UK FHS
Historická sociologie
(LS 2010)

Analýza kvantitativních dat:

1. Popisné statistiky a testování hypotéz

Jiří Šafr

jiri.safr(zavináč)seznam.cz

vytvořeno 29. 6. 2009, **poslední aktualizace 25. 5. 2010**

Dva základní typy statistiky

1. **Popisná statistika**: metody pro zjišťování a sumarizaci informací → grfy, tabulky, popisné chrakteristiky (průměr, rozptyl percentily,..)
   Příklad:
2. **Inferenční statistika** (statistická indukce): metody pro přijímání a měření spolehlivosti závěrů o populaci založených na informacích získaných z jejího výběru (odhad parametru na základě výběru z populace)

Proces analýzy dat musíme promyslet již ve stadiu plánování dotazníku (modelu vztahů a hypotéz).

Základní pojmy

- Populace
- Základní soubor
- Výběrový soubor (vzorek)
- Datový soubor

- Třídění dat (jedno a vícestupňové)
- Absolutní četnost
- Relativní (poměrná) četnost
- Kumulativní četnost
- Distribuce: hodnoty proměnné nebo charakteristiky a jejich výskyt

Typy znaků – proměnných

Nominální

- Kategorie jsou rovnocenné (na úrovni jmen)
- př.: pohlaví, jména, typ rodiny, barva vlasů, profese

Pořadové (ordinální)

- Kategorie lze seřadit do hierarchie
- Lze se ptát: vyšší/nižší apod., ale ne o kolik
  př.: spokojenost, stupeň souhlasu
- **Intervalové**
- číselné proměnné
  lze se ptát větší/ menší a o kolik
  př.: věk, příjem, počet dětí

Typy proměnných - úrovně měření

[Babbie 1995: 137]

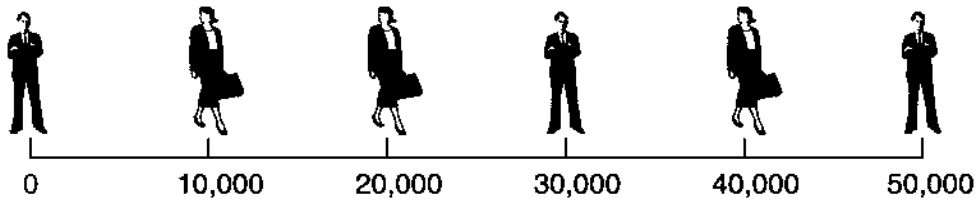
Nominální př: Pohlaví

žena muž

Ordinální př: Religiozita "Jak důležité je pro vás náboženství?"

Intervalové př.: IQ

Poměrové př.: Příjem



Znaky / proměnné

- **Kardinální**:
- A) **intervalové** – nemají přirozený počátek: obsahový smysl má rozdíl ale nikoliv podíl
  Příklad: „Dnes je o 10 st. C tepleji“, ale ne „o 25% tepleji.“ / IQ nemá nulu
- B) **poměrové** – mají přirozený počátek (0 má význam), tudíž lze uvažovat i podíl.
  Příklad: „nulové“ i „dvojnásobné tržby“

Standardizace: odstranění původní metriky

- Z – skóry: průměr X=0 a StD =1

  Odchylka od průměru / směrodatnou odchylkou:

$$z = \frac{x - \mu}{\sigma}$$

- → umožňuje porovnat znaky s odlišnou metrikou.
- Přímá standardizace

- **Rozptyl** = střední hodnota kvadrátů odchylek od střední hodnoty
- **Směrodatná odchylka** = odmocnina z rozptylu náhodné veličiny
- **Výběrová směrodatná odchylka**
- odmocninu z **výběrového** rozptylu)

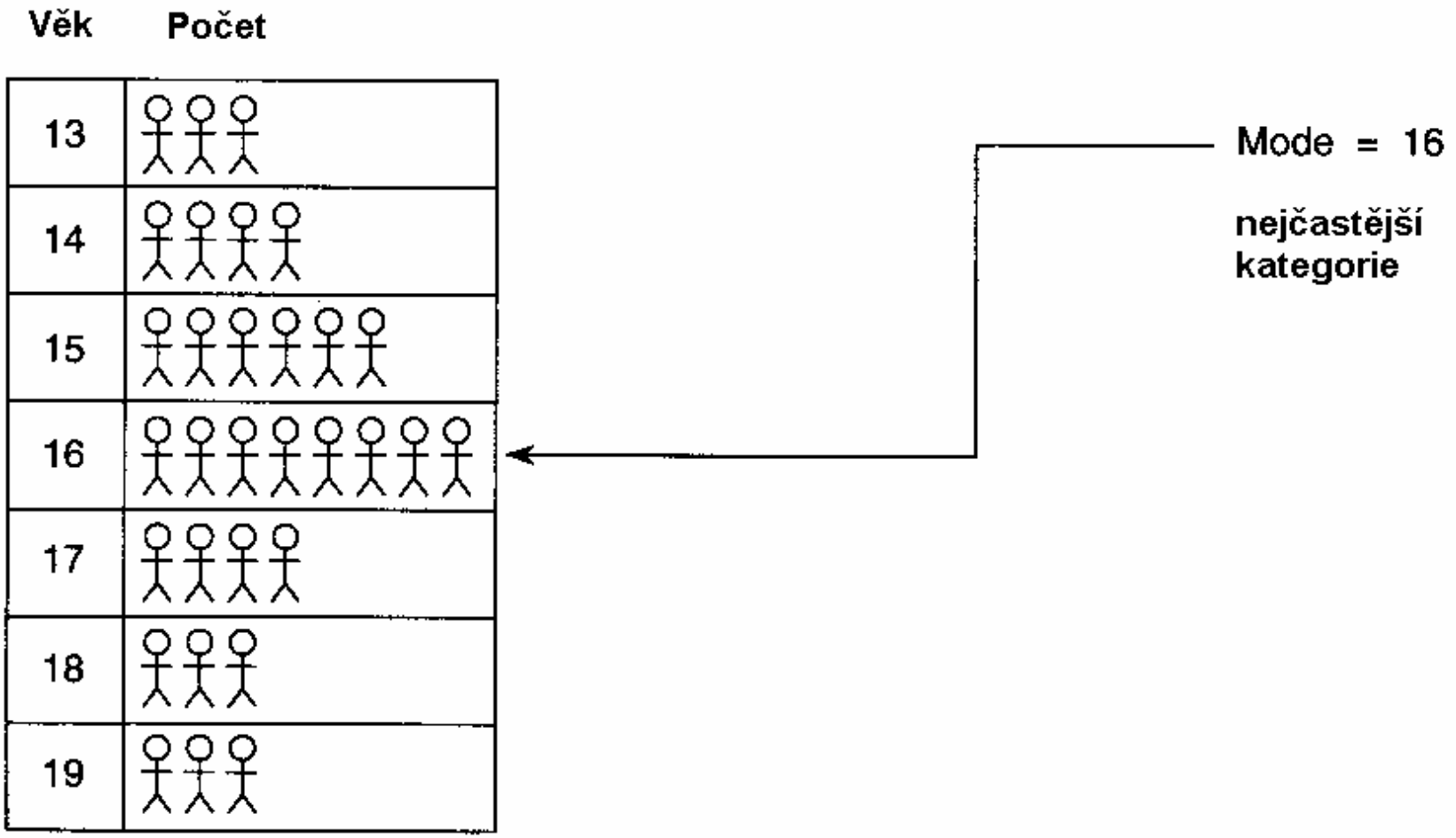
Jednoduché popisné statistiky

Střední hodnoty:

- nominální znaky → **modus**
- ordinální znaky → **medián**
  (aritmetický průměr)
- intervalové znaky → **aritmetický průměr**

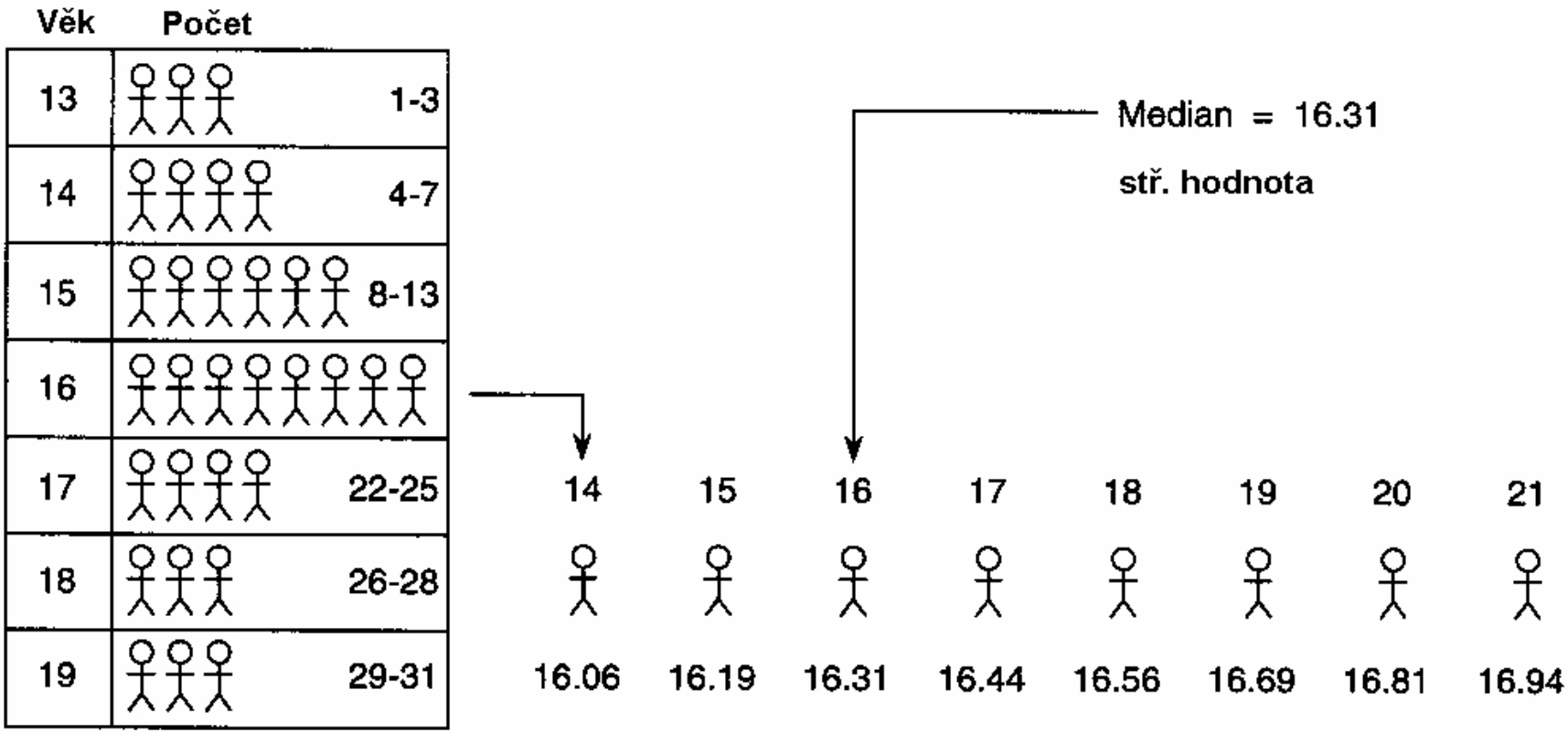
- **Modus** = kategorie s největší četností
- **Medián** = hodnota, která je ve prostředku všech pozorování seřazených podle hodnoty
- **Aritmetický průměr** = součet hodnot dělený počtem pozorování

Modus



[Babbie 1995]

Medián



[Babbie 1995]

Průměr

Věk	Počet
13	
14	
15	
16	
17	
18	
19	

$13 \times 3 = 39$

$14 \times 4 = 56$

$15 \times 6 = 90$

$16 \times 8 = 128$

$17 \times 4 = 68$

$18 \times 3 = 54$

$19 \times 3 = 57$

492 (celkem) $\div$ 31 (poč. případů) $= 15.87$

Mean = 15.87

aritmetický průměr

[Babbie 1995]

Charakteristiky variability

Udávají koncentraci nebo rozptýlení kolem střední hodnoty. Ukazují na „kvalitu“ průměru.

Rozptyl = součet kvadratických odchylek od průměru dělený rozsahem výběr zmenšeným o 1.

- **Směrodatná odchylka** = odmocnina z rozptylu.

Směrodatná odchylka je míra rozptýlení hodnot od průměrné (střední) hodnoty.

Výpočet směrodatné odchylky

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Příklad. Máme pozorování:

7 2 5 4 3 1 8 2 6 2

Součet řady = 40; **n** = 10; **průměr** = 40/10 = **4**

Odchylky:

3 -2 1 0 -1 -3 4 -2 2 -2

součet odchylek je 9 – 9 = **0**

čtverce odchylek:

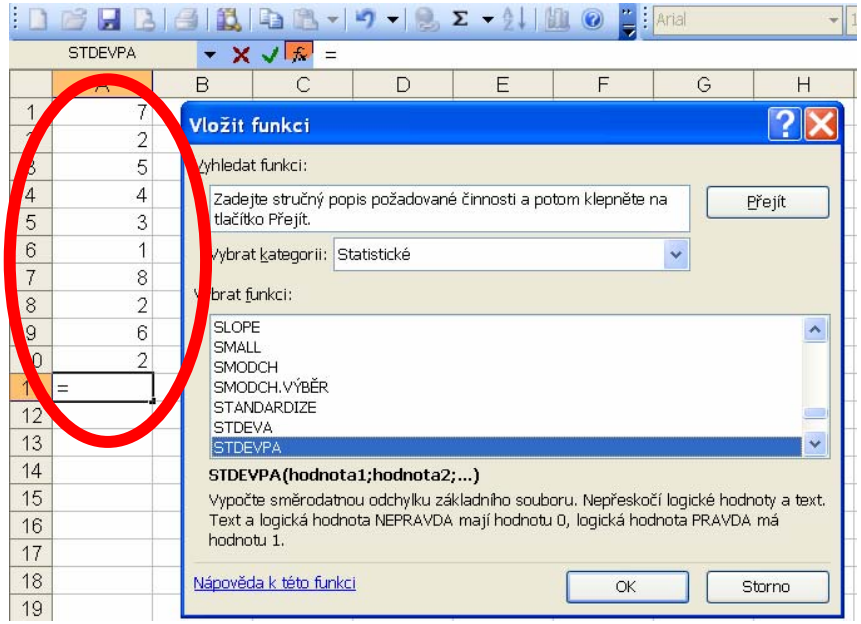
9; 4; 1; 0; 1; 9; 16; 4; 4; 4

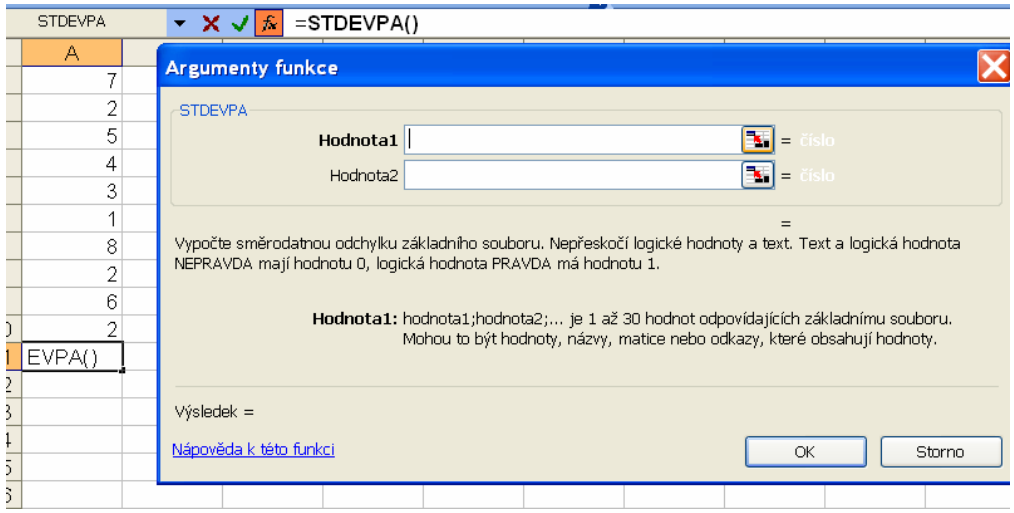
součet čtverců odchylek = 52

průměrná čtvercová odchylka tj. **rozptyl** = 52/10= 5,2

směrodatná odchylka (odmocnina z rozptylu) = 2,28

Směrodatná odchylka v Excelu





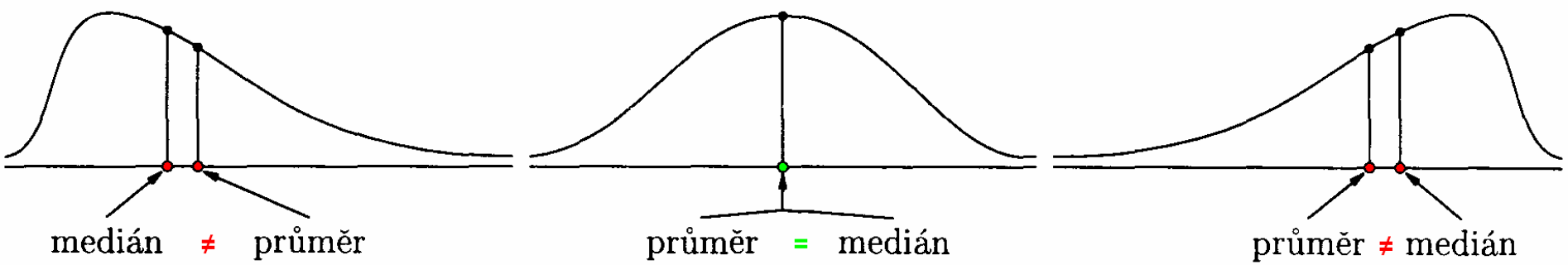
STDEVPA pro **základní** soubor
STDEVA pro **výběrový** soubor

Další popisné statistiky

- Minimum / maximum
- Rozpětí
- Kvantily: dolní a horní kvartil
- Koeficienty šikmosti

Na co si dát v datech pozor

Vzájemná poloha průměru a mediánu



(a) vpravo sešikmené (b) symetrické (c) vlevo sešikmené

Přesnost měření

je funkcí **celkové chyby měření** = jak se rozchází naměřené a skutečné výsledky, má dvě složky

a) Nevýběrová chyba (nonsampling error)

faktory uvnitř i vně metodiky výzkumu obtížně zjistitelné: chybně formulované otázky, nezastihneme všechny vybrané respondenty doma, lidé nechtějí odpovídat, neříkají pravdu,….

b) Výběrová chyba (sampling error)
výsledky ve vzorku se lišší od cílové populace, lze statisticky vyčíslit

Intervaly spolehlivosti

Tolerance chyb (margin of error)
suma všech možných výběrových chyb, která kvantifikuje nejistotu výsledků měření → pravděpodobnostní interval -/+ (např. 95% interval spolehlivosti určuje rozpětí kolem naměřené hodnoty)

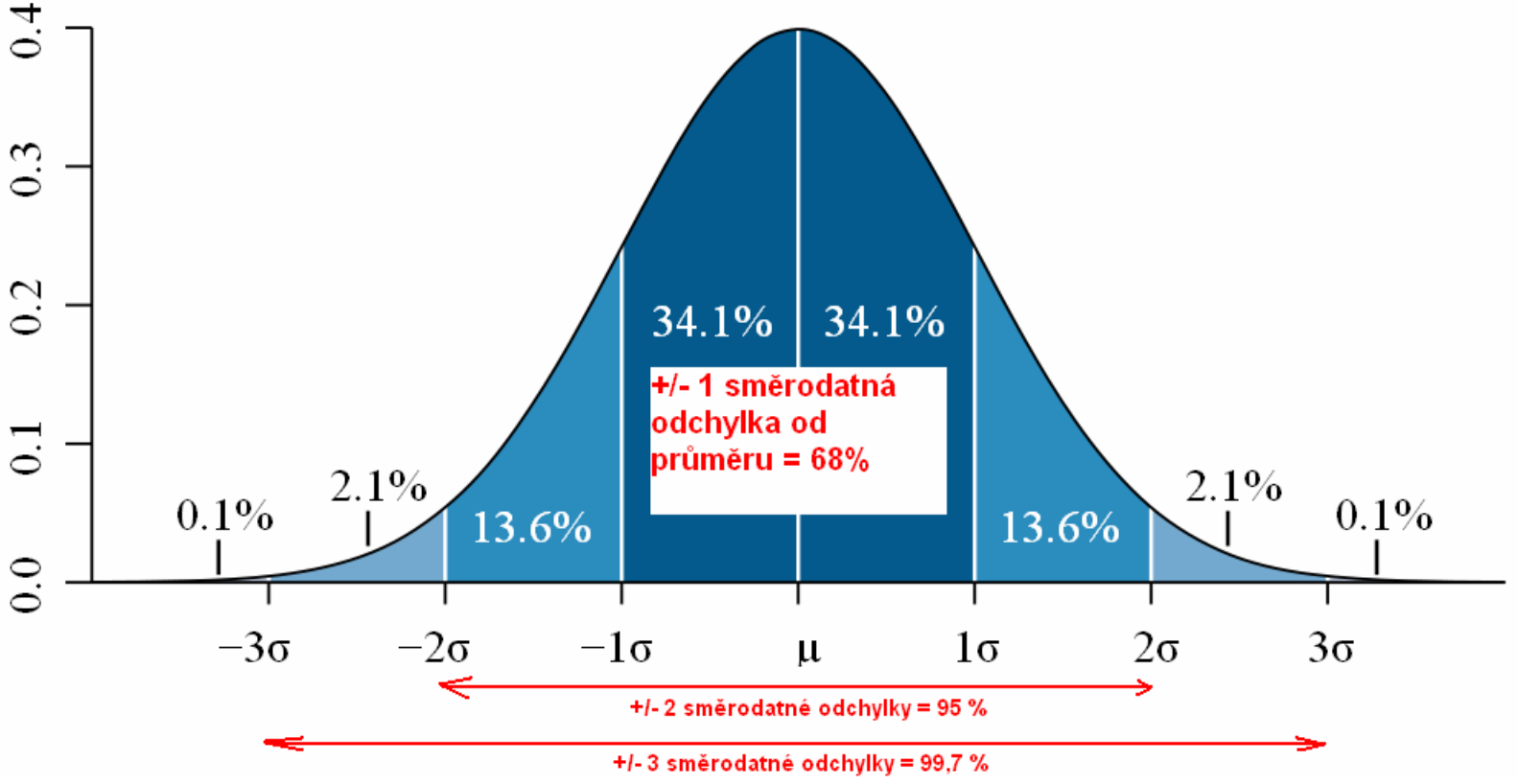
ovlivněno: velikostí výběru, metoda výběru, velikost populace

95 % (konfidenční) interval spolehlivosti

→ jsme si jistí, že naše výběrová data z 95 % budou obsahovat skutečnou hodnotu v celé populaci

Směrodatná odchylka a (konfidenční) interval spolehlivosti

- Normální rozložení



http://www.stat.tamu.edu/~west/applets/ci.html

Odhad parametrů v populaci na základě výběrového vzorku

- Standardní chyba průměru

StD Error (of mean) **s.e.** = $\sqrt{\frac{s^2}{n}}$

kde s^2 je rozptyl (ve výběrovém vzorku)

95 % konfidenční interval pro výběrový průměr
= X ± C * s.e.
kde C = 1,96 (pro 95 % CI)

Výpočet konfidenčního intervalu výběrového průměru

- Hypotetická **populace**

  Průměr v celé populaci μ = 8

jednotky	A	B	C	D	E	F
hodnoty	2	6	8	10	10	12

- Náhodný **výběr** 2 jednotek (např. respondentů)
  A (=2) a **D** (=10)
- Průměr ve výběru X = (2+10)/2 = 6
- Rozptyl ve výběru 4

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \quad \rightarrow \quad \frac{(2-6)^2 + (10-6)^2}{2-1} = 32$$

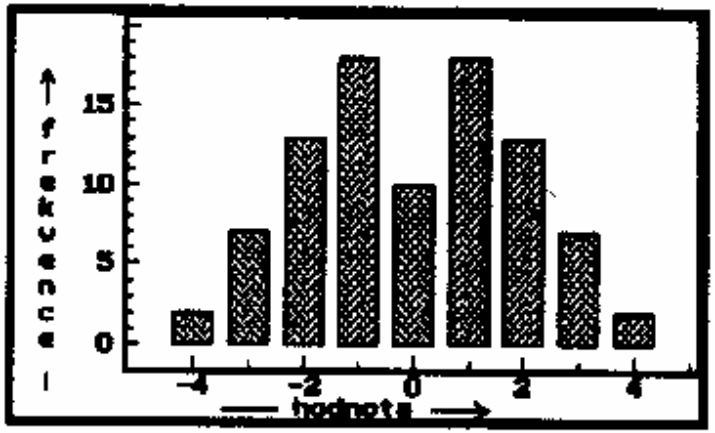
$$s = \sqrt{s^2} \rightarrow \sqrt{32} = 5{,}7 \qquad \text{s.e.} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{32}{2}} = 4$$

CI = X ± 1,96 * 4 = **6 ± 7,84** → **-1,84 až 13,84**

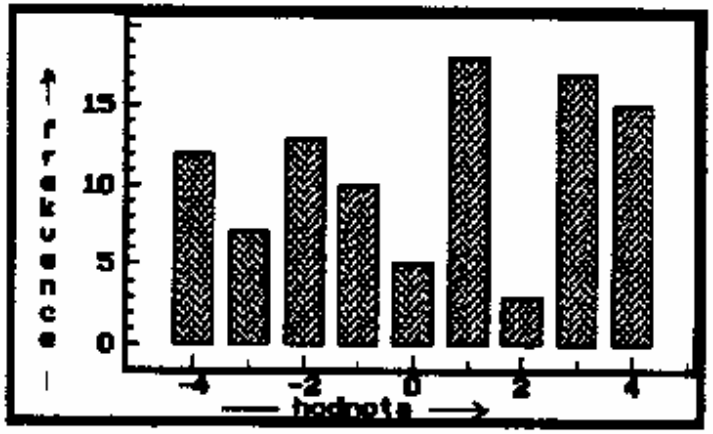
Vlastnosti rozdělení znaků

Symetrie, variabilita

symetrie

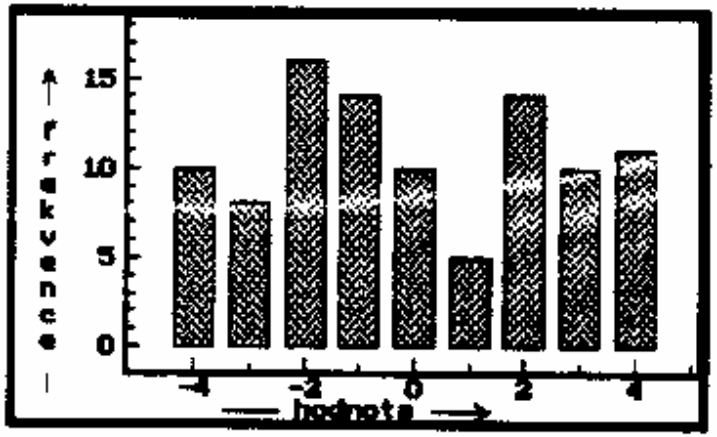


Obr I.A.4a Symetrické rozdělení

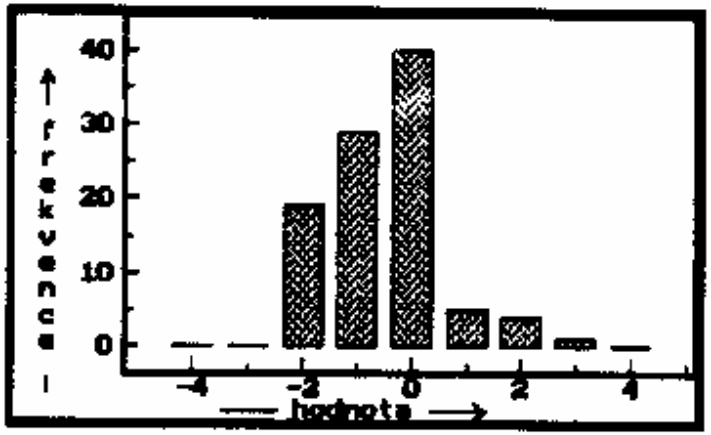


Obr. I.A.4b Asymetrické rozdělení

variabilita



Obr. I.A.4c Velká variabilita

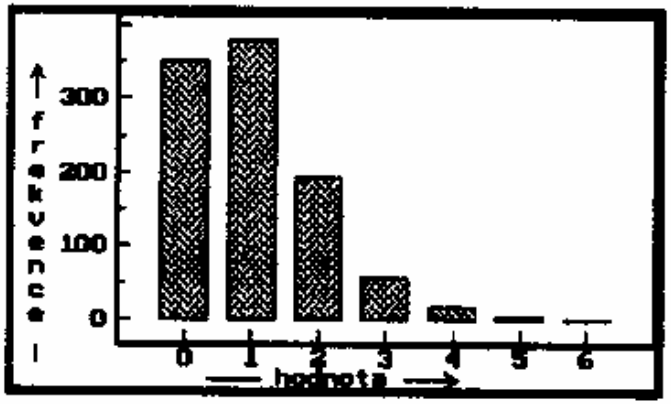


Obr. I.A.4d Malá variabilita

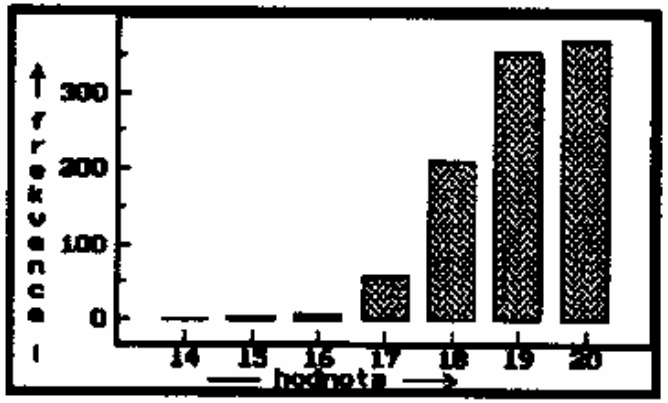
[Hanousek, Charamza 1992: 21]

Šikmost a špičatost

šikmost

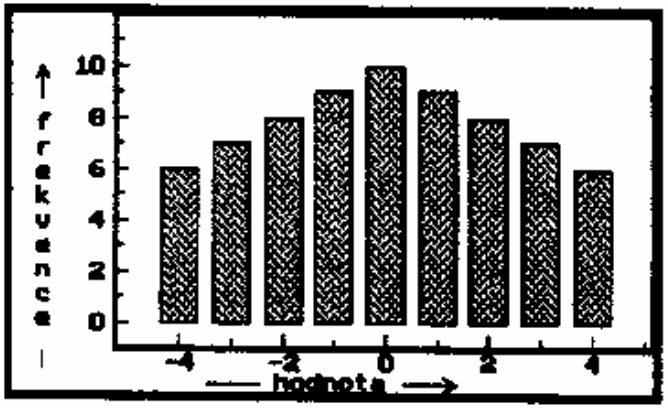


Obr. I.A.5a Kladně sešikmené

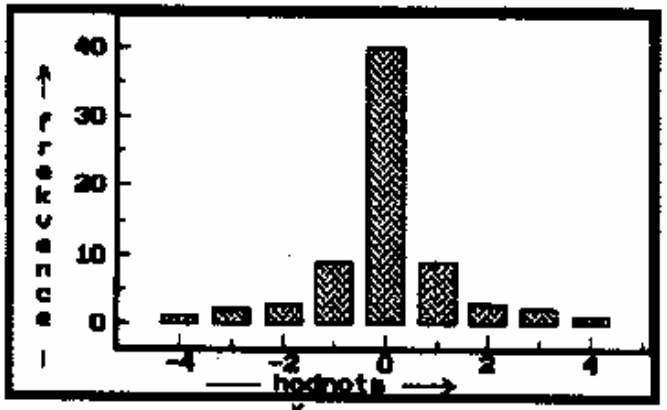


Obr. I.A.5b Záporně sešikmené

špičatost



Obr. I.A.5c Ploché rozdělení



Obr. I.A.5d Špičaté rozdělení

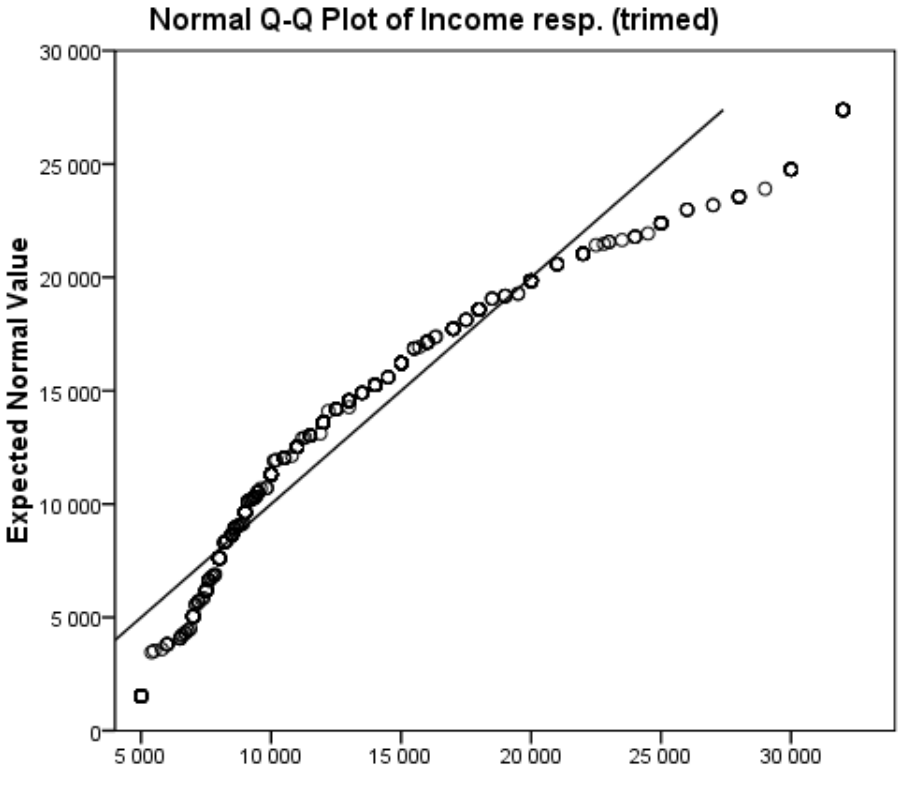
[Hanousek, Charamza 1992: 21]

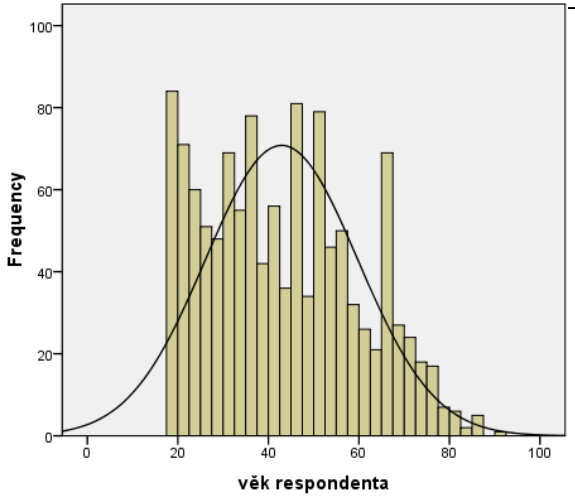
Ověření normality rozložení dat

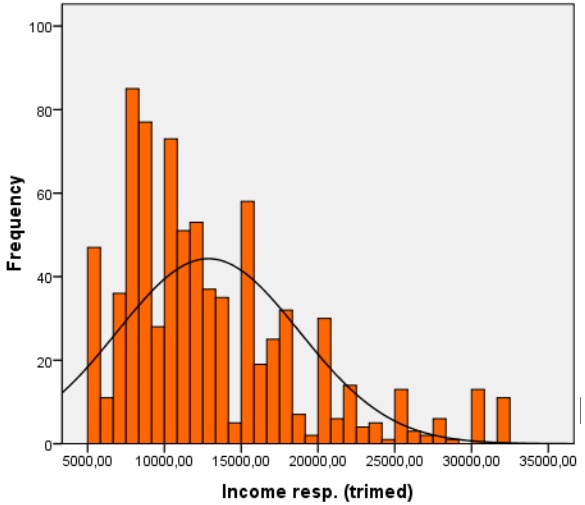
- **Q-Q graf (quantile-quantile):** ukazuje kvantily pozorované distribuce proměnné proti kvantilů zvolené distribuční funkce
- **Normálně rozložená data → přímkový charakter**

  v SPSS: Analyze, Descriptive statistics, Q-Q plots
- **Kolmogorov-Smirnov test:** H0 = data jsou normálně rozložena, **Pozor: nízké! *p* (< 0,05) → distribuce dat se signifikantně lišší od normální distribuce.**
  v SPSS: Analyze, Nonparametric Tests, 1-Sample K-S...
- **Porušení normality rozložení**
  → rekódování, transformace (např. logaritmická), použití neparametrických metod

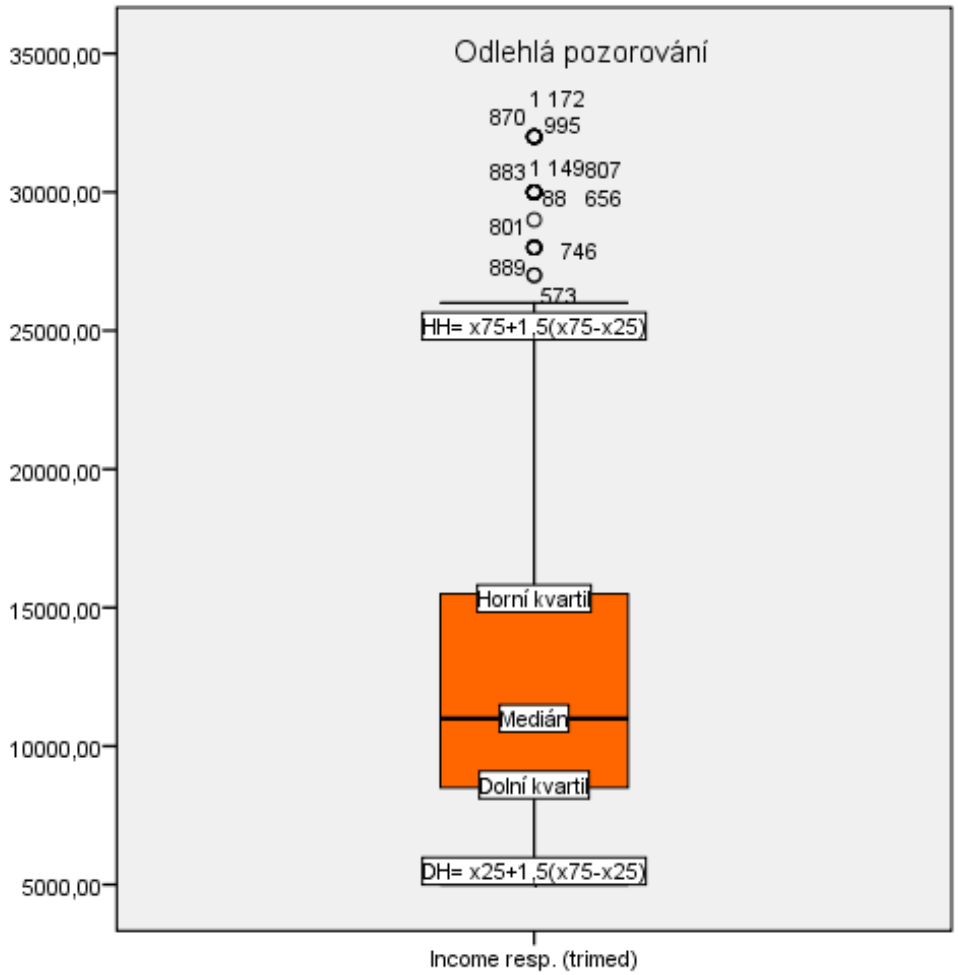
Rozložení četností a Q-Q graf

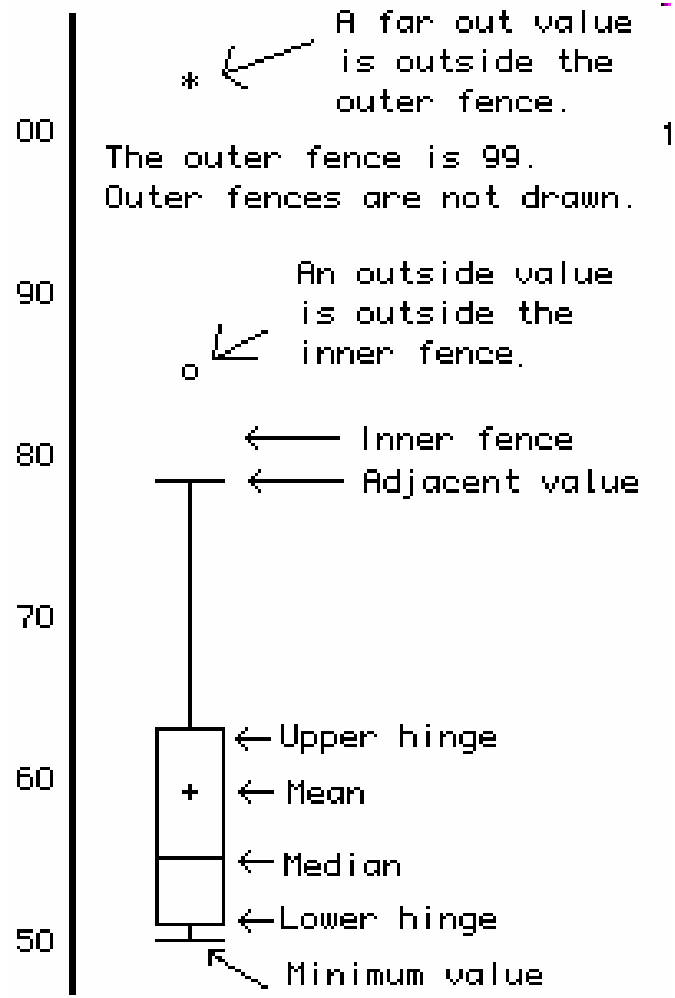






Boxplot – vousaté kabičky: vizualizace distribuce





KVARTILY dělí statistický soubor na desetiny: *dolní Q*0,25 (Q1) a *horní Q*0,75 (Q3)

Interkvartilové rozpětí:

HH = horní kvartil + 1,5 násobku interkvartilového rozpětí

DH = dolníkvartil + 1,5 násobku interkvartilového rozpětí

Testování hypotéz

Vstupní poznámka

Vícerozměrná analýza

Vztahy mezi dvěma a více proměnnými

Testování hypotéz

Statistická hypotéza H0: „žádný rozdíl“ (variabilita v datech je náhodná) → testem hodnotíme sílu dokladu proti tomuto předpokladu

H1: alternativní, platí, když neplatí H0 „existence rozdílů / závislosti“

- **Hladina významnosti α** = pravděpodobnost, že zamítneme H0, ačkoliv ona platí. → „míra naší ochoty smířit se s výskytem chyby“. Obvykle 0,05 či 0,01, což je ale pouze konvence.
- **Hodnota významnosti *p*** - pravděpodobnost realizace hodnoty testovací statistiky, pokud platí H0. **Dosažená hladina hodnoty $p \leq \alpha$ ukazuje na neplatnost H0.**

K testování hypotéz podrobněji viz [Hendl 2006: 176-188]

Testování hypotéz

- *p*-hodnoty nevypovídají nic o síle evidence → jsou závislé na velikosti výběru
- Nezamítnutí H0 neznamená její důkaz.
- Jednostranné testy (test zda hodnota leží napravo/nalevo, tj. vyšší /nižší, od očekávané hodnoty) **Dvoustranné testy**: odchylky od H01 bez ohledu na směr
- Chyba I druhu → H0 platí, ale my jí zamítneme
- Chyba II: druhu → H0 neplatí, ale my jí nezamítneme (přijmeme)

Statistické testy:

Z-test → porovnání průměrů, známe směrod. odchylku populace

T-test → porovnání průměrů, stejné rozptyly neznáme směrod. odchylku populace

F-test → porovnání rozptylů

Neparametrické: **Chí-kvadrát**, Komolgorův-Smirnovův rozdělení ve 2populacích, Mann-Whitney test (dvouvýběrový t-test Mediánu ve dvou subpopulacích) Wilkoxnův, …

Statistická indukce a testování hypotéz

→ **zobecňování výsledků z výběrového souboru na základní soubor**

Při tom musí být splněny předpoklady:

- velkého náhodného výběru (n > 30)
- z dostatečně velké populace (min 100x větší než plánovaný vzorek),
- musí jít o výběr, pro celou populaci (census) nedává smysl

Podrobně viz [Soukup, Rabušic 2007].

Testování hypotéz: **dvouvýběrový T-test**

Testujeme rozdíl v průměru mezi dvěma podskupinami.

Group Statistics

	s30 Pohlaví	N	Mean	Std. Deviation	Std. Error Mean
prijem	1 muž	365	12,93	7,507	,393
	2 žena	486	9,06	5,420	,246

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
prijem	Equal variances assumed	45,970	,000	8,733	849	,000	3,870	,443	3,001	4,740
	Equal variances not assumed			8,350	632,052	,000	3,870	,464	2,960	4,781

1. Levenův test rovnosti rozptylů

2. T-test o „rovnosti průměru mezi podskupinami“.
Nulová hypotéza předpokládá, že průměry se v podskupinách (zde pohlaví) v celé populaci neliší, tedy že jsou způsobeny náhodou.

Test v principu neříká nic jiného, než že **riziko zobecnění výsledku z našeho náhodného výběru na celý základní soubor je pod 5 %**.

Při interpretaci výsledků proto vždy sledujte věcnou významnost.
Např. je rozdíl v průměrném příjmu mezi muži a ženami 3870 Kč substantivní?

Testování hypotéz: **One-way ANOVA**

→ obecnější test pro dvě a více kategorií nezávislého znaku, včetně porovnání rozdílu mezi podskupinami.

Descriptives

prijem

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1 ZŠ	107	6,48	4,871	,471	5,54	7,41	0	28
2 VYUČ	360	10,57	4,809	,253	10,07	11,07	0	38
3 SŠ	325	11,66	7,783	,432	10,82	12,51	0	55
4 VŠ	51	14,04	9,342	1,308	11,41	16,67	0	45
Total	843	10,68	6,682	,230	10,23	11,13	0	55

ANOVA

prijem

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2785,592	3	928,531	22,383	,000
Within Groups	34805,639	839	41,485		
Total	37591,231	842			

H0 nepřijímáme (Sig. < 0,05): alespoň jedna kategorie nezávislého znaku (vzdělání) se liší od ostatních. Které se odlišují? → Post hoc test.

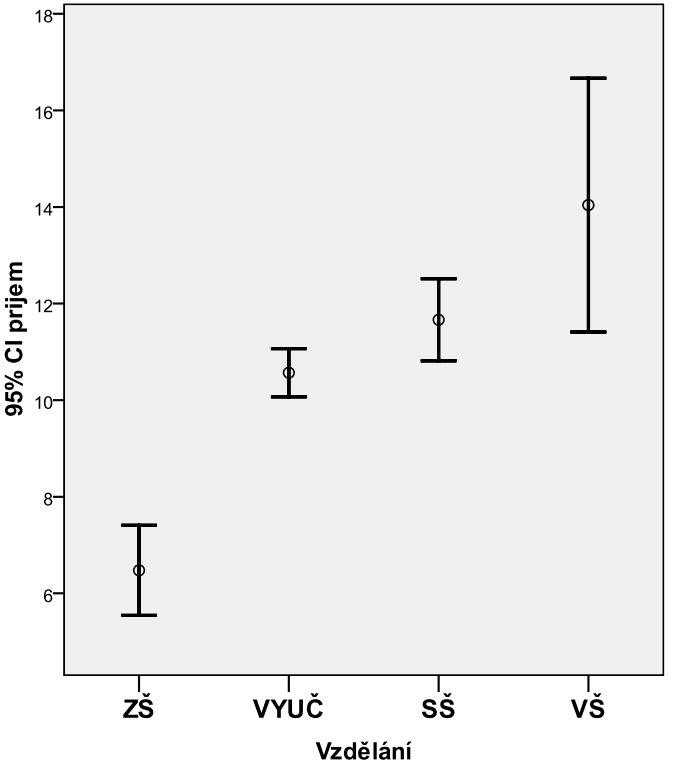
Testování hypotéz: **One-way ANOVA – Post hoc test** (Bonferroniho korekce)

Multiple Comparisons

prijem
Bonferroni

(I) vzd4 Vzdělání	(J) vzd4 Vzdělání	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1 ZŠ	2 VYUČ	-4,091*	,709	,000	-5,97	-2,22
	3 SŠ	-5,188*	,718	,000	-7,09	-3,29
	4 VŠ	-7,563*	1,096	,000	-10,46	-4,66
2 VYUČ	1 ZŠ	4,091*	,709	,000	2,22	5,97
	3 SŠ	-1,097	,493	,158	-2,40	,21
	4 VŠ	-3,471*	,964	,002	-6,02	-,92
3 SŠ	1 ZŠ	5,188*	,718	,000	3,29	7,09
	2 VYUČ	1,097	,493	,158	-,21	2,40
	4 VŠ	-2,375	,970	,087	-4,94	,19
4 VŠ	1 ZŠ	7,563*	1,096	,000	4,66	10,46
	2 VYUČ	3,471*	,964	,002	,92	6,02
	3 SŠ	2,375	,970	,087	-,19	4,94

*. The mean difference is significant at the 0.05 level.



Post hoc test pro podskupiny ukazuje, které skupiny nezávislé proměnné se v průměrech liší (s 95 % jistotou, že tomu tak je i v populaci).

zde: ZŠ od všech ostatních; VYUČ od ZŠ a VŠ

Viz též graf s intervaly spolehlivosti

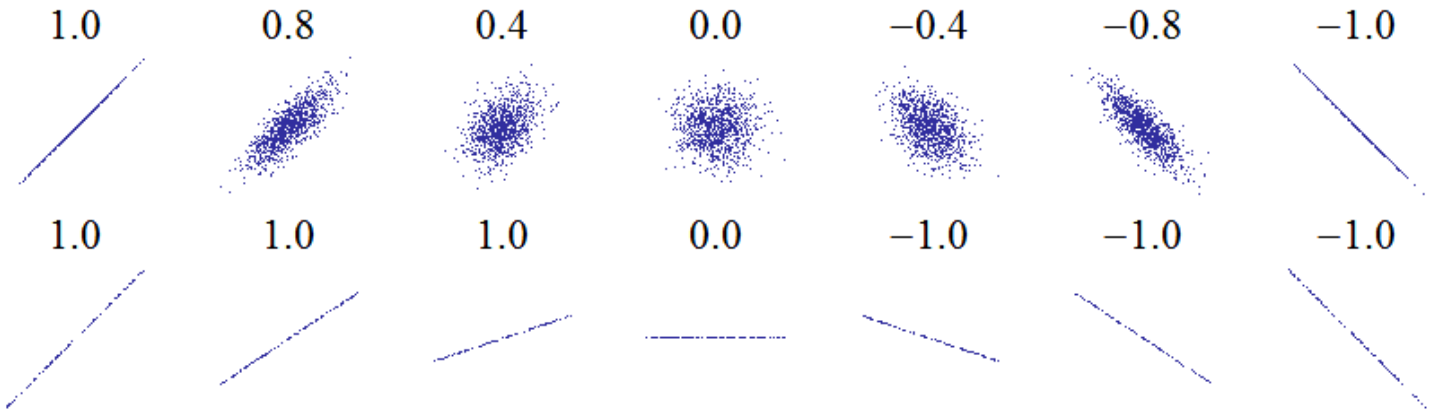
Korelace

- **Korelační koeficient – Pearsonův**
  pro číselné znaky (s normálním rozdělením)

$$r = \frac{1}{n-1} \Sigma \frac{x - \bar{x}}{s_x} \cdot \frac{y - \bar{y}}{s_y}$$

Sx a Sy směrodatné odchylky

1 = přímá závislost
0 = žádná statisticky zjistitelná **lineární** závislost → *i při nulovém korelačním koeficientu na sobě veličiny mohou záviset* !
−1 = nepřímá závislost: čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků,



Korelace: test hodnoty v populaci

- Je třeba pomocí T-testu otestovat, zda je korelace přítomná i v populaci (základním souboru).
- Testujeme, zda se jeho hodnota ve výběru lišší od populační hodnoty.
- H0: korelace v základním souboru je nulová (je způsobená náhodou) r = 0

$$t = r * \sqrt{\frac{n-2}{1-r^2}}$$

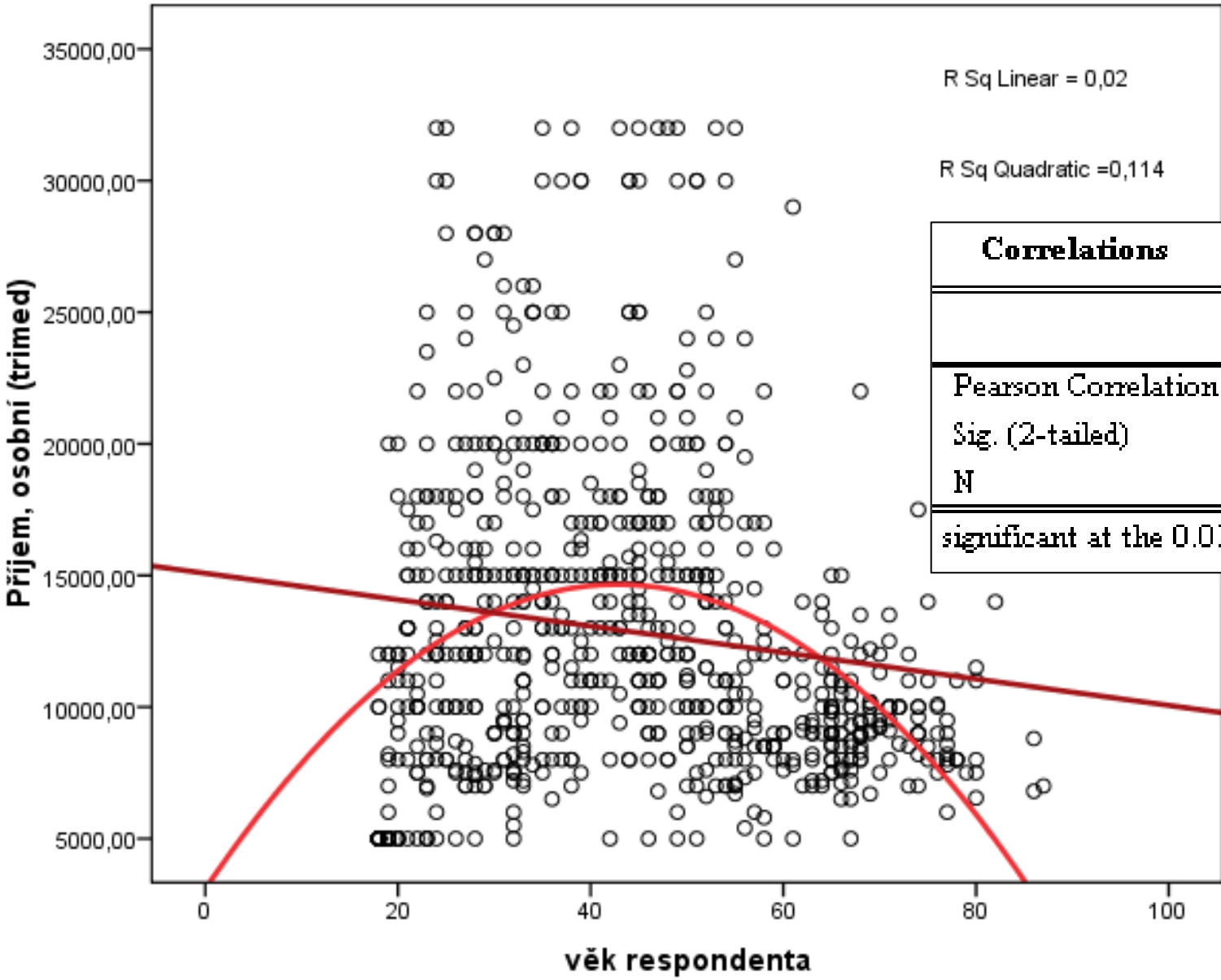
- Porovnáme s tabulkovou hodnotou (dle stupňů volnosti) na hladině významnosti, např. (oboustranný test). Je-li tabulkové $\mathbf{t}_{0,05} > \mathbf{t}$ pak H0 nezamítáme
  → hodnota r není významně rozdílná od 0; korelace je v populaci nulová.

Korelace a vysvětlená variance

r	r^2
1.00	1.00
.90	.81
.80	.64
.70	.49
.60	.36
.50	.25
.40	.16
.30	.09
.20	.04
.10	.01
.0	.0

- Umocněním r – korelačního koeficientu dostaneme
  Rsq – koeficient determinence.
- Ten nám říká kolik variance znaku X jsme vysvětlili pomocí znaku Y

Korelace: věk a příjem; Scatterplot



Correlations

	Incom_R Income resp. (trimed)
Pearson Correlation	-,142**
Sig. (2-tailed)	,000
N	788

significant at the 0.01 level (2-tailed).

Pořadová korelace: pro ordinální proměnné

- **Spermanův korelační koeficient Rho**
- +1 = úplná shoda pořadí jednotek podle obou znaků
- **Kendallovo Tau**
- ve srovnání s Pearsonovým r, ale i Spearmanovým Rho několik výhod → větší citlivost na některé nelineární vztahy.
  Více k porovnání korelačních koeficientů viz [Hendl 2004: 259-262].

Asociace nominálních znaků: Kontingenční koeficient

- Analogie korelačního koeficientu (ten je pro kardinální/ordinální znaky)
  → míra těsnosti závislosti.
- Výsledek není kontingenčních tabulkách v intervalu (0,1) → různé korekce
  CC je rozšíření Phi pro >2x2 tabulky.
-

$$CC = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

V SPSS: Analyze, Descriptive Statistics, Crosstabs; vložit Row a Column variables; → Statistics; → Contingency Coefficient / Phi & Cramer‘s V

Pořadová korelace: př. Soc. Distance

	CR	USA_rank14	GER_rank14	CR_rank14
1	Majitel/ka / vedoucí obchodu	6	6	2
2	Ředitel/ka podniku	4	5	3
3	Lékař/ka	1	1	1
4	Učitel/ka ZDŠ	5	2	6
5	Truhlář/ka	7	9	8
6	Automechanik/čka	9	4	5
7	Dělník/ce v továrně	10	12	11
8	Řidič/ka nákladního auta	12	11	10
9	Prodavač/ ka v hypermarketu	11	10	9
10	Nekvalif. stavební dělník/ce	13	13	13
11	Metař/ metařka	14	14	14
12	Mistr/ová v továrně	8	8	12
13	Univerzitní profesor/ka	3	7	7
14	Projektant/ konstruktér/ka	2	3	4

Correlations

			USA_rank14 USA-RANK 14	GER_rank14 GER-Rank 14
Kendall's tau_b	CR_rank14 CR-rank 14	Correlation Coefficient	,648**	,714**
		Sig. (2-tailed)	,001	,000
		N	14	14
Spearman's rho	CR_rank14 CR-rank 14	Correlation Coefficient	,824**	,873**
		Sig. (2-tailed)	,000	,000
		N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Korelace: parciální korelace

- kontrolovaný vliv 3 proměnné

Parciální korelace pro X,Y/U s kontrolou vlivu U
(platí i pro neparametrické korelace, např. Spearman)

$$\tau_{(X,Y)/U} = \frac{\tau_{X,Y} - \tau_{X,U} * \tau_{Y,U}}{\sqrt{(1-\tau^2_{X,U}) * (1-\tau^2_{Y,U})}}$$

Příklad: korelace příjmu a věku, při kontrole vlivu vzdělání („čistý“ efekt)

věk-příjem	R x,y	-0,14
věk-vzdělání	R x,u	-0,10
příjem-vzdělání	R y,u	0,33

x - věk
y - příjem
u - vzdělání

R x,y/u = $\frac{-0,11}{0,94}$ **=** **-0,12**

Analýza rozptylu

Jednoduchá analýza rozptylu
One-way ANOVA

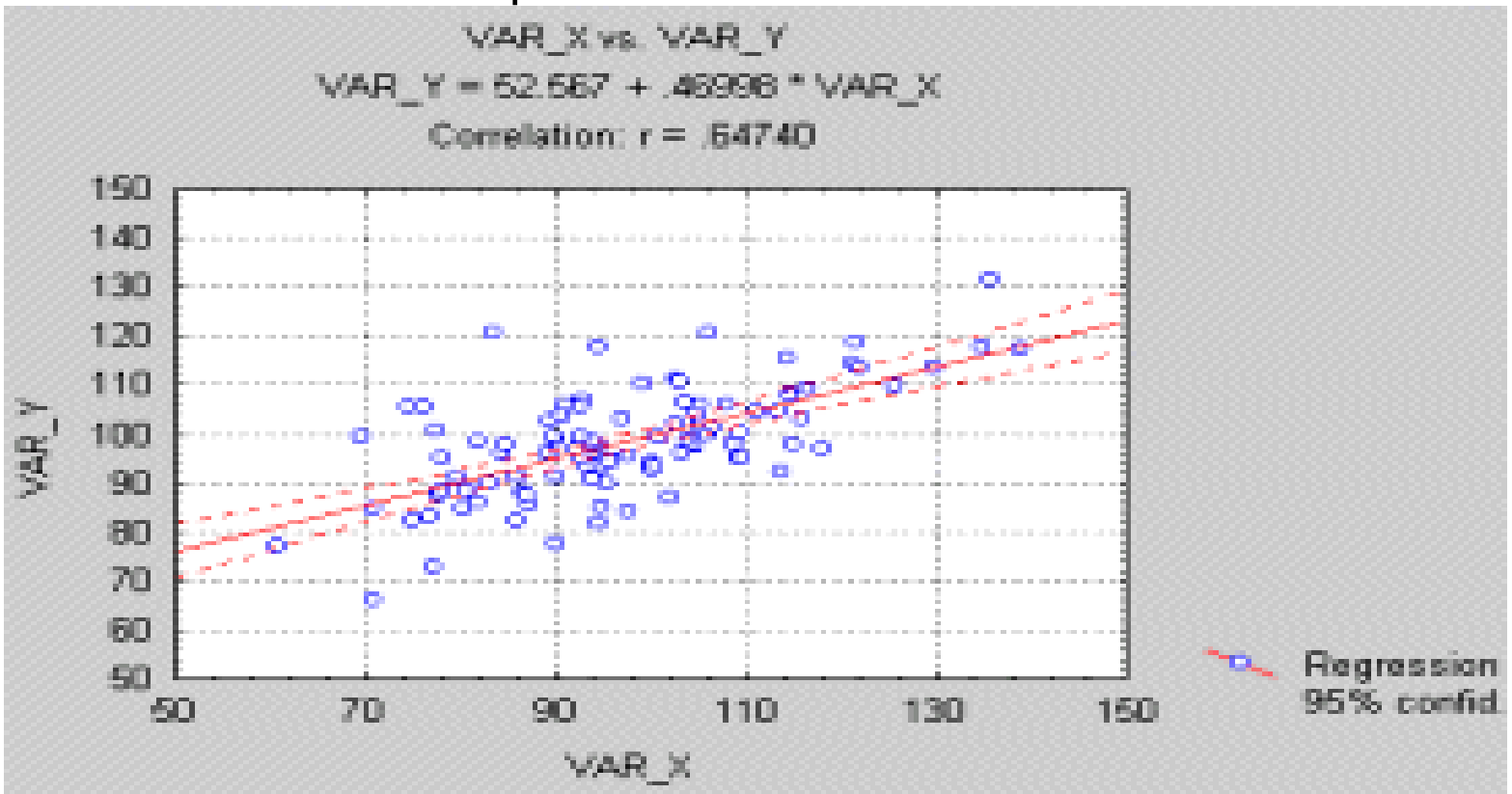
- Proměnná nominální (ordinální) x kardinální
- Rozdílnost rozptylu číselné proměnné podle kategorií nominálního znaku
- Založena na F-statistice

Lineární regrese

Odhadujeme hodnotu závislého znaku (y) na základě znalosti jiných veličin - nezávisle proměnných (x, ….).

y = a + bx

y = hodnota závislé,
a = konstanta (typická závislé při nejnižší hodnotě nezávislé,
b = regresní koeficient „o kolik vzroste Y, když se x změní o jednotku“,
x = hodnota nezávislé proměnné

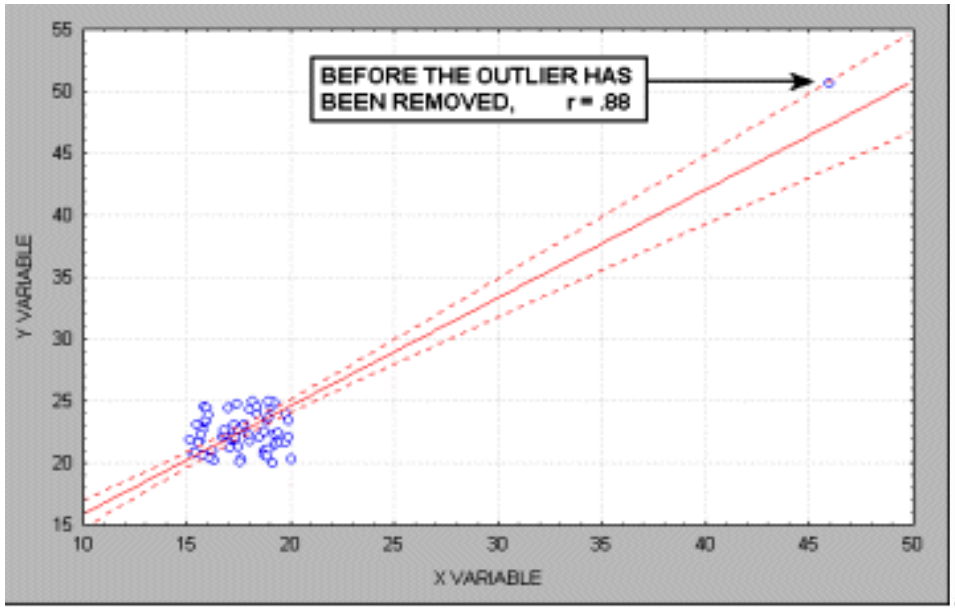


Na co si dát pozor

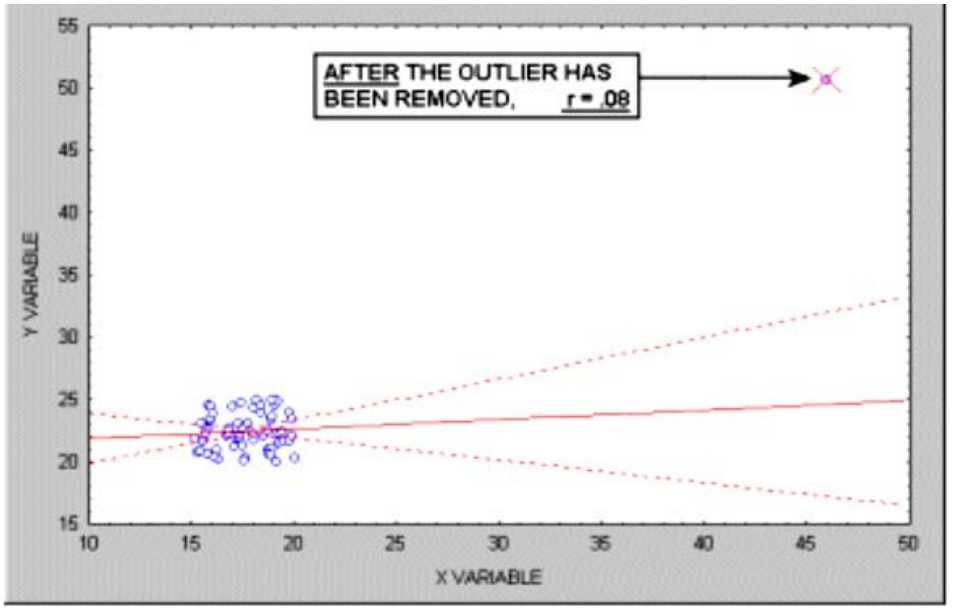
Vícerozměrná analýza

Odlehlá pozorování (outliers)

R = 0,88



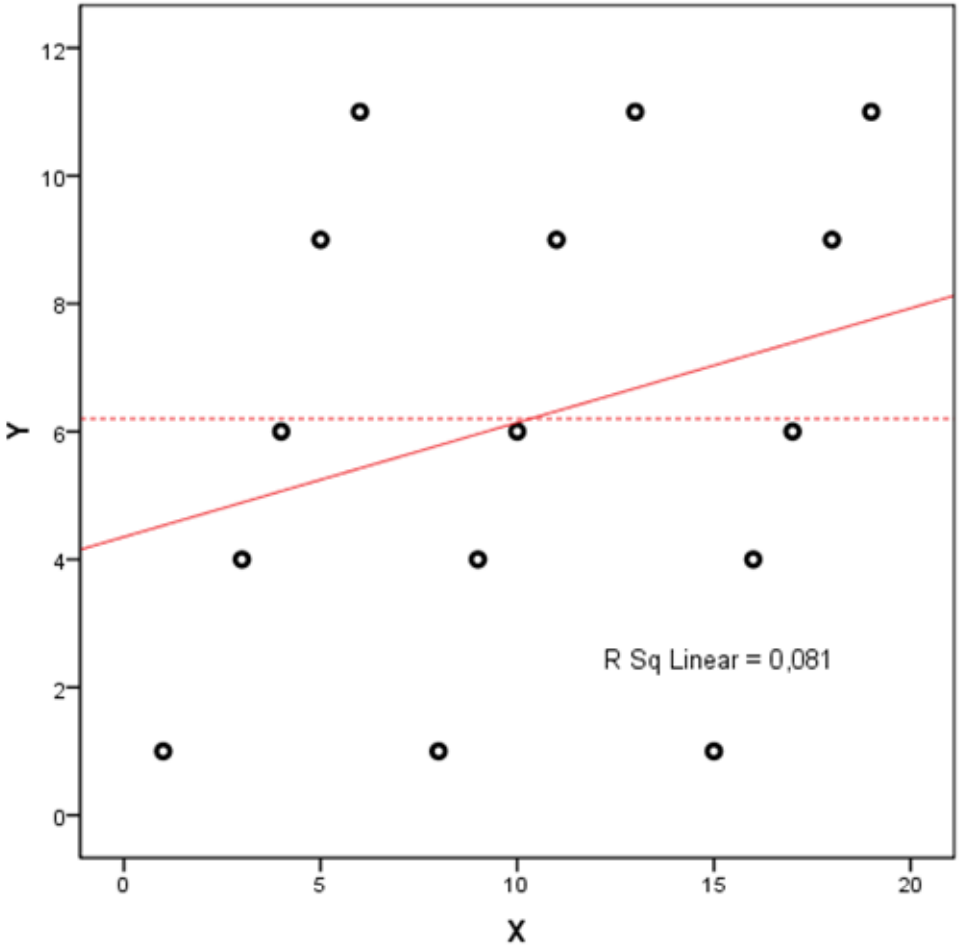
R = 0,08

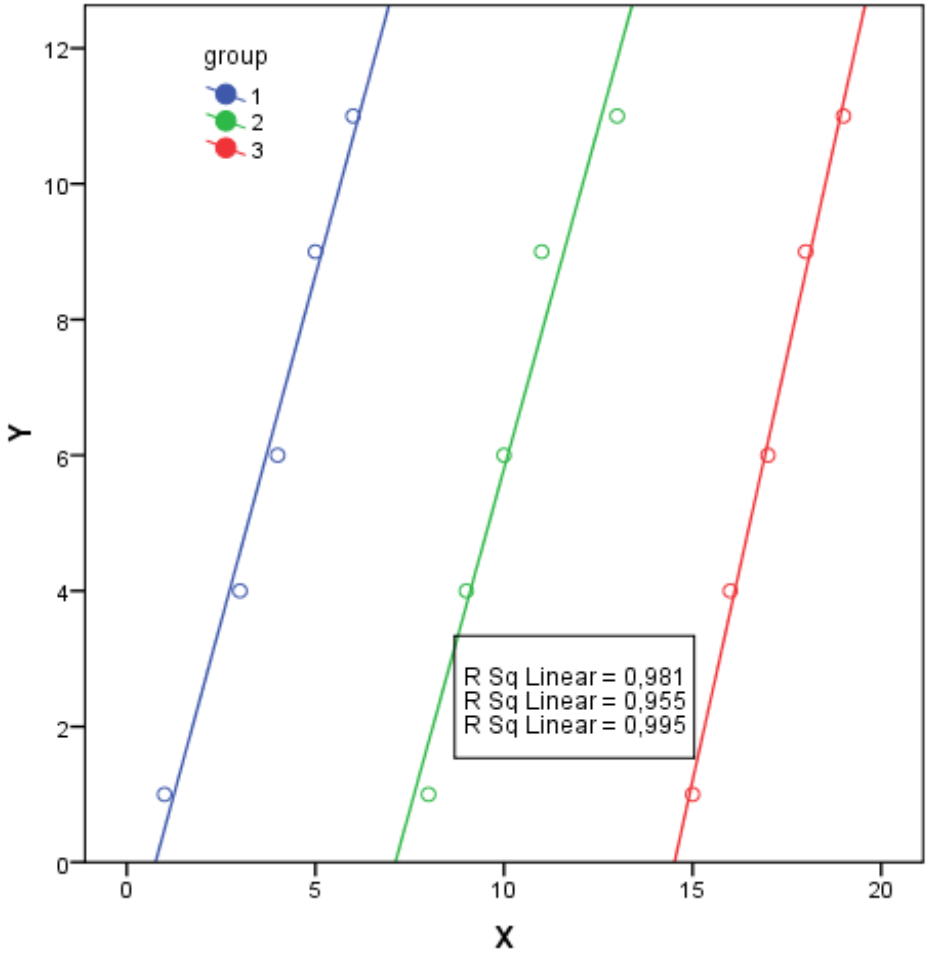


Téměř všech rozptyl byl vnesen pouze jedním pozorováním.
Outliers mohou významně ovlivnit vztah dvou (a více) znaků!

Vždy nejprve zjistit odlehlá pozorování → Scatterplot

Konfigurace v datech na základě podskupin



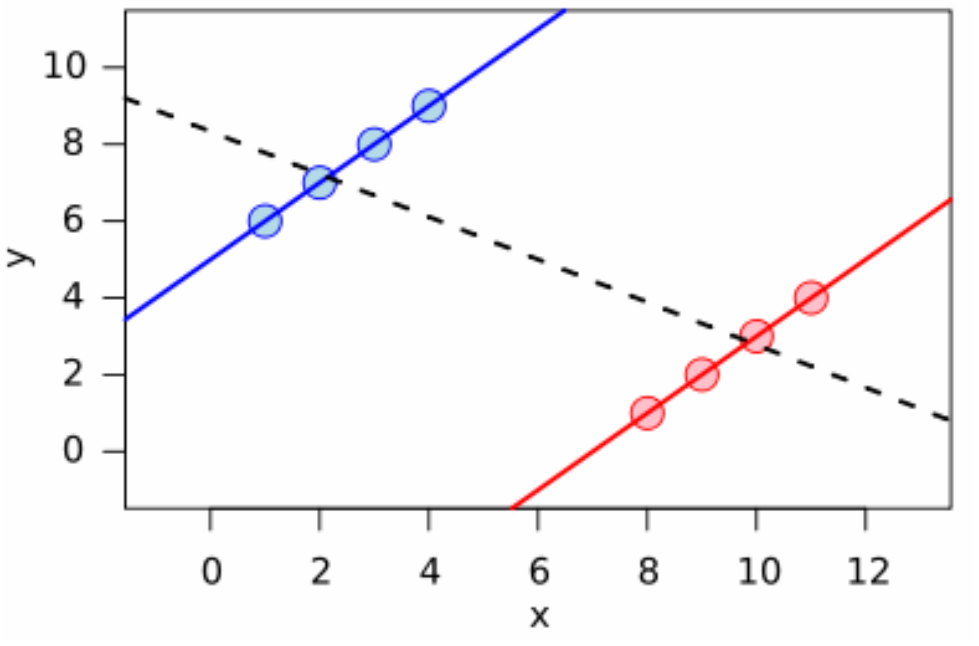


[Disman 1993: 210-211]

- Pozor korelační koeficient ukazuje jen na míru souvislosti,
  ale **neříká nic o kauzalitě** – směru působení mez dvěma znaky.

Simpsonův / reversal paradox – špatná inference z agregovaných dat

- Obrácení závislosti (směru působení) v konntingenční tabulce způsobeného působením třetí proměnné.
- Hrozí při agregaci dat.



V agregovaných datech (černá čára) je negativní souvislost

V oddělených podskupinách (modrá a červená čára) je ovšem pozitivní trend

Neparametrické testy (Non-parametric Tests)

- Parametrické metody předpokládají: **náhodný výběr**, **normální rozdělní** (distribuce znaku), **velké výběry** z populace, známé (shodné) rozptyly v sub/populacích, z nichž byl proveden výběr
- **Neparametrické metody**:
  - nezávislé na rozdělní
  - méně citlivé na odchylky extrémních hodnot
  - i pro výběry velmi malého rozsahu
  - vhodné pro nominální i ordinální znaky
- Ale dochází častěji k chybnému nezamítnutí nepravdivé H0.
- Chí-kvadrát testy,

Kategoriální data

Kontingenční tabulka

Kontingenční tabulka

Statistické míry a testování

- **Nezávislost** = oba znaky navzájem neovlivňují v tom, jakých konkrétních hodnot nabývají
- **Homogenita** (shodnost struktury) = očekávané četnosti jsou v políčcích každého řádku ve stejném vzájemném poměru bez ohledu na konkrétní volbu řádku
- → **test dobré shody** = porovnání očekávaných četností v jednotlivých polích tabulky - za předpokladu, že hodnoty obou sledovaných znaků na sobě nezávisí - a skutečných četností.
- Pokud hypotéza nezávislosti (resp. homogenity) platí, má testová statistika přibližně rozdělení chí kvadrát o (r-1)(s-1) stupních volnosti. Hodnota testové statistiky se tedy porovná s kritickou hodnotou (kvantilem) příslušné hladiny významnosti.

Kontingenční tabulka

- Pro použití testů založených na testu dobré shody (test nezávislosti nebo homogenity) je třeba, aby se v tabulce vyskylo **méně než 20 % políček, v nichž by očekávané četnosti byly menší než 5.** V případě, že se tak stane, můžeme zvážit transformaci — sloučení některých méně obsazených kategorií (např. "ano" a "spíše ano").

Kontingenční tabulka

- Statistika chí kvadrát nevypovídá nic o síle vztahu, pouze zamítá/nezamítá nulovou hypotézu o závislosti nebo homogenitě na dané hladině významnosti alfa.
- Pro zjištění síly vztahu →
  - koeficienty (obdobné korelaci: CC),
  - podíl šancí (OR),
  - u ordinálních veličin koef. dle pořadí.

  Odlišné testy pro nominální a ordinální proměnné (jedna / obě).

Chí-kvadrát testy: test dobré shody

- Test pro homogenitu distribucí mezi kategoriemi znaku/ů
- Pro nominální znaky (i ordinální a kardinállí)
- Nevyžaduje znalost předschozího rozdělení znaku
- Očekávané frekvence
- Odpovídá na otázku, zda jsou rozíly mezi empirickými (pozorovanými - f_O) četnostmi a teoretickými (očekávanými -f_E) četnostmi náhodné nebo ne.

$$\chi^2_{obt} = \sum \frac{(f_O - f_E)^2}{f_E}$$

- Počet stupňů volnosti df = (r-1) (s-1)

  r = počet řádků s = počet sloupců v tabulce

Chí-kvadrát test: příklad

- Pozorované četnosti kategorií

velmi nízký	střední	vysoký	
5	10	9	$\sum n_i = n = 24.$

očekávané (teoretické) četnosti = 24 : 3 = 8.

H0: počet respondentů je ve všech kategoriích stejný

5	10	9	n_i
8	8	8	$\tilde{n}_i$

$$\chi^2 = \sum \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}$$

$$\chi^2 = \frac{(5-8)^2}{8} + \frac{(10-8)^2}{8} + \frac{(9-8)^2}{8} = 1{,}74$$

Chí-kvadrát test: příklad

- Určení stupňů volnosti df = k - 1 – r
- k - počet kategorií r - počet parametrů předp. rozdělní
- Kritický bod z tabulky statist významnosti pro Alpha 0,05
- Pokud vypočítaná X < X kritická → nelze odmítnout H0 (= četnosti jsou mezi kategoriemi stejné).

Chí-kvadrát test: příklad: Kouření marihuany u žáků 9 a 12 třídy.

TABLE 6.14 Marijuana Use

	Grade		
	9th	12th	Total
None	42	33	75
Marijuana	23	22	45
Total	65	55	120

Because 45 out of 120 of the total sample (9th- and 12th-graders) used marijuana during the past year, the proportion for the total sample is 45/120 = .375. The ex-

Chí-kvadrát test: příklad:

TABLE 6.15 Observed and Expected Frequencies for Marijuana Use

	Grade		
	9th	12th	Total
None	42 (40.625)	33 (34.375)	75
Marijuana	23 (24.375)	22 (20.675)	45
Total	65	55	120

NOTE: Expected frequencies are in parentheses.

$$\chi^2_{obt} = \sum \frac{(f_o - f_e)^2}{f_e}$$

$df = (\text{Rows} - 1)(\text{Columns} - 1), \quad df = (2 - 1)(2 - 1) = (1)(1) = 1.$

Chí-kvadrát test: příklad

TABLE 6.16 Computation of χ^2_{obt}

Observed (f_O)	Expected (f_E)	$\lvert f_o - f_e \rvert - 0.5$	$(\lvert f_o - f_e \rvert - 0.5)^2$	$\frac{(\lvert f_O - f_E \rvert - 0.5)^2}{f_e}$
42	40.625	.875	0.765625	0.019
33	34.375	.875	0.765625	0.022
23	24.375	.875	0.765625	0.031
22	20.625	.875	0.765625	0.037

NOTE: $\chi^2_{obt} = 0.019 + 0.022 + 0.031 + 0.037 = 0.109$.

For $df = 1$ and $\alpha = .05$, the critical value for χ^2 is 3.84. Our calculated value (χ^2_{obt}) was 0.109. Because the obtained (calculated) value did not exceed the critical value, we would not reject the null hypothesis at $\alpha = .05$.

Chíkvadrát **kritický** z tabulek > Chíkvadrát **dosažený** (naměřený)

→ Ho nelze zamítnout = homogenita mezi kategoriemi

Chí-kvadrát test: změna v čase

Teoretickou četností zde není poměrové rozložení ale hodnota z předchozí etapy.

Je podle vašeho názoru nabídka kulturních žánrů v našem městě dostatečná?

	Ano	Neví	Ne
Epirická četnost (2010)	65	28	6,7
Teoretická četnost (2007)	60	34	6

$$\chi^2 = \sum \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}$$

Chí-kvadr	**1,53**
tabulková hodnota (pro 5 %)	5,99

Vypočítaná hodnota Chisq je **menší než tabulková-kritická hodnota**. Platí H0 o "nerozdílu„ (rozdíl v četnostech je způsoben náhodnými faktory).

Adjustovaná residua
Znaménkové schéma

- CROSSTABS: Adj. standardised (v SPSS / PSPP)

Adjustovaná residua

- Residuum v daném políčku tabulky (=pozorovaná (observed) minus očekávaná (expected) hodnota) dělený odhadem vlastní standardní chyby. Odpovídající standardizovaný residuál je vyjádřen v jednotkách směrodatné odchylky nad nebo pod průměrem.

Znaménkové schéma → jednoduchá vizualizace

- 'kde abs(z) >= **3.29** nahradí +++ resp. ---,
- 'kde abs(z) >= **2.58** nahradí ++ resp. --,
- 'kde abs(z) >= **1.96** nahradí + resp. -.

Dodatek: uděláno středa

- Descriptives
- Explore – outliers, median, zešikmení,…

Grafy:

- Konfidenční intervaly pro sadu proměnných 8 x různá spokojenost – porovnání (seřazení) mezi nimi
- Konfidenční intervaly pro kategorie proměné příjme x vzdělání

Webové nástroje pro analýzu

Index of On-line Stats Calculators

http://www.physics.csbsju.edu/stats/Index.html

- **Exact r×c Contingency Table:**
  http://www.physics.csbsju.edu/stats/exact_NROW_NCOLUMN_form.html
- **Statistical Calculations**
- http://statpages.org/
- **R. Webster West applets**
  http://www.stat.tamu.edu/~west/
  http://www.stat.tamu.edu/~west/ph/

Učebnice:

Interstat - hypertextová interaktivní učebnice statistiky pro ekonomy
http://www.stahroun.me.cz/interstat/

Statnotes: Topics in Multivariate Analysis, by G. David Garson
http://faculty.chass.ncsu.edu/garson/PA765/index.htm

StatSoft - Elektronická učebnice statistiky (anglicky)
http://www.statsoft.cz/page/index2.php?pg=navigace&nav=31
http://www.statsoft.com/textbook/

Nejprve se ptej, k čemu analýza tvá má sloužit, potom teprv výběrem metody dej se soužit.

[Hanousek, Charamza 1992 : 61