

Cvičení 10: Hodnocení závislosti dvou náhodných veličin

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. Vypočítejte Cramérův koeficient. Simultánní četnosti znázorněte graficky.

Návod:

Otevřeme datový soubor oci_vlasy.sta o 12 případech a třech proměnných (OCI, VLASY, CETNOST).

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

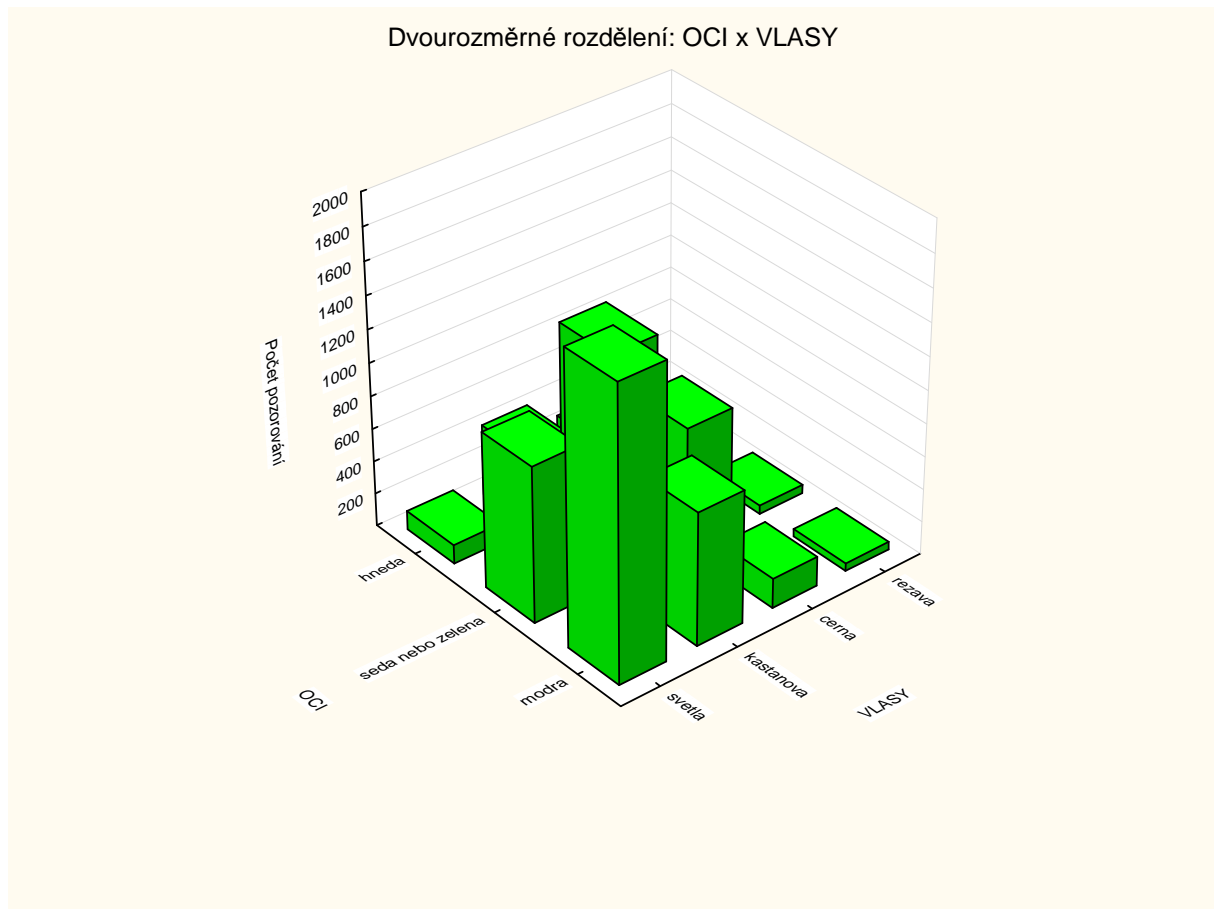
Podmínky dobré aproximace jsou splněny. Všechny teoretické četnosti jsou větší než 5. Nyní budeme testovat hypotézu o nezávislosti proměnných OCI, VLASY.

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. Graf lze natáčet pomocí volby Zorný bod.



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.

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 vypočítejte podíl šancí a sestrojte 95% asymptotický interval spolehlivosti pro logaritmus podílu šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti 0,05 hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Návod:

V našem případě podíl šancí vypočteme ručně: $OR = \frac{ac}{bd} = \frac{20 \cdot 20}{30 \cdot 30} = \frac{4}{9} = 0,4\bar{4}$.

Dolní a horní mez intervalu spolehlivosti pro logaritmus podílu šancí 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 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: $-1,61108 < \ln op < -0,01078$ 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.

Úkoly k samostatnému řešení:

Příklad 1.: 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

Vypočtete a interpretujte podíl šancí.

Výsledek: OR = 1,1̄.

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. Na hladině významnosti 0,05 testujte pomocí Fisherova přesného testu, že hodnocení úrovně časopisu nezávisí na pohlaví respondenta. Vypočtete Cramérův koeficient.

Výsledek: 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.

Další výsledky máme v tabulce:

Statist.	Statist. : hodnoceni(2) x pohlavi(2) (Tabulka13)		
	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.

Úkol 4.: Testování nezávislosti ordinálních veličin

12 různých softwarových firem nabízí speciální programové vybavení pro vedení účetnictví. Jednotlivé programy byly posouzeny odbornou komisí složenou z počítačových odborníků a komisí složenou z profesionálních účetních. Úkolem bylo doporučit vhodný program na základě stanovení pořadí jednotlivých programů. Výsledky posouzení:

Produkt firmy číslo	1	2	3	4	5	6	7	8	9	10	11	12
Pořadí dle odborníků	6	7	1	8	4	2,5	9	12	10	2,5	5	11
Pořadí dle účetních	4	5	2	10	6	1	7	11	8	3	12	9

Vypočtete Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou komisí jsou nezávislá.

Návod:

Testujeme vlastně nulovou hypotézu, že koeficient pořadové korelace je roven nule proti oboustranné alternativě.

Otevřeme datový soubor hodnoceni_programu.sta o dvou proměnných X (hodnocení 1. komise), Y (hodnocení 2. komise) a 12 případech.

Statistiky – Neparametrické statistiky – Korelace – OK – vybereme Vytvořit detailní report - Proměnné X, Y – OK – Spearmanův koef. R. Dostaneme tabulku

Dvojice proměnných	Spearmanovy korelace (Tabulka1)			
	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	12	0,714537	3,229806	0,009024

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,7145, testová statistika se realizuje hodnotou 3,2298, odpovídající p-hodnota je 0,009024, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení dvou komisí ve prospěch oboustranné alternativy.

Upozornění: Systém STATISTICA používá při testování hypotézy o pořadové nezávislosti veličin X, Y asymptotickou variantu testu bez ohledu na rozsah náhodného výběru. Pokud rozsah výběru nepřesáhne 20, měli bychom systém STATISTICA použít jen k výpočtu r_s a testování bychom měli provést pomocí tabelované kritické hodnoty. V našem případě pro $n = 12$ a $\alpha = 0,05$ je kritická hodnota 0,5804. Vidíme, že nulovou hypotézu zamítáme na hladině významnosti 0,05, protože $0,7145 \geq 0,5804$.

Příklad k samostatnému řešení: Bylo sledováno 10 žáků. Na základě psychologického vyšetření byli tito žáci seřazeni podle nervové lability (čím byl žák labilnější, tím dostal vyšší pořadí R_i). Kromě toho sledování žáci dostali pořadí Q_i na základě svých výsledků v matematice (nejlepší žák v matematice dostal pořadí 1). Výsledky jsou uvedeny v tabulce:

Pořadí R_i	1	2	3	4	5	6	7	8	9	10
Pořadí Q_i	9	3	8	5	4	2	10	1	7	6

Vypočtete Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že nervová lability a výsledky v matematice jsou nezávislé.

Výsledek: $r_s = -0,127$, H_0 nezamítáme na hladině významnosti 0,05.

Úkol 5.: Testování nezávislosti intervalových a poměrových veličin

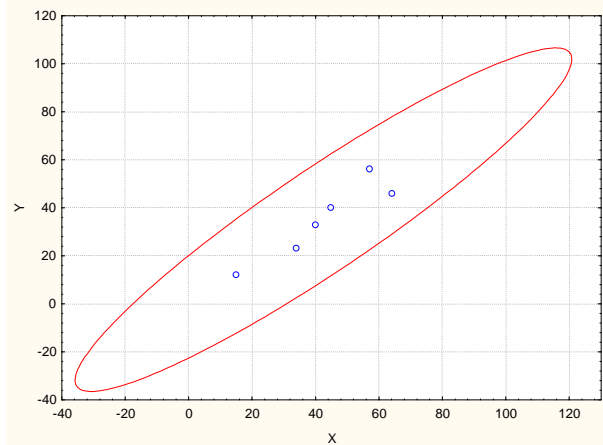
Zjišťovalo se, kolik mg kyseliny mléčné je ve 100 ml krve matek prvorodiček (veličina X) a u jejich novorozenců (veličina Y) těsně po porodu. Byly získány tyto výsledky:

Číslo matky	1	2	3	4	5	6
x_i	40	64	34	15	57	45
y_i	33	46	23	12	56	40

Nakreslete dvourozměrný tečkový diagram, vypočtete výběrový korelační koeficient a na hladině významnosti 0,05 testujte hypotézu o nezávislosti výsledků obou měření.

Návod: Otevřeme datový soubor kyselina_mlecna.sta. Zobrazíme dvourozměrný tečkový diagram, s jehož pomocí posoudíme dvourozměrnou normalitu dat.

Grafy – Bodové grafy – vypneme lineární proložení - Proměnné X, Y – OK – Details - Elipsa normální – OK. Ve vzniklém grafu upravíme měřítka na vodorovné a svislé ose:



Testování hypotézy o nezávislosti: Statistika – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Korelace (Tabulka3)											
Označ. korelace jsou významné na hlad. $p < ,05000$											
(Celé případy vynechány u ChD)											
Prom. X & prom. Y	Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	42,50000	17,39828									
Y	35,00000	15,89969	0,934832	0,873912	5,265339	0,006232	6	-1,30823	0,854311	6,696994	1,022943

Ve výstupní tabulce je mj. hodnotu výběrového korelačního koeficientu R_{12} ($r=0,9348$), tzn. že mezi X a Y existuje silná přímá lineární závislost), hodnota testové statistiky ($t = 5,2653$) a p-hodnotu pro test hypotézy o nezávislosti ($p=0,006232$), H_0 tedy zamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že mezi oběma koncentracemi existuje závislost.

Příklad k samostatnému řešení: V náhodném výběru 10 dvoučlenných domácností byl zjišťován měsíční příjem (veličina X, v tisících Kč) a vydání za potraviny (veličina Y, v tisících Kč).

x_i	15	21	34	35	39	42	58	64	75	90
y_i	3	4,5	6,5	6	7	8	9	8	9,5	10,5

Vypočtete výběrový koeficient korelace. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti veličin X, Y. (Data jsou uložena v souboru příjem_vydání.sta).

Výsledek: $r_{12} = 0,9405$, H_0 zamítáme na hladině významnosti 0,05.