Měření statistické závislosti, korelace, regrese

Prof. RNDr. Jana Zvárová, DrSc.

1

MĚŘENÍ ZÁVISLOSTI

Cílem statistické analýzy v epidemiologii bývá nejen stanovit, zda onemocnění závisí na výskytu rizikového faktoru, ale rovněž vyjádřit **STUPEŇ ZÁVISLOSTI**.
Stupeň závislosti však stanovujeme i mezi příznaky a nemocí či mezi příznaky samotnými.

STUPEŇ ZÁVISLOSTI (KORELACI) vyjadřujeme pomocí různých **MĚR STATISTICKÉ ZÁVISLOSTI**, ke kterým patří i **KORELAČNÍ KOEFICIENTY**.
Obecně požadujeme:

$$0 \leqq |\text{míra statistické závislosti}| \leqq 1$$

NEZÁVISLOST — **FUNKČNÍ ZÁVISLOST**

Obecné principy

závislost ..... vzájemná souvislost měřených znaků

- **funkční** závislost x **statistická** závislost
- nástroje pro **měření závislosti**
  - lineární regrese
  - korelace

} kvantitativní znaky

3

Korelace a regrese

- síla (těsnost) závislosti dvou náhodných veličin: **korelace**
  - symetrický vztah obou veličin
  - neslouží k předpovědi
- způsob (tvar) závislosti náhodné veličiny na jiné veličině: **regrese**
  - možnost předpovědi
- příklad: výška otce, výška jeho syna (v dospělosti)
  - **korelace:** jak těsně spolu souvisejí ? populace - všechny dvojice (otec, syn)
  - **regrese**: lze z výšky otce odhadnout výšku syna ? řada populací - synové otců vysokých 170 cm, 171 cm ...

4

Korelace

- kvantifikace síly lineární závislosti mezi **dvěma kvantitativnímu veličinami**

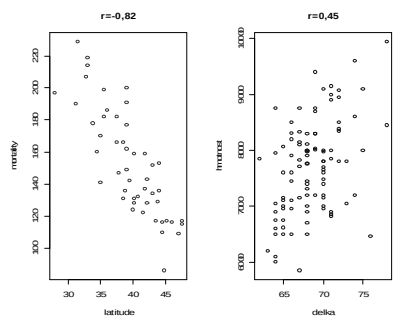
(Pearsonův) korelační koeficient:

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- důležité je **znaménko** a **velikost korelačního koeficientu**
- korelace **neznamená** příčinnost.
- hodnoty posuzujte **kriticky**

5

Příklady



6

Měření závislosti pro kvantitativní znaky

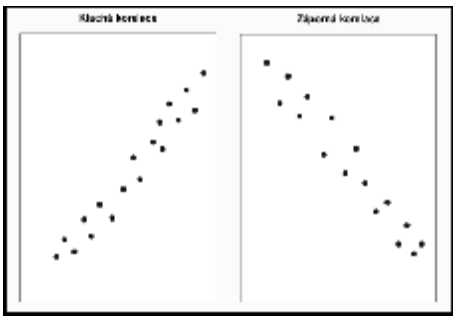
Kromě stupně závislosti, který vyjadřuje korelační koeficient, se často snažíme zjistit i typ závislosti.

Orientačně můžeme typ závislosti posoudit z **bodového grafu**.

Typ závislosti určuje křivka, kterou můžeme empirickými body proložit.

7

Korelace



8

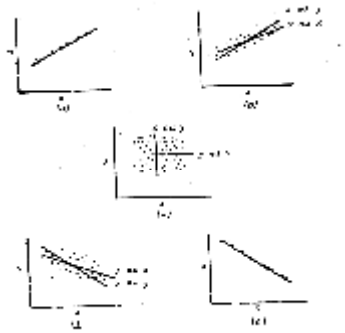
Pearsonův korelační koeficient

- měří sílu lineární závislosti spojitých veličin
- vždy platí: $-1 \le \rho_{X,Y} \le 1$
- v případě normálního rozdělení platí:
  nezávislost $X, Y \Leftrightarrow \rho_{X,Y} = 0$
- odhad pomocí $r_{X,Y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$
- nezávislost zamítáme, pokud $|t| \ge t_{1-\alpha/2}(n-2)$, kde

$$t = \frac{r}{\sqrt{1 - r_{X,Y}^2}} \sqrt{n-2}$$

9

Grafy



10

Lineární regrese

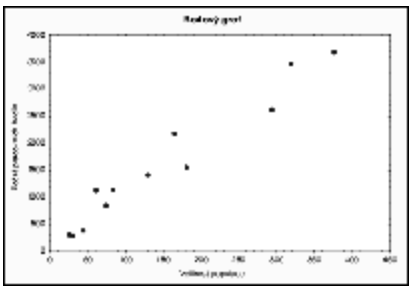
- kvalifikace lineárního vztahu mezi **dvěma kvantitativnímu veličinami**

Př.: Analyzujeme data o počtu pracovních hodin za měsíc v anesteziologické službě v závislosti na velikosti spádové oblasti.

Nemocnice	Počet pracovních hodin	Spádová populace (v tis.)
1	304,37	25,50
2	2616,32	294,30
3	1139,12	83,70
4	285,43	30,70
5	1413,77	129,80
6	1555,68	180,80
7	383,78	43,40
8	2174,27	165,20
9	845,30	74,30
10	1125,28	60,80
11	3462,60	319,20
12	3682,33	376,20

11

Lineární regrese - motivační příklad



12

Lineární regrese - regresní přímka

Regresní přímka:

$$y_i = a + b\ x_i + e_i, \qquad i = 1,...,n$$

a ..... absolutní člen (*intercept*)
b ..... směrnice (*slope*)
e ..... náhodná chyba

Př. (pokr.):

pracovní doba = $\alpha + \beta$ velikost populace + ε

13

Lineární regrese - odhad parametrů

Odhady hodnot **parametrů** α a β se určují *metodou nejmenších čtverců*.

Princip **metody nejmenších čtverců**:

Za odhad parametrů α a β se berou taková čísla a a b, pro která výraz

$$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

nabývá minimální hodnoty.

Zde $\hat{y}_i = a + b\,x_i$ je vyhlazená hodnota y_i. Rozdíl $y_i - \hat{y}_i$ se nazývá i-té *reziduum*. Tzv. *reziduální rozptyl* je pak zaveden jako

$$s^2 = \frac{S_e}{n-2}$$

14

Lineární regrese - výpočet odhadů α a β

Odhady parametrů a ***a*** *b:*

$$a = \bar{y} - b\,\bar{x} \qquad b = \frac{s_{xy}}{s_x^2}$$

Pomocné výpočty

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

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

(s_{xy} je odhad **kovariance** veličin X a Y)

15

Lineární regrese - interpretace výsledků

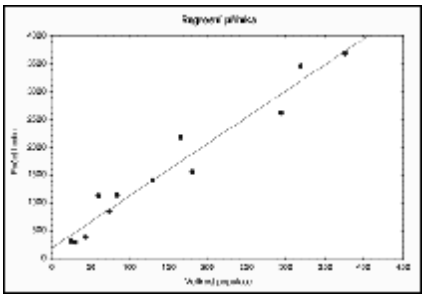
Př.: Obdrželi jsme rovnici

pracovní doba = 180,658 + 9,429 * velikost populace

- výsledek je třeba interpretovat pouze v rozsahu pozorovaných dat
- odhadnuté parametry závisejí na použitých datech
- můžeme zjistit intervalové odhady skutečných parametrů

16

Graf odhadnuté regresní přímky



17

Koeficient determinace

Koeficient determinace:

$$R^2 = r^2$$

- měření síly závislosti mezi proměnnými X a Y
- míra vhodnosti modelu - určuje část **variability** Y **vysvětlenou** pomocí modelu **lineární regrese**

! $(1-R^2) \cdot 100$ % variability Y nelze vysvětlit variabilitou X

18