STATISTIKA

Sázíte-li ve Sportce, je to hazard.
Sázíte-li se, že vám v kartách přijdou tři postupky po sobě, je to zábava.
Vsadíte-li se, že cena plynu stoupne o 10 %, je to podnikání. Vidíte ten rozdíl?

Martin Sebera, FSpS MU, 12.2.2014

Pravidla výzkumu z pohledu analýzy dat

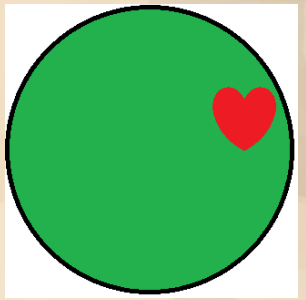
1. příprava výzkumného šetření je nejdůležitější část
2. sběr a analýza dat slouží k zamítnutí/nezamítnutí předem stanovených úkolů práce a hypotéz (explorační vs. konfirmační přístup)
3. vždy mít na paměti věcné hledisko výzkumu, zejména v souvislosti s interpretací statistických výsledků

Role statistiky

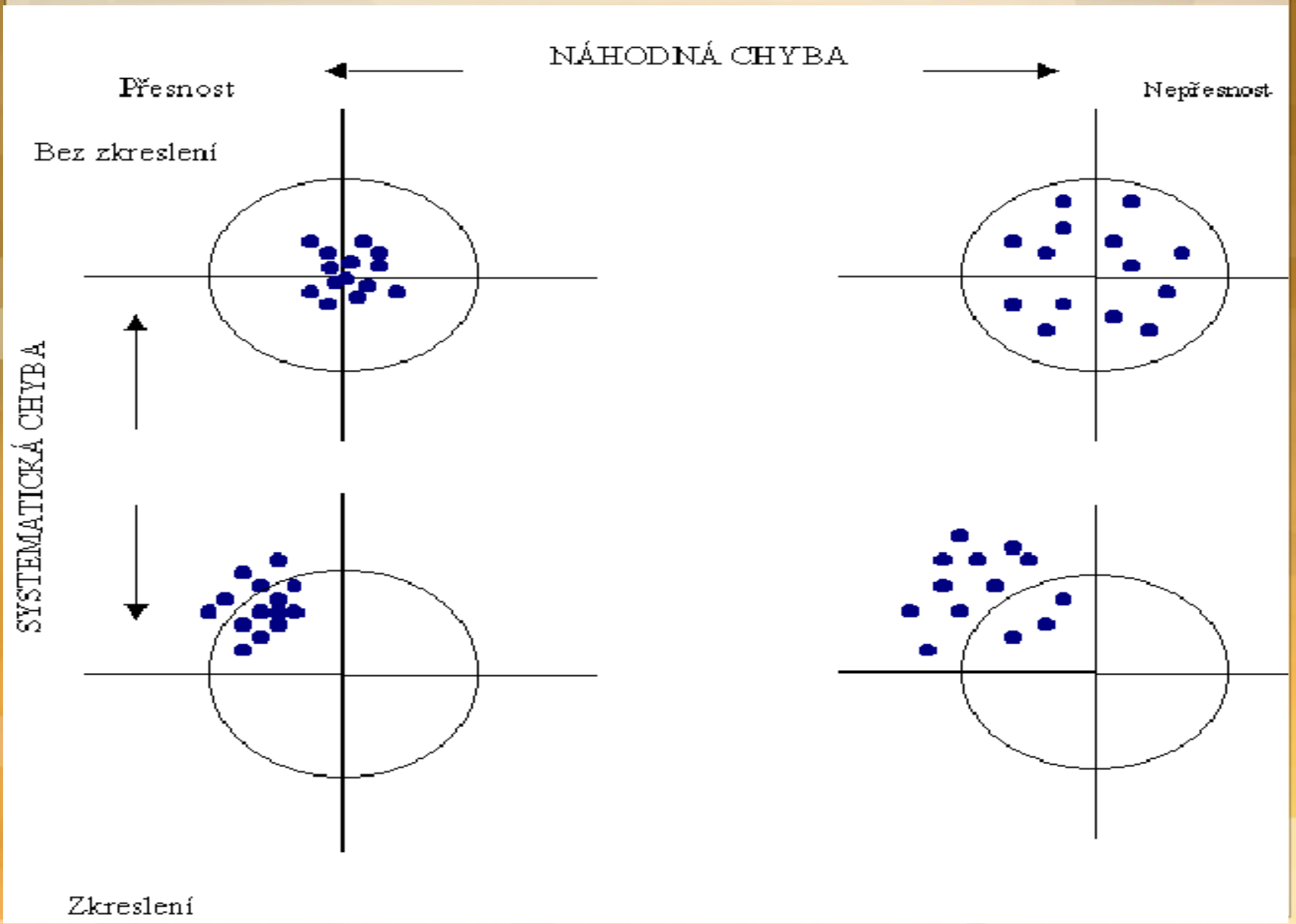
- Porozumění a zkoumání hromadných jevů
- Zjišťování zákonitostí
- V kvantitativním výzkumu (deduktivní princip) – pojítko mezi teorií a výzkumem
- Zpracování, popsání a analyzování dat

Základní pojmy

- Základní a výběrový soubor a jeho rozsah (N)
- Výběr:
  - náhodný (každý prvek má stejnou pravděpodobnost výběru - losování)
  - systematický (n-tý objekt, n<N)
  - stratifikovaný (náhodný výběr ve skupinách)



Náhodná a systematická chyba



Typy proměnných

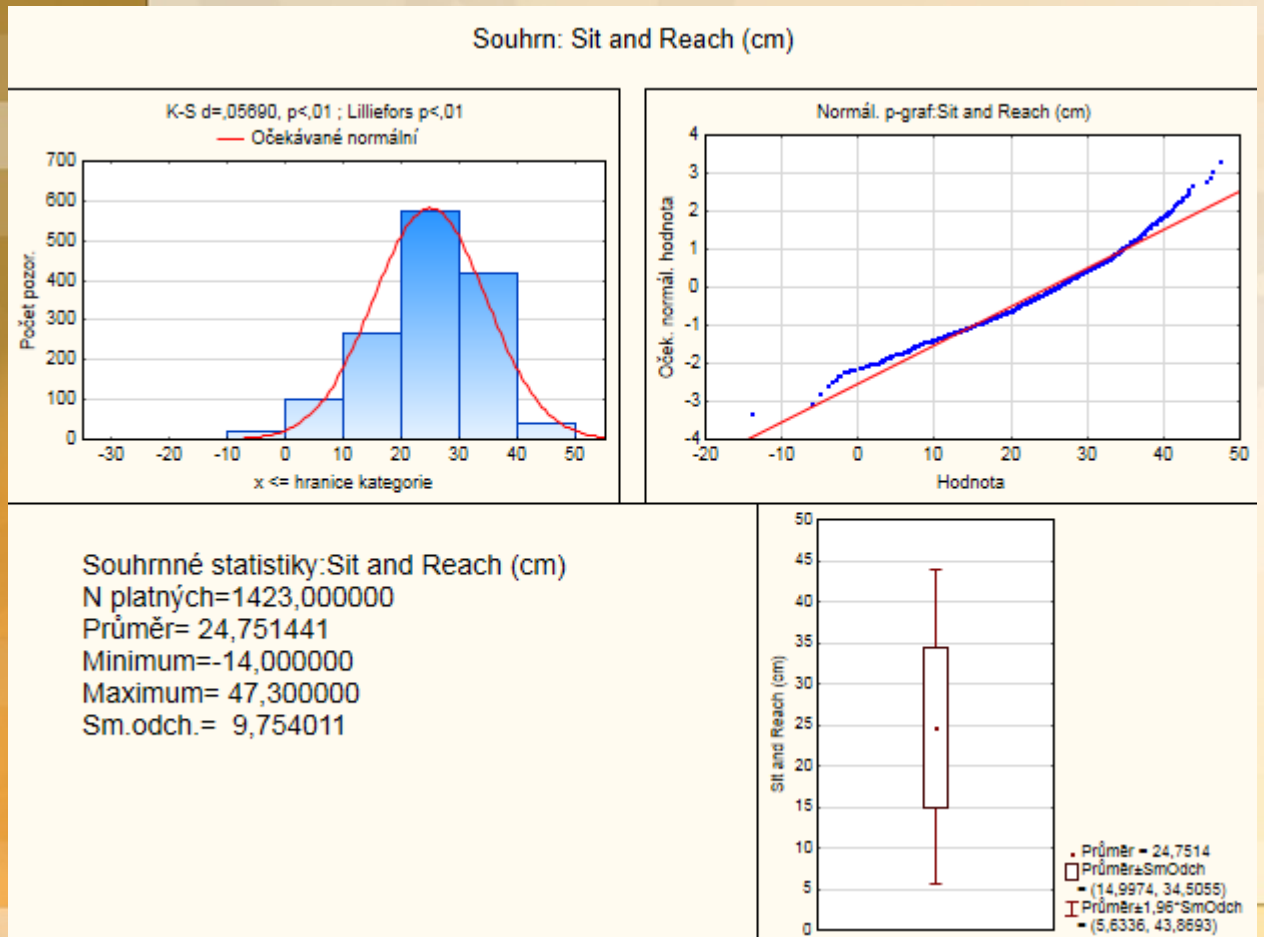
- Nominální (text, číselné kódy; hodnoty jsou různé; nelze provádět aritmetické operace)
- Ordinální (lze seřadit; většinou se převede na čísla),
- Intervalová (lze říct o kolik je hodnota větší)
- Poměrová (lze říct kolikrát je hodnota větší)
- Spojité X Diskrétní

Škály (měřítka, stupnice)

- Nominální (temperament; národnost)
- Ordinální (školní známky, bodování v slopestyle; relace =, ≠, >, <,),
- Metrické
  - Intervalová (lze říct o kolik je hodnota větší)
  - Poměrová (lze říct kolikrát je hodnota větší)
  - Př. teplota, čas, hmotnost, …

První náhled na data – popisná statistika

- průměr, sm. odchylka, medián, kvartily aj.
- četnosti: absolutní, relativní, kumulativní
- grafy: krabicový, histogram

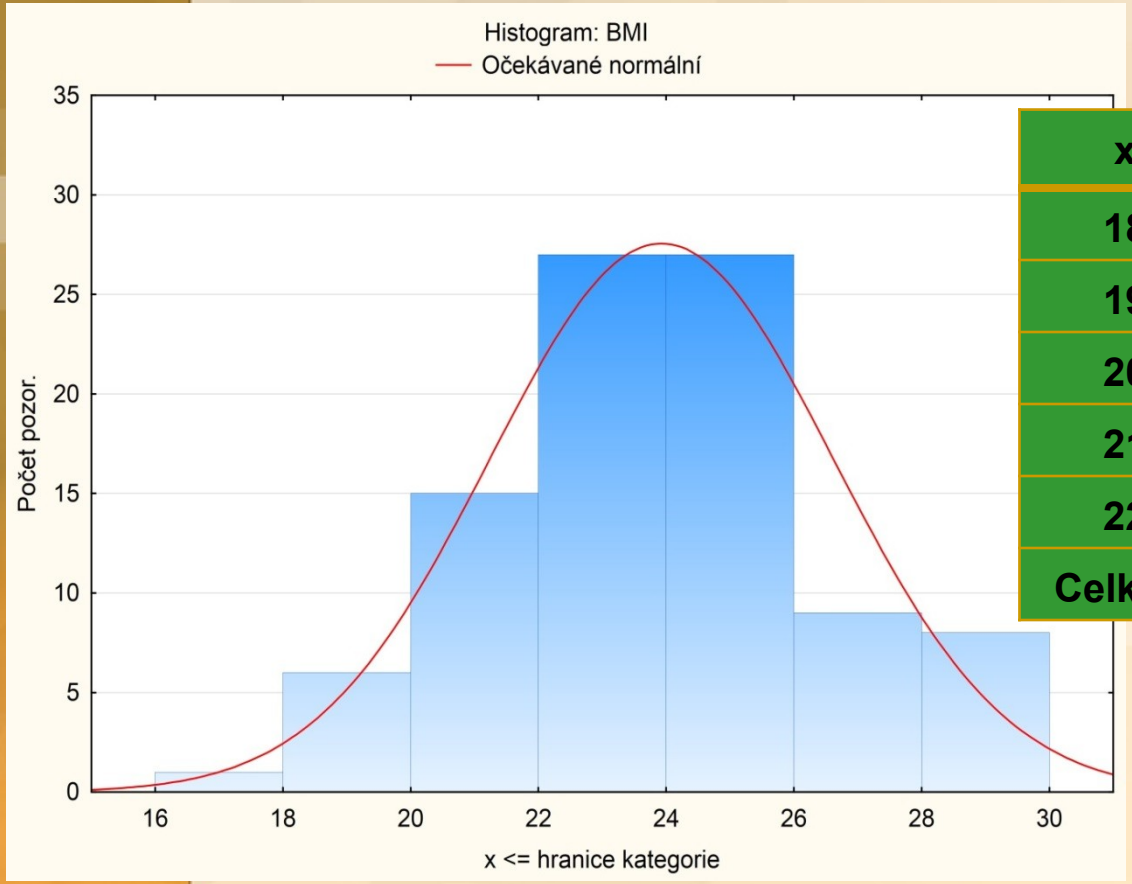


Proč?

- **chybná měření, extrémy**
- **homogenitu souboru**
- **chybějící data**

Intervalové rozložení četností

BMI:
18 19 19 20 20 20 20 20 20 20
20 21 21 21 21 21 21 22 22 22



N – rozsah souboru
n_i – absolutní četnost
r_i – relativní četnost
N_i – kumulativní absolutní četnost
F_i – kumulativní relativní četnost

x	n_i	r_i	N_i	F_i
18	1	0,05 (= 1/20)	1	0,05
19	2	0,10 (= 2/20)	3	0,15
20	8	0,40 (= 8/20)	11	0,55
21	6	0,30 (= 6/20)	17	0,85
22	3	0,15 (= 3/20)	20	1,00
Celkem	20	1,00		

lze usuzovat na některé vlastnosti,
záleží na počtu intervalů

Základní statistické charakteristiky

- **Míry střední hodnota**
  - **Aritmetický a geometrický průměr, modus, medián**
- **Míry variability**
  - **variační rozpětí, kvantily, rozptyl, směrodatná odchylka, variační koeficient**
- **ztrácíme mnoho cenných informací o původních datech**
  - **1; 10; 22** **průměr 11** **SD 10,53** **n = 3**
  - **11; 11; 11** **průměr 11** **SD 0** **n = 3**

Časté chyby při statistických výpočtech

- Uvedení průměru bez směrodatné odchylky SD a bez N
- Procenta
  - Regulovaná složka stoupla o 200 %, silová zlevnila o 20 %. Jak se změnila celková cena?
  - Regul: 100,- Kč → 300,- Kč **původní cena 3100,- Kč**
  - Silová: 3000,- Kč → 2700,- Kč **nová cena: 3000,- Kč**
  - Nejen procenta, ale i z jakých základů se počítají
- snížení platu o 30 % a jeho následné zvýšení o 30 %
  - při původním platu 100 Kč je plat po snížení 70 Kč (-30 %), ale po následném zvýšení o 30 % pouze 91 Kč.

Testování hypotéz, koncept věcné vs. statistické významnosti

Postup testování hypotéz → poměrně jasný a jednoduchý.

- Vytvoříme hypotézu H_0, o které předpokládáme, že platí. Proti ní postavíme alternativu (H_A). Sesbíráme data. Najdeme věrohodný aparát, který konstatuje, zda domněnka platí nebo ne → statistický test.
- **chyba 1. druhu** se značí α a nazývá se **hladina významnosti**. Výraz **1 - α** se nazývá **spolehlivost**
- **chyba 2. druhu** se značí β. Výraz **1 - β** se nazývá **síla testu**
- Obvyklé hodnoty spolehlivost: 0,95 nebo 0,99;
- síla testu např. 0,8 → volíme např. hladinu významnosti $\alpha = 0{,}05$ nebo 0,01.

Testování hypotéz		výsledek testu	
		hypotéza H_0 platí	hypotéza H_A platí
reálná situace	hypotéza H_0 platí	správné rozhodnutí	chyba 1. druhu značí se α
	hypotéza H_A platí	chyba 2. druhu značí se β	správné rozhodnutí

Koncept věcné významnosti

Alternativou k statistické významnosti je posuzování tzv. věcné významnosti (effect size). Lze ji stanovit jako:

- minimální hodnotu v absolutních hodnotách znamenající věcnou významnost
- minimální vysvětlené procento rozptylu (relativní zhodnocení podílu ostatních faktorů – koeficient ω^2)

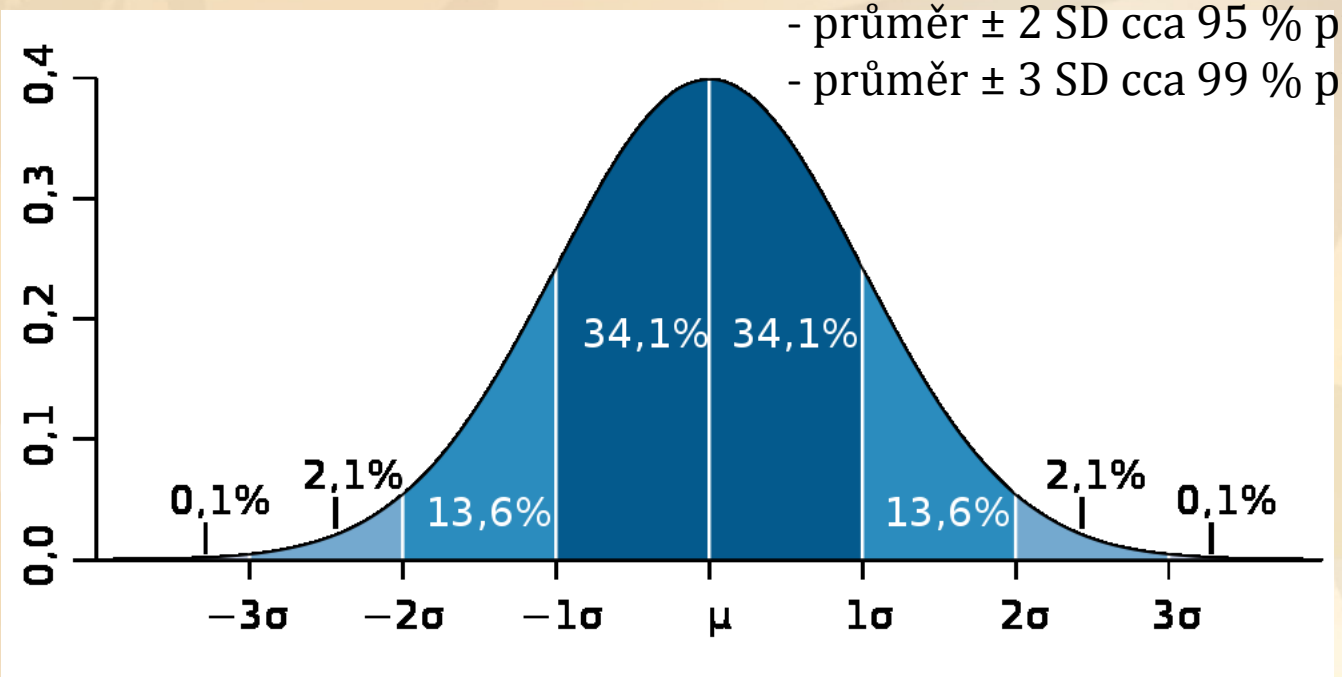
Pro jednotlivé testy lze v literatuře nalézt mnoho tzv. koeficientů věcné významnosti. **Jednou z výhod konceptu věcné významnosti je nezávislost na počtu měření N**.

	koeficient	hodnocení efektu
Korelační koeficient r	r^2 koeficient determinace	malý (nízký) efekt: r = 0,10 – 0,30 střední efekt: r = 0,31 – 0,70 velký (výrazný) efekt: r = 0,71 – 1
t-test	Cohenovo d	d = 0,20 malý efekt d = 0,50 střední efekt d = 0,80 velký efekt

Normalita

- Kolmogorov-Smirnov a Shapiro-Wilks test
- Proč?
  rozhodnutí, zda použít parametrické nebo neparametrické testy

Pro normální rozložení platí:
- průměr ± 1 SD cca 68 % případů
- průměr ± 2 SD cca 95 % případů
- průměr ± 3 SD cca 99 % případů

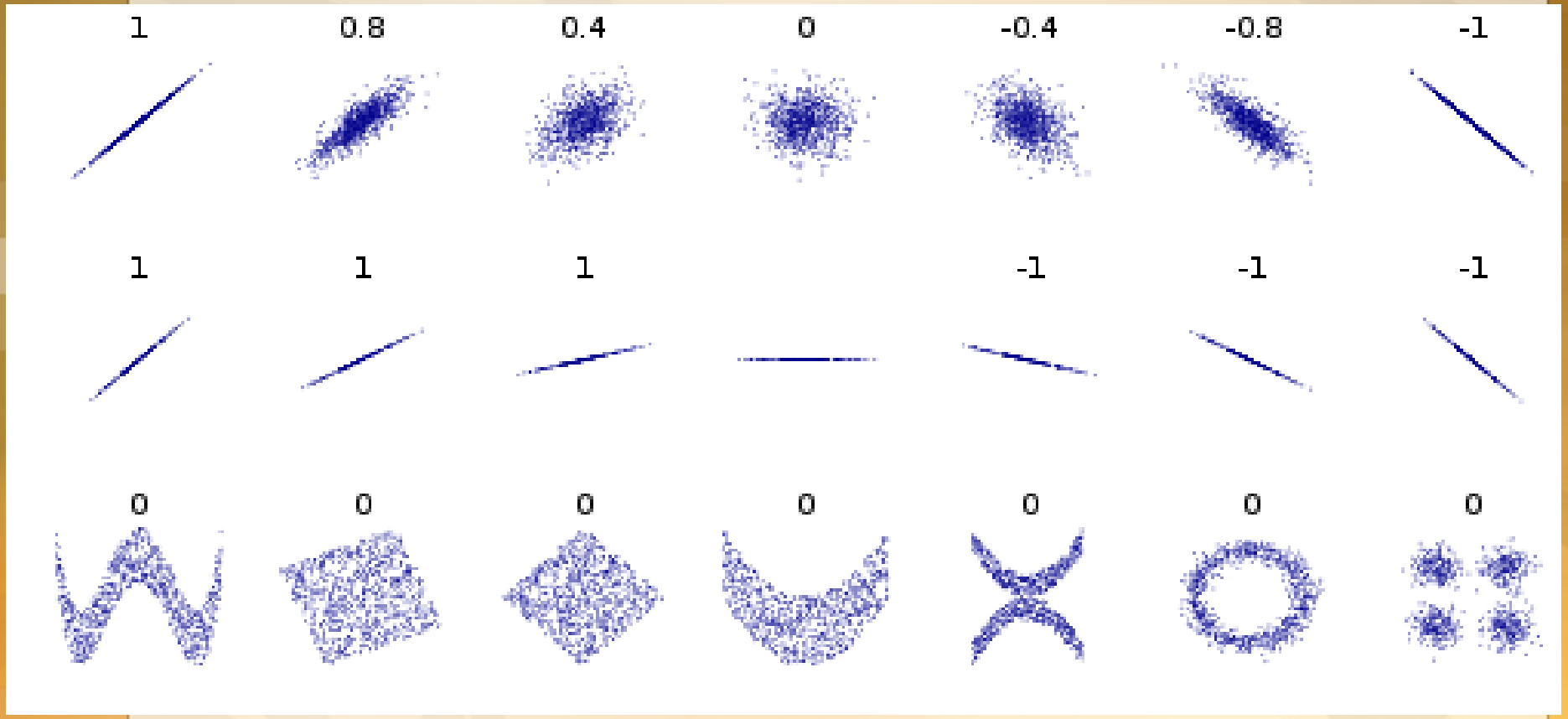


Korelace ANEB korelace není kauzalita

- = vzájemný vztah mezi veličinami proměnnými, jevy (dostatečně velký rozsah)
- Úkol: zjistit závislost a popsat ji
- Př. 3 proměnné:
  - BMI
  - % fat
  - WHR

Korelační koeficient

- *R*: <-1 do 1>
- Omezení:
  - předpokládá 2-rozměrné norm.rozdělení
  - měří pouze vztahy lineární
  - nerozeznává, která proměnná je závislá a která nezávislá. Nelze rozhodnout o příčinnosti vztahu mezi proměnnými
- interpretace → dodatečné koeficienty, např. index determinace r^2
- *Pearsonův, neparametrický Spearmonův*
- *jednoduchý, parciální, mnohonásobný*


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Příklad

	% fat	WHR	BMI
% fat	1	0,36	0,41
WHR	0,36	1	0,85
BMI	0,41	0,85	1

Nejvyšší **jednoduchý** korelační koeficient je mezi proměnnými BMI a WHR a to 0,85. Celkem vysvětluje 72,2 % procent celkové variability mezi těmi to proměnnými. K číslu 72,2 % jsme dospěli pomocí koeficientu determinace ($r^2 = 0,85^2 = 0,722$).

T-testy

- Testy o rovnosti středních hodnot dvou výběrů
- Jaký konkrétní t-test vybrat?
- varianta testu bude
  - parametrická (závislé, nezávislé soubory)
  - neparametrická (Wilcoxonův - závislé, Mann-Whitneyův test nezávislé hodnoty
- Statistická vs. věcná významnost

T-test

T-test - příklad

- **Cohenovo d** $d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$ $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
  - d > 0,8 → velký efekt
  - *d* z intervalu 0,5 – 0,8 → střední efekt
  - d < 0,2 → malý efekt
- **d = 0,44**
- **rozdíl mezi oběma disciplínami je i věcně i statisticky významný.**

Statistický rozcestník ANEB co s daty

1. příprava výzkumného šetření je nejdůležitější část
2. sběr a analýza dat slouží k zamítnutí/nezamítnutí předem stanovených úkolů práce a hypotéz
3. vždy mít na paměti věcné hledisko výzkumu, zejména v souvislosti s interpretací statistických výsledků

CHCI S DATY PROVÉST	ZPŮSOB	UMOŽNÍ MI ZJISTIT
První náhled na data	Základní popisná statistika • průměr, směrodatná odchylka, rozptyl, N, medián, kvartily a další míry polohy a variability • tabulky četností: absolutní, relativní, kumulativní • grafy: krabicový, histogram, bodový	• chybná měření, extrémy • homogenitu souboru • chybějící data • trend v datech
Otestovat normalitu	• Kolmogorov-Smirnov test, Shapiro-Wilks test	• rozhodnutí, zda použít parametrické nebo neparametrické testy
Zjistit, zda výběry/skupiny jsou shodné nebo ne	• 2 skupiny/proměnné: t-testy • 3 a více skupin/proměnných: Analýza rozptylu (ANOVA) ○ T-testy i ANOVA má svou parametrickou i neparametrickou variantu!	• konstatovat statisticky nebo věcně (size of effect) významný rozdíl Př. došlo ke zlepšení výbušné síly po intervenci ?(pretest-posttest) Př. která ze dvou tréninkových metod je úspěšnější? Př. mezi kterými skupinami je statisticky významný rozdíl Př. byl zkoumán výsledný čas v motorickém testu v závislosti na typu suplementace sportovce (faktor A) a na způsobu tréninku (faktor B)
Zjistit závislost více proměnných (spojité)	• korelace, index determinace • faktorová analýza	• těsnost lineárního vztahu mezi proměnnými • může existovat jasný vztah ale nelineární, který nezachytíme pomocí korelace nebo faktorové analýzy • korelace neznamená kauzalitu!!! Př. závisí výkon v běhu na 100 m s výkonem do skoku do dálky? Př. závisí ekonomika běhu na povrchu?
Zjistit závislost více proměnných (kategoriální-např. dotazník)	• test nezávislosti chí-kvadrát v kontingenční tabulce • vícerozměrné kontingenční tabulky - asociační stromy • shluková analýza • regresní a klasifikační stromy (CART, CHAID)	• sílu a směr vztahu Př. závisí bolestivost zad na věku a způsobu zaměstnání? Př. mezi kterými proměnnými z dotazníku existuje nejsilnější vazba?
Redukovat velký počet vstupních dat	• faktorová analýza • analýza hlavních komponent	• zda za naměřenými daty není nějaká latentní struktura (POZOR na interpretaci) Př. lze 10 disciplín desetiboje popsat menším počtem faktorů?
Vysvětlit závislou proměnnou několika nezávislými, provést předpověď	• lineární regrese • regresní a klasifikační stromy (CART, CHAID) • časové řady • neuronové sítě	• příspěvek jednotlivých nezávislých proměnných k popisu proměnné závislé Př. Popsat trend výkonnosti v atletických disciplínách a provést předpověď výkonů na olympiádě v Riu 2016

Zdroje:

- Cyhelský, L., Kahounová, J., & Hindls, R. (2001). *Elementární statistická analýza*. (2. dopl. vyd., 318 s.) Praha: Management Press.
- Hendl, J. (2006). *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. (Vyd. 2., opr., 583 s.) Praha: Portál.
- Meloun, M., & Militký, J. (1998). *Statistické zpracování experimentálních dat*. (2. vyd., xxi, 839 s.) Praha: East Publishing.

- Sebera, M. *Vícerozměrné statistiky*, 2013
- Zvonař, M., Pavlík, J ., Sebera, M., Vespalec, T. & Štochl, J. *Vybrané kapitoly z antropomotoriky*. Brno: Masarykova univerzita, 2010.