

STATISTIKA

přednáška 1

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

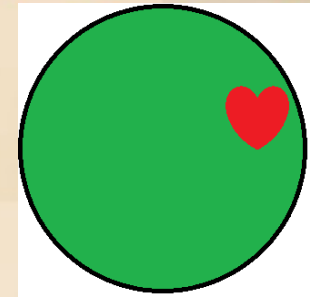
- 1. příprava výzkumného šetření je nejdůležitější část**
- 2. sběr a analýza dat slouží k zamítnutí/nezamítnutí předem stanovených úkolů práce a hypotéz (explorační vs. konfirmační přístup)**
- 3. vždy mít na paměti věcné hledisko výzkumu, zejména v souvislosti s interpretací statistických výsledků**

Role statistiky

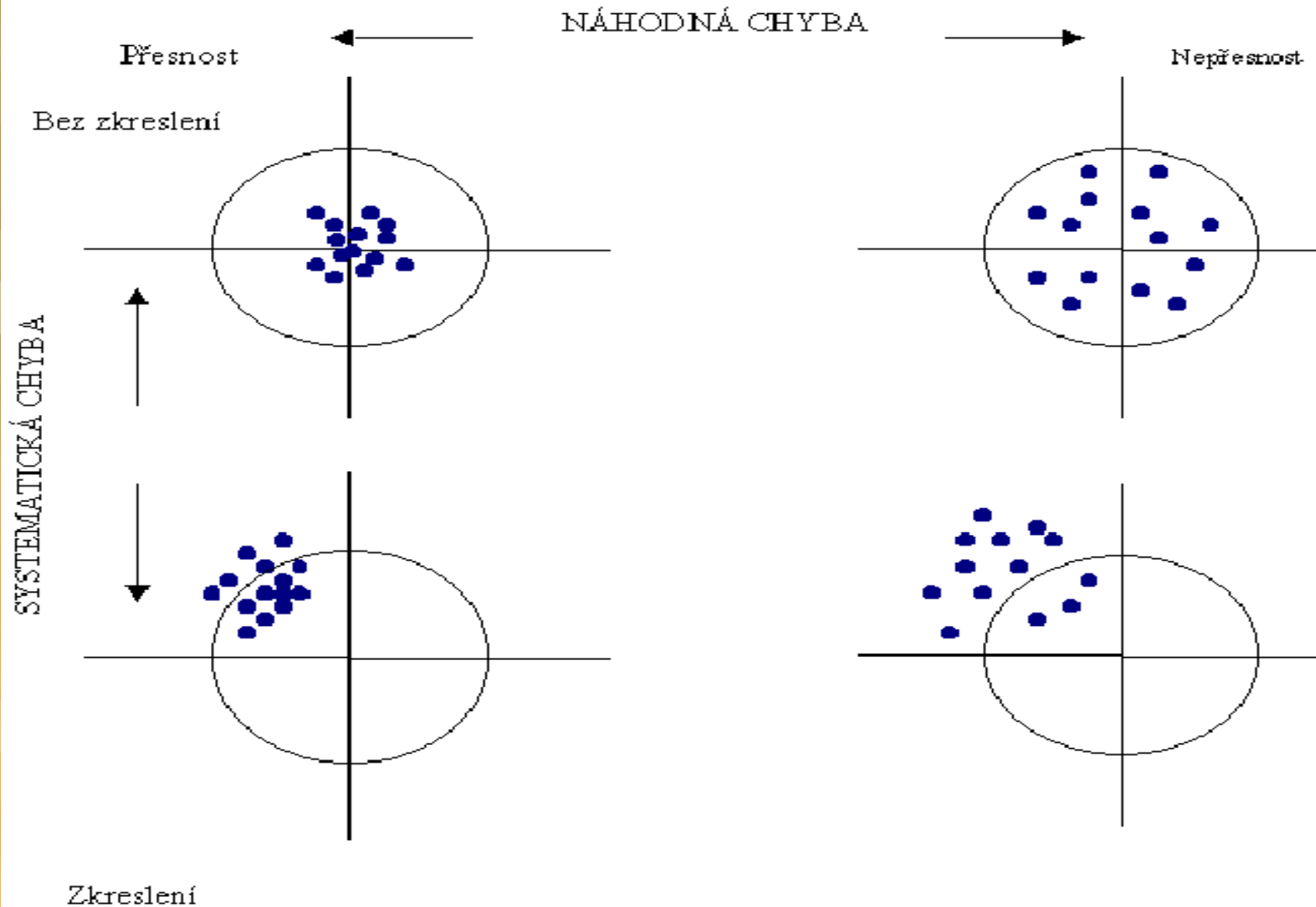
- Porozumění a zkoumání hromadných jevů
- Zjišťování zákonitostí
- V kvantitativním výzkumu (deduktivní princip) – pojitko mezi teorií a výzkumem
- Zpracování, popsání a analyzování dat

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**

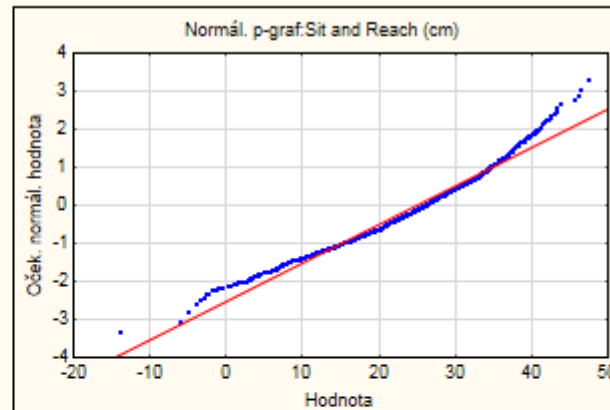
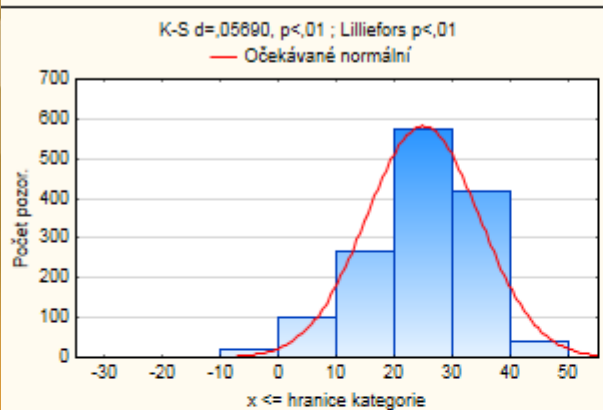
Škály (měřítka, stupnice)

- **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

Souhrn: Sit and Reach (cm)



Souhrnné statistiky: Sit and Reach (cm)

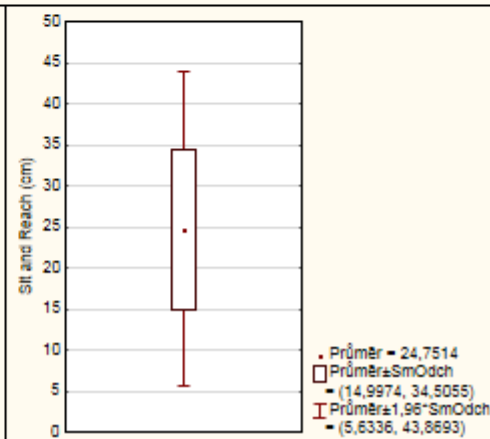
N platných=1423,000000

Průměr= 24,751441

Minimum=-14,000000

Maximum= 47,300000

Sm.odch.= 9,754011



Proč?

- chybná měření, extrémny
- 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

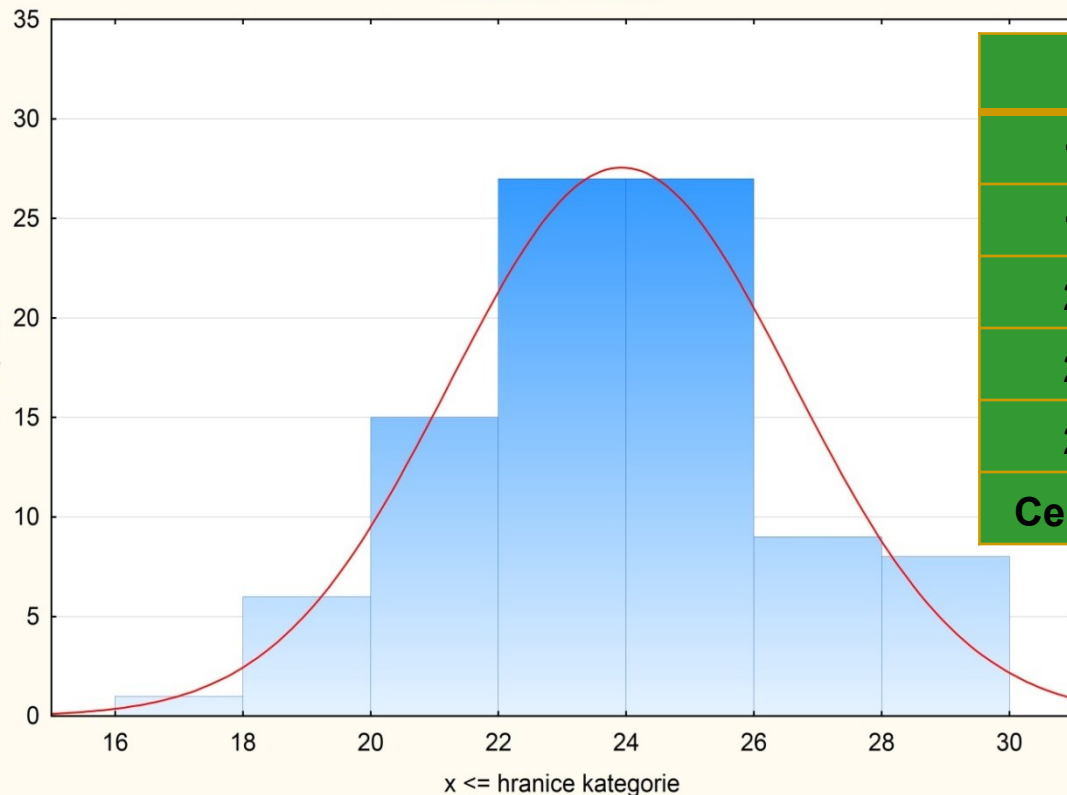
n_i – absolutní četnost

r_i – relativní četnost

N_i – kumulativní absolutní četnost

F_i – kumulativní relativní četnost

Histogram: BMI
— Očekávané normální



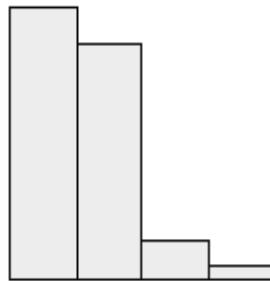
x	n_i	r_i	N_i	F_i
18	1	0,05 (= 1/20)	1	0,05
19	2	0,10 (= 2/20)	3	0,15
20	8	0,40 (= 8/20)	11	0,55
21	6	0,30 (= 6/20)	17	0,85
22	3	0,15 (= 3/20)	20	1,00
Celkem	20	1,00		

Ize usuzovat na některé vlastnosti, záleží na počtu intervalů

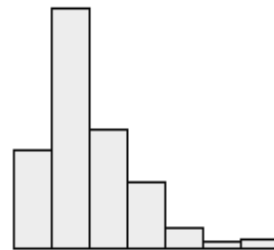
**Sturgersovo
pravidlo**



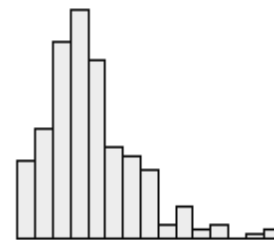
$$I = 1 + 3,3 \log n$$



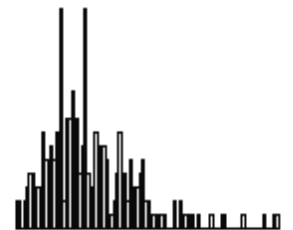
4 intervaly



8 intervalů



15 intervalů



100 intervalů

**Dixonovo
pravidlo**



$$I = 10 \log n$$

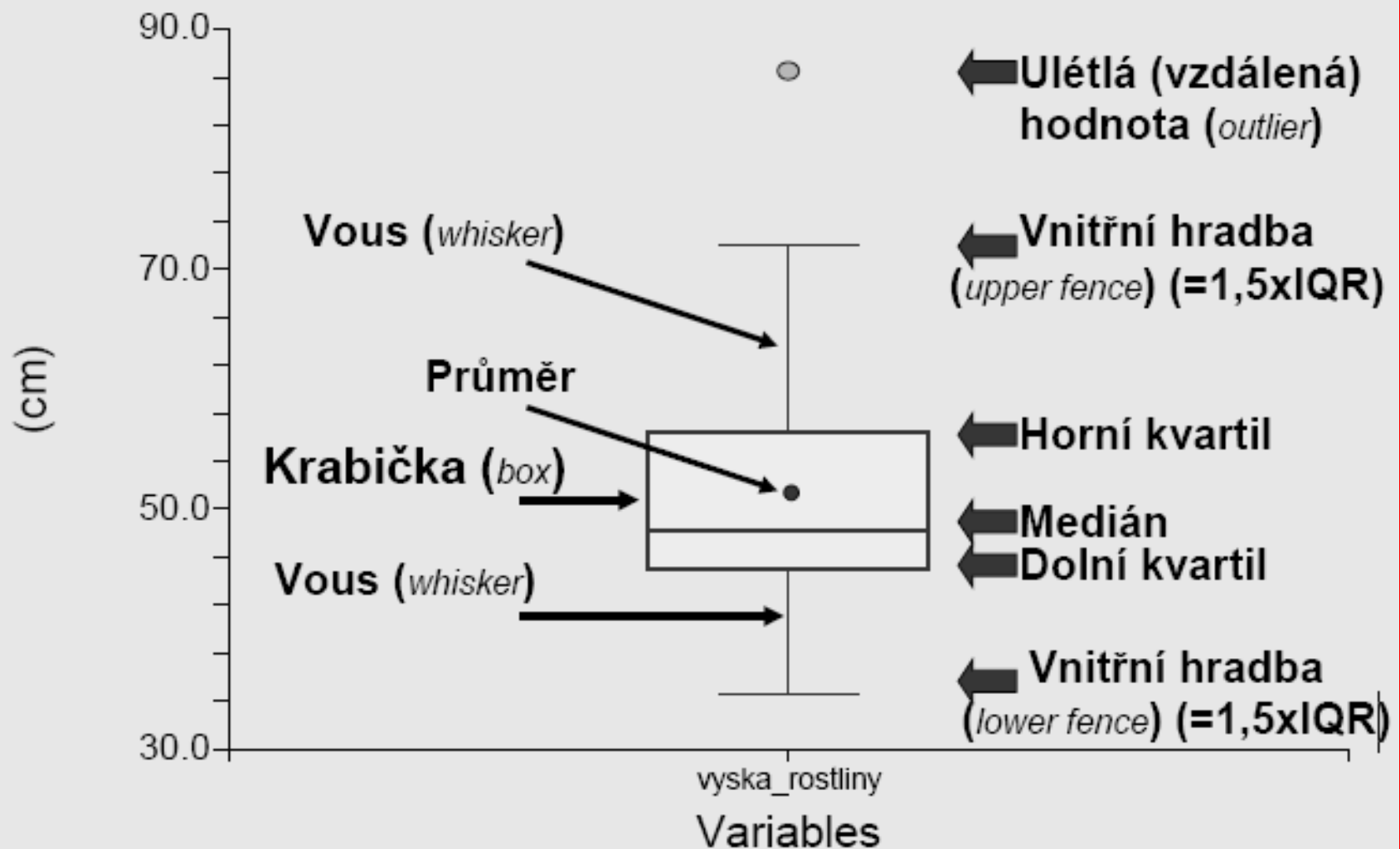


Základní statistické charakteristiky

- **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**

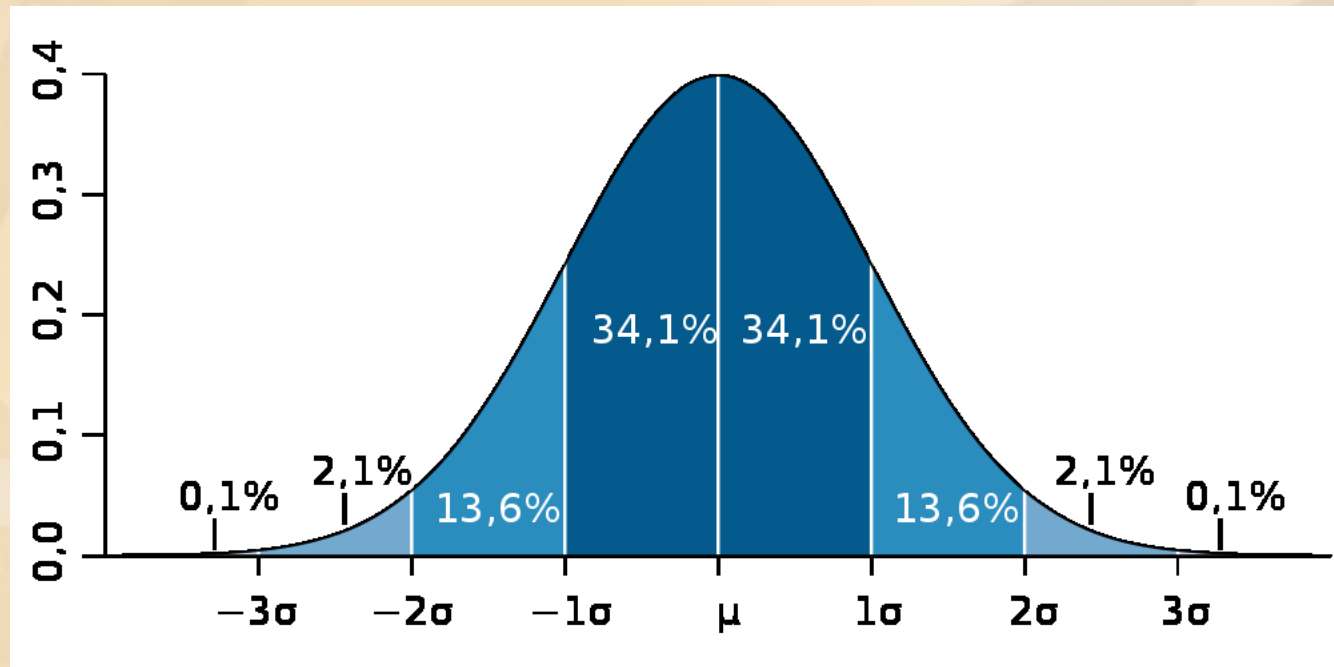
– 1; 10; 22	průměr 11	SD 10,53	n = 3
– 11; 11; 11	průměr 11	SD 0	n = 3

Box Plot



Normalita

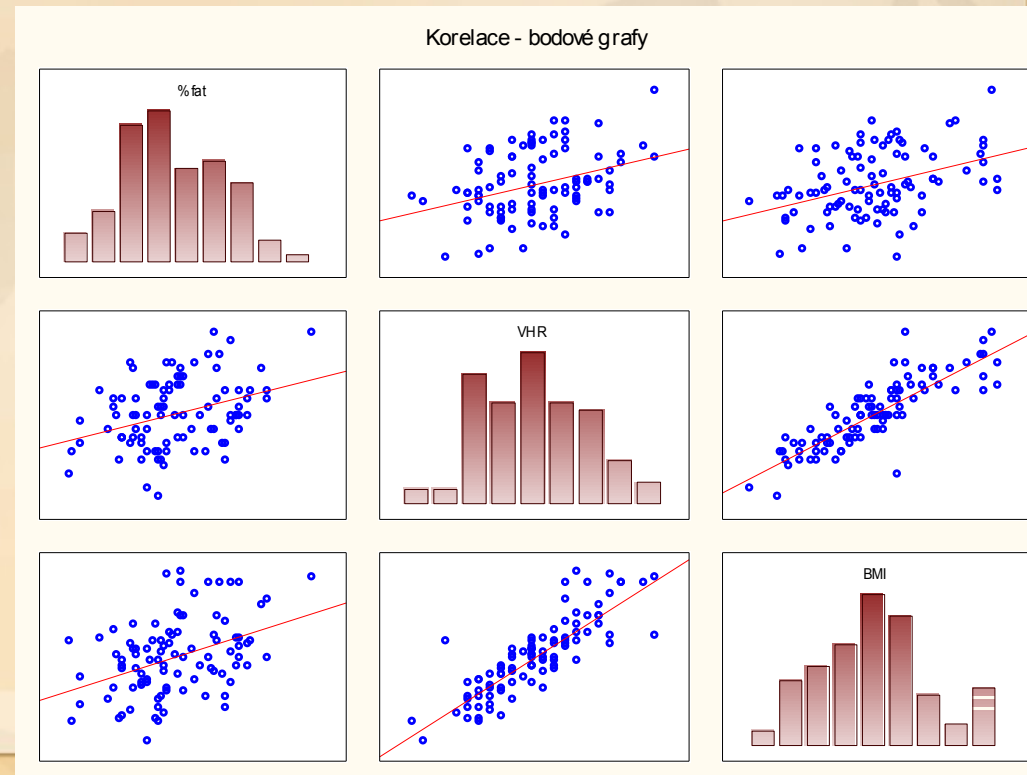
- Kolmogorov-Smirnov a Shapiro-Wilks test
- **Proč?**
rozhodnutí, zda použít parametrické nebo
neparametrické testy



Korelace ANEB korelace není kauzalita

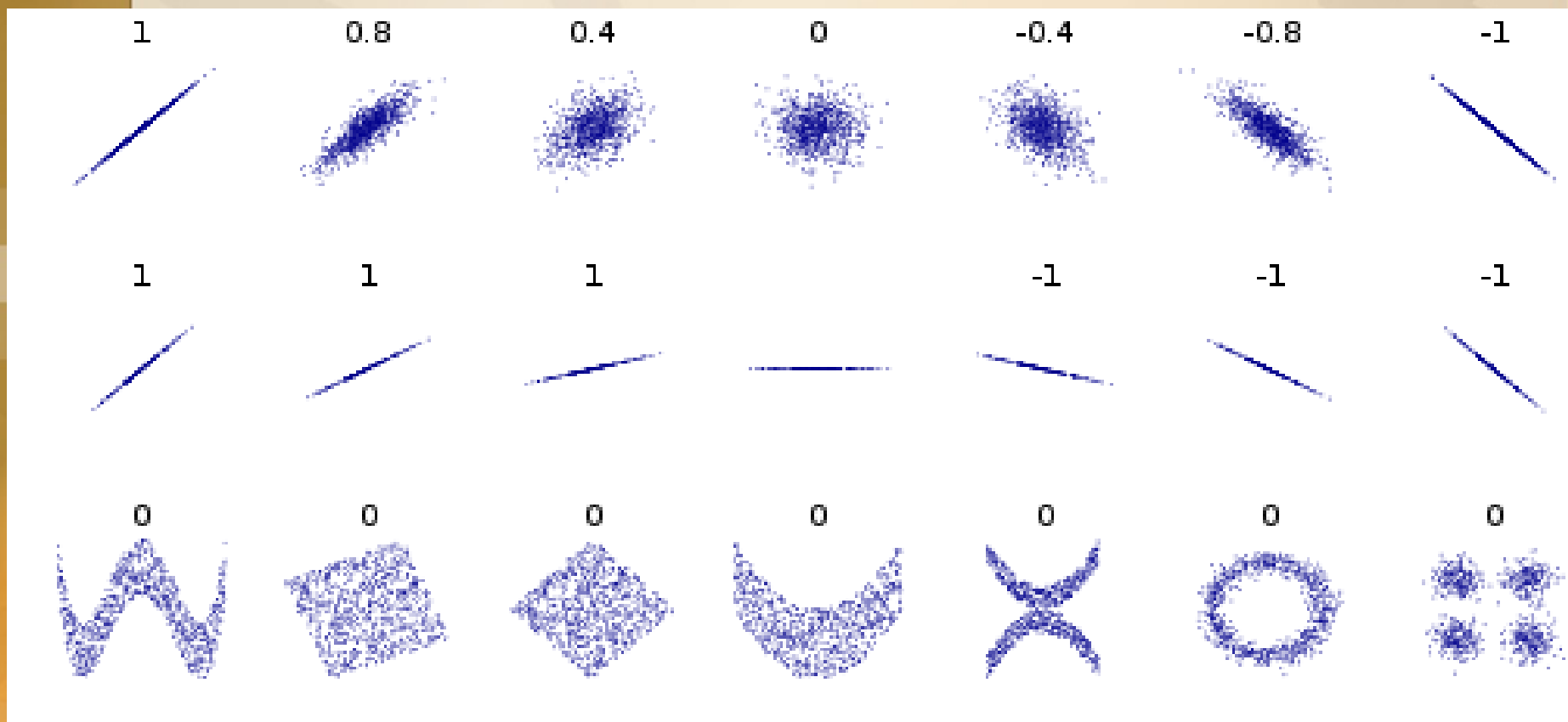
- = vzájemný vztah mezi veličinami proměnnými, jevy (dostatečně velký rozsah)
- Úkol: zjistit závislost a popsat ji
- Příklad 3 proměnné:

- BMI
- % fat
- WHR



Korelační koeficient

- **R : $\langle -1 \text{ do } 1 \rangle$**
- **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 Spearmanův***
- ***jednoduchý, parciální, mnohonásobný***



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 ($r^2 = 0,85^2 = 0,722$).

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 PROVĚST	ZPŮSOB	UMOŽNÍ MI ZJISTIT
První náhled na data	<p>Základní popisná statistika</p> <ul style="list-style-type: none"> • průměr, směrodatná odchylka, rozptyl, N, medián, <u>kvartily</u> a další míry polohy a variability • tabulky četností: absolutní, relativní, kumulativní • grafy: krabicový, histogram, bodový 	<ul style="list-style-type: none"> • chybná měření, extrémny • homogenitu souboru • chybějící data • trend v datech
Otestovat normalitu	<ul style="list-style-type: none"> • <u>Kolmogorov-Smirnov test</u>, <u>Shapiro-Wilks test</u> 	<ul style="list-style-type: none"> • rozhodnutí, zda použít parametrické nebo neparametrické testy
Zjistit, zda výběry/skupiny jsou shodné nebo ne	<ul style="list-style-type: none"> • 2 skupiny/proměnné: t-testy • 3 a více skupin/proměnných: Analýza rozptylu (ANOVA) <ul style="list-style-type: none"> ○ T-testy i ANOVA má svou parametrickou i <u>neparametrickou</u> variantu! 	<ul style="list-style-type: none"> • konstatovat statisticky nebo věcně (<u>size of effect</u>) významný rozdíl Př. <u>došlo ke zlepšení výbušné síly po intervenci?</u> (<u>pretest-posttest</u>) Př. <u>kteřá ze dvou tréninkových metod je úspěšnější?</u> Př. <u>mezi kterými skupinami je statisticky významný rozdíl</u> Př. <u>byl zkoumán výsledný čas v motorickém testu v závislosti na typu <u>suplementace sportovce (faktor A)</u> a na <u>způsobu tréninku (faktor B)</u></u>
Zjistit závislost více proměnných (spojité)	<ul style="list-style-type: none"> • korelace, index determinace • faktorová analýza 	<ul style="list-style-type: none"> • těsnost lineárního vztahu mezi proměnnými • může existovat jasný vztah ale nelineární, který nezachytíme pomocí korelace nebo faktorové analýzy • korelace neznamená kauzalitu!!! Př. <u>závisí výkon v běhu na 100 m s výkonem do skoku do dálky?</u> Př. <u>závisí ekonomika běhu na povrchu?</u>
Zjistit závislost více proměnných (kategoriální-např. dotazník)	<ul style="list-style-type: none"> • test nezávislosti chí-kvadrát v kontingenční tabulce • vícerozměrné kontingenční tabulky - asociační stromy • shluková analýza • regresní a klasifikační stromy (CART, CHAID) 	<ul style="list-style-type: none"> • sílu a směr vztahu Př. <u>závisí bolestivost zad na věku a způsobu zaměstnání?</u> Př. <u>mezi kterými proměnnými z dotazníku existuje nejsilnější vazba?</u>
Redukovat velký počet vstupních dat	<ul style="list-style-type: none"> • faktorová analýza • analýza hlavních komponent 	<ul style="list-style-type: none"> • zda za naměřenými daty není nějaká latentní struktura (POZOR na interpretaci) Př. <u>lze 10 disciplín desetiboje popsat menším počtem faktorů?</u>
Vysvětlit závislou proměnnou několika nezávislými, provést předpověď	<ul style="list-style-type: none"> • lineární regrese • regresní a klasifikační stromy (CART, CHAID) • časové řady • neuronové sítě 	<ul style="list-style-type: none"> • příspěvek jednotlivých nezávislých proměnných k popisu proměnné závislé Př. <u>Popsat trend výkonnosti v atletických disciplínách a provést předpověď výkonů na olympiádě v Riu 2016</u>

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

DVA NÁHODNÉ VÝBĚRY

NEZÁVISLÉ

ZÁVISLÉ

t-test pro
nezávislé výběry

t-test pro
závislé výběry

F-test

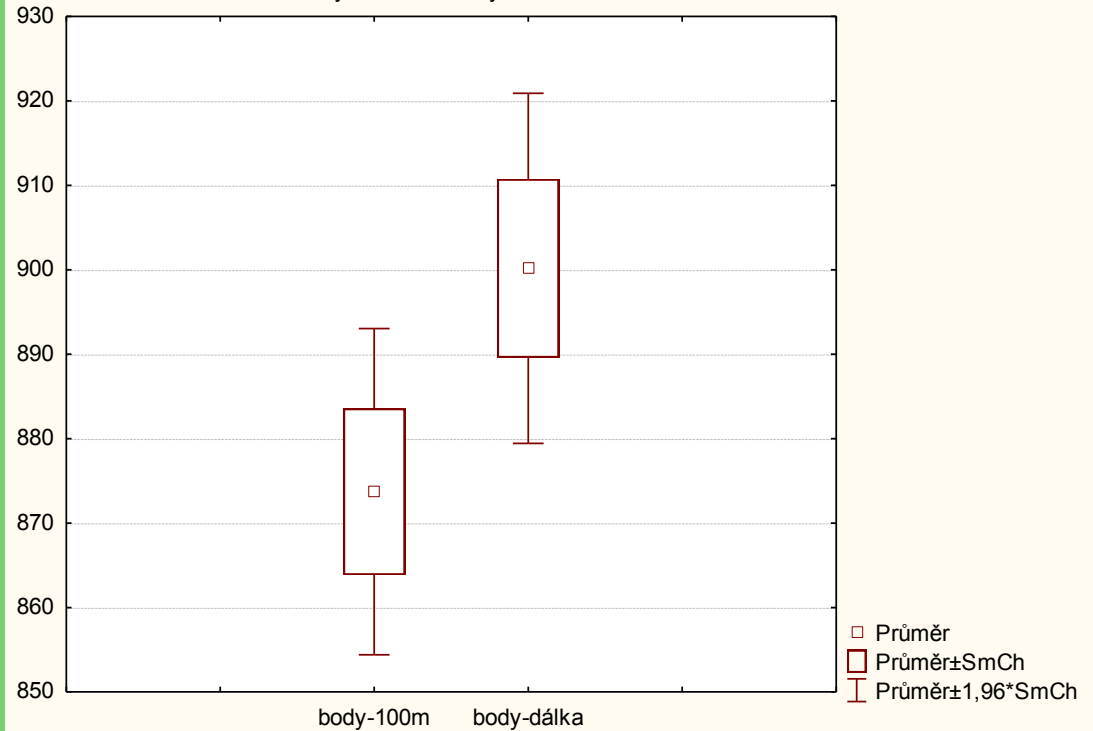
homogenní
rozptyl
 $s_1^2 = s_2^2$

heterogenní
rozptyl
 $s_1^2 \neq s_2^2$

t-test pro
homogenní
rozptyl

t-test pro
heterogenní
rozptyl

Krabicový graf
body-100m vs. body-dálka



T-test - příklad

Proměnná	t-test pro závislé vzorky (Tabulka1) Označ. rozdíly jsou významné na hlad. $p < ,05000$							
	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p
body-100m	873,71	58,32						
body-dálka	900,17	62,56	35	-26,46	66,16	-2,37	34	0,02

- **Cohenovo d**

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- $d > 0,8 \rightarrow$ velký efekt
- d z intervalu $0,5 - 0,8 \rightarrow$ střední efekt
- $d < 0,2 \rightarrow$ malý efekt

- **$d = 0,44$**

- **rozdíl mezi oběma disciplínami je i věcně i statisticky významný.**

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 (v tisku)
- Sebera, M. *Časové řady*, 2013 (v tisku)
- Zvonař, M., Pavlík, J., Sebera, M., Vespalec, T. & Štochl, J. *Vybrané kapitoly z antropomotoriky*. Brno: Masarykova univerzita, 2010.