Téma 10: Analýza závislosti dvou nominálních veličin

Úkol 1.: **Testování nezávislosti nominálních veličin**
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:
Načteme soubor oci_vlasy.sta. V proměnné CETNOST jsou uloženy zjištěné četnosti dvojic barev očí a vlasů, proměnná OCI má varianty 1 (modrá), 2 (šedá nebo zelená), 3 (hnědá) a proměnná VLASY má varianty 1 (světlá), 2 (kaštanová), 3 (černá), 4 (rezavá).
Pomocí STATISTIKY je možno lehce ověřit splnění podmínek dobré aproximace (viz Poznámka 11.2.2.1): teoretické četností mají být aspoň v 80% případů větší než 5 a ve zbylých 20% případů nemají klesnout pod 2.
Statistiky – Základní statistiky/tabulky – OK – Kontingenční tabulky - Specif. Tabulky – List 1 OCI, List 2 VLASY – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Očekávané četnosti. Dostaneme kontingenční tabulku teoretických četností:

Souhmná tab.: Očekáv ané četnosti (oci_v lasy .sta)
Četnost označených buněk > 10
Pearsonův chí-kv. : 1088,15, sv=6, p=0,00000

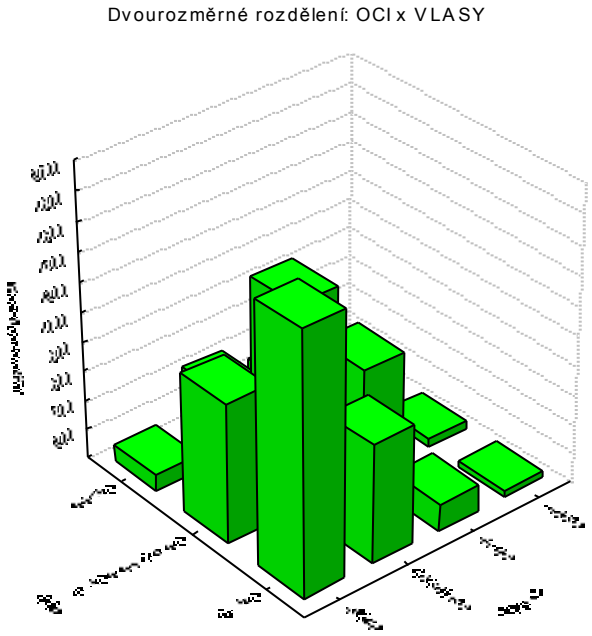
OCI	VLASY světlá	VLASY kaštanov á	VLASY černá	VLASY rezav á	Řádk. součty
modrá	1167,259	1085,976	500,902	47,8622	2802,000
šedá nebo zelená	1304,731	1213,875	559,895	53,4990	3132,000
hnědá	357,010	332,149	153,202	14,6388	857,000
Vš.skup.	2829,000	2632,000	1214,000	116,0000	6791,000

Vidíme, že všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny. V záhlaví tabulky je uvedena hodnota testové statistiky K = 1088,15, vypočtené podle vzorce (11.1.) (Pearsonův chí-kv.:1088,149) s počtem stupňů volnosti (sv = 6, protože r = 3, s = 4, tedy (r – 1)(s – 1) = 6) a odpovídající p-hodnotou (p = 0,0000). p-hodnota je velmi blízká 0, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o nezávislosti barvy očí a barvy vlasů.
Hodnotu testové statistiky a Cramérův koeficient dostaneme také tak, že na záložce Možnosti zaškrtneme Pearsonův & M-V chí kvadrát, Cramérovo V, na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

Statist.	Statist. : OCI (3) x VLASY(4) (oci_vlasy.sta)		
	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		

Cramérův koeficient svědčí o poměrně slabém vztahu mezi barvou očí a barvou vlasů.

Pro grafické znázornění četností se vrátíme do Výsledky: kontingenční tabulky – Detailní výsledky – 3D histogramy. Graf lze natáčet pomocí Zorný bod – Automatická rotace.



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

pohlaví	typ nápoje	
	A	B
muž	20	30
žena	30	20

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

Návod:
Načteme datový soubor napoje_A_B.sta. Proměnná POHLAVI nabývá hodnoty 1 pro muže, 2 pro ženu, proměnná TYP NAPOJE má hodnotu 1 pro typ A a hodnotu 2 pro typ B, proměnná CETNOST obsahuje zjištěné četnosti pro dvojice pohlaví a typ nápoje.
Statistiky – Základní statistiky/tabulky – OK – Kontingenční tabulky - Specif. Tabulky – List 1 POHLAVI, List 2 TYP NAPOJE – OK, zapneme proměnnou vah četnost – OK, Výpočet –

na záložce Možnosti zaškrtneme Fisher exakt. – Detailní výsledky – Detailní dvou-rozm.tabulky.

	Statist. : POHLAVI(2) x TYP NAPOJE(2) (napoje_A_B.sta)		
Statist.	Chí-kv adr.	sv	p
Pearsonův chí-kv.	4,000000	df=1	p=,04550
M-V chí-kv adr.	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

Komentář: Ve výstupní tabulce je mimo jiné uvedena p-hodnota pro oboustranný (Fisherův přesný, 2-stranný). Ta je 0,07134. Protože p-hodnota je větší než 0,05, nezamítáme na hladině významnosti hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Úkol 3.: **Podíl šancí**
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 podíl šancí a sestrojte 95% asymptotický interval spolehlivosti pro logaritmus podíl šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti 0,05 hypotézu, že přežití nezávisí na léčení.

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 bodový odhad teoretického podílu šancí $o\rho$. Pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro logaritmus teoretického podílu šancí lze na asymptotické hladině významnosti α testovat hypotézu o nezávislosti nominálních veličin X a Y. Asymptotický 100(1-α)% interval spolehlivosti pro přirozený lo-

garitmus teoretického 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 nezahrne interval spolehlivosti 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

V našem případě podíl šancí vypočteme ručně. $OR = \frac{ac}{bd} = \frac{5 \cdot 4}{3 \cdot 6} = \frac{20}{18} = \frac{10}{9} = 1,\overline{1}$. (Protože podíl šancí je větší než 1, je zřejmě výhodnější se nechat léčit.)
Dolní a horní mez intervalu spolehlivosti pro $\ln o\rho$ zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a dvou případech.
Do Long Name proměnné DM napíšeme vzorec pro dolní mez:
=log(10/9)-sqrt(1/5+1/3+1/6+1/4)*VNormal(0,975;0;1)
a analogicky do Long Name proměnné HM napíšeme vzorec pro horní mez:
=log(10/9)+sqrt(1/5+1/3+1/6+1/4)*VNormal(0,975;0;1)

	1 DM	2 HM
1	-1,80498	2,015697

Komentář: Zjistili jsme, že $-1,805 < \ln o\rho < 2,016$ s pravděpodobností přibližně 0,95. Protože tento interval spolehlivosti obsahuje 0**,** nelze na asymptotické hladině významnosti 0,05 zamítnout hypotézu, že přežití nezávisí na léčení.

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:

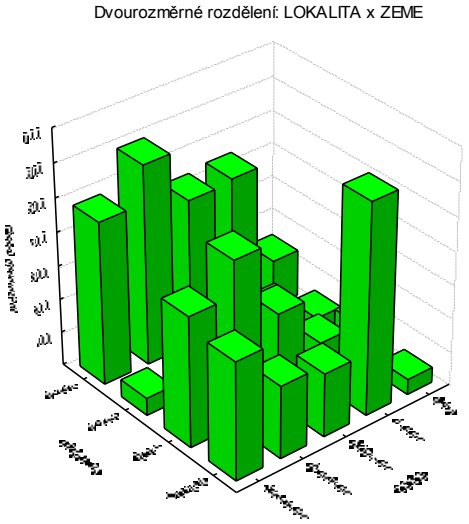
	Kam:				
Odkud:	Slovensko	Rakousko	Německo	Polsko	USA
Ostrava	350	216	189	626	46
Brno	387	489	274	126	115
Plzeň	52	83	264	132	51
Praha	484	594	737	447	141

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.

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

$$\ln d = -0{,}39468 - 0{,}298 \cdot 1{,}96 = -0{,}979, \; \ln h = -0{,}39468 + 0{,}298 \cdot 1{,}96 = 0{,}189$$

Protože interval (-0,979; 0,189) 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.