časopisu na pohlaví respondenta.
Další výsledky máme v tabulce:

	Statist. : hodnoceni(2) x pohlavi(2) (Tabulka13)		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	1,760835	df=1	p=,18452
M-V chí-kvadr.	1,752654	df=1	p=,18555
Yatesův chí-kv.	1,386184	df=1	p=,23905
Fisherův přesný, 1-str.			p=,11967
2-stranný			p=,23131
McNemarův chí-kv. (A/D)	17,76316	df=1	p=,00003
(B/C)	,5697674	df=1	p=,45035
Fí pro tabulky 2 x 2	,0938306		
Tetrachorická korelace	,1507792		
Kontingenční koeficient	,0934202		

Fisherův přesný test poskytl p-hodnotu 0,23131, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti hodnocení úrovně časopisu na pohlaví respondenta. Cramérův koeficient je 0,0938, což svědčí o zanedbatelné závislosti mezi sledovanými veličinami.

Příklad 3.: Na hladině významnosti 0,05 testujte hypotézu o nezávislosti pedagogické hodnosti a pohlaví a vypočtěte Cramérův koeficient vyjadřující intenzitu závislosti pedagogické hodnosti na pohlaví, jsou-li k dispozici následující údaje:

pohlaví	pedagogická hodnost		
	odb. asistent	docent	profesor
muž	32	15	8
žena	34	8	3

Výsledek: Podmínky dobré aproximace jsou splněny, pouze jediná teoretická četnost klesne po 5. Testová statistika K nabývá hodnoty 3,5, p = 0,1739, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti pedagogické hodnosti a pohlaví. Cramérův koeficient: V = 0,187.

Příklad 4.: 18 mužů onemocnělo určitou chorobou. Někteří z nich se léčili, jiní ne. Někteří se uzdravili, jiní zemřeli. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

přežití	léčení	
	ano	ne
ano	5	3
ne	6	4

Vypočtěte a interpretujte podíl šancí. Pomocí intervalu spolehlivosti pro podíl šancí testujte na asymptotické hladině významnosti 0,05 hypotézu, že přežití nezávisí na léčení proti tvrzení, že léčení zvyšuje šance na přežití.

Výsledek: $\mathrm{OR} = 1{,}\bar{1}$, nulovou hypotézu nezamítáme asymptotické hladině významnosti 0,05, protože levostranný 95% asymptotický interval spolehlivosti pro logaritmus podílu šancí je (-1,49785; ∞).