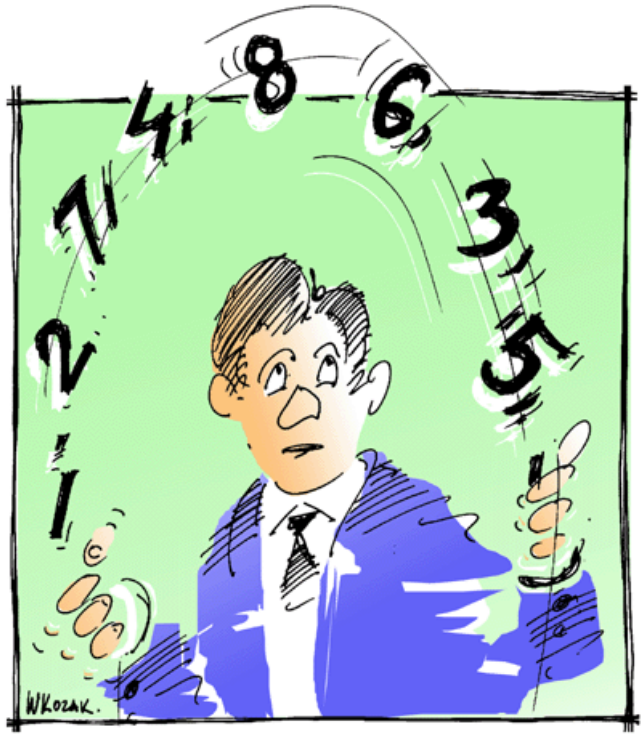
Základy statistiky

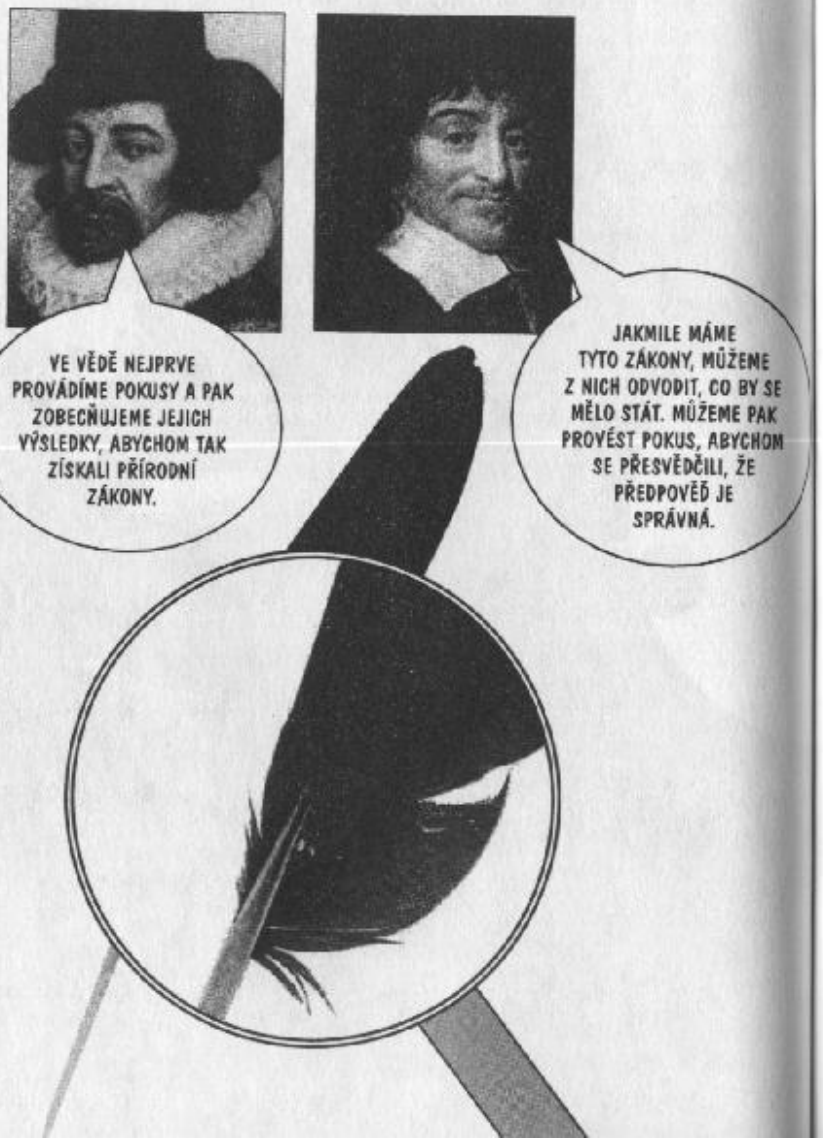


Definice

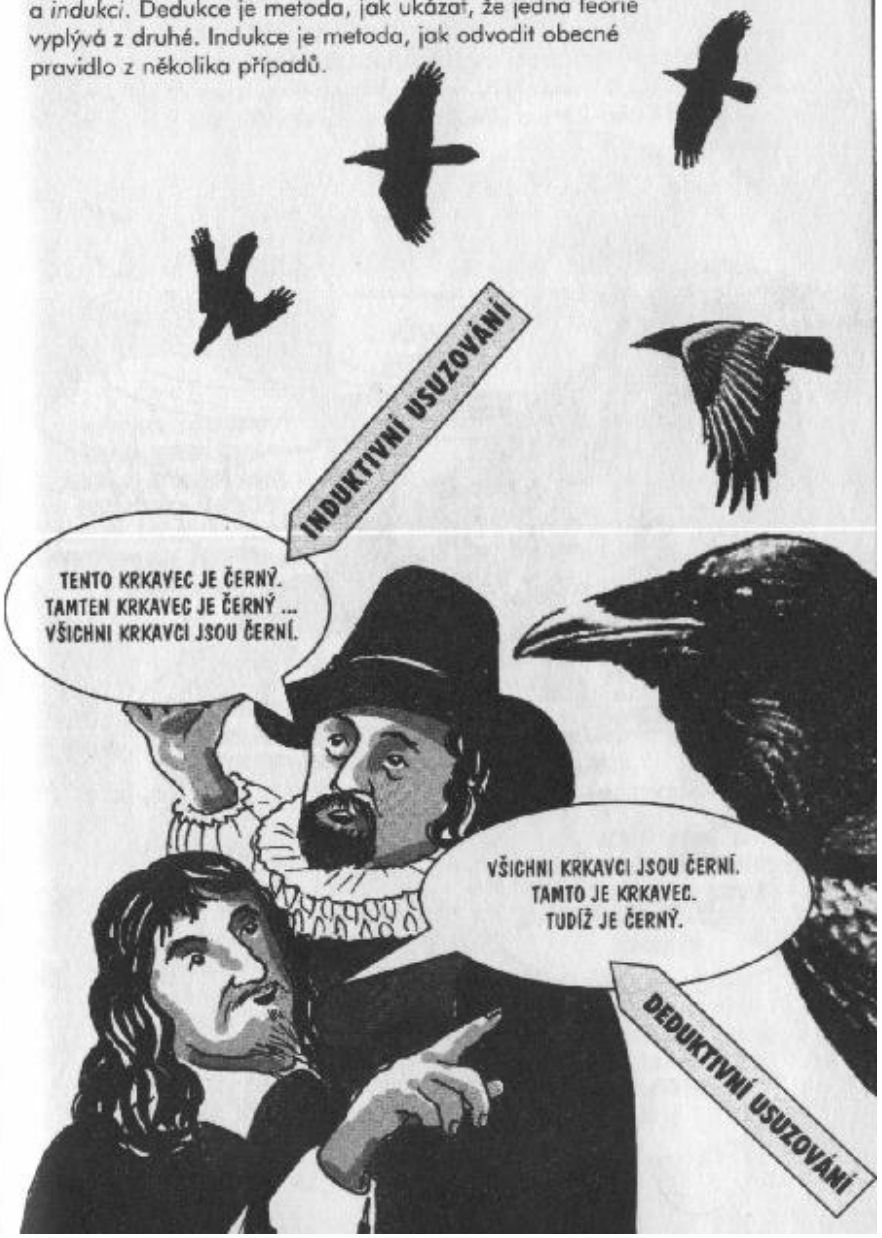
- Statistika - věda
- Statistika - statisticky vyjádřené šetření
- Statistika je věda, která nám dává návod, jak pracovat s daty obsahujícími náhodnou složku a jak odlišit zákonitosti od variability
- Deduktivní vs. Induktivní myšlení

Metoda dedukce a indukce

Galileovy metody se staly součástí metodologie vědy, kterou vypracovali filozofové **Francis Bacon** (1561–1626) a René Descartes.



Descartes a Bacon představují dvě formy usuzování – *dedukci* a *indukci*. Dedukce je metoda, jak ukázat, že jedna teorie vyplývá z druhé. Indukce je metoda, jak odvodit obecné pravidlo z několika případů.



Původ slova „statistika“

- Slovo statistika má stejný původ jako slovo stát
- Statistika vychází jako matematická věda především z počtu pravděpodobnosti a teorie her.
- Studuje převážně tak zvané hromadné jevy

Co je statistika ?

- V současné době se bez znalosti základů statistiky neobejdeme – variabilita v biol. oborech
- Správné plánování experimentů
- Správný design experimentů
- Snadná manipulace a demagogie se sebranými daty

Statistika jako věda

- Soubor postupů užívaných při sběru, zpracování a interpretaci dat směřujících ke zlepšení rozhodování
- Soubor metod, které nám umožňují činit rozumná rozhodnutí v případě nejistoty.

Obsah a význam statistiky

Lékaři i výzkumní pracovníci v biologii se často domnívají že hlubší znalosti statistické metodologie nejsou nezbytné.

Důvodů, proč si myslíme, že je statistika významná a důležitá, je hned několik

- Statistika je v určitém smyslu jazykem pro shromažďování dat, manipulaci s nimi a jejich kvanitativní manipulaci – lékař dělá v podstatě totéž.
- Otázky, které lékař klade jsou mnohdy statistického charakteru (jaké léky, kolik nemocných…).
- Exploze výpočetní techniky, která zasáhla do zdravotnictví už i u nás, umožňuje také laikům zpracování dat pomocí náročných a donedávna prakticky neproveditelných statistických postupů.
- V publikovaných článcích s biomedicínskou tématikou je statistika nezbytná.

Pokus vs. Šetření

Statistika

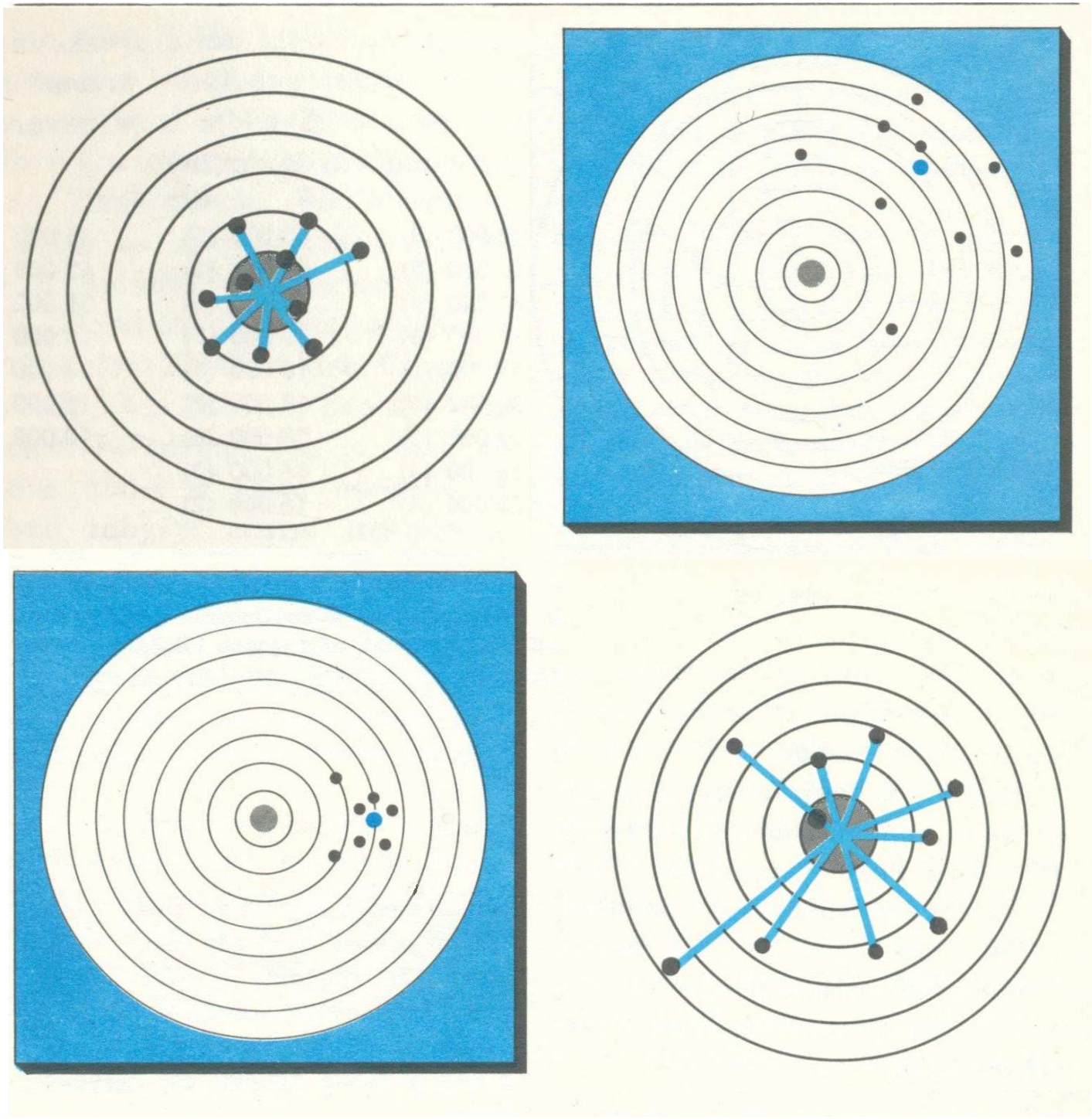
- popisná
  - základní charakteristika získaných dat (volební preference např.)
- vyčerpávající šetření
- analytická, induktivní
  - charakterizace určitého vzorku populace, ze které usuzujeme na vlastnosti celého základního souboru
- Výběr (výzkumy veř. mínění)
- Vztah mezi základním souborem a výběrem

Statistika se zabývá variabilitou měření

- Metodologická, přesnost měření
- Časová, v rámci individua = intraindividuální variabilita
- Interindividuální variabilita = populační

Statistika opakovaných **měř**ení

- Sledujeme správnost a přesnost měření
- Měření
  - Správné a přesné
  - Správné a nepřené
  - Nesprávné a přesné
  - Nesprávné a nepřesné



Zpracování naměřených dat

- Kontrola konzistence dat
- Zobrazení dat
- Testy normality
- (Vyřazení výsledků ovlivněných velkou chybou)
- Odhad středních hodnot a variability

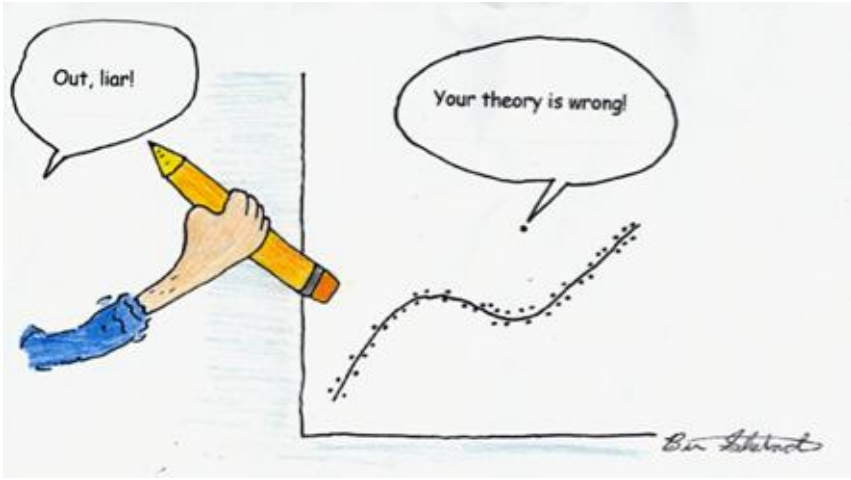
Typy biologických dat

- Data na poměrové stupnici (výška rostliny, váha potkana..)
- Data na intervalové stupnici (např. stupně teploty)
- Data na ordinální stupnici (školní klasifikace, klasifikace zdraví..)
- Data na nominální stupnici (barva, příslušnost ke druhu, umístění hnízda..)

Kvantitativní data

- Diskrétní data (např. počet pacientů)

- Spojitá data (výška, hmotnost apod.)



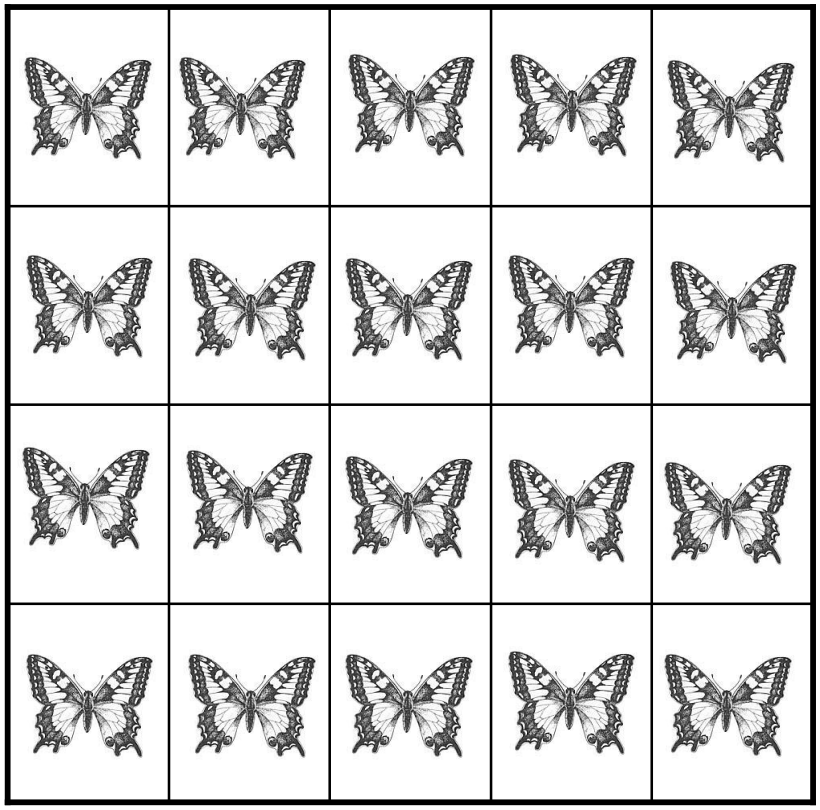
Sběr dat

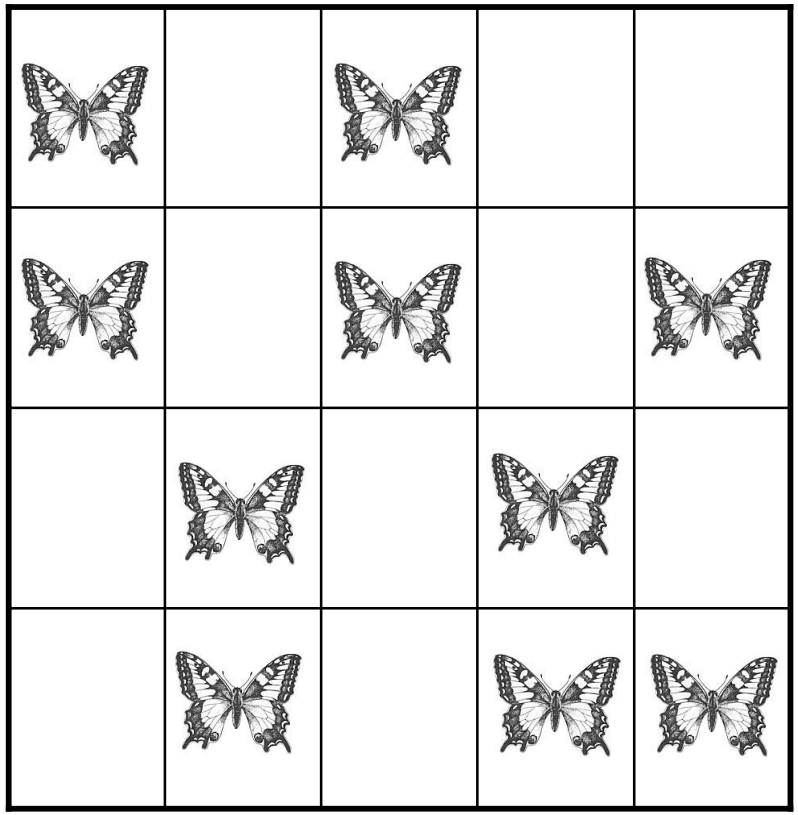
- data
  - kvalitativní
    - kategoriální, nominální (např. pohlaví) à potřeba kódování (např. muž 0; žena 1)
  - kvantitativní
    - diskrétní x kontinuální (spojitá)
    - ordinální (např. známky ve škole 1,2,3,4,5 – umožňuje seřadit podle velikosti)
    - intervalová
    - poměrová

Základní data a náhodný výběr

- Základní soubor (větší až potenciálně nekonečná skupina individuí)
- Náhodný výběr – každé individuum základního souboru má stejnou a nezávislou šanci, že bude vybráno
- Výběrové šetření (charakterizovat základní soubor na základě výběru)

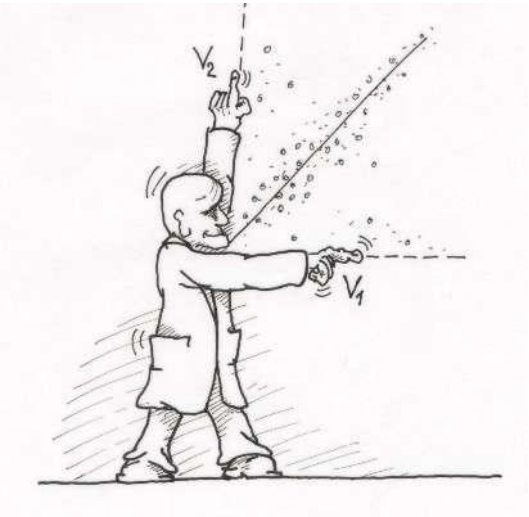
Populace a výběr



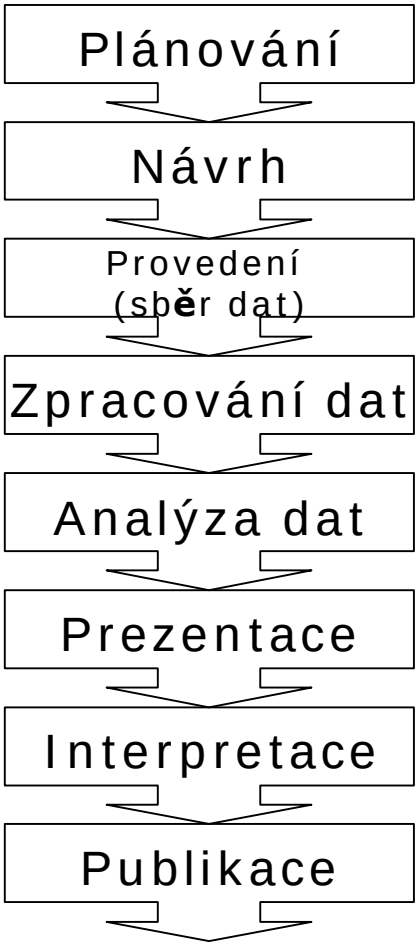


Pokud zkoumaný výběr dobře odráží strukturu celého zkoumaného souboru, nazýváme jej reprezentativním výběrem.

- Representativní výběr
- Za určitých předpokladů můžeme závěry z výběrů vztáhnout na celou populaci
- Kvantitativní znaky vs. Kvalitativní znaky



Obecné schéma dílčích stádií výzkumného projektu



Plánování a návrh výzkumného projektu – statistické hledisko

- Nemůžeme studovat celou populaci, která nás zajímá – vhodný výběr
- Musíme přesně formulovat cíle a účel výzkumu
- Musíme vymezit pojmy a metody pro: studovanou populaci, sledované znaky, sběr dat a statistickou analýzu

Sběr dat

- dostupnost dat
- úplnost dat
- spolehlivost dat
- cena dat

Úvahy zahrnuté v plánování experimentu!!!!

Sběr dat

- měřítka
  - přímo naměřená hodnota
  - intervalové (o kolik?)
  - poměrové (kolikrát?)

Sběr dat

- Databáze
  - záznam: nositel znaku
  - pole: znaky/proměnné

	Pole 1	Pole 2	Pole 3	Pole 4	Pole 5
Záznam 1					
Záznam 2					
Záznam 3			Data		
Záznam 4					

Sběr dat

- Vztah základní soubor x výběr
  - každý prvek základního souboru musí mít stejnou pravděpodobnost, že se stane prvkem výběru!!!!
- Definice výběrových kritérií / kritérií exkluze
- Opakovatelnost výběru

Zobrazení dat

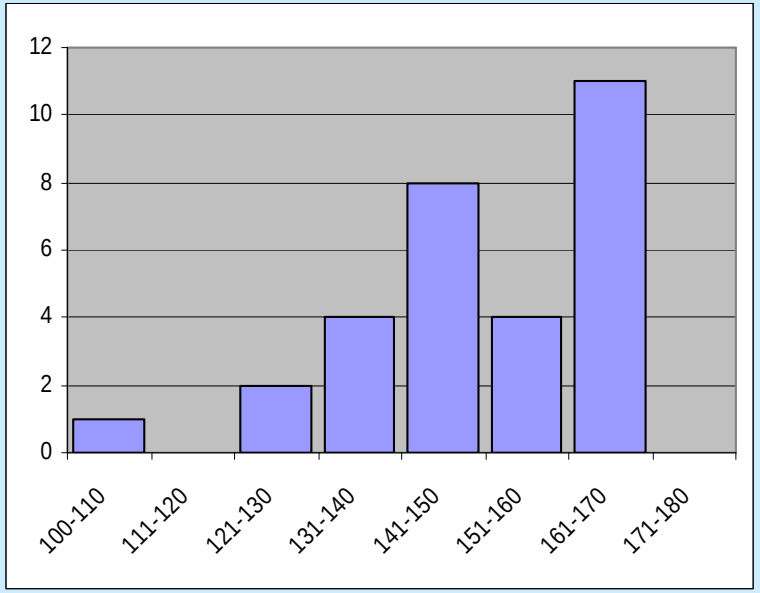
- Tabulky absolutních četností
- Relativní četnost
  - porovnání zastoupení jednotlivých kategorií mezi různě velikými skupinami
  - vyjádření struktury, vztahu části k celku
  - indexy pro porovnání vývoje v čase (pevný základ a zřetězený index)

Zobrazení dat

- tabulka, četnostní tabulka, histogram četností)

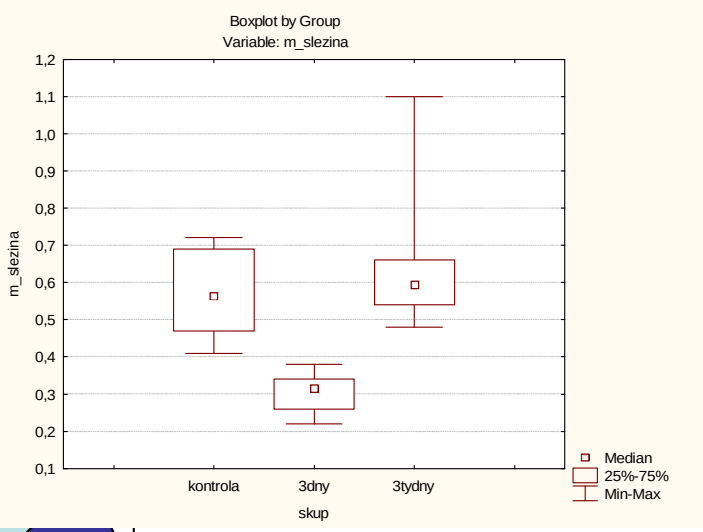
originální data	setříděná data
115	<100: 0
135	100-110: 1
120	111-120: 0
140	121-130: 2
125	131-140: 4
130	141-150: 8
150	151-160: 4
145	161-170: 11
.	>171: 0
.	
.	

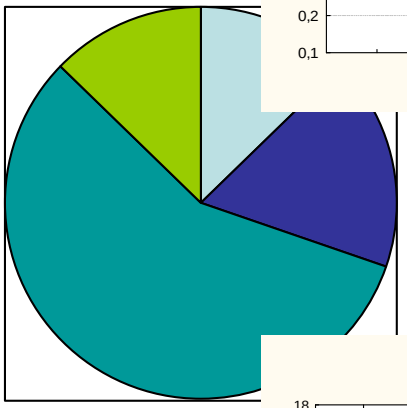
histogram

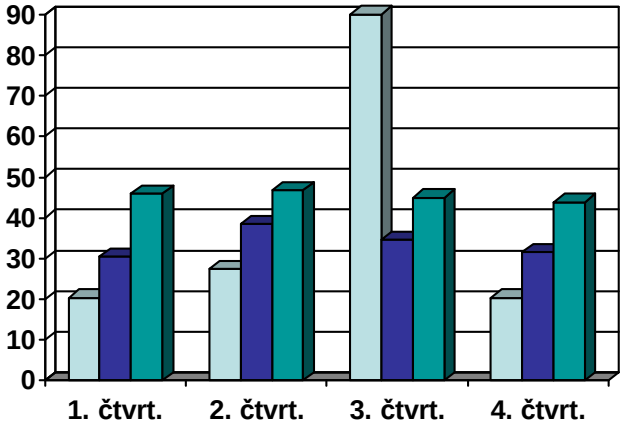


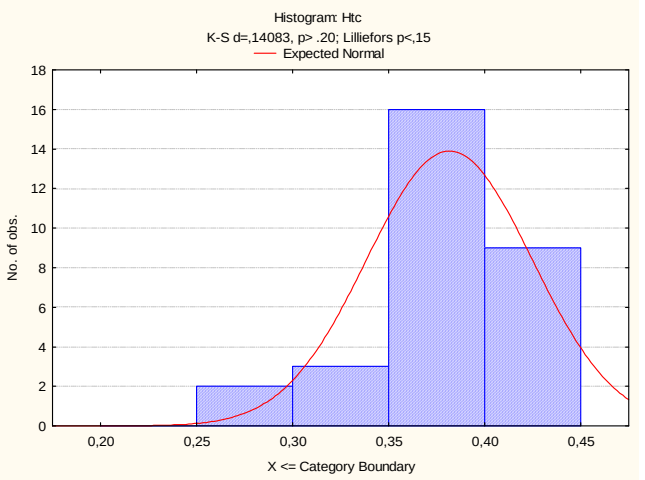
Zobrazení dat

- histogram
- box and whisker plot
- sloupcový graf
- koláčový graf



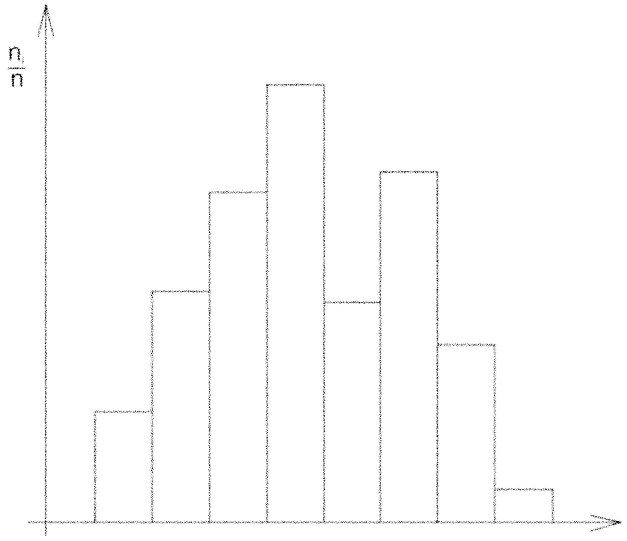






Histogram

- je graf kdy na vodorovnou osu znázorníme třídy a na svislou osu četnosti či relativní četnosti. Často se používá ve tvaru, kdy se hodnota odpovídající tříde znázorní jako sloupec s intervalem třídy jako základnou a výška je dána četností.

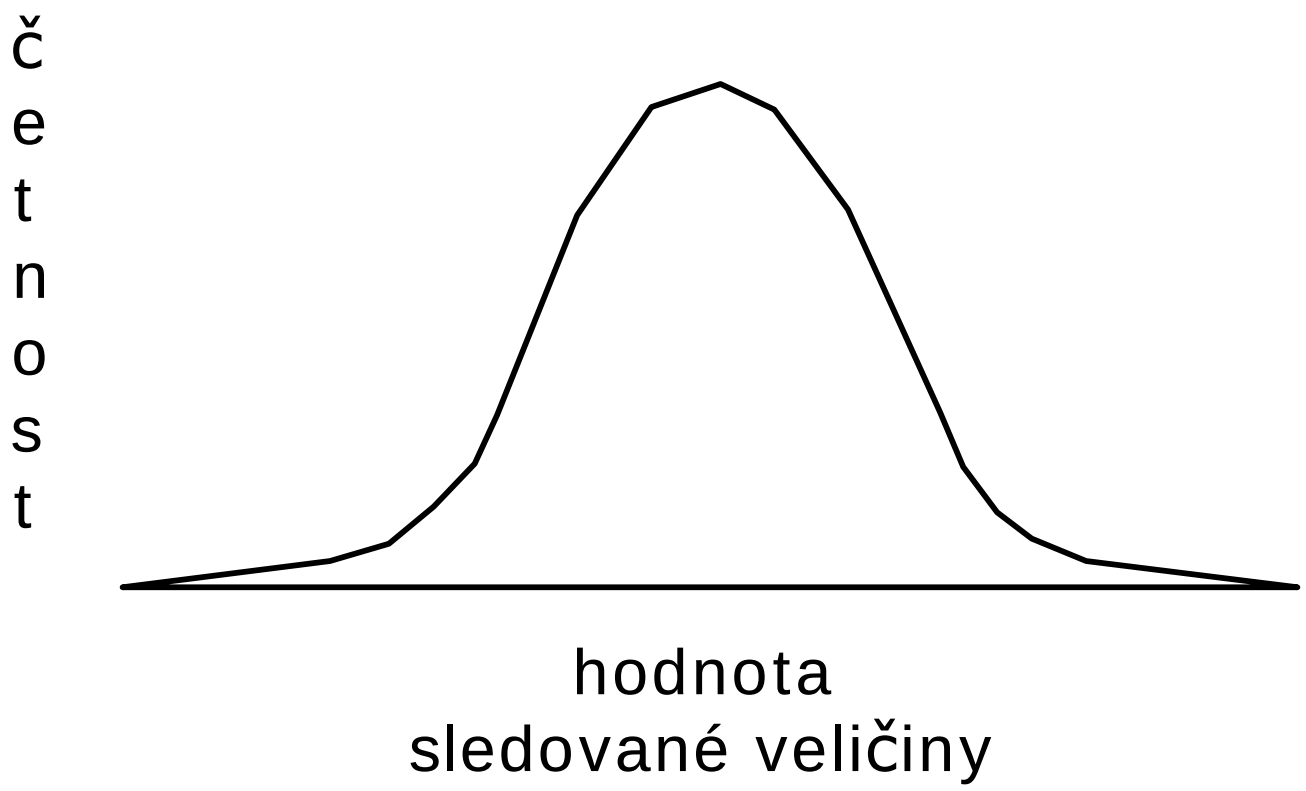


Analýza, interpretace a prezentace výsledků

- Využíváme metod popisné a induktivní statistiky
- Statistické třídění – jednostupňové, vícestupňové
- Absolutní četnost
- Konstrukce statistických tabulek
- Grafické znázornění – typy grafů

Publikace výsledků výzkumu

- Většinou recenzované časopisy
- Nekvalitní a špatně navržené výzkumy nalezneme téměř všude
- Jak vypadá struktura článku
- Důležité je zmínit, co daná studie přinesla nového


četnost
hodnota
sledované veličiny

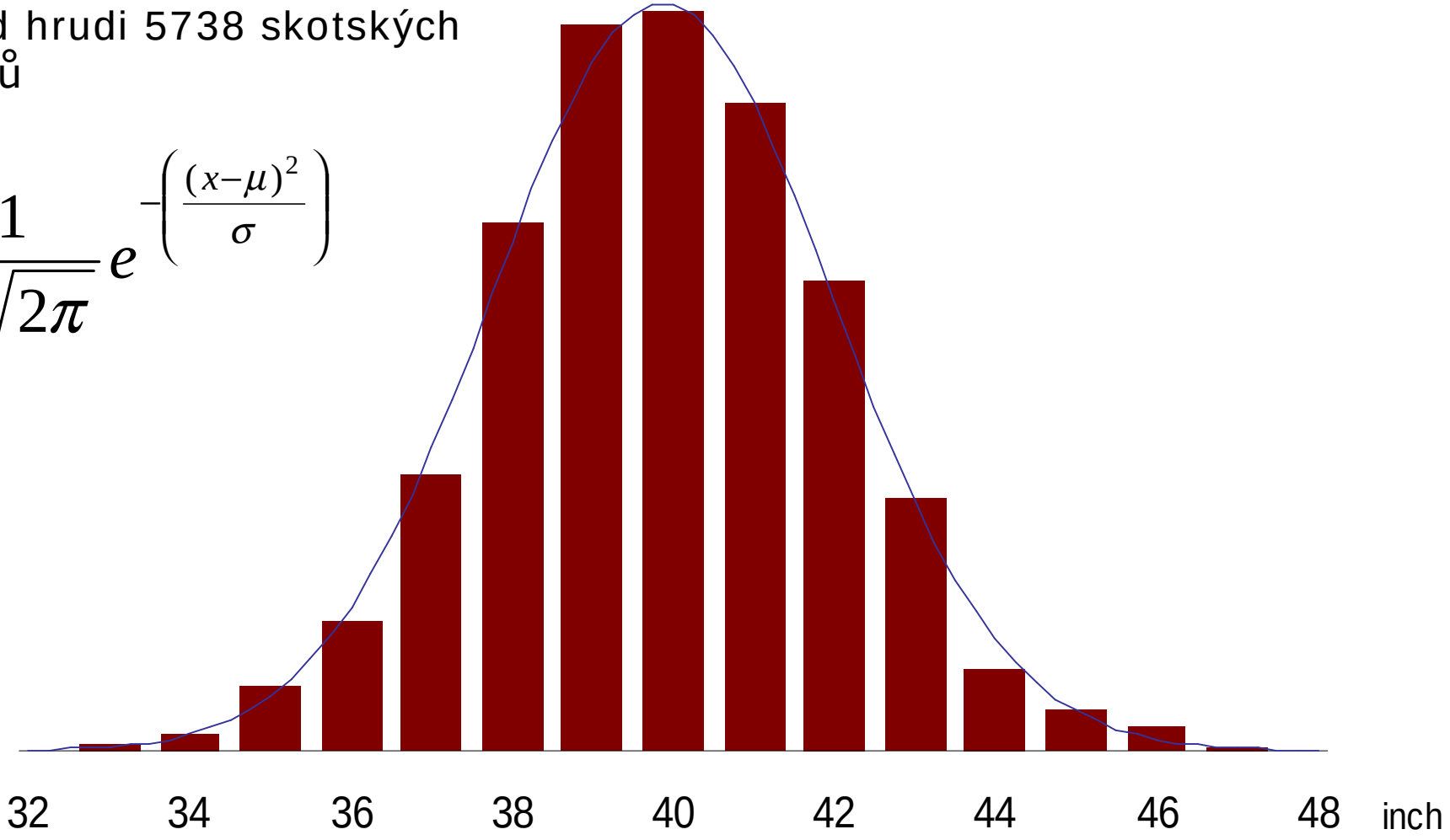
Normální rozložení (Gaussovo)

Abraham de Moivre 1733

Quételet

obvod hrudi 5738 skotských vojáků

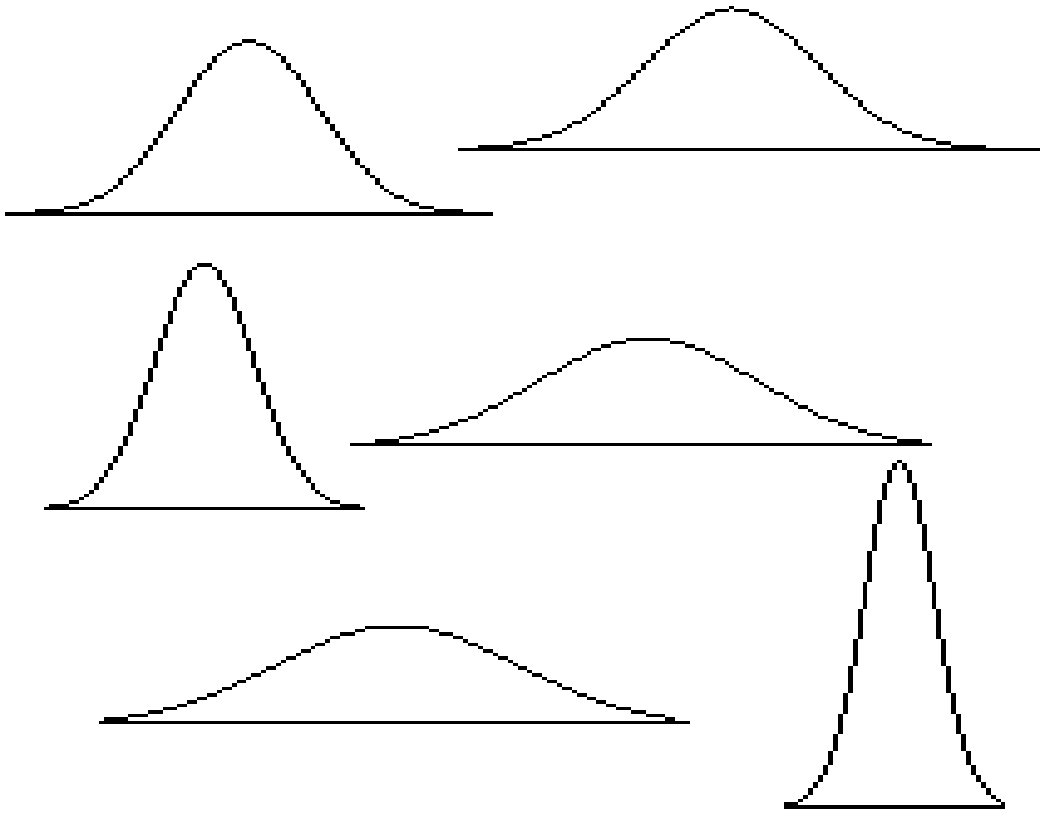
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)^2}{\sigma}\right)}$$



Popis dat

- Distribuce
  - normální
  - Poissonova
  - binomická
- Testy normality

Normální rozložení



Popis dat

- míry polohy
  - průměr (μ)
  - medián (= 50 percentil, frekvenční střed)
  - modus (= nejčastější hodnota)

Popis dat

- míry variability
  - min-max (=rozsah, range)
  - kvantily (horní 25%, dolní 75%)
  - směrodatná odchylka (SD, σ)
  - rozptyl (σ^2)

Statistika a lékař

- „sběratel“ dat
- „konzument“ výsledků

Základní veličiny

1. Rozsah souboru *(n)*: počet prvků v souboru
2. Aritmetický průměr ($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{neboli} \quad \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

3. Medián: prostřední člen v řadě naměřených hodnot uspořádaných podle velikosti
4. Modus: nejčastěji se vyskytující hodnota v daném souboru (výskyt dvou nebo více hodnot stejně často = bimodální, event. polymodální soubor)
5. Rozptyl *(s2, σ2)*: součet druhých mocnin odchylek od průměru dělený rozsahem souboru *(n)*, v případě výběrového rozptylu rozsahem souboru zmenšeným o 1 *(n-1)*.

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \qquad s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

6. Směrodatná odchylka *(s, σ)*: kladná odmocnina z rozptylu

$$s = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \qquad s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

7. Střední chyba průměru: směrodatná odchylka dělená odmocninou z *n*

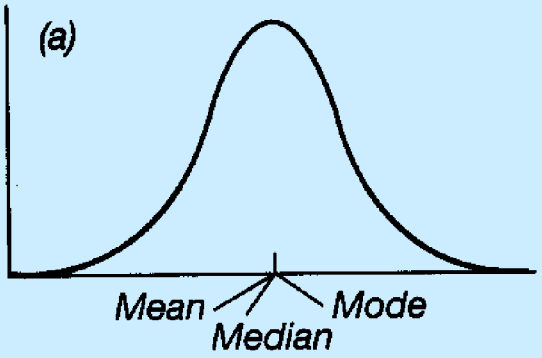
$$= \frac{s}{\sqrt{n}}$$

Příklady

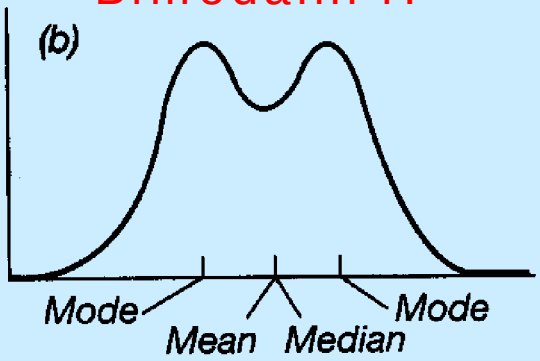
- Vypočtěte průměr následujích výsledků vyšetření: 39, 42, 73, 67, 24, 55.
- Co je modus v následujících výsledcích zjišťování krevních skupin: A, 0, 0, B, B, AB, A, A, 0, 0, 0, AB, B, 0, B, A, 0, AB, 0, 0, B, 0, A?
- Co je mediánem následujících výsledků hodnocení závažnosti průběhu onemocnění, přičemž A je nejlehčí a F je nejtěžší průběh: C, E, B, D, A, A, B, F, C, C, D?
- Co je mediánem následujících výsledků vyšetření: 61, 49, 35, 74, 53, 82?

Vztah mezi modusem, mediánem a průměrem v případě kvantitativních dat

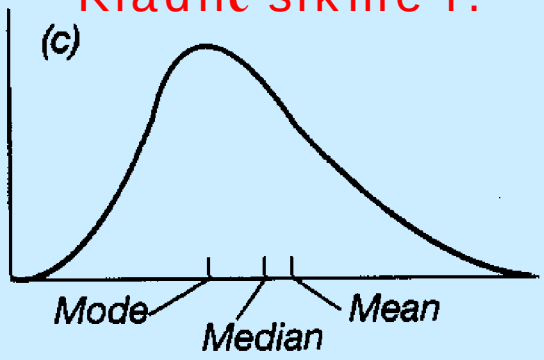
Unimodální rozdělení



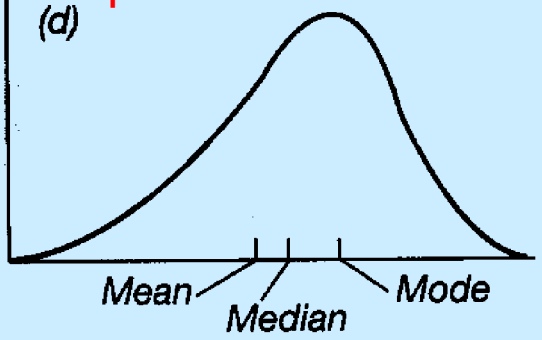
Bimodální r.



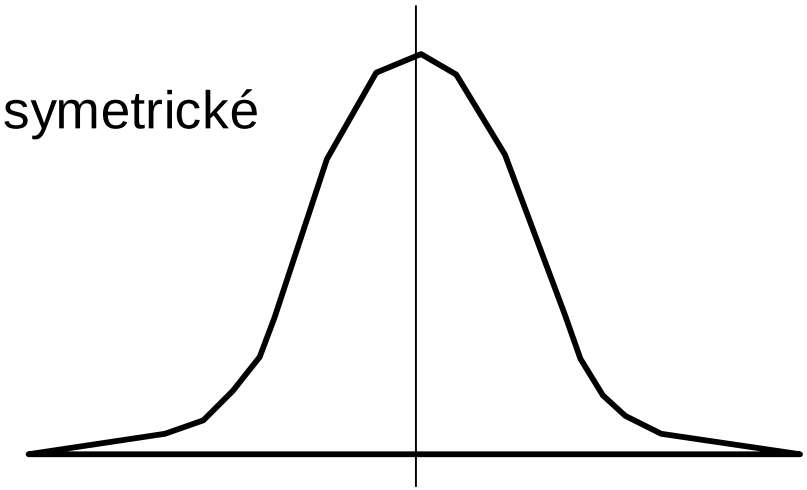
Kladně šikmé r.



Záporně šikmé r.

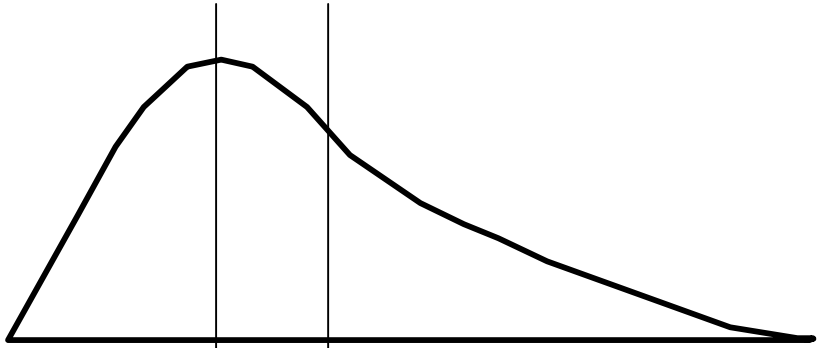


symetrické



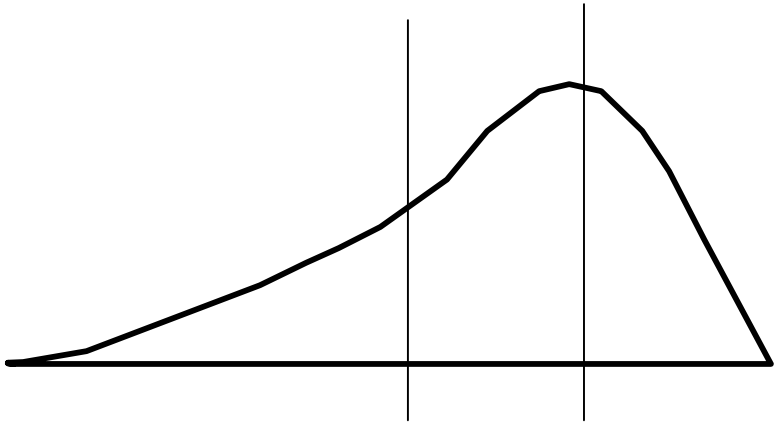
průměr
=medián
=modus

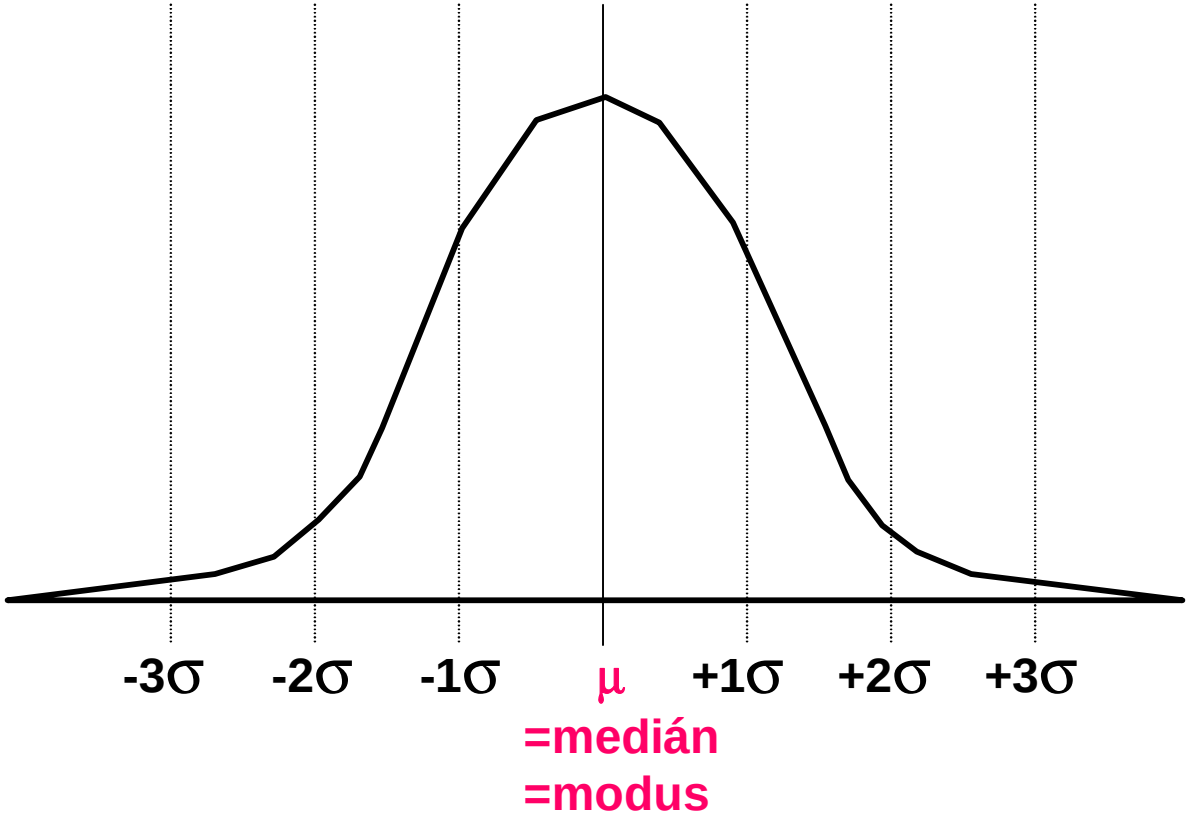
asymetrická

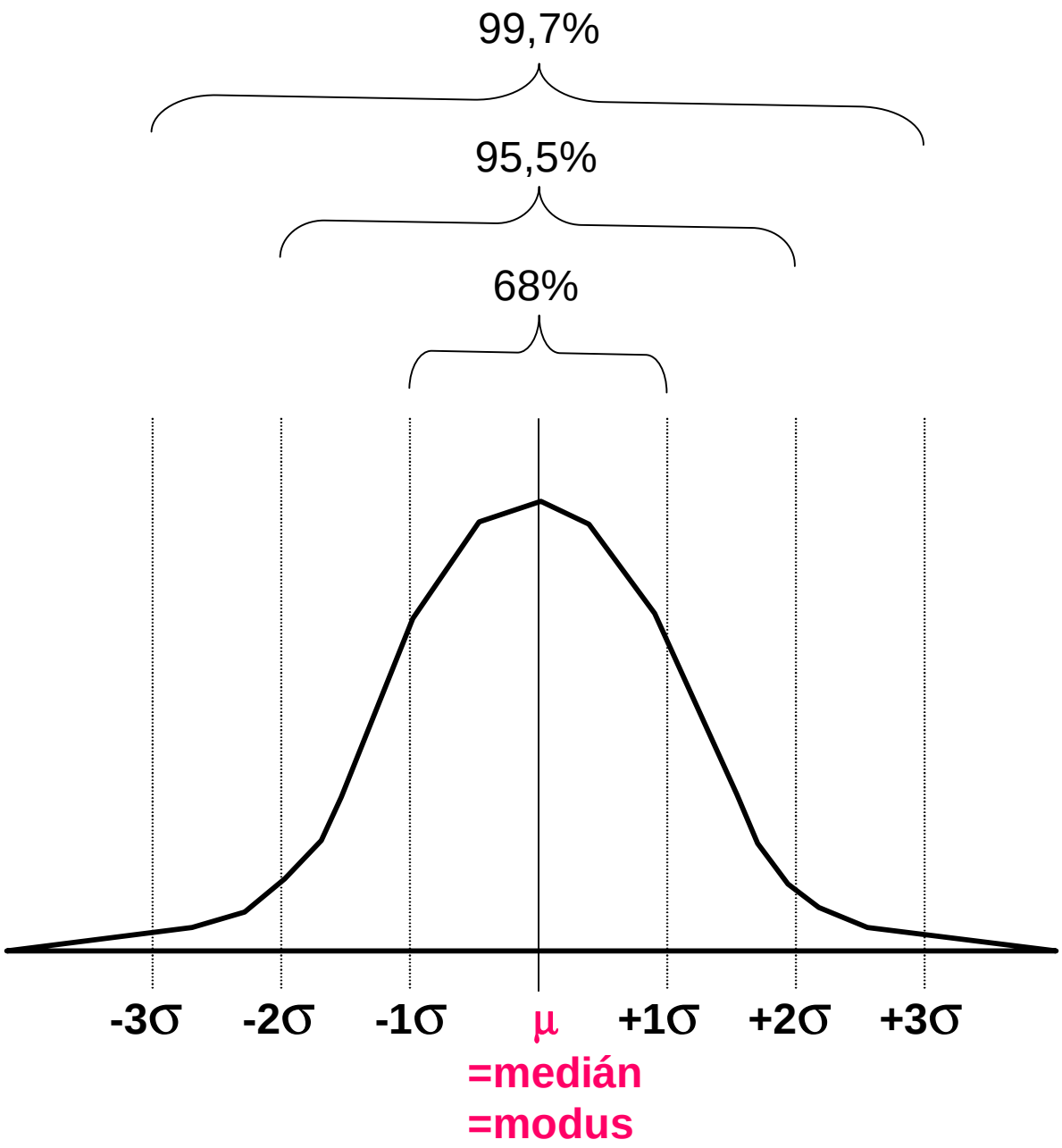


medián

průměr



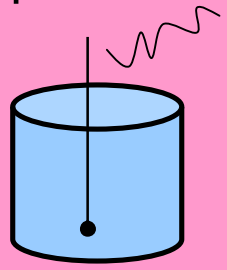

-3σ
-2σ
-1σ
μ
+1σ
+2σ
+3σ
=medián
=modus


99,7%
95,5%
68%
-3σ
-2σ
-1σ
μ
+1σ
+2σ
+3σ
=medián
=modus

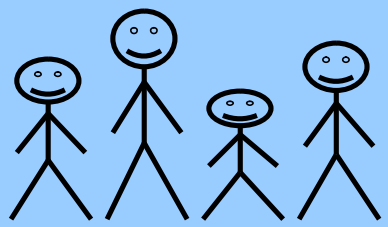
Variabilita - příčiny

proměnlivost biologických společenstev
mezipopulační rozdíly
rasové rozdíly
= BIODIVERZITA

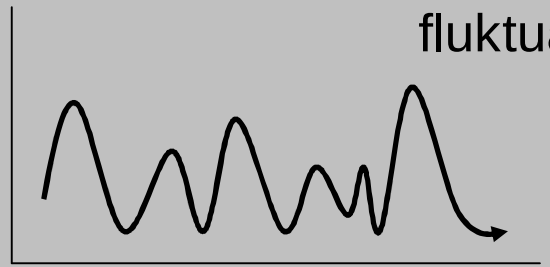
opakovaná měření, např. teploty
18,2°C
18,5°C
19,1°C
18,7°C



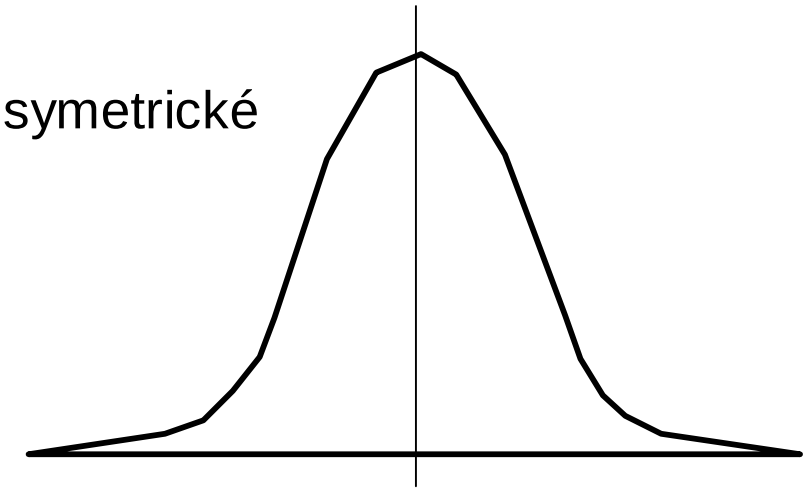
variabilita výšky v populaci
180cm
175cm
165cm
157cm



časová proměnlivost
fluktuace

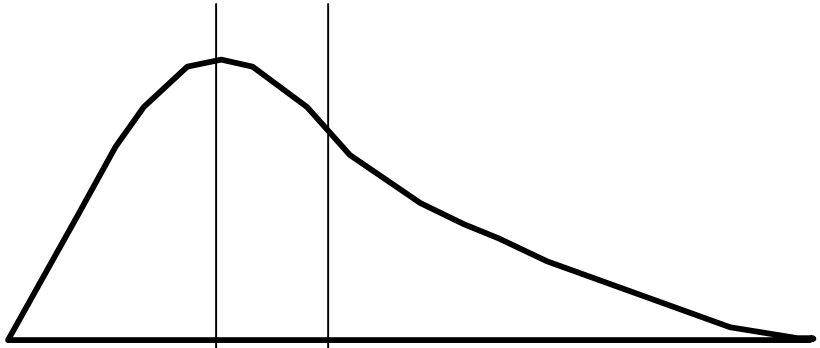


symetrické



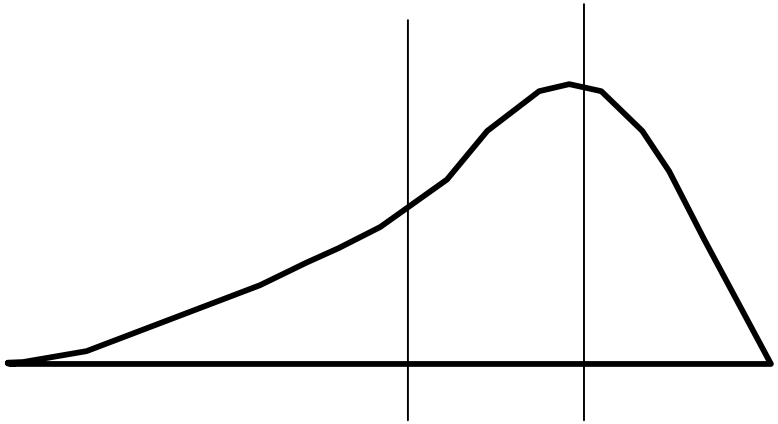
průměr
=medián
=modus

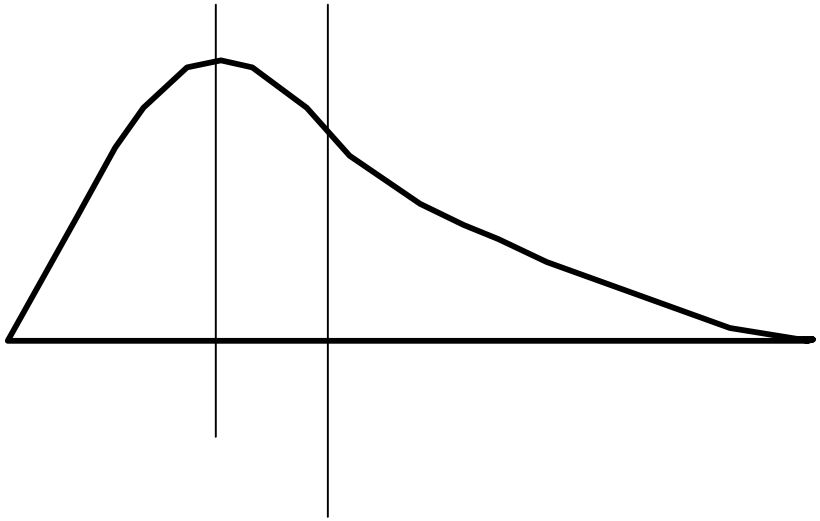
asymetrická



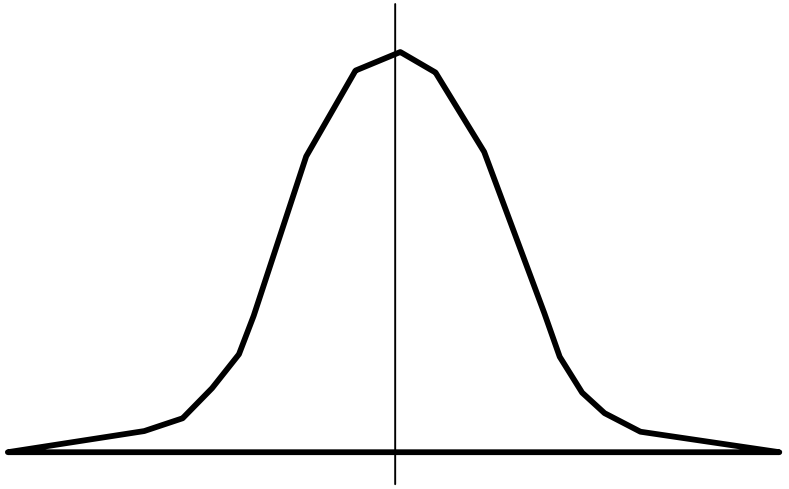
medián

průměr





Transformace dat



Statistická indukce

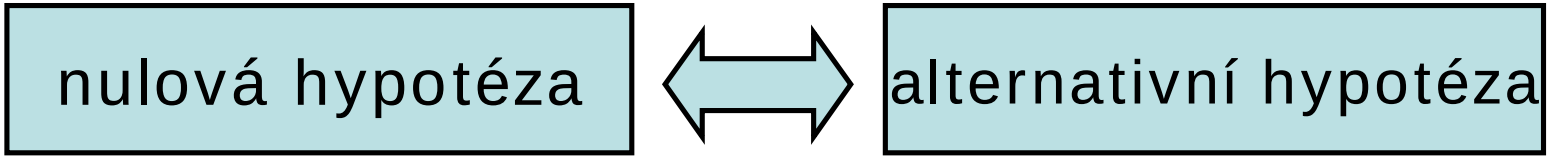
- základní soubor (populace)
  - soubor prvků, o kterém chceme statistickými metodami něco zjistit
- výběr
  - reprezentativní část dané populace (zákl. souboru), která má sloužit k odvození závěrů platných pro celou populaci

Odhady parametrů rozložení

- Výběrové charakteristiky
  - průměr $\bar{x}$, směrodatná odchylka s
- Vztahujeme na základní soubor
  - průměr μ, směrodatná odchylka σ

Testování hypotéz

- porovnání výběrového souboru a teorie o základním souboru
- porovnání dvou základních souborů na základě porovnání dvou výběrů



Dosažená hladina významnosti

- Poté co zformulujeme nulovou hypotézu a nasbíráme data, spočteme pravděpodobnost, s jakou bychom mohli obdržet pozorovaná data nebo data stejně, či ještě více odporující nulové hypotéze, *za předpokladu, že je nulová hypotéza pravdivá*.
- Tato pravděpodobnost se nazývá dosažená hladina významnosti a značí se *p*.

Dosažená hladina významnosti

!!!**Č**ím menší je p, tím neudržitelně**j**ší **č**ili méně **důvě**ryhodná je nulová hypotéza!!!

Vysoká hladina významnosti

- Jestliže porovnáme např. dvě léčby a dostaneme vysoké *p*, pak můžeme tvrdit, že taková data, jako jsou naše bychom mohli dostat celkem často i v případě, že platí nulová hypotéza.
- Nelze proto vyloučit, že nulová hypotéza je pravdivá – tj. že obě léčby jsou stejně efektivní.

Nízká hladina významnosti

- Je-li *p* velmi malé, pak se nulová hypotéza zdá být téměř nemožnou, protože naše data by mohla sotva kdy vzniknout pouze náhodou kdyby platila nulová hypotéza.
- Můžeme tedy tvrdit se značnou spolehlivostí, že nulová hypotéze není pravdivá a jedna léčba je prokazatelně lepší než druhá.
- Hladina významnosti – 5% (p=0.05)

Významnost statistického testu

Test není statisticky významný – hypotézu nezamítáme

– pozorované odchylky od hypotézy je možno vysvětlit pouhou náhodou důvodem může být i to, že rozdíl je tak malý, že na jeho prokázání nestačí použitý rozsah souboru.

Test je statisticky významný – hypotézu zamítáme

– pozorované odchylky od hypotézy není možno vysvětlit pouhou náhodou odchylka od hypotézy je tak velká, že při opakování šetření bychom s velkou pravděpodobností hypotézu opět zamítli

P-hodnota – vypočtená pravděpodobnost chyby α, kdybychom na základě našich dat hypotézu zamítli. Slouží k provedení testu porovnáním se zvoleným α.

Jaký je vlastně princip konstrukce testu?

1. Vytvoříme testovanou hypotézu kterou chceme ověřit a altrernativní („širokou“) hypotézu, o jejíž platnosti nepochybujeme.
2. Porovnáme zda je rozdíl mezi skutečností a hypotézou vysvětlitelný pouhou náhodou.

Jak?

3. Porovnáme model alternativní hypotézy s testovaným modelem.
4. Převedeme data do tvaru nějaké statistické „normy“ (t-, F-, χ^2-, ... rozložení), která nám umožní test dokončit

Chyba 1. a 2. typu

decision	in reality H_0 valid	in reality H_0 false
H_0 rejected	**type I error (α)**	correct
H_0 accepted	correct	**type I error (β)**

Postup při testování hypotéz

- vyslovení hypotéz
- volba testu
- volba pravděpodobnosti chyby zamítnutí, hladiny významnosti α
- výpočet
- zamítnutí/nezamítnutí nulové hypotézy

Statistické testy

testy	nepárové	párové
parametrické (pro normální nebo téměř normální rozložení)	• t-test nezávislý (klasický t-test, two-sample)	• t-test závislý (one-sample)
neparametrické (pro jiné než normální rozložení)	• Mann-Whitney (=Wilcoxon nezávislý) • mediánový test	• Wilcoxon závislý • znaménkový test
	srovnání parametru mezi 2 skupinami objektů	srovnání parametru u stejných objektů v časové sousloednosti

Regresní a korelační analýza

- Sleduje závislost dvou proměnných
- Zprostředkovaná korelace

Kontingenční tabulky

- Chi-square
- Fischer exact test

Mnohorozměrná analýza dat

- Shluková analýza