Hodnocení závislosti

STAT metody pro posouzení závislosti – jiné pro:

- **kvantitativní znaky**
- **kvalitativní znaky**

→ závislost funkční x statistická

Příklad (1)

Posuďte vztah mezi obsahem kyseliny mléčné v krvi matky a novorozence těsně po porodu (mg/100ml).

matka	**novorozenec**
x	**y**
39,0	31,8
6,5	34,5
41,1	33,7
43,0	43,0
33,5	21,0
11,2**x**	9,0**x**
40,2	32,6
50,9	32,0
66,5**x**	48,7
54,7	48,2
66,4	62,4
64,7	64,7**x**
56,8	6,8
40,9	40,9

Příklad (2)

Sestrojte bodový graf.

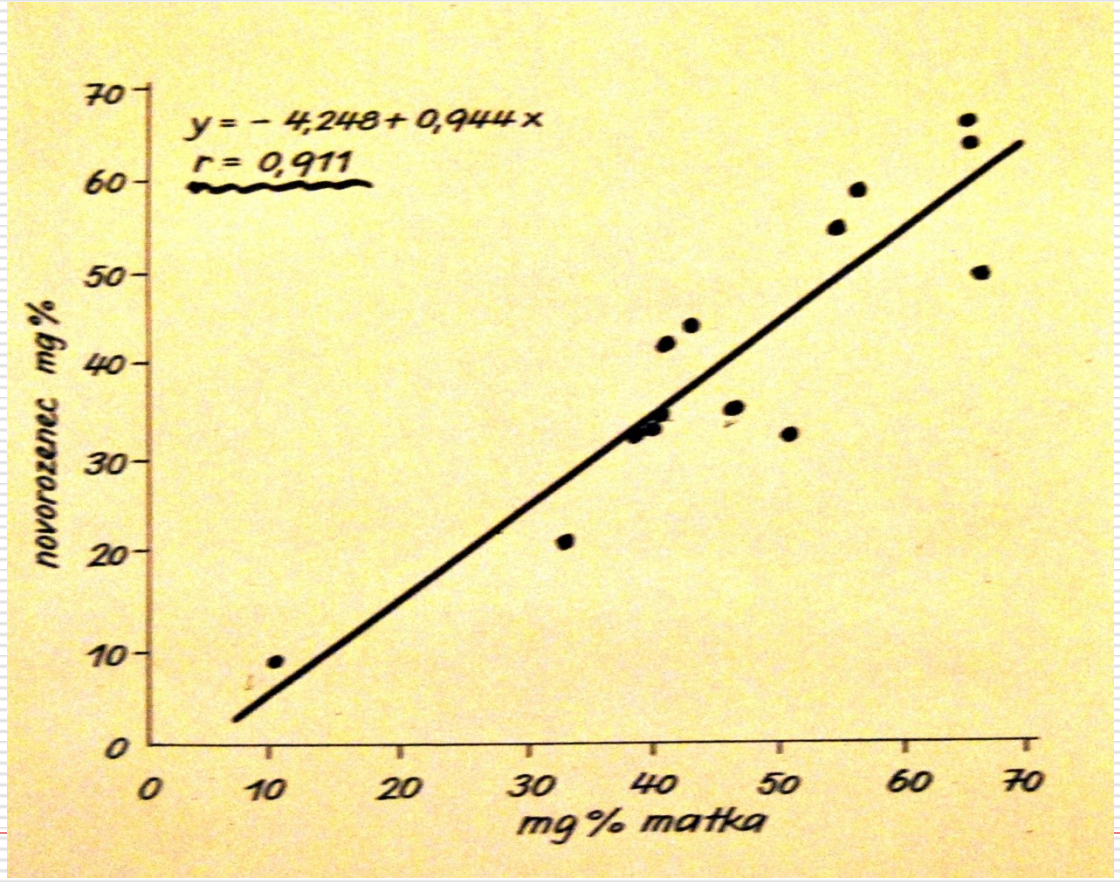
$m_x = 46{,}81$ $m_z = 39{,}95$

$s_x = 14{,}40$ $s_z = 14{,}94$

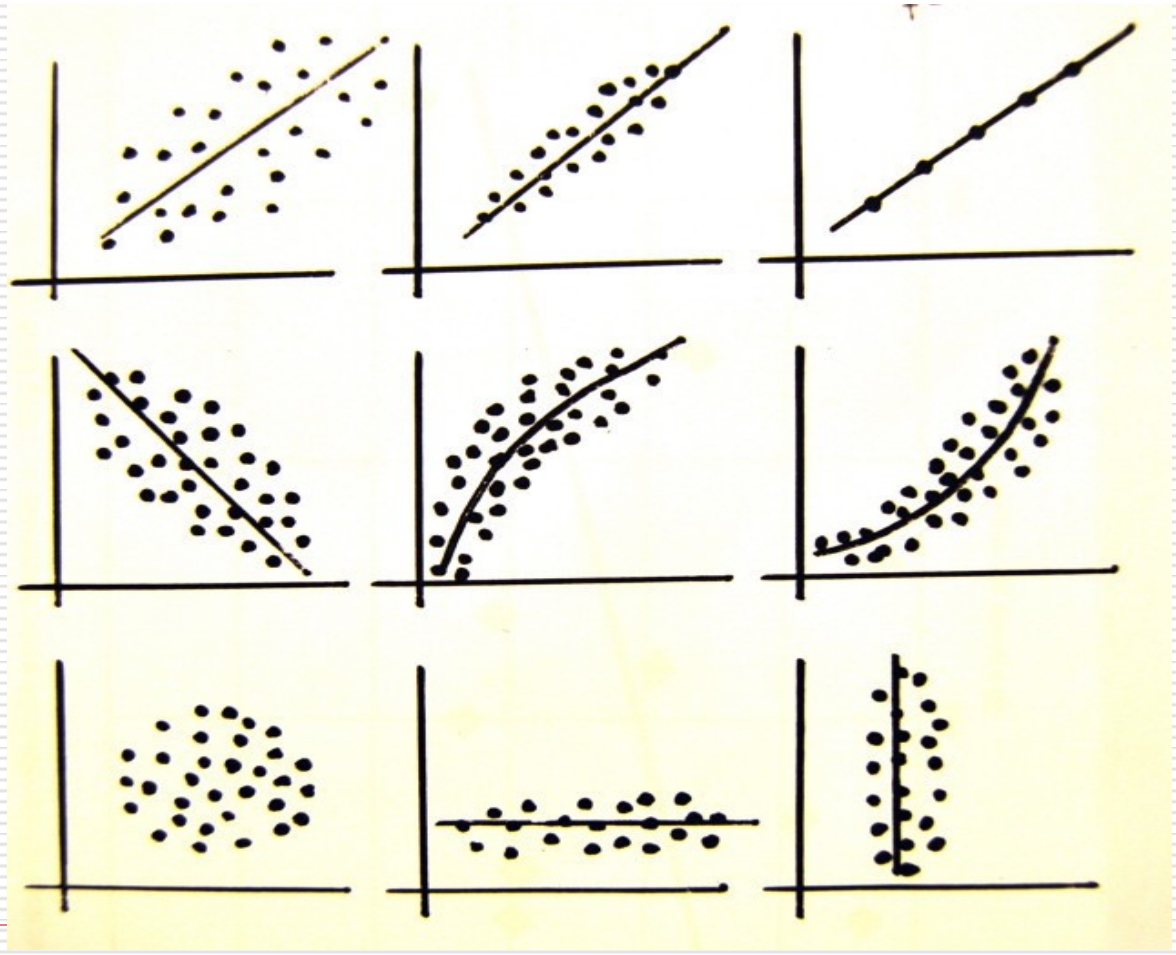
$$\sum(x_i - m_x)(y_i - m_y) = 2\,742{,}49$$

Bodový graf

Závislost mezi obsahem kyseliny mléčné u novorozence a matky těsně po porodu.



Bodový graf



1. Typ závislosti (funkce)
2. Směr (přímá, nepřímá)
3. Těsnost (rozptyl bodů)

Lineární závislost

→ měří se **korelačním koeficientem ρ** (parametr); je to nejlepší míra těsnosti.

Vlastnosti: $-1 \leq \rho \leq 1$

$\rho = 0$ → veličiny jsou nezávislé

$\rho = 1$ → funkční závislost (přímá, nepřímá)

ρ je kladné v případě přímé statistické závislosti

ρ je záporné v případě nepřímé stat.závislosti

Nelineární závislost

Pro hodnocení nelineární závislosti používáme:

a) Transformace – příklady

1) **$y = 1/x$** místo závislosti veličin **x** a **y** se studuje lineární závislost veličiny **x** a **$z = 1/y$**

2) **$y = ax^b \rightarrow \log y = \log a + b \log x$**

místo nelineární závislosti **x** a **y** se studuje lineární závislost veličin **log x** a **log y**

b) Pořadový korelační koeficient (Spearmanův, Kendallův)

Korelační koeficient (1)

Ve výběru se počítá tzv. **výběrový korelační koeficient r**, který je nejlepším odhadem neznámého korelačního koeficientu **ρ**

Mějme **n** dvojic dat (x_i , y_i) i = 1, 2, … n, pak

$$r = \frac{\sum (x_i - m_x)(y_i - m_y)}{n \cdot s_x \cdot s_y}$$

kde m_x, s_x → průměr a směrodatná odchylka veličiny X

m_y, s_y → průměr a směrodatná odchylka veličiny Y

Korelační koeficient (2)

! **r** je výběrová charakteristika, která má povahu náhodné veličiny

→ mění výběr od výběru

→ je zatížen náhodnou chybou **SE**, která je dána vztahem

$$SE = \frac{1 - \rho^2}{\sqrt{n - 1}}$$

Pro velké výběry (**n > 50**) má **r** normální rozdělení, jeho vlastnosti můžeme využít pro hodnocení závislosti.

Hodnocení významnosti r

1) $H_0 \equiv \rho = 0$ → veličiny jsou nezávislé
2) $H_A \equiv \rho \neq 0$ → veličiny jsou závislé
3) Za platnosti H_0 chyba

$$SE = \frac{1 - \rho^2}{\sqrt{n - 1}}$$

u-test (pro n > 50)!!!

4)

$$u = r\sqrt{n - 1}$$

→ kritické hodnoty: 1,96; 2,58

Pro malá n **kritické hodnoty** (viz skripta str. 28)

Příklad 1:

Zhodnoťte závislost obsahu kyseliny mléčné v krvi novorozence a matky těsně po porodu (viz naměřené hodnoty v úvodu).

Příklad 2:

Zhodnoťte závislost kojenecké úmrtnosti a podílu živě narozených dětí s porodní hmotností do 2 500g:

a) ve 14 okresech Jmk (r = 0,429)

b) ve 76 okresech ČR (r = 0,471)

Příklad 3

V souboru 225 jednoletých brněnských chlapců byl sledován vztah mezi tělesnou délkou a hmotností. Výpočtem jsme zjistili $r = 0{,}648$.

Zhodnoťte závislost pomocí u-testu i pomocí intervalu spolehlivosti.

Interpretace korelačního koeficientu

$100 . r^2$ udává procento variability náhodné veličiny **Y,** která připadá na vrub lineární závislosti veličiny **Y** na veličině **X.**

Příklad: Jestliže těsnost vztahu mezi hmotností a tělesnou délkou jednoletých chlapců vyjadřuje korelační koeficient **r = 0,648**, pak **42%** celkové variability hmotnosti jednoletých chlapců připadá na vrub závislosti na délce. Znamená to, že variabilita vah jednoletých chlapců určité délky by byla o **42%** nižší než variabilita celková (pro chlapce všech délek).

Regresní analýza

Pokud je závislost těsná (r – hodně velké), je vhodné vyjádřit ji pomocí tzv. regresní přímky ve tvaru

$$y = a + bx$$

Regresní koeficienty:

b = $r\,(s_y/s_x)$ → sklon přímky

a = $m_y - b\,m_x$ → úsek na ose y

Regresní analýza – viz příklad v úvodu

Vypočítejte regresní koeficienty a sestavte regresní funkci pro závislost mezi obsahem kyseliny mléčné u novorozence a matky těsně po porodu.

Regresní analýza - příklad

V souboru 76 okresů ČR byla zjištěna závislost mezi podílem dětí s nízkou porodní hmotností (X) a kojeneckou úmrtností (Y), kterou lze vyjádřit rovnicí:

$$y = 4{,}139 + 0{,}942x.$$

Vypočítejte, jaká by byla kojenecká úmrtnost v okrese, kde na 100 živě narozených připadá 7 dětí s nízkou porodní hmotností.

Nelineární závislost (1)

Spearmanův koeficient pořadové korelace

1) Nejprve seřadíme všechny hodnoty veličiny **X** dle velikosti a označíme je pořadovými čísly.
2) Pak seřadíme všechny hodnoty veličiny **Y** dle velikosti a označíme je pořadovými čísly.
3) Pro každou dvojici hodnot **x**, **y** stanovíme jejich rozdíl **d**
4) Spearmanův koeficient pořadové korelace vypočítáme ze vztahu:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Nelineární závislost (2)

$\mathbf{r_s}$ nabývá hodnot od -1 do 1, opět platí, že když:

$\mathbf{r_s} = 0$ → nezávislost

$\mathbf{r_s} = 1$ → přímou funkční závislost

$\mathbf{r_s} = -1$ → nepřímou funkční závislost

Hodnocení $\mathbf{r_s}$**:** Čím více se hodnota blíží ± 1, tím větší je těsnost vztahu

Nelineární závislost (3)

TEST VÝZNAMNOSTI

Absolutní hodnota r_s se porovná s kritickými hodnotami Spearmanova koeficientu pořadové korelace:

- $|r_s| \geq$ k.h. $\rightarrow$ zamítáme H_0
- $|r_s| <$ k.h. $\rightarrow$ nezamítáme H_0

List 1 - okresy

okresy Jmk	x por.hmotnost do 2500g na 100 ŽN	y KÚ 1990-94
Blansko	4,10	7,70
Brno-město	6,20	9,69
Brno-venkov	5,00	9,33
Břeclav	4,40	6,40
Hodonín	4,20	7,77
Jihlava	4,60	8,98
Kroměříž	5,30	6,27
Prostějov	5,50	11,22
Třebíč	4,80	6,88
Uh.Hradiště	4,70	8,92
Vyškov	5,20	9,09
Zlín	5,10	7,92
Znojmo	5,70	7,64
Žďár n/S.	4,60	6,39

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

d_i = rozdíl pořadí

Postup při hodnocení závislosti kvantitativních veličin

1) Udělat bodový graf, tím získáme rozumnou vizuální představu o **typu závislosti.**
2) Pro určení síly lineární závislosti je vhodný **Pearsonův korelační koeficient r** (-1; +1). Kladné hodnoty svědčí pro **přímou** závislost , záporné pro **nepřímou**.
3) Zhodnotit významnost korelačního koeficientu. Sílu závislosti posoudit podle velikosti **r.**
4) Korelace neznamená příčinnost. Nerozhoduje, která veličina je závislá, která nezávislá.
5) Nemůže-li se empirickými body proložit přímka, je třeba použít:

   - transformace

   - pořadový Spearmanův korelační koeficient

Hodnocení závislosti kvalitativních znaků

- východiskem je kontingenční tabulka

	ALERGIE+	ALERGIE-	CELKEM
MUŽI	21	84	105
ŽENY	19	176	195
CELKEM	40	260	300

- je založeno na **srovnání empirických a teoretických četností**
- **empirická četnost –** rozdělení lidí podle pohlaví a alergie, jak bylo skutečně zjištěno ve výběrovém souboru
- **teoretická četnost** – jaké by bylo rozdělení lidí ve výběrovém souboru podle pohlaví a alergie, kdyby šlo o jevy nezávislé

Hodnocení závislosti kvalitativních znaků

1. Stanovení hypotéz

H_0 – mezi empirickými a teoretickými četnostmi není statisticky významný rozdíl, zjištěné rozdíly nejsou natolik velké, aby nemohly být způsobeny náhodou:

H_A - mezi empirickými a teoretickými četnostmi je statisticky významný rozdíl, zjištěné rozdíly jsou natolik velké, že nemohou být způsobeny náhodou:

2. Hladina významnosti

$\alpha = 5\%$ nebo $\alpha = 1\%$

3. Výběr testu

- chí-kvadrát test (χ^2)

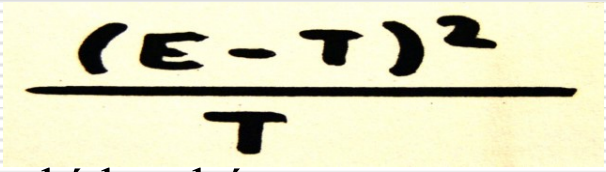
Hodnocení závislosti kvalitativních znaků

4. Podmínky pro použití testu

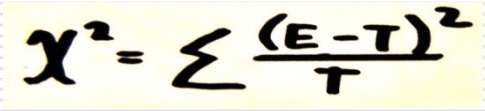
Všechny **teoretické četnosti** musí být **větší než 5.**

5. Výpočet testovací charakteristiky chí-kvadrát

1. Pro každé políčko vypočítáme teoretickou četnost
2. Pro každé políčko vypočítáme rozdíl mezi empirickou (E) a teoretickou četností (T) podle vzorečku:



3. Součet vypočítaných rozdílů je hodnota chí-kvadrátu:



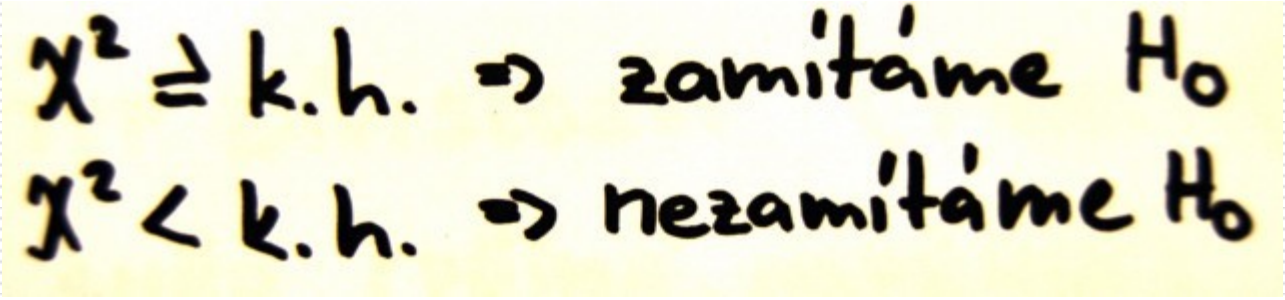
Hodnocení závislosti kvalitativních znaků

6. **Srovnání s kritickými hodnotami**

   Chí-kvadrát srovnáme s příslušnými **kritickými hodnotami** chí-kvadrát rozdělení:

   - Kritické hodnoty určujeme z tabulek podle zvolené hladiny významnosti a tzv. stupňů volnosti.

7. **Zamítáme nebo nezamítáme nulovou hypotézu**



8. **Interpretace výsledků**

Příklad (1):

Pro čtyřpolní tabulku (typu 2x2) můžeme veličinu chí 2 počítat jednodušeji

→ postup viz následující příklad

Tabulka: Vztah mezi způsobem výživy a výskytem novorozeneckého ikteru u 210 novorozenců

Příklad (2):

způsob výživy	výskyt ikteru +	-	součet
A_1	61	49	110
A_2	85	15	100
součet	146	64	210

Kritické hodnoty

Tab. 2. Kritické hodnoty Pearsonova korelačního koeficientu
(pro rozsah výběru n a hladinu významnosti 0,05 a 0,01)

n	0,05	0,01	n	0,05	0,01	n	0,05	0,01
3	0,9969	0,9999	14	0,5324	0,6614	25	0,3961	0,5052
4	0,9500	0,9900	15	0,5140	0,6411	30	0,3610	0,4629
5	0,8783	0,9587	16	0,4973	0,6226	35	0,3338	0,4296
6	0,8114	0,9172	17	0,4822	0,6055	40	0,3120	0,4026
7	0,7545	0,8745	18	0,4683	0,5897	45	0,2940	0,3801
8	0,7067	0,8343	19	0,4555	0,5751	50	0,2787	0,3610
9	0,6664	0,7977	20	0,4438	0,5614	60	0,2542	0,3301
10	0,6319	0,7646	21	0,4329	0,5487	70	0,2352	0,3060
11	0,6021	0,7348	22	0,4227	0,5368	80	0,2199	0,2864
12	0,5760	0,7079	23	0,4132	0,5256	90	0,2072	0,2702
13	0,5529	0,6835	24	0,4044	0,5151	100	0,1966	0,2565

Tab. 3. Kritické hodnoty Spearmanova koeficientu pořadové korelace
(pro rozsah výběru n a hladinu významnosti 0,05 a 0,01)

N	0,05	0,01	n	0,05	0,01	n	0,05	0,01
			11	0,6091	0,7545	21	0,4351	0,5545
			12	0,5804	0,7273	22	0,4241	0,5426
			13	0,5549	0,6978	23	0,4150	0,5306
			14	0,5341	0,6747	24	0,4061	0,5200
5	0,9000	-	15	0,5179	0,6536	25	0,3977	0,5100
6	0,8286	0,9429	16	0,5000	0,6324	26	0,3894	0,5002
7	0,7450	0,8929	17	0,4853	0,6152	27	0,3822	0,4915
8	0,6905	0,8571	18	0,4716	0,5975	28	0,3749	0,4828
9	0,6833	0,8167	19	0,4579	0,5825	29	0,3685	0,4744
10	0,6364	0,7818	20	0,4451	0,5684	30	0,3620	0,4665

Tab. 4. Kritické hodnoty rozdělení χ^2 pro počet stupňů volnosti f a hladinu významnosti 0,05 a 0,01

f	0,05	0,01	f	0,05	0,01
1	3,84	6,63	11	19,68	24,73
2	5,99	9,21	12	21,03	26,22
3	7,81	11,35	13	22,36	27,69
4	9,49	13,28	14	23,69	29,14
5	11,07	15,09	15	25,00	30,58
6	12,59	16,81	16	26,30	32,00
7	14,07	18,48	17	27,59	33,31
8	15,51	20,09	18	28,87	34,81
9	16,92	21,67	19	30,14	36,19
10	18,31	23,21	20	31,41	37,57