13. Měření závislosti

- Nezávisle proměnná *X* má řídit závisle proměnnou *Y* => lineární regrese (y = a + bx)
- Měření síly lineární závislosti => Pearsonův korelační koeficient (normální rozložení)
- Závislost kvalitativních znaků => (neparametrický) Spearmanův korelační koeficient

Pearsonův korelační koeficient

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

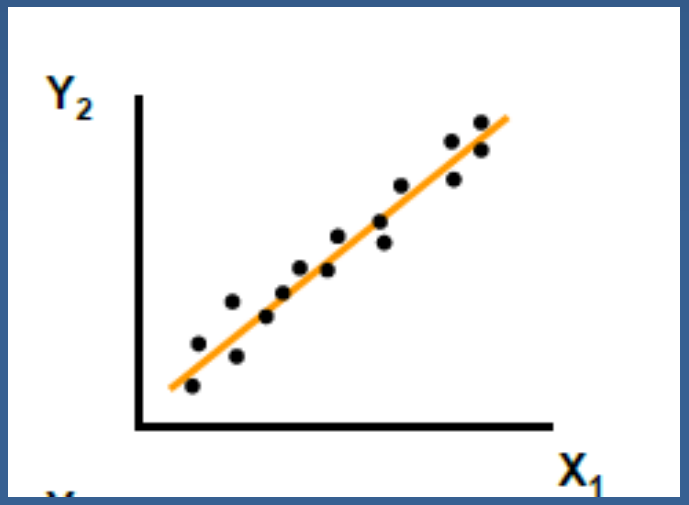
- Kde s_{xy} je kovariance

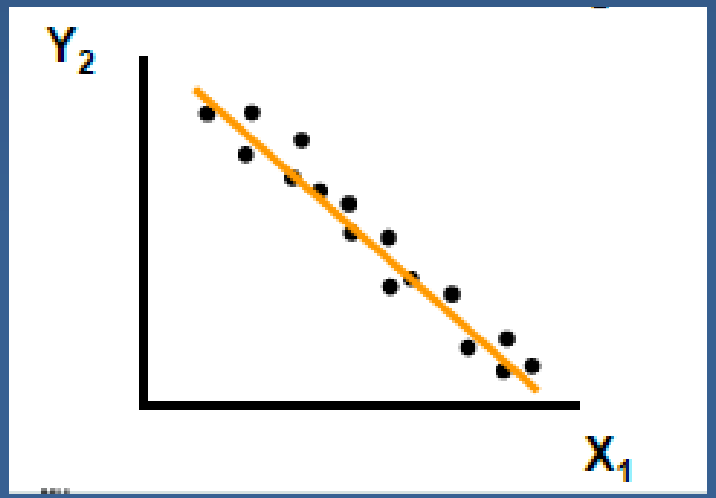
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- s_x^2, s_y^2 jsou výběrové rozptyly
- Hodnoty od -1 do 1
- ± 1 pokud všechny body $[x_i, y_i]$ leží na přímce
- 0 => veličiny jsou nezávislé

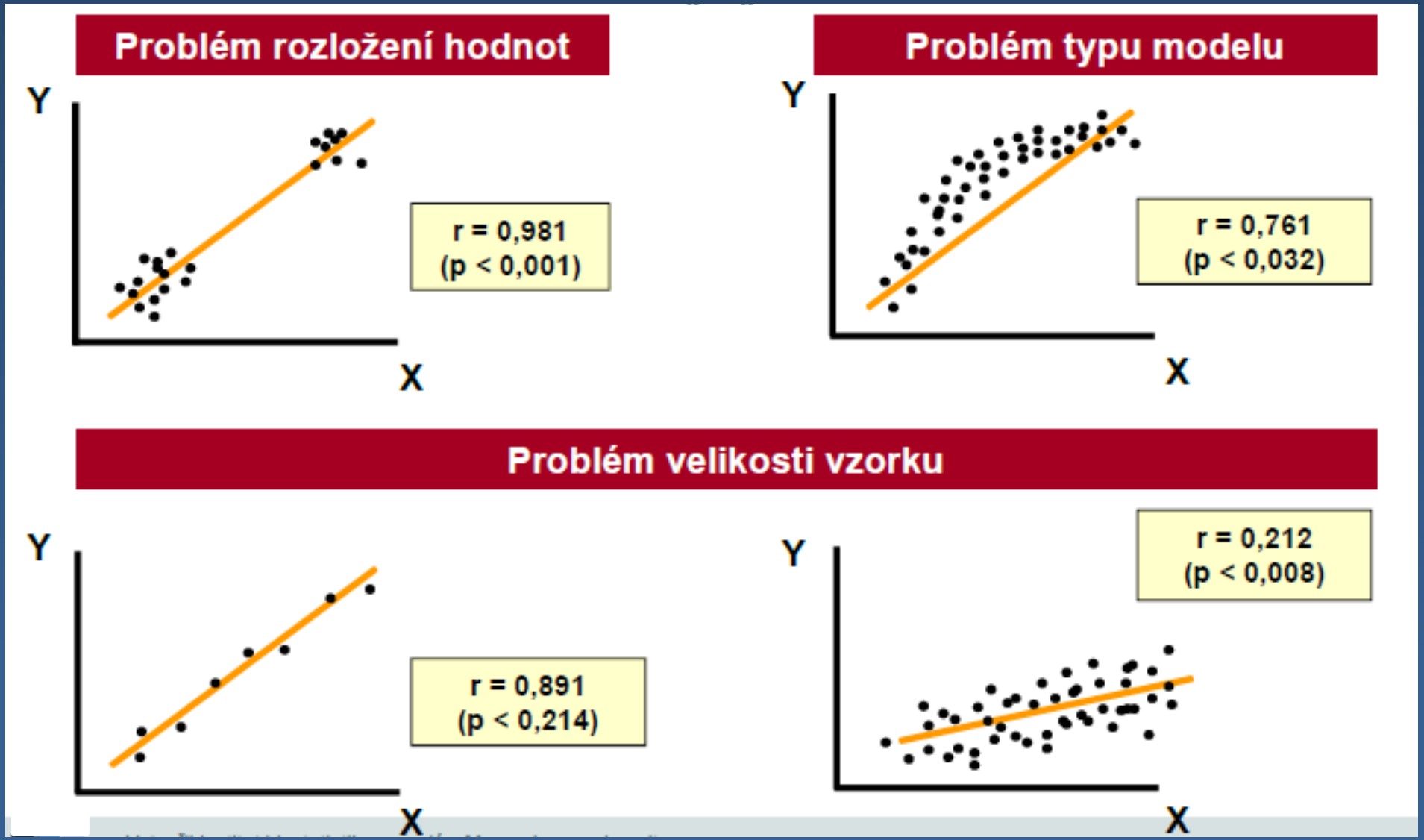
Pearsonův korelační koeficient

- Pokud závislost není lineární, může *r* vyjít 0 a přitom jsou veličiny funkčně závislé
- +*r* => obě veličiny *X* a *Y* zároveň rostou nebo obě zároveň klesají
- -*r* => jedna veličina roste zatímco druhá klesá





Korelace v grafech



Korelace

- Znaménko - směr
- Velikost – nashromáždění bodů kolem přímky
- *p* – k testu nulové hypotézy $r = 0$ (mezi *X* a *Y* neexistuje žádný lineární vztah

Pearsonův korelační koeficient

- Síla lineární závislosti mezi dvěma spojitými veličinami
- Doplnit bodovým grafem
- Znaménko
- Velikost
- Korelace neznamená příčinnost
- Pro velký rozsah výběru vyjde i malé *r* statisticky významně

Pearsonův korelační koeficient - příklad

Během 8 let se zaznamenávaly průměrné doby slunečního svitu ve vegetačním období pšenice a její hektarové výnosy.

Rok	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Doba slunečního svitu [h]	4,1	3,9	3,5	3,8	4,2	4,1	4,1	3,9
Výnos pšenice [q]	1,2	1,1	0,5	0,9	1,0	1,3	1,0	1,0

Předpokládáme, že doba slunečního svitu a výnos pšenice má normální rozložení

$$\bar{x} = 3{,}95 \qquad s_x^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2{,}52$$

$$\bar{y} = 1{,}00 \qquad s_y^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 2{,}80$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 2{,}17$$

$$r = \frac{s_{xy}}{\sqrt{s_x^{\,2} s_y^{\,2}}}$$

$$= \frac{2{,}17}{\sqrt{2{,}25 \cdot 2{,}80}} = 0{,}817$$

Spearmanův korelační koeficient

- Neparametrický
- Založený na pořadí
- *Q* – pořadí podle první veličiny *X*
- *R* – pořadí podle druhé veličiny *Y*
- S rostoucím *X* vzrůstá *Y* =>shodné pořadí
- S rostoucím *X* klesá *Y* =>opačná pořadí
- Nezávislé veličiny => náhodně zpřeházená pořadí

Spearmanův korelační koeficient

- Diference pořadí $d_i = Q_i - R_i$

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}$$

- Test korelačního koeficientu: když $|r_S|$ > kritická hodnota => zamítáme H_0 o nezávislosti
- Pro $n > 30$ je testovou statistikou: $Z = r_s\sqrt{n-1}$
- $Z \sim N(0,1)$ => H_0 zamítáme pro $|Z| \geq z_{1-\frac{\alpha}{2}}$

Spearmanův korelační koeficient- příklad

Na interním oddělení bylo léčeno na určitou chorobu deset pacientů. Za čtyři týdny po ukončení hospitalizace byli pozváni ke kontrolní prohlídce, při níž se mimo jiných údajů sledovala též sedimentace červených krvinek. Závažnost klinického průběhu všech deseti pacientů lze zhruba vyjádřit na základě uspořádání podle závažnosti zjištěných klinických příznaků do řady, jejíž členy označíme jedničkou (nejlehčí průběh) až desítkou (nejtěžší průběh). Označme pacienty velkými písmeny A, B, C atd. Jejich hodnocení podle závažnosti klinického průběhu, zjištěné hodnoty sedimentace a pořadí podle výše sedimentace jsou uvedeny v tabulce. Ověřte na 5% hladině významnosti, zda hodnota sedimentace nezávisí na klinickém průběhu onemocnění.

Pacient	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Pořadí podle klinického průběhu (Q)	4	6	1	5	10	2	7	3	9	8
Sedimentace	19	22	26	30	33	23	28	21	59	39
Pořadí podle sedimentace (R)	1	3	5	7	8	4	6	2	10	9
Diference pořadí (d=Q-R)	3	3	-4	-2	2	-2	1	1	-1	-1
Čtverec diference d_2	9	9	16	4	4	4	1	1	1	1

$$r_S = 1 - \frac{6 \cdot 50}{10(100-1)} = 1 - \frac{300}{990} = 0{,}697 \qquad r_S = 0{,}697 > 0{,}6364$$

=> Zamítáme H_0 o nezávislosti

14. Analýza rozptylu

ANOVA

- Analýza vlivu jedné či více kategoriálních proměnných (tři a více úrovní) na kvantitativní nebo ordinální proměnnou
- Nelze použít tři dílčí t-testy!
- Opakované testování neoprávněně zvyšuje pravděpodobnost chyby prvního druhu

Počet testů	**Pst výskytu aspoň jedné chyby I. druhu**
3	0,14
10	0,40
45	0,90
c	$1-(1-\alpha)^c$

ANOVA - předpoklady

- Každý z výběrů pochází z populace s normálním rozložením se stejnou směrodatnou odchylkou (testujeme, zda střední hodnoty (průměry) jsou všechny shodné nebo se liší)
- Náhodný výběr z každé populace
- Nezávislá pozorování

ANOVA - princip

- Celkový rozptyl závisle proměnné rozdělíme do dvou částí
- 1. variabilita uvnitř skupin – odlišnost hodnot v rámci skupiny od skupinového průměru
- 2. variabilita mezi skupinami – jak se navzájem liší průměry – porovnání s celkovým průměrem
- Pokud neexistuje rozdíl, variabilita mezi skupinami i v rámci skupin popisuje stejný jev (stejný populační rozptyl)
- Porovnání F testem

ANOVA

$$S_T = S_A + S_e$$

- S_T – celková variabilita – součet rozdílů pozorovaných hodnot a celkového průměru umocněných na druhou
- S_A – variabilita mezi skupinami - vážený součet druhých mocnin rozdílů každého skupinového průměru a celkového průměru
- S_e – variabilita uvnitř skupin – součet druhých mocnin rozdílů hodnot a příslušného skupinového průměru

ANOVA

- df_T – počet pozorování – 1
- df_A – počet skupin - 1
- df_e – počet pozorování – počet skupin

$$F = \frac{S_A / df_A}{S_E / df_E}$$

ANOVA

- H_0: mezi populačními průměry není žádný rozdíl
- H_1: aspoň dva populační průměry se od sebe liší (nezjistíme však, které to jsou)
- Za platnosti H_0 je čitatel F statistiky (zhruba) stejně velký jako jmenovatel => $F \approx 1$
- Kritická hodnota: kvantil F rozložení $F_{1-\alpha}(df_A, df_e)$
- Nutno ověřovat předpoklady modelu
- Metody mnohonásobného porovnávání (Bonferroniho, Scheffého, Tukeyova metoda, …)

ANOVA - příklad

Byly zjišťovány váhy u tří plemen králíků chovaných za standardních podmínek. Od každého plemene bylo chováno 5 kusů (tj. $n_i=5$ pro všechna i). Cílem bylo zjistit, zda se za daných podmínek budou váhy tří porovnávaných plemen lišit. Na konci pokusu byly zjištěny následující váhy (v kg):

Plemeno 1: 3,3,4,5,5

Plemeno 2: 4,4,5,6,6

Plemeno 3: 5,5,6,7,7

Z toho spočteme průměry: $\overline{X}_1 = 4, \overline{X}_2 = 5, \overline{X}_3 = 6$, celkový průměr $\overline{X} = 5$. Počet stupňů volnosti v každé skupině je čtyři, tzn $DF_e=4+4+4=12$, $DF_G=3-1=2$. Součet čtverců odchylek uvnitř skupin je podle 8.1 $SS_e=(3-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(5-4)^2+(4-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(6-5)^2+(5-6)^2+(5-6)^2+(6-6)^2+(7-6)^2+(7-6)^2=12$, součet čtverců odchylek mezi skupinami je podle 8.4 $SS_G= 5x(4-5)^2 + 5x(5-5)^2 + 5x(6-5)^2= 10$. Po vydělení součtu čtverců počtem stupňů volnosti dostáváme $MS_G= 10/2 = 5$, $MS_e = 12/12 =1$. $F= MS_G / MS_e = 5/1 = 5$. Protože kritická hodnota F pro jednostranný test na pětiprocentní hladině významnosti při 2 a 12 stupních volnosti je 3.89, zamítáme nulovou hypotézu o rovnosti středních hodnot na 5%-ní hladině významnosti.

Lepš (1996: 80)

Zadání k projektům ke zkoušce

- Krátký úvod do problematiky datového souboru
- Analýza odlehlých a nesprávných hodnot, zdokumentování nakládání s takovými údaji
- Popisná statistika všech (většiny) proměnných v datovém souboru
- - kategoriální proměnné: tabulka s absolutními a relativními četnostmi
- - proměnné poměrového typu: základní popisná statistika => parametrické (průměr, sm. odchylka, IS) x neparametrické (medián, rozsah, kvartily, percentily, …)
- Grafy
- Formulace H0 a HA
- Ověření předpokladů vybraného testu
- Popisná statistika hodnocených proměnných dle testované hyp.
- Výsledek a závěr testování
- Shrnutí použitých metod a dosažených výsledků