

1. Závislost dvou kategoriálních proměnných

A, B - dva kvalitativní znaky

$A : A_1, \dots, A_r, \quad r \geq 2, \quad B : B_1, \dots, B_s, \quad s \geq 2$

Kontingenční tabulka (Contingency table)

– roztrídění n jednotek statistického souboru podle variant *dvou kvalitativních* znaků A a B do $r.s$ tříd

r – počet řádků, s – počet sloupců

n_{ij} - sdružené četnosti udávají počet jednotek, u nichž se vyskytla kombinace variant A_i a A_j

$n_{i.}, n_{.j}$ - marginální (okrajové) četnosti

$$n_{i.} = \sum_{j=1}^s n_{ij}, \quad n_{.j} = \sum_{i=1}^r n_{ij}$$

- Kontingenční tabulka

$A \setminus B$	B_1	B_2	\dots	B_j	\dots	B_s	Σ_j
A_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2s}	$n_{2\cdot}$
\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
A_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{is}	$n_{i\cdot}$
\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
A_r	n_{r1}	n_{r2}	\dots	n_{rj}	\dots	n_{rs}	$n_{r\cdot}$
Σ_i	$n_{\cdot 1}$	$n_{\cdot 2}$	\dots	$n_{\cdot j}$	\dots	$n_{\cdot s}$	n

n_{ij} - sdružené četnosti

$n_{i\cdot}, n_{\cdot j}$ - marginální (okrajové) četnosti

- Chí-kvadrát test nezávislosti
(*The Chi-square independence test*)

A, B - dva kvalitativní znaky

$A : A_1, \dots, A_r, \quad r \geq 2, \quad B : B_1, \dots, B_s, \quad s \geq 2$

Náhodný pokus: výsledkem jedna z variant $A_i B_j$

Pravděpodobnost, že nastane kombinace $A_i B_j$:

$$\pi_{ij}, \quad i = 1, \dots, r, \quad j = 1, \dots, s, \quad \sum_{i=1}^r \sum_{j=1}^s \pi_{ij} = 1$$

If A, B nezávislé

$$\pi_{ij} = \pi_{i.} \pi_{.j}$$

$$\pi_{i.} = \sum_{j=1}^s \pi_{ij}, \quad \pi_{.j} = \sum_{i=1}^r \pi_{ij}$$

π_{ij} - sdružené psti (dvourozměrné)

$\pi_{i.}, \pi_{.j}$ - příslušné marginální psti

Chí-kvadrát test nezávislosti

H_0 : A a B jsou statisticky nezávislé

- **Předpoklady**

P1. $n_{ij}^o \geq 1 \quad \forall i = 1, \dots, r; j = 1, \dots, s$

P2. Nejvýše 20% $n_{ij}^o < 5$

- **Testová statistika: za platnosti H_0**

$$(G) \quad \chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_{ij}^o)^2}{n_{ij}^o} \approx \chi^2[\nu], \quad \nu = (r - 1)(s - 1)$$

n_{ij} – pozorované četnosti

n_{ij}^o – hypotetické četnosti

$$n_{ij}^o = \frac{n_{i.} \cdot n_{.j}}{n}, \quad i = 1, \dots, r, \quad j = 1, \dots, s$$

$$n_{i.} = \sum_{j=1}^s n_{ij}, \quad n_{.j} = \sum_{i=1}^r n_{ij}$$

- **Kritický obor:**

$$W_\alpha = \{\chi^2 : \chi^2 \geq \chi_{1-\alpha}^2(\nu)\}, \quad \nu = (r - 1)(s - 1)$$

Postup: (Chí-kvadrát test nezávislosti)

Předpoklady

P1. $n_{ij}^o \geq 1 \quad \forall i = 1, \dots, r; j = 1, \dots, s$

P2. Nejvýše 20% $n_{ij}^o < 5$

1. Formulujte H_0 a H_A
2. Vypočtete hypotetické četnosti n_{ij}^o
3. Ověřte předpoklady P1 a P2. Pokud nejsou splněny, test by neměl být použit.
4. Zvolte hladinu významnosti α .
5. Určete kritickou hodnotu:
 $\chi_{1-\alpha}^2[\nu]$, $\nu = (r - 1)(s - 1)$

Kritický obor:

$$W_\alpha = \{\chi^2 : \chi^2 \geq \chi_{1-\alpha}^2[\nu]\}, \quad \nu = (r - 1)(s - 1)$$

6. Vypočtete hodnotu testové statistiky

$$\chi_c^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_{ij}^o)^2}{n_{ij}^o}$$

7. If $\chi_c^2 \in W_\alpha$, zamítněte H_0 , jinak nezamítejte
8. Formulujte slovně závěr.

- Kontingenční koeficienty

(koeficienty kontingence)

– míry závislosti dvou kategoriálních proměnných, hodnotí intenzitu závislosti

- *Pearsonův kontingenční koeficient*

$$G_P = \sqrt{\frac{\chi^2}{n + \chi^2}} \in \langle 0, 1 \rangle$$

- *Cramérův kontingenční koeficient*

$$G_C = \sqrt{\frac{\chi^2}{nh}} \in \langle 0, 1 \rangle, \quad h = \min(r - 1, s - 1)$$

χ^2 je hodnota testové statistiky

Interpretace:

”čím je hodnota blíže jedné, tím je závislost silnější ”

Příklad (*Chí-kvadrát test nezávislosti*)

Máte k dispozici náhodný výběr 1367 absolventů vysokých škol, rozdělený následujícím způsobem:

Pohlaví	Stupeň vysokoškolského vzdělání			
	Bc	Mgr	Dr	Celkem
Muž	534	144	22	700
žena	515	141	11	667
Celkem	1049	285	33	1367

Rozhodněte, zda stupeň vzdělání závisí na pohlaví. Testujte na 5% hladině významnosti.

2. Závislost kvantitativní a kategoriální proměnné

• Jednofaktorová analýza rozptylu

Předpoklady:

$y_{i1}, y_{i2}, \dots, y_{in_i}$ – realizace k nezávislých náhodných výběrů z $N(\mu_i, \sigma)$, $i = 1, 2, \dots, k$.

Tabulka ANOVA (analýzy rozptylu)

Zdroj variability	SS	Df	MS	F
meziskupinový	$S_{y \cdot m}$	$k - 1$	$\frac{S_{y \cdot m}}{k - 1}$	$\frac{S_{y \cdot m} / (k - 1)}{S_{y \cdot v} / (n - k)}$
vnitroskupinový	$S_{y \cdot v}$	$n - k$	$\frac{S_{y \cdot v}}{n - k}$	
Celkový	S_y	$n - 1$		

$$S_{y \cdot m} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2, \quad S_{y \cdot v} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$S_y = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = S_{y \cdot m} + S_{y \cdot v}$$

$$n = \sum_{i=1}^k n_i, \quad \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$$

Test hypotézy o rovnosti středních hodnot

$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ versus $H_1 : \text{non } H_0$

- **Testová statistika:** za platnosti H_0

$$F = \frac{S_{y.m}/(k-1)}{S_{y.v}/(n-k)} \sim F[k-1, n-k]$$

$F[\nu_1; \nu_2]$ – Fisherovo-Snedecorovo rozdělení c ν_1 a ν_2 stupni volnosti

- **Kritický obor:**

$$F \geq F_{1-\alpha}(k-1, n-k)$$

Příklad: (*Závislost kvantitativní a kategoriální proměnné*)

Předpokládejme náhodný výběr 5 domácností pro každou ze 3 úrovní vzdělání hlavy rodiny, jejichž příjmy v roce 1985 byly následující (v tisících dolarů):

Poslední ukončené vzdělání	Příjmy				
Základní vzdělání	17	20	10	15	13
Střední škola	22	25	26	27	30
Vysoká škola	45	41	38	46	50

Rozhodněte, zda se příjmy domácností liší podle stupně vzdělání hlavy rodiny. Testujte na 5% hladině významnosti.

3. Závislost dvou kvantitavních proměnných

- Lineární korelace

”Existuje pozitivní korelace mezi výdaji za reklamu a prodejem výrobků.”

”IQ a spotřeba alkoholu nejsou korelované.”

- *Korelační koeficient* r_{xy} :

popisná míra síly lineárního (přímkového) vztahu mezi dvěma proměnnými.

Korelační koeficient dvou proměnných x a y :

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{yx}}{s_y s_x} = r_{yx} \in \langle -1, 1 \rangle$$

s_x, s_y – směrodatné odchylky veličin x resp. y

$$s_x = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}, \quad s_y = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

s_{xy} – kovariance

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

○ Interpretace korelačního koeficientu

$r_{xy} = 1$ – *přímá* lineární závislost x a y

$r_{xy} = -1$ – *nepřímá* lineární závislost x a y

$r_{xy} = 0$ – závislost x a y *není lineární*

r_{xy} blízké -1 nebo 1 – *silná lineární závislost*

r_{xy} blízké nule – *slabá lineární závislost* x a y

$r_{xy} > 0$ – *kladně lineárně korelované* veličiny

$r_{xy} < 0$ – *záporně lineárně korelované* veličiny

- Některá úskalí při používání lineárního k. k.

r_{xy} – popisuje *sílu lineární závislosti* mezi x a y

Používat pouze tehdy, když bodový diagram naznačuje, že data jsou *soustředěna kolem přímky*.

- Korelace není příčinnost

Veličiny mohou být silně korelované, to však neznamená, že je mezi nimi vztah *příčinný*.

Příklad:

V tabulce jsou uvedena data týkající se počtu hodin, které každý z osmi náhodně vybraných studentů (veličina x) věnoval přípravě na test z matematiky, který se měl uskutečnit za 14 dní a počet bodů získaných za test (veličina y).

x	10	15	12	20	8	16	14	22
y	92	81	84	74	85	80	84	80

Silně záporně korelované veličiny ($r_{xy} = -0.779$)
– neznamená to, že větší počet hodin věnovaný přípravě na test je příčinou horšího výsledku testu.

Dvě veličiny mohou být silně korelované z toho důvodu, že obě jsou vázány s jinými veličinami, tzv. *skryté veličiny*, které jsou příčinou variability veličin, které zkoumáme.

• Test hypotézy o koeficientu korelace

Předpoklad: X, Y – náhodné veličiny

Jestliže $\rho_{xy} = 0 \implies$

$$T = \frac{r_{xy}}{\sqrt{\frac{1-r_{xy}^2}{n-2}}} \sim t(n-2)$$

Test: $H_0 : \rho_{yx} = 0$ versus $H_1 : \rho_{yx} \neq 0$

◇ Testovací statistika:

$$T = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \sqrt{n-2} \sim t(n-2)$$

◇ Kritický obor:

$$|T| > t_{1-\frac{\alpha}{2}}(n-2)$$

$t_{1-\frac{\alpha}{2}}(n-2)$ – 100(1 - $\alpha/2$)% kvantil Studentova t -rozdělení s $(n-2)$ stupni volnosti

Příklad:

V padesátých letech došlo k úniku radioaktivního odpadu ze skládky v Hanfordu do řeky Columbia River. V devíti okrscích níže po proudu řeky bylo zjišťováno vystavení radioaktivitě X . Současně se sledovala úmrtnost na rakovinu Y (úmrtí na 100tisíc obyvatel za rok v letech 1959-1964).

Zjištěné údaje jsou v následující tabulce:

okrsek	1	2	3	4	5	6	7	8	9
X	8.3	6.4	3.4	3.8	2.6	11.6	1.2	2.5	1.6
Y	210	180	130	170	130	210	120	150	140

Poskytují nám údaje dostatek argumentů pro to, abychom udělali na 1% hladině významnosti závěr, že vystavení radioaktivitě a úmrtnost na rakovinu jsou kladně lineárně korelované?