12. Testování nezávislosti náhodných veličin

12.1. Motivace (vysvětlení pojmů nominální, ordinální, intervalová a poměrová náhodná veličina, motivace k testování nezávislosti)

12.2. Definice (definice kontingenční tabulky)

12.3. Věta (věta o testové statistice K)

12.4. Poznámka (podmínky dobré aproximace)

12.5. Definice (definice Cramérova koeficientu, význam jeho hodnot: 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.)

12.6. Příklad: V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází a typ školy, na kterou se hlásí. Výsledky jsou zaznamenány v kontingenční tabulce:

Typ školy	Sociální skupina				$n_{j.}$
	I	II	III	IV	
univerzitní	50	30	10	50	140
technický	30	50	20	10	110
ekonomický	10	20	30	50	110
$n_{.k}$	90	100	60	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérův koeficient.

Řešení:

Nejprve vypočteme všech 12 teoretických četností:

$$\frac{n_{1.}n_{.1}}{n}=\frac{140\cdot 90}{360}=35,\ \frac{n_{1.}n_{.2}}{n}=\frac{140\cdot 100}{360}=38{,}9,\ \frac{n_{1.}n_{.3}}{n}=\frac{140\cdot 60}{360}=23{,}3,\ \frac{n_{1.}n_{.4}}{n}=\frac{140\cdot 110}{360}=42{,}8,$$

$$\frac{n_{2.}n_{.1}}{n}=\frac{110\cdot 90}{360}=27{,}5,\ \frac{n_{2.}n_{.2}}{n}=\frac{110\cdot 100}{360}=30{,}6,\ \frac{n_{2.}n_{.3}}{n}=\frac{110\cdot 60}{360}=18{,}3,\ \frac{n_{2.}n_{.4}}{n}=\frac{110\cdot 110}{360}=33{,}6,$$

$$\frac{n_{2.}n_{.1}}{n}=\frac{110\cdot 90}{360}=27{,}5,\ \frac{n_{2.}n_{.2}}{n}=\frac{110\cdot 100}{360}=30{,}6,\ \frac{n_{2.}n_{.3}}{n}=\frac{110\cdot 60}{360}=18{,}3,\ \frac{n_{2.}n_{.4}}{n}=\frac{110\cdot 110}{360}=33{,}6.$$

Vidíme, že podmínky dobré aproximace jsou splněny, všechny teoretické četnosti převyšují číslo 5.

Nyní dosadíme do vzorce pro testovou statistiku K:

$K=\frac{(50-35)^2}{35}+\frac{(30-38{,}9)^2}{38{,}9}+\ldots+\frac{(50-33{,}6)^2}{33{,}6}=76{,}84$, r = 3, s = 4, $\chi^2_{0,95}(6)=12{,}6$. Protože $K \geq 12{,}6$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient: $V=\sqrt{\frac{76{,}4}{360\cdot 2}}=0{,}3267$. Hodnota Cramérova koeficientu svědčí o tom, že mezi veličinami X a Y existuje středně silná závislost.

12.7. Definice (definice čtyřpolní kontingenční tabulky)

12.8. Věta (věta o testové statistice K pro čtyřpolní tabulky)

12.9. Poznámka: U čtyřpolní KT lze rovněž použít následující podmínky dobré aproximace: $a + b > 5$, $c + d > (a + c)/3$.

12.10. Příklad: U 135 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

přijetí	dojem		$n_{j.}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{.k}$	56	69	125

Řešení:
Ověříme splnění podmínek dobré aproximace:
$a + b = 28 > 5$, $c + d = 97 > (a + c)/3 = 56/3 = 18,66$ – v pořádku
Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{125 \cdot (17 \cdot 58 - 11 \cdot 39)^2}{28 \cdot 97 \cdot 56 \cdot 69} = 3,6953$$

Kritický obor: $W = \langle \chi^2_{0,95}(1), \infty) = \langle 3,841, \infty)$.

Protože testová statistika se nerealizuje k kritickém oboru, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

12.11. Definice (definice podílu šancí)

12.12. Věta (asymptotický interval spolehlivosti pro podíl šancí a jeho využití k testování hypotézy o nezávislosti)

12.13. Příklad: Pro údaje z příkladu 12.10. vypočtěte a interpretujte podíl šancí, sestrojte 95% asymptotický interval spolehlivosti pro podíl šancí a s jeho pomocí testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.
Řešení:

$OR = \frac{ad}{bc} = \frac{17 \cdot 58}{11 \cdot 39} = 2,298$. Podíl šancí nám říká, že uchazeč, který zapůsobil na komisi dobrým dojmem, má asi 2,3 x větší šanci na přijetí než uchazeč, který zapůsobil špatným dojmem. Provedeme další pomocné výpočty:

$\ln OR = 0,832,$

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}} = 0,439,\ u_{0,975} = 1,96$$

Dosadíme do vzorců pro meze asymptotického intervalu spolehlivosti pro podíl šancí:

$$\ln d = \ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} = 0,832 - 0,439 \cdot 1,96 = -0,028$$

$$\ln h = \ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} = 0,832 + 0,439 \cdot 1,96 = 1,692$$

Po odlogaritmování dostaneme:

$d = e^{-0,028} = 0,972, h = e^{1,692} = 5,433$

Protože interval (0,972; 5,433) obsahuje číslo 1, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

12.14. Poznámka: Pro čtyřpolní tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako Fisherův faktoriálový test. (Je popsán např. v knize K. Zvára: Biostatistika, Karolinum, Praha 1998.) Jestliže p-hodnota pro tento test $\leq \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.

12.15. Definice (definice Spearmanova koeficientu pořadové korelace, význam jeho hodnot)

12.16. Věta (věta o testování hypotézy o pořadové nezávislosti veličin X, Y)

12.17. Věta (asymptotická varianta testu)

12.18. Příklad: Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku. Postupovali tak, že nejvyšší pořadí dostal nejtěžší případ.

Číslo pacienta	1	2	3	4	5	6	7
Hodnocení 1. lékaře	4	1	6	5	3	2	7
Hodnocení 2. lékaře	4	2	5	6	1	3	7

Vypočtěte Spearmanův koeficient r_S a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.

Řešení:

$$r_S = 1 - \frac{6}{7\left(7^2 - 1\right)}\left[(4-4)^2 + (1-2)^2 + (6-5)^2 + (5-6)^2 + (3-1)^2 + (2-3)^2 + (7-7)^2\right] = 0,857.$$

Kritická hodnota: $r_{S,0,95}(7) = 0,745$. Protože $0,857 \geq 0,745$, nulovou hypotézu zamítáme na hladině významnosti 0,05.

12.19. Definice (definice Pearsonova koeficientu korelace)

12.20. Věta (věta o vlastnostech koeficientu korelace)

12.21. Definice (definice výběrového koeficientu korelace)

12.22. Poznámka: Vlastnosti Pearsonova koeficientu korelace uvedené v 13.3. se přenášejí i na výběrový koeficient korelace.

12.23. Věta (věta o koeficientu korelace dvourozměrného normálního rozložení)

12.24. Věta (testování hypotézy o nezávislosti)

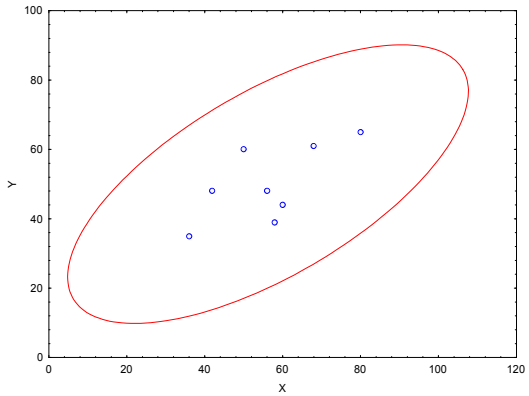
12.25. Příklad: Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Na hladině významnosti 0,05 testujte hypotézu, že výsledky obou testů nejsou kladně korelované.

Řešení:

Nejprve se musíme přesvědčit, že uvedené výsledky lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení. Lze tak učinit orientačně pomocí dvourozměrného tečkového diagramu. Tečky by měly vytvořit elipsovitý obrazec, protože vrstevnice hustoty dvourozměrného normálního rozložení jsou elipsy.



Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti.

Testujeme H_0: $\rho = 0$ proti pravostranné alternativě H_1: $\rho > 0$.

Výpočtem zjistíme: $R_{12} = 0{,}6668$, $T = 2{,}1917$. V tabulkách najdeme $t_{0,95}(6) = 1{,}9432$. Kritický obor: $W = \langle 1{,}9432; \infty)$. Protože $T \in W$, hypotézu o neexistenci kladné korelace výsledků z 1. a 2. testu zamítáme na hladině významnosti 0,05.

12.26. Věta (test o porovnání koeficientu korelace s danou konstantou)

12.27. Příklad: U 600 vzorků rudy byl stanoven obsah železa dvěma analytickými metodami s výběrovým koeficientem korelace 0,85. V literatuře se uvádí, že koeficient korelace těchto dvou metod má být 0,9. Na asymptotické hladině významnosti 0,05 testujte hypotézu H_0: $\rho = 0{,}9$ proti H_1: $\rho \neq 0{,}9$.

Řešení:

$$Z = \frac{1}{2}\ln\frac{1+0{,}85}{1-0{,}85} = 1{,}2562\,,\ U = \left(1{,}2562 - \frac{1}{2}\ln\frac{1+0{,}9}{1-0{,}9} - \frac{0{,}9}{2(600-1)}\right)\sqrt{600-3} = -5{,}2976\,,\ u_{0{,}975} = 1{,}96,\ W = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$$

. Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

12.28. Věta (test o porovnání dvou koreficientů korelace)

12.29. Příklad: Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých jedinců činil výběrový korelační koeficient mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že korelační koeficienty v obou skupinách se neliší.

Řešení:

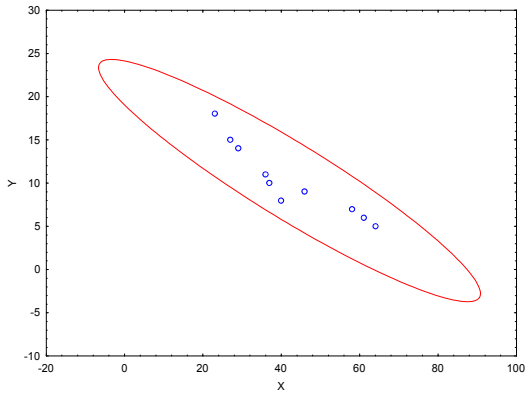
$Z = \frac{1}{2}\ln\frac{1+0{,}65}{1-0{,}65} = 0{,}7753,\ Z^* = \frac{1}{2}\ln\frac{1+0{,}37}{1-0{,}37} = 0{,}3884,\ U = \frac{0{,}7753-0{,}3884}{\sqrt{\frac{1}{100-3}+\frac{1}{142-3}}} = 2{,}9242,\ u_{0{,}975}$ = 1,96, $W = \langle -\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$. Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

12.30. Věta (věta o asymptotickém intervalu spolehlivosti pro koeficient korelace)

12.31. Příklad: Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi počtem dní absence za rok (veličina Y) a věkem pracovníka (veličina X). Proto náhodně vybral údaje o 10 pracovnících.

Č.prac.	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Za předpokladu, že uvedené údaje tvoří číselné realizace náhodného výběru rozsahu 10 z dvourozměrného normálního rozložení, vypočtěte výběrový korelační koeficient a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny. Sestrojte 95% asymptotický interval spolehlivosti pro skutečný korelační koeficient ρ.
Řešení:
Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu.



Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.
Testujeme H_0: ρ = 0 proti H_1: ρ ≠ 0. Vypočítáme R_{12} = -0,9325, tedy mezi věkem pracovníka a počtem dnů pracovní neschopnosti existuje silná nepřímá lineární závislost. Testová statistika: T = -7,3053, kvantil $t_{0{,}975}(8)$ = 2,306, kritický obor $W = \langle -\infty, -2{,}306\rangle \cup \langle 2{,}306, \infty)$. Jelikož $T \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y. Vypočítáme $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}} = \frac{1}{2}\ln\frac{1-0{,}9325}{1+0{,}9325} = -1{,}6772$. Meze 95% asymptotického intervalu spolehlivosti pro ρ jsou $\operatorname{tgh}\left(-1{,}6772 \pm \frac{1{,}96}{\sqrt{7}}\right)$, tedy -0,9842 < ρ < -0,7336 s pravděpodobností přibližně 0,95.