

# Biostatistika



**Opakování**  
**Modelová rozložení náhodné veličiny**  
**Normální rozložení dat**  
**Základy testování hypotéz**

# Opakování



## Základy popisné statistiky Vizualizace dat

# Opakování



1. Co jsou **kvalitativní** a **kvantitativní data**?
2. Jaký je rozdíl mezi **spojitými** a **diskrétními daty**?
3. Uveďte **příklady binárních / nominálních / ordinálních dat**.
4. Uveďte **příklady spojitých a diskrétních dat**.
5. Jakými charakteristikami **popisujeme kvalitativní data**?
6. Jakými charakteristikami **popisujeme kvantitativní data**?
7. Jak správně **vizualizujeme kvalitativní data**?
8. Jak správně **vizualizujeme kvantitativní data**?

# Modelová rozložení

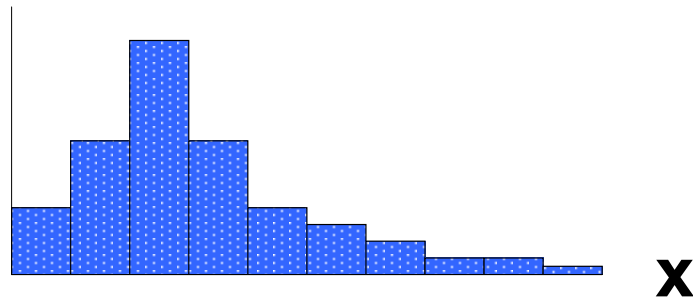


**Parametry rozložení**  
**Přehled modelových rozložení**  
**Logaritmicko-normální rozložení**

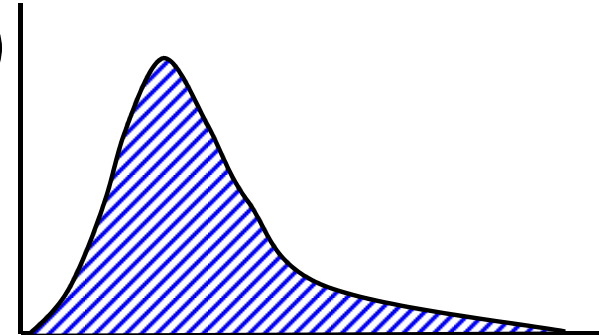
# Výběrové rozložení hodnot lze modelově popsat, a definovat tak pravděpodobnost výskytu $X$



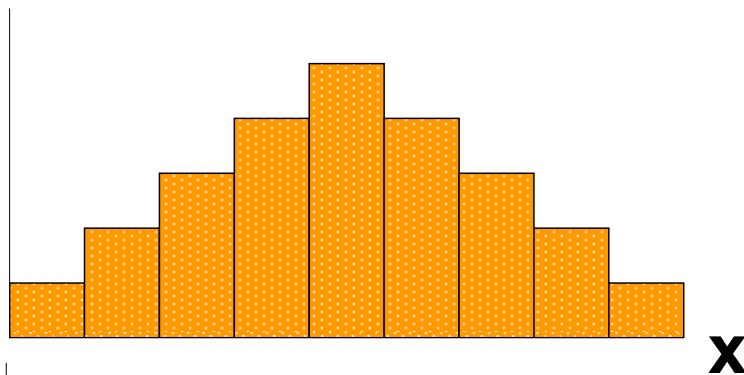
$f(x)$



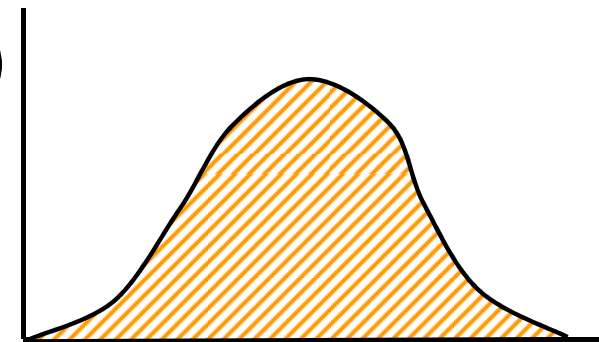
$\varphi(x)$



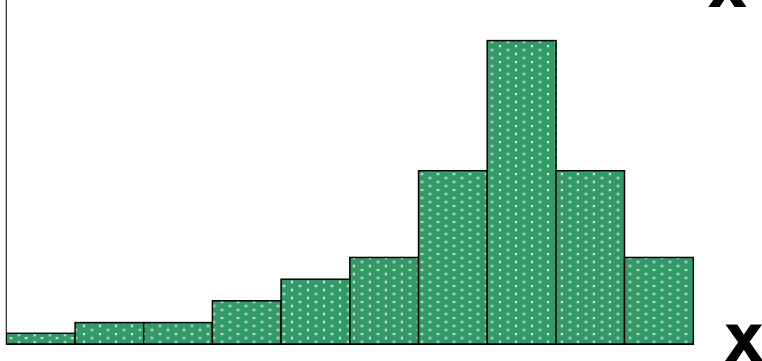
$f(x)$



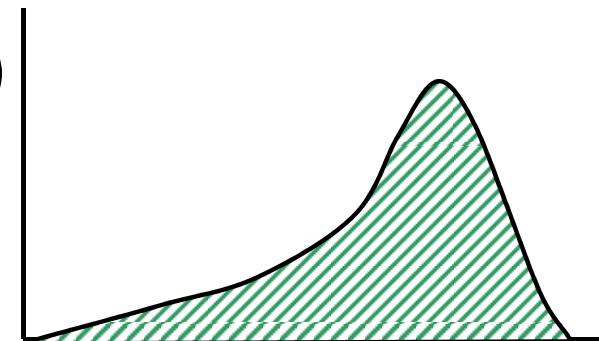
$\varphi(x)$



$f(x)$



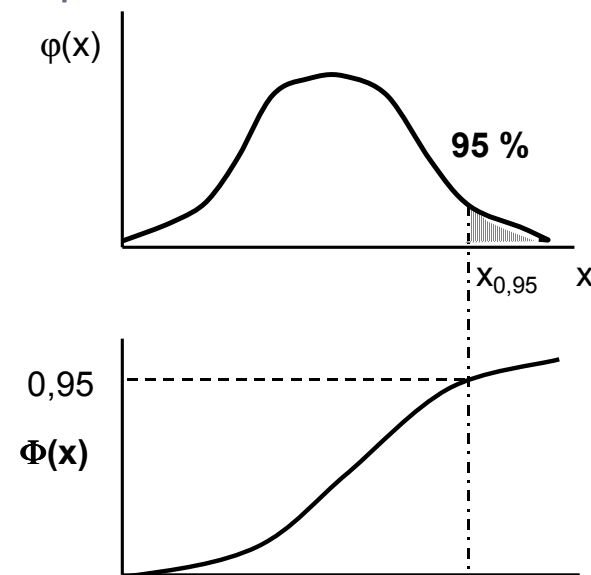
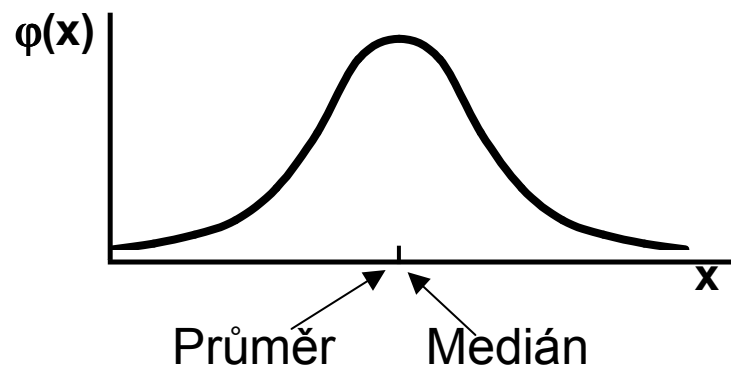
$\varphi(x)$



# Parametry rozložení



- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
  - **Středu** (medián, průměr, geometrický průměr)
  - **Šířky rozložení** (rozsah hodnot, rozptyl, směrodatná odchylka)
  - **Tvaru rozložení** (skewness, kurtosis)
  - **Kvantily rozložení** – kolik % řady dat leží nad a pod kvantilem



# Stručný přehled modelových rozložení I.

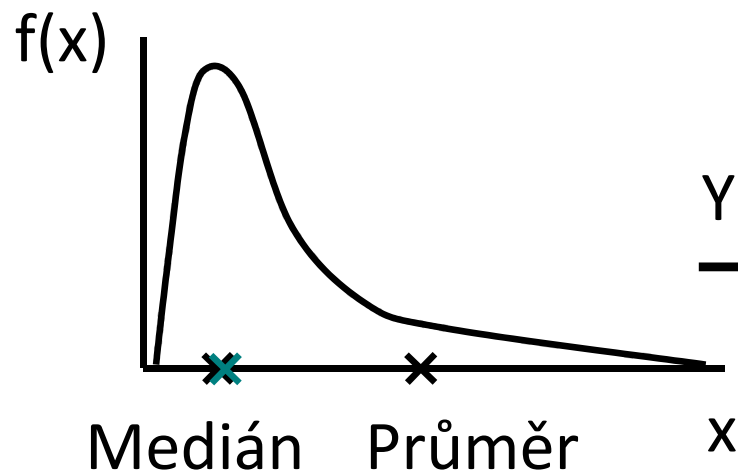
Rozložení	Parametry	Stručný popis
<b>Normální</b>	Průměr ( $\mu$ ) Rozptyl ( $\sigma^2$ )	Symetrická funkce popisující intervalovou hustotu četnosti; nejpravděpodobnější jsou průměrné hodnoty znaku v populaci.
<b>Log-normální</b>	Medián Geometrický průměr Rozptyl ( $\sigma^2$ )	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
<b>Weibullovo</b>	$\alpha$ - parametr tvaru $\beta$ - parametr rozsahu hodnot	Změnou parametru $\alpha$ lze modelovat distribuci doby přežití, např. stresovaného organismu. Rozložení využívané i jako model k odhadu $LC_{50}$ nebo $EC_{50}$ u testů toxicity.
<b>Rovnoměrné</b>	Medián Geometrický průměr Rozptyl ( $\sigma^2$ )	Všem hodnotám proměnné přiřazena stejná pravděpodobnost.
<b>Triangulární</b>	$f(x) = [b - \text{ABS}(x - a)] / b^2$ $a - b < x < a + b$	Pravděpodobnostní funkce pro typ rozložení, kdy jsou střední hodnoty výrazně pravděpodobnější než hodnoty okrajové.
<b>Gamma</b>	Parametry distribuční funkce: $\alpha$ - parametr tvaru $\beta$ - parametr rozsahu hodnot	Umožňuje flexibilní modelování distribučních funkcí nejrůznějších tvarů. Např. $\chi^2$ rozložení je rozložení typu Gamma. Gamma rozložení s $\alpha = 1$ je známo jako exponenciální rozložení.

# Stručný přehled modelových rozložení II.

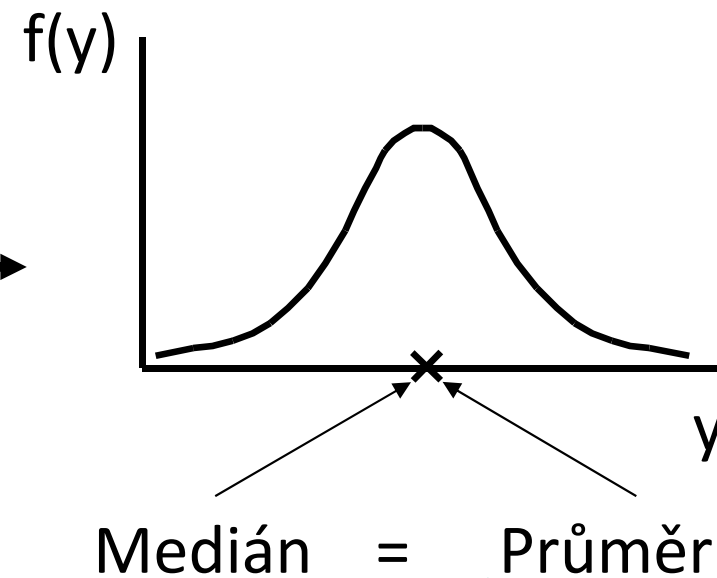
Rozložení	Parametry	Stručný popis
<b>Beta</b>	Parametry distribuční funkce: $\alpha$ - parametr tvaru $\beta$ - parametr rozsahu hodnot	Pravděpodobnostní funkce pro proměnnou omezenou rozsahem do intervalu [0; 1]. Je matematicky komplikovanější, ale velmi flexibilní při popisu změn hodnot proměnné v ohraničeném intervalu.
<b>Studentovo</b>	Stupně volnosti - uvažuje velikost vzorku Průměr Rozptyl	Simuluje normální rozložení pro menší vzorky čísel. Pro větší soubory ( $n > 100$ ) se limitně blíží k normálnímu rozložení.
<b>Pearsonovo (Chí-kvadrát)</b>	Stupně volnosti - uvažuje velikost vzorku	Slouží především k porovnání četností jevů ve dvou a více kategoriích. Používá se k modelování rozložení odhadu rozptylu normálně rozložených dat.
<b>Fisher-Snedecorovo</b>	Dvojí stupně volnosti - uvažuje velikost dvou vzorků	Používá se k testování - F test pro porovnání dvou výběrových rozptylů, ANOVA atd.



# Log-normální rozložení lze jednoduše transformovat



$$Y = \ln [X]$$



EXP ( $\bar{Y}$ ) = Geometrický průměr  $X$



$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n}$$



# Normální rozložení



Normální rozložení

Pravidlo 3 sigma

Parametry normálního rozložení

Vizuální ověření normality dat

# Normální rozdělení



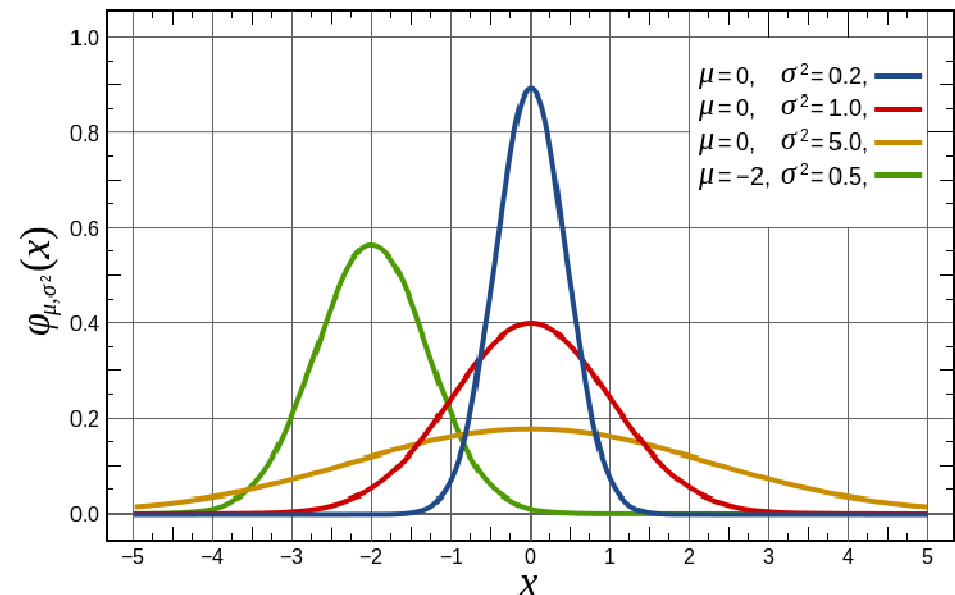
- Nejklasičtějším modelovým rozložením, od něhož je odvozena celá řada statistických analýz, je tzv. **normální rozložení**, známé též jako **Gaussova křivka**.
- Popisuje rozdělení pravděpodobnosti spojité náhodné veličiny: např. výška v populaci, chyba měření...

- Je kompletně popsáno dvěma parametry:

**$\mu$  – střední hodnota**

**$\sigma^2$  – rozptyl**

Označení:  **$N(\mu, \sigma^2)$**

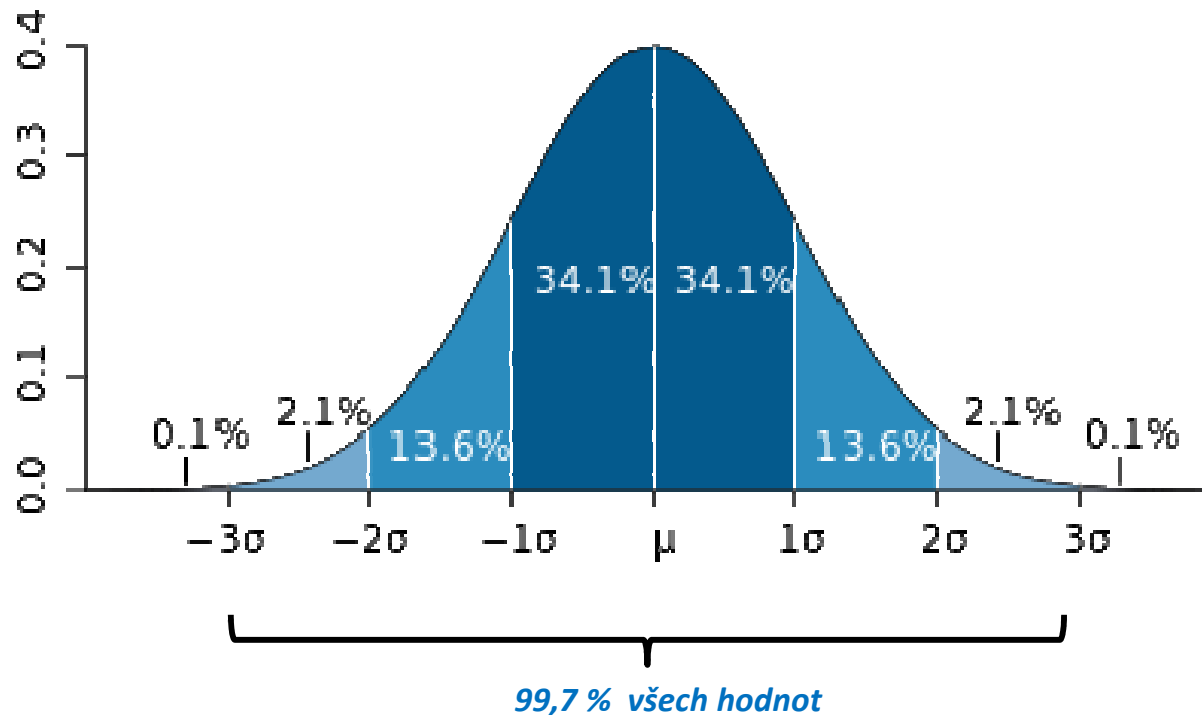


- Normalita je klíčovým předpokladem řady statistických metod
- Pro ověření normality existuje řada testů a grafických metod

# Pravidlo 3 sigma



- V rozmezí  $\mu \pm 3\sigma$  by se mělo vyskytovat 99,7 % všech hodnot

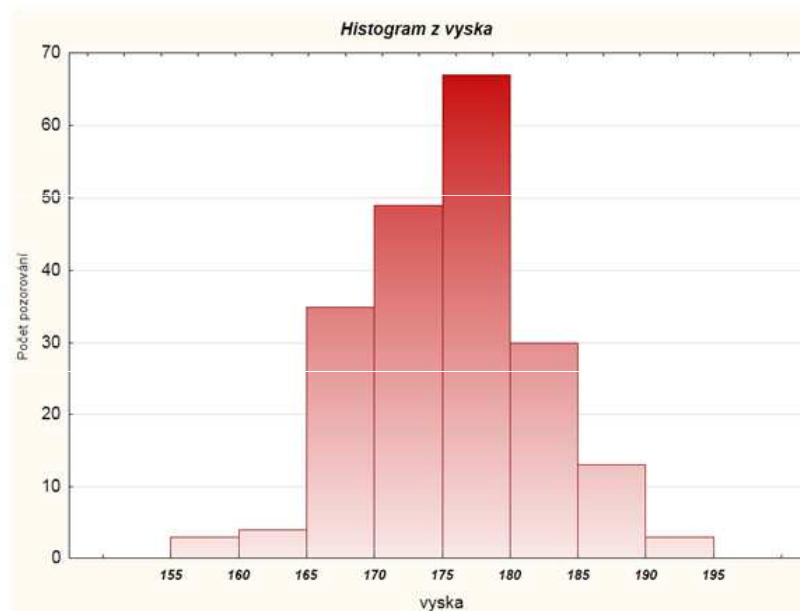
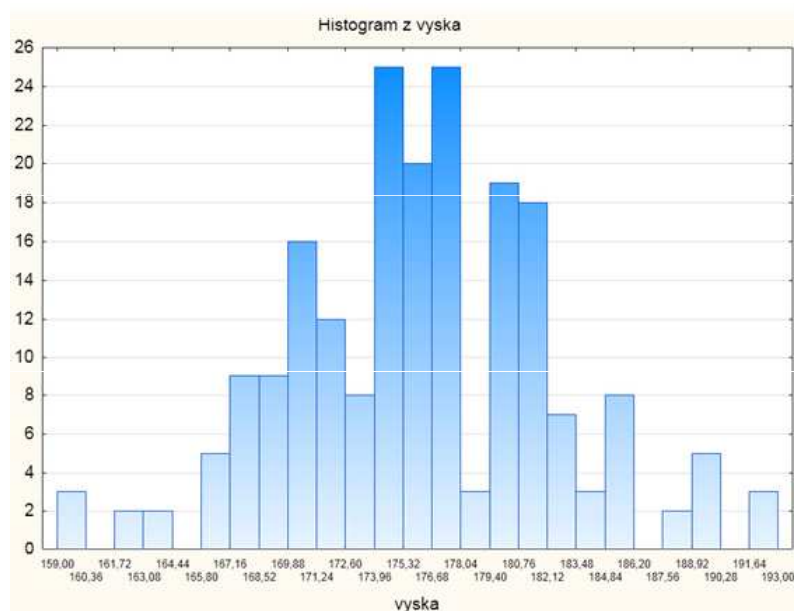


- Použití: zhodnotíme tvar rozdělení (pouze orientačně) a přítomnost odlehlých hodnot

# Vizuální ověření normality



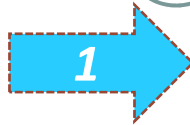
- Pro hodnocení tvaru rozložení lze využít histogram (nevýhoda: nutné určit „vhodný“ počet sloupců)



- Další možnosti:
  1. **Q-Q graf** (kvantil-kvantilový graf)
  2. **P-P graf** (pravděpodobnostně-pravděpodobnostní graf)
  3. **N-P graf** (normální-pravděpodobnostní graf)

# Řešení v softwaru Statistica

V menu *Graphs* zvolíme **2D**



- Scatterplots w/Box Plots...
- Normal Probability Plots...
- Quantile-Quantile Plots...**
- Probability-Probability Plots...

Quantile-Quantile Plots

Quick | Advanced | Appearance | Categorized | Options 1 | Options 2

Variables:  
none

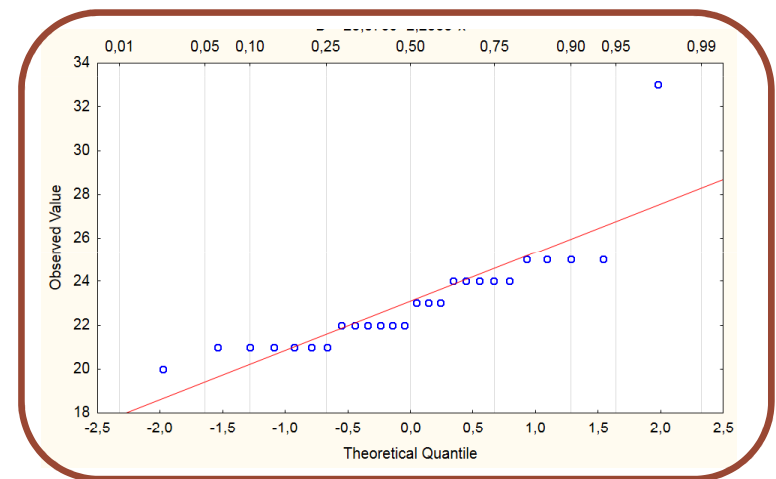
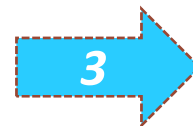
Distribution:  
Normal  
Beta

Plot layout  
 Multiple plots in one graph

Do not assign average ranks to tied observations

Výběr rozdělení

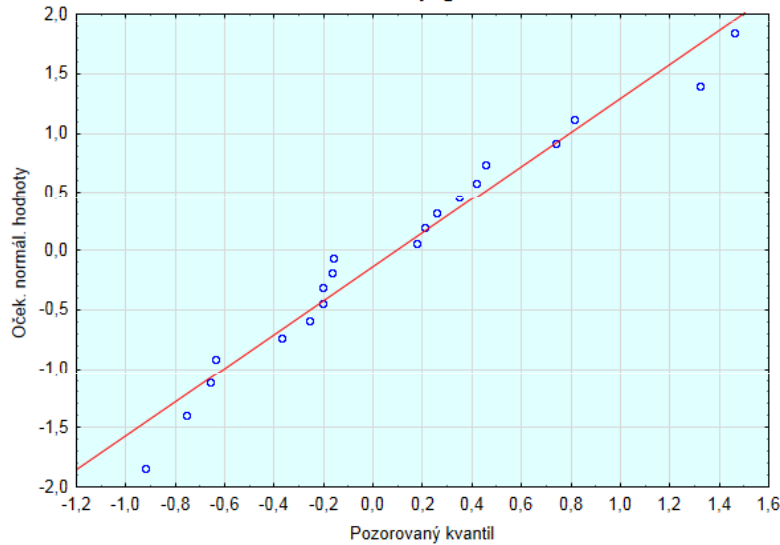
- V případě, že máme v datech několik stejných hodnot, je vhodné odškrtnout Neurčovat průměrnou pozici svázaných pozorování



# Rozdíl mezi N-P, Q-Q, P-P grafem



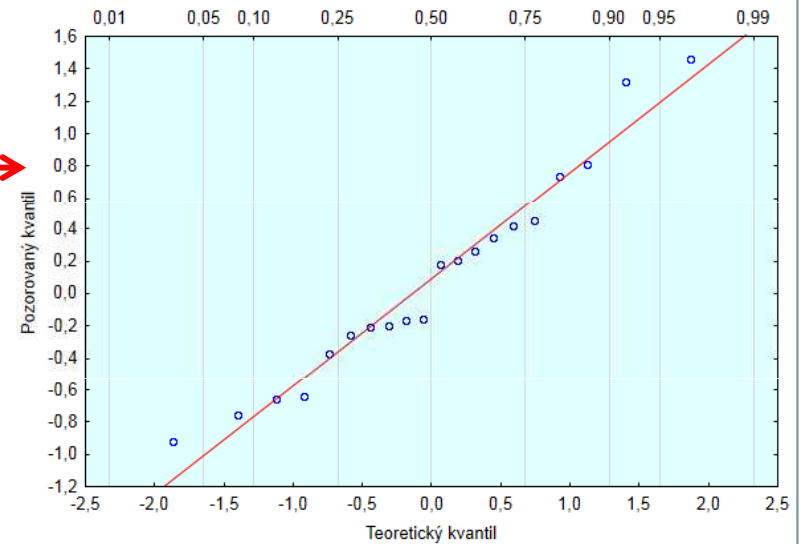
Normální p-graf



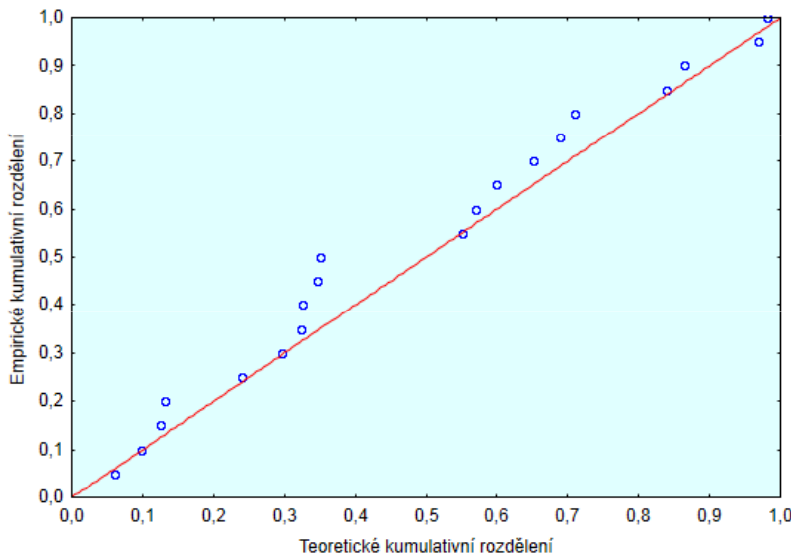
???

- Pouze výměna os
- Znázorněn pozorovaný a teoretický kvantil

Graf Q-Q



Graf P-P



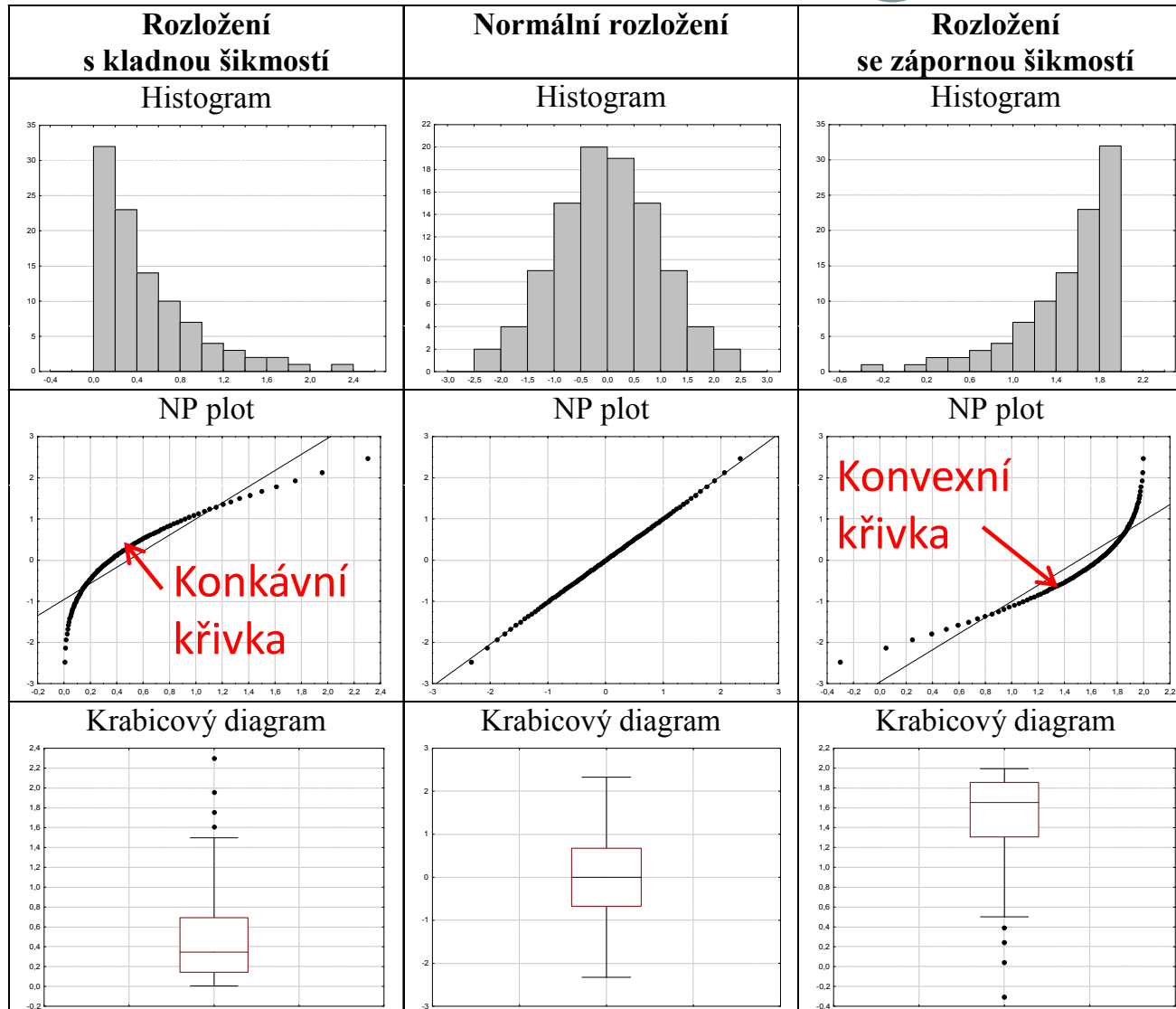
- Vykresleno kumulativní rozdělení

**PAMATUJ:**

**Pocházejí-li data z normálního rozložení, pak body budou ležet okolo přímky**



# Jak se projeví asymetrie dat v diagnostických grafech?



Výukové materiály: Výpočetní statistika,  
RNDr. Marie Budíková, Dr., 2011



# Základy testování hypotéz



**Princip statistického testování hypotéz**

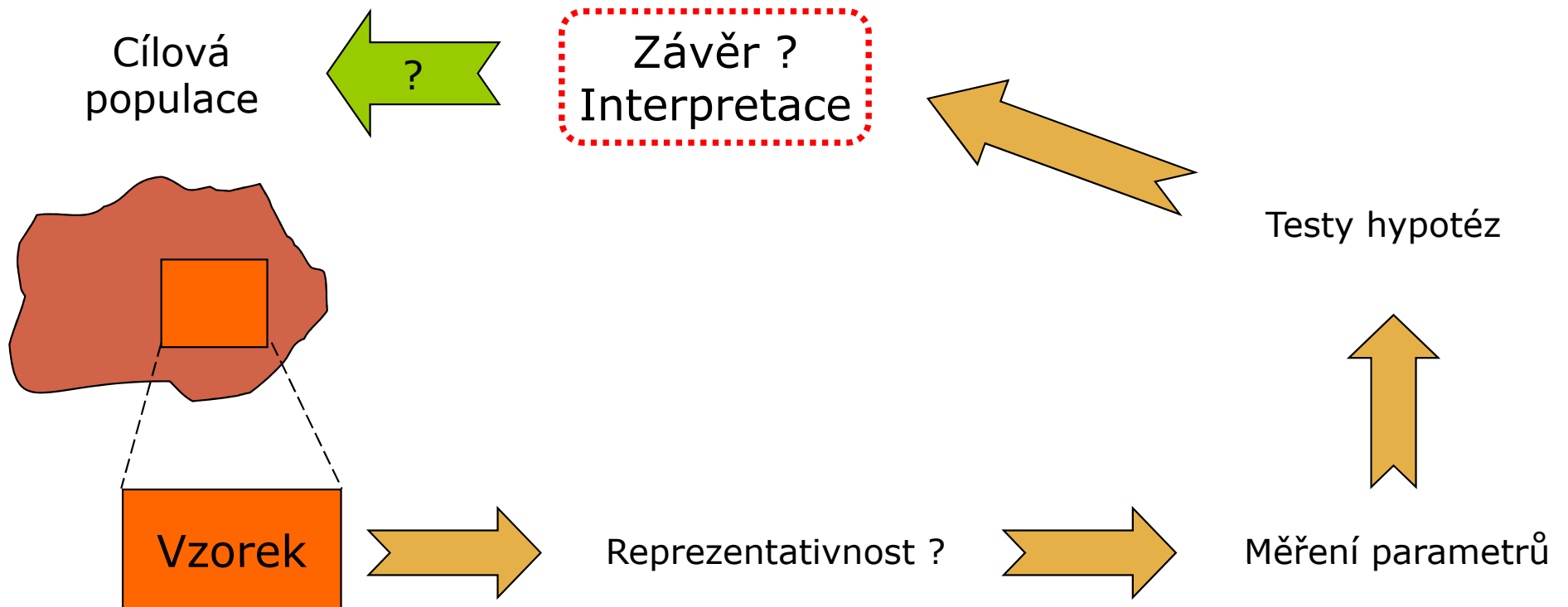
**Pojmy statistických testů**

**Normalita dat a její význam pro testování**

**Ověření normality dat pomocí testu**

# Princip testování hypotéz

- Formulace hypotézy
- Výběr cílové populace a z ní reprezentativního vzorku
- Měření sledovaných parametrů
- Použití odpovídajícího testu → závěr testu
- Interpretace výsledků



# Statistické testování – základní pojmy



➤ Nulová hypotéza  $H_0$

$H_0$ : sledovaný efekt je nulový

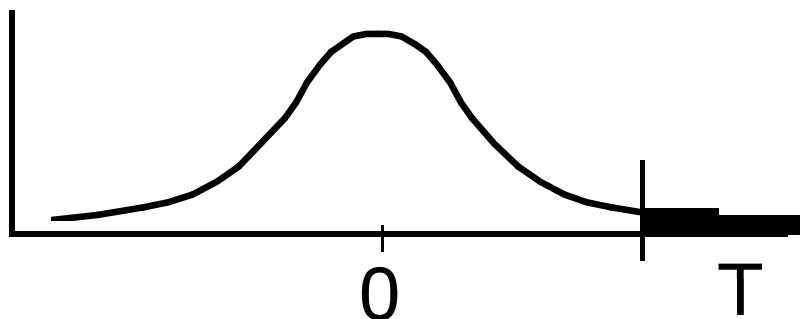
➤ Alternativní hypotéza  $H_A$

$H_A$ : sledovaný efekt je různý mezi skupinami

➤ Testová statistika

$$\text{Testová statistika} = \frac{\text{Pozorovaná hodnota} - \text{Očekávaná hodnota}}{\text{Variabilita dat}} * \sqrt{\text{Velikost vzorku}}$$

➤ Kritický obor testové statistiky



Statistické testování odpovídá na otázku, zda je pozorovaný rozdíl náhodný či nikoliv. K odpovědi na otázku je využit statistický model – testová statistika.

# Hypotézy – příklady



## 1. Je systolický tlak větší u pacientů s hypertenzí než u zdravých lidí?

Označme střední hodnotu systolického tlaku u pacientů symbolem  $\theta_1$  a střední hodnotu systolického tlaku u zdravých lidí  $\theta_2$ .

**Nulová hypotéza:**  $H_0 : \theta_1 = \theta_2$  Není rozdíl v sys. tlaku u pacientů a kontrol.

**Alternativní hypotéza:**  $H_1 : \theta_1 > \theta_2$  Sys. tlak vyšší u pacientů než u kontrol.

## 2. Je průměrná hodnota IQ u pacientů s Alzheimer. chorobou menší než průměrná hodnota celé populace? Označme střední hodnotu IQ u pacientů symbolem $\theta_1$ a u celé populace symbolem $\theta_0$ .

**Nulová hypotéza:**  $H_0 : \theta_1 = \theta_0$

**Alternativní hypotéza:**  $H_1 : \theta_1 < \theta_0$

## 3. Liší se objem hipokampu u pacientů s Alzheimer. chorobou (AD), pacientů s mírnou kognitivní poruchou (MCI) a zdravých lidí (CN)?

Označme střední hodnotu objemu hipokampu u jednotlivých skupin symboly  $\theta_{AD}$ ,  $\theta_{MCI}$ ,  $\theta_{CN}$ .

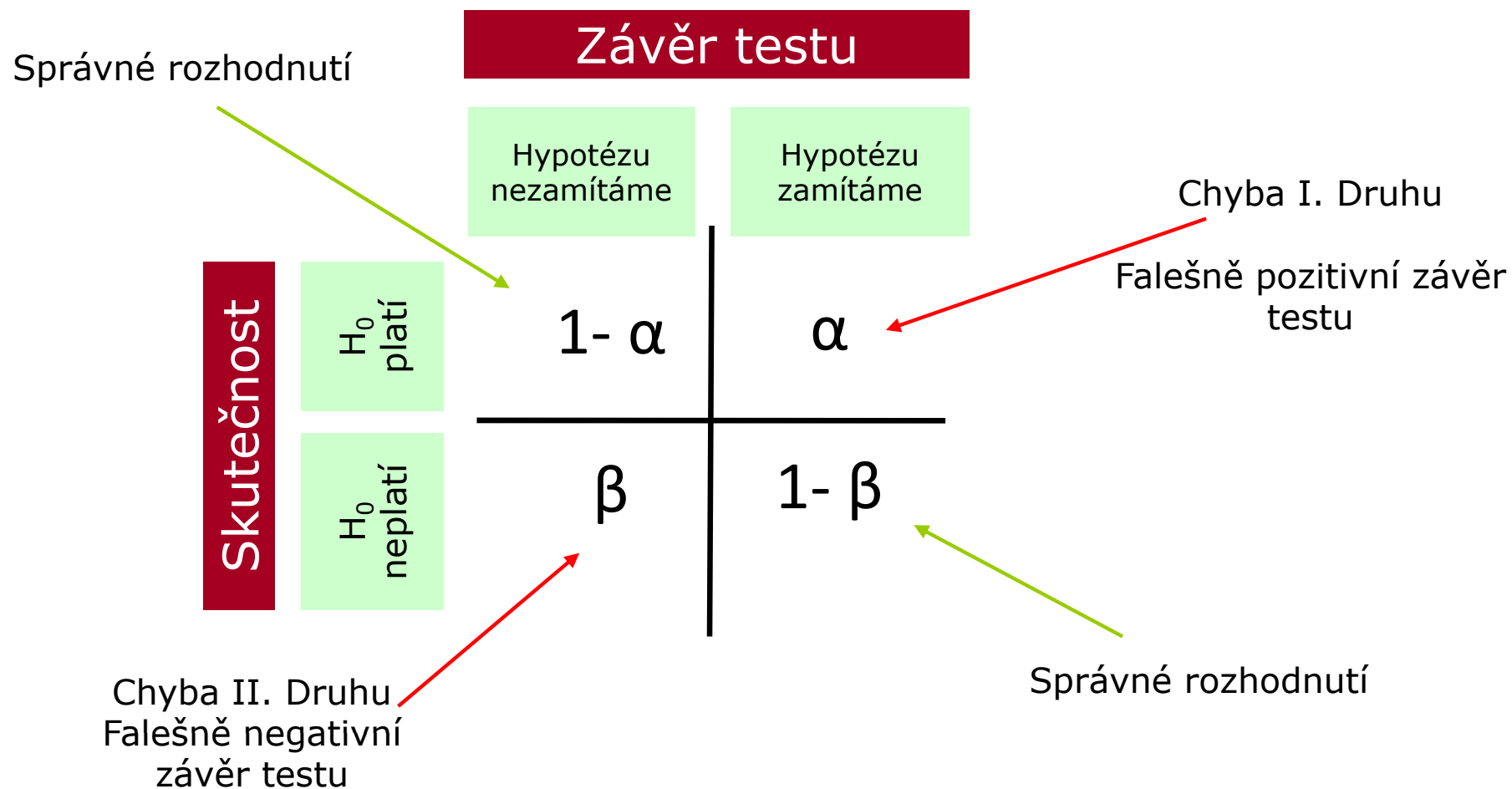
**Nulová hypotéza:**  $H_0 : \theta_{AD} = \theta_{MCI} = \theta_{CN}$

**Alternativní hypotéza:**  $H_1 : \text{Nejméně jedno } \theta \text{ je odlišné od ostatních.}$

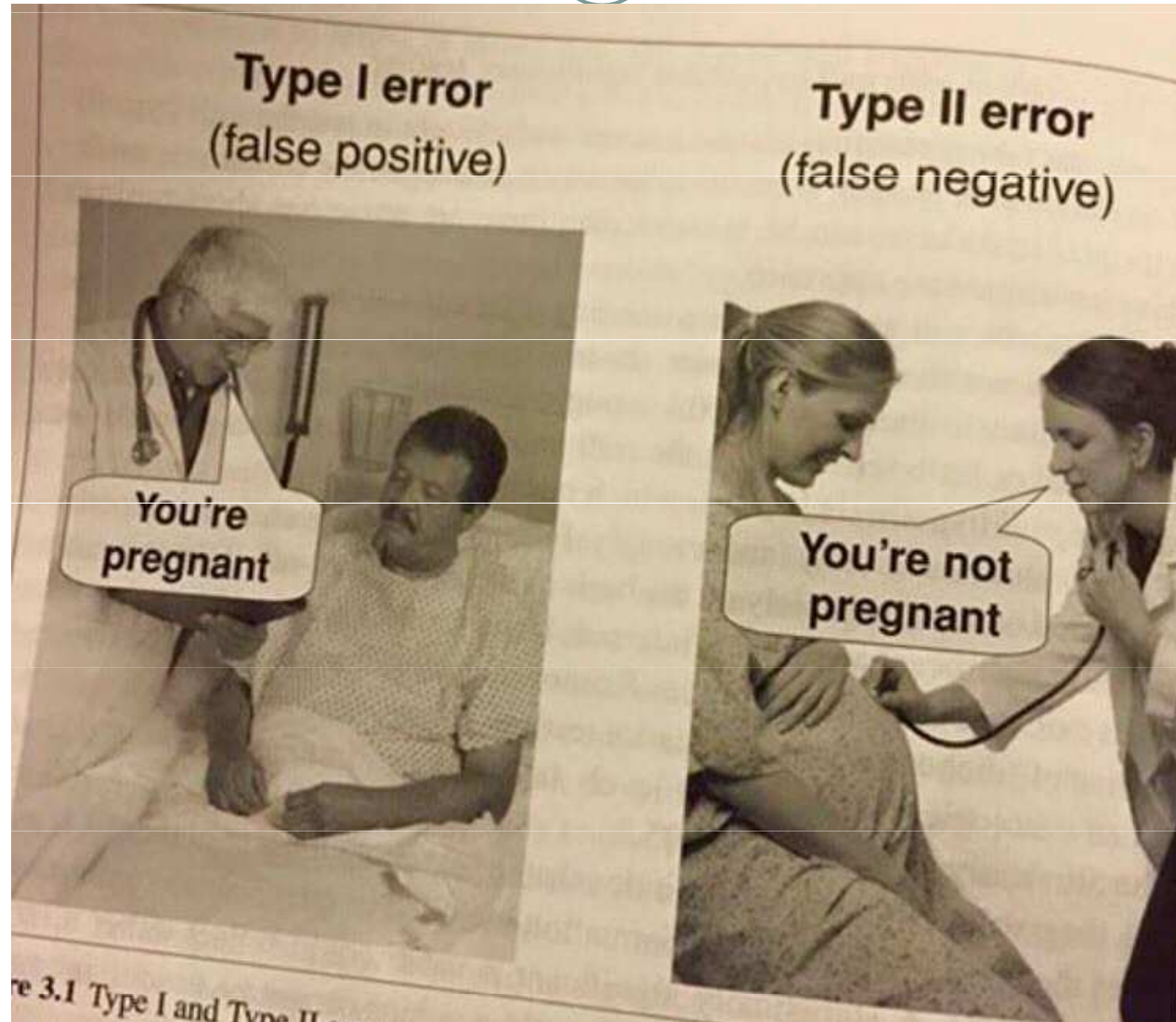
# Možné chyby při testování hypotéz



- I přes dostatečnou velikost vzorku a kvalitní design experimentu se můžeme při rozhodnutí o zamítnutí/nezamítnutí nulové hypotézy dopustit chyby.



# Možné chyby při testování hypotéz



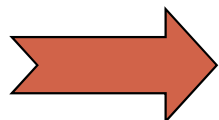
re 3.1 Type I and Type II

# Význam chyb při testování hypotéz



## Pravděpodobnost chyby 1. druhu

$\alpha$

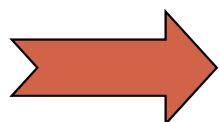


Pravděpodobnost nesprávného zamítnutí nulové hypotézy, **hladina významnosti**



## Pravděpodobnost chyby 2. druhu

$\beta$

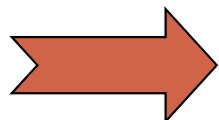


Pravděpodobnost nerozpoznání neplatné nulové hypotézy



## Síla testu

$1-\beta$



Pravděpodobnostně vyjádřená schopnost rozpoznat neplatnost hypotézy

# Způsoby testování



Testování  $H_0$  proti  $H_A$  na hladině významnosti  $\alpha$  můžeme provést třemi různými způsoby:

1. **Kritický obor** (označení  $W$ ), neboli obor zamítnutí  $H_0$ ,
2. **Interval spolehlivosti**,
3. **P-hodnota**.



# Způsoby testování: P-hodnota



Významnost hypotézy hodnotíme dle získané tzv. **p-hodnoty**, která vyjadřuje pravděpodobnost, s jakou číselné realizace výběru podporují  $H_0$ , je-li pravdivá.

P-hodnotu porovnáme s  $\alpha$  (**hladina významnosti**, stanovujeme ji na 0,05, tzn., že připouštíme 5% chybu testu, tedy, že zamítneme  $H_0$ , ačkoliv ve skutečnosti platí).

P-hodnotu získáme při testování hypotéz ve statistickém softwaru.

- Je-li p-hodnota  $\leq \alpha$ , pak  $H_0$  zamítáme na hladině významnosti  $\alpha$  a přijímáme  $H_A$ .
- Je-li p-hodnota  $> \alpha$ , pak  $H_0$  nezamítáme na hladině významnosti  $\alpha$ .

P-hodnota vyjadřuje pravděpodobnost za platnosti  $H_0$ , s níž bychom získali stejnou nebo extrémnější hodnotu testové statistiky.

# Parametrické vs. neparametrické testy



## Parametrické testy

- Mají předpoklady o rozložení vstupujících dat (např. normální rozložení)
- Při stejném N a dodržení předpokladů mají vyšší sílu testu než testy neparametrické
- **Pokud nejsou dodrženy předpoklady parametrických testů, potom jejich síla testu prudce klesá a výsledek testu může být zcela chybný a nesmyslný**

## Neparametrické testy

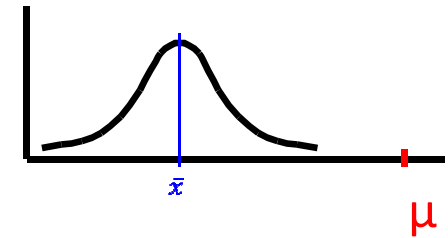
- Nemají předpoklady o rozložení vstupujících dat, lze je tedy použít i při asymetrickém rozložení, odlehlých hodnotách, či nedetekovatelném rozložení
- Snížená síla těchto testů je způsobena redukcí informační hodnoty původních dat, kdy neparametrické testy nevyužívají původní hodnoty, ale nejčastěji pouze jejich pořadí

# One-sample vs. two-sample testy



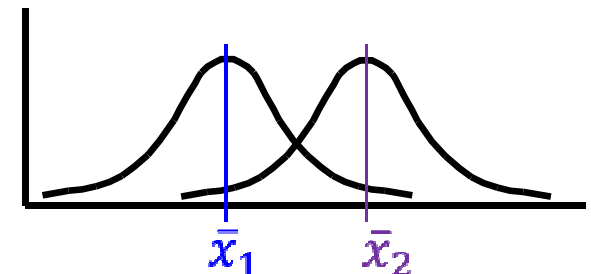
## Jednovýběrové testy (one-sample)

- Srovnávají jeden vzorek (one sample, jednovýběrové testy) s referenční hodnotou (popřípadě se statistickým parametrem cílové populace)
- V testu je tedy srovnáváno rozložení hodnot (vzorek) s jediným číslem (referenční hodnota, hodnota cílové populace)
- Otázka položená v testu může být vztažena k průměru, rozptylu, podílu hodnot i dalším statistickým parametrům popisujícím vzorek



## Dvouvýběrové testy (two-sample)

- Srovnávají navzájem dva vzorky (two sample, dvouvýběrové testy)
- V testu jsou srovnávány dvě rozložení hodnot
- Otázka položená v testu může být opět vztažena k průměru, rozptylu, podílu hodnot i dalším statistickým parametrům popisujícím vzorek
- Kromě testů pro dvě skupiny hodnot existují samozřejmě i testy pro více skupin dat

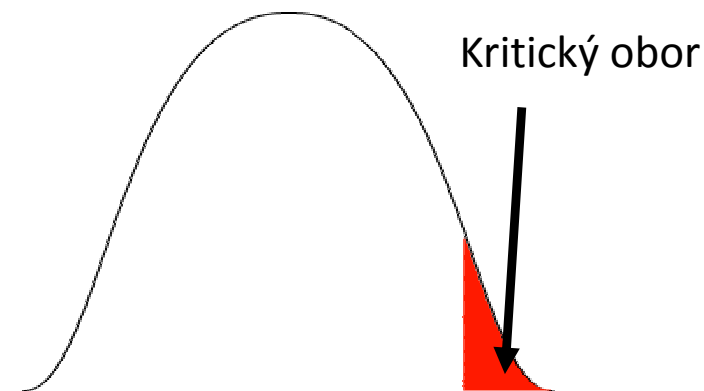


# One-tailed vs. two-tailed testy



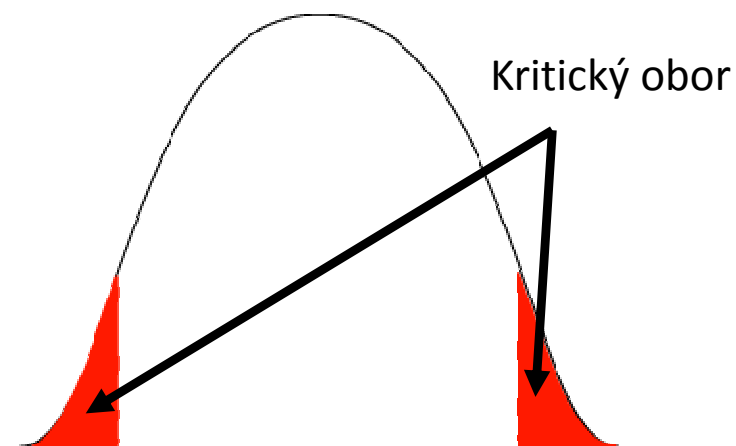
## Jednostranné testy (one-tailed)

- Hypotéza testu je postavena asymetricky, tedy ptáme se na větší než/ menší než
- Test může mít pouze dvojí výstup – jedna z hodnot je větší (menší) než druhá a všechny ostatní případy



## Oboustranné testy (two-tailed)

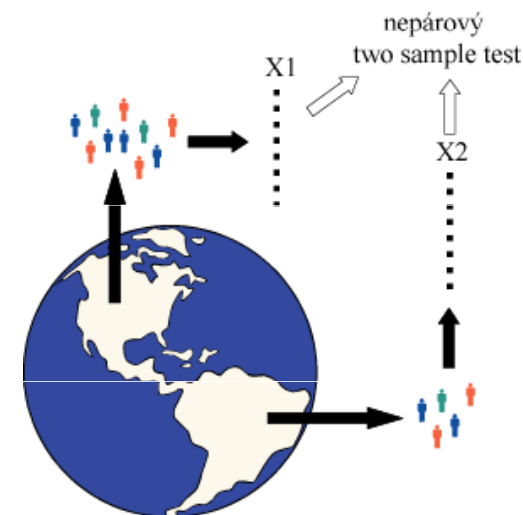
- Hypotéza testu se ptá na otázku rovná se/nerovná se
- Test může mít trojí výstup – menší - rovná se – větší než
- Situace nerovná se je tedy souhrnem dvou možných výstupů testu (menší+větší)



# Nepárový vs. párový design

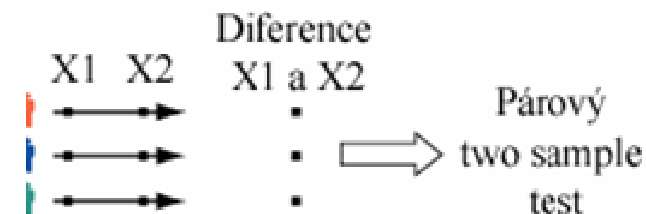
## Nepárový design

- Skupiny srovnávaných dat jsou na sobě zcela nezávislé (též nezávislý, independent design), např. lidé z různých zemí, nezávislé skupiny pacientů s odlišnou léčbou atd.
- Při výpočtu je nezbytné brát v úvahu charakteristiky obou skupin dat



## Párový design

- Mezi objekty v srovnávaných skupinách existuje vazba, daná např. člověkem před a po operaci, reakce stejného kmene krys atd.
- Vazba může být buď přímo dána, nebo pouze předpokládána (v tom případě je nutné ji ověřit)
- Test je v podstatě prováděn na diferencích skupin, nikoliv na jejich původních datech

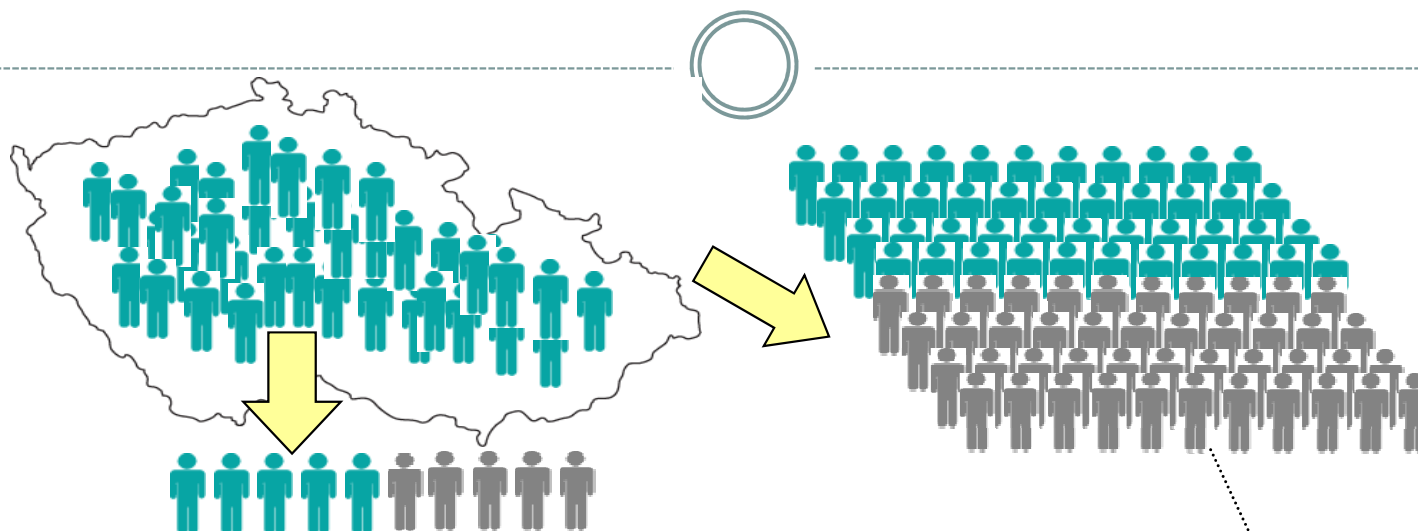


# Důležité poznámky k testování hypotéz



- **Nezamítnutí nulové hypotézy neznamená automaticky její přijetí!** Může se jednat o situaci, kdy pro zamítnutí nulové hypotézy nemáme dostatečné množství informace.
- **Dosažená hladina významnosti testu** (ať už 0,05, 0,01 nebo 0,1) **nesmí být slepě brána jako hranice pro existenci / neexistenci testovaného efektu.**
- **Malá p-hodnota nemusí znamenat velký efekt.** Hodnota testové statistiky a p-hodnota mohou být ovlivněny velkou velikostí vzorku a malou variabilitou pozorovaných dat.
- **Na výsledky testování musí být nahlíženo kriticky** – jedná se o závěr založený „pouze“ na jednom výběrovém souboru.
- **Statistická významnost** indikuje, že pozorovaný rozdíl není náhodný, ale nemusí znamenat, že je významný i ve skutečnosti. Důležitá je i **praktická (klinická) významnost.**

# Vliv velikosti vzorku na výsledky testování



Dvě skupiny pacientů s nepatrným rozdílem v dané charakteristice, který ale není klinicky významný.

$$n_1 = 10, n_2 = 10$$
$$p = 0,797$$

$$n_1 = 100, n_2 = 100$$
$$p = 0,140$$

$$n_1 = 1000, n_2 = 1000$$
$$p < 0,001$$

**Statistická významnost způsobená velkým N**

# Statistické testy a normalita dat



- Normalita dat je jedním z předpokladů tzv. **parametrických testů** (testů založených na předpokladu nějakého rozložení) – např. *t*-testy
- Pokud data nejsou normální, neodpovídají ani modelovému rozložení, které je použito pro výpočet (*t*-rozložení), a test tak může lhát
- Řešením je tedy:
  - Transformace dat za účelem dosažení normality jejich rozložení
  - **Neparametrické testy** – tyto testy nemají žádné předpoklady o rozložení dat

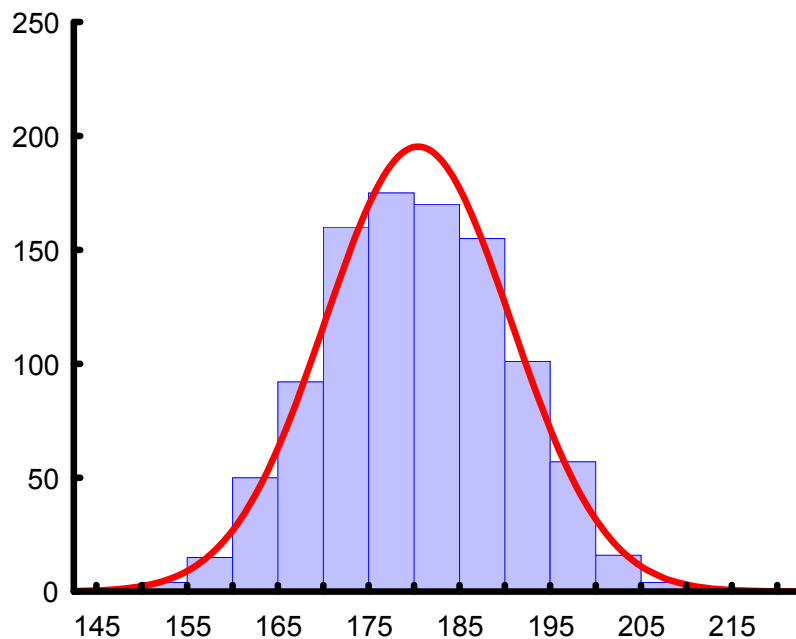
Typ srovnání	Parametrický test	Neparametrický test
2 skupiny dat nepárově:	Nepárový dvouvýběrový <i>t</i> -test	Mannův-Whitneyův test
2 skupiny dat párově:	Párový <i>t</i> -test	Wilcoxonův test, znaménkový test
Více skupin nepárově:	ANOVA (analýza rozptylu)	Kruskalův- Wallisův test
Korelace:	Pearsonův koeficient	Spearmanův koeficient



# Testy normality



- Testy normality pracují s nulovou hypotézou, že není rozdíl mezi zpracovávaným rozložením a normálním rozložením. Vždy je ovšem dobré prohlédnout si i histogram, protože některé odchylky od normality, např. bimodalitu některé testy neodhalí.



## •Chí-kvadrát test dobré shody

V testu dobré shody jsou data rozdělena do kategorií (obdobně jako při tvorbě histogramu), tyto intervaly jsou normalizovány (převedeny na normální rozložení) a podle obecných vzorců normálního rozložení jsou k nim dopočítány očekávané hodnoty v intervalech, pokud by rozložení bylo normální. Pozorované normalizované četnosti jsou poté srovnány s očekávanými četnostmi pomocí  $\chi^2$  testu dobré shody. Test dává dobré výsledky, ale je náročný na  $n$ , tedy množství dat, aby bylo možné vytvořit dostatečný počet tříd hodnot.

## •Kolmogorovův - Smirnovův test

Tento test je často používán, dokáže dobře najít odlehlé hodnoty, ale počítá spíše se symetrií hodnot než přímo s normalitou. Jde o neparametrický test pro srovnání rozdílu dvou rozložení. Je založen na zjištění rozdílu mezi reálným kumulativním rozložením (vzorek) a teoretickým kumulativním rozložením. Měl by být počítán pouze v případě, že známe průměr a směrodatnou odchylku hypotetického rozložení, pokud tyto hodnoty neznáme, měla by být použita jeho modifikace – Lilieforsův test.

## •Shapirův-Wilkův test

Jde o neparametrický test použitelný i při velmi malých  $n$  (10) s dobrou silou testu, zvláště ve srovnání s alternativními typy testů, je zaměřen na testování symetrie.

# Společné cvičení



# Společné cvičení – ověřování normality dat



1. Načtěte si do programu STATISTICA soubor **03\_spolecne\_cviceni\_pacienti.sta**.
2. Vypište základní popisné statistiky pro proměnné *Leukocyty*, *Výška* a *Náklady za hospitalizaci* pro celý soubor pacientů.

## Normální rozdělení – proměnná *Leukocyty*:

3. Ověřte normalitu proměnné *Leukocyty* pomocí:
  - histogramu (*Nápověda: Graphs – Histogram*),
  - krabicového grafu (*Nápověda: Graphs – 2D – Box Plots*),
  - diagnostických grafů (Q-Q grafu, N-P grafu a P-P grafu) (*Nápověda: Graphs – 2D – Quantile-Quantile Plots / Normal Probability Plots / Probability-Probability Plots*),
  - Shapirova-Wilkova testu nebo Lilieforsovy modifikace Kolmogorovova-Smirnovova testu (*Nápověda: lze provést čtyřmi způsoby:*
    - 1) v nastavení histogramu: záložka *Advanced* → *Statistics*: vybereme test;
    - 2) v nastavení N-P grafu: záložka: *Quick* → *Statistics*: zaškrtneme test;
    - 3) v menu *Basic statistics* → *Frequency tables* → záložka *Normality* → vybereme test a klikneme na *Tests for Normality*);
    - 4) v menu *Basic statistics* → *Descriptive statistics* → záložka *Normality* → vybereme test a klikneme na *Histograms*)
4. Podívejte se, jak vypadají jednotlivé diagnostické grafy v případě normálního rozdělení.

# Společné cvičení – ověřování normality dat



## Normální rozdělení s odlehlou hodnotou – proměnná *Výška*:

5. Ověřte normalitu proměnné *Výška* pomocí:

- histogramu,
- krabicového grafu,
- diagnostických grafů (Q-Q grafu, N-P grafu a P-P grafu),
- Shapirova-Wilkova testu / Lilieforsovy modifikace Kolmogorovova-Smirnovova testu.

6. Jak se projeví odlehlá hodnota v grafech?

7. Zkopírujte proměnnou *výška* (nebo vytvořte pomocí vzorce) do nové proměnné a vymažte v této nové proměnné odlehlou hodnotu (*nápověda: seřadte si data podle proměnné výška: karta Data → Sort → vložíme proměnnou výška*). Ověřte, zda se po vynechání odlehlé hodnoty data řídí normálním rozložením.

Odlehlou hodnotu (řádek 16, hodnota 100, nahraďte hodnotou 144).

# Společné cvičení – ověřování normality dat



## Logaritmicko-normální rozdělení – proměnná *Náklady za hospitalizaci*:

9. Vykreslete histogram proměnné *Náklady za hospitalizaci*. Proložte histogram nejdříve normálním rozložením, poté log-normálním rozložením.

10. Dále ověřte normalitu dat pomocí:

- diagnostických grafů (Q-Q grafu, N-P grafu a P-P grafu),
- Shapirova-Wilkova testu / Lilieforsovy modifikace Kolmogorovova-Smirnovova testu.

11. Jak se výsledky liší ve srovnání s daty, která se řídí normálním rozdělením?

12. Transformujte proměnnou *Náklady za hospitalizaci* pomocí přirozeného logaritmu do nové proměnné (*nápověda: Data → Transforms: LogNaklady=Log(v10)*).

13. Ověřte normalitu dat nové proměnné *LogNaklady* pomocí:

- histogramu, krabicového grafu, diagnostických grafů (Q-Q grafu, N-P grafu a P-P grafu), Shapirova-Wilkova testu / Lilieforsovy modifikace Kolmogorovova-Smirnovova testu.

14. Vypočtěte průměr a medián proměnné *Náklady za hospitalizaci*. Podívejte se na histogram proměnné *Náklady za hospitalizaci* a zhodnoťte vztah průměru a mediánu.

# Samostatné cvičení



# Samostatné cvičení – ověřování normality dat



1. Načtěte si do programu STATISTICA data **03\_spolecne\_cviceni\_pacienti.sta** (pokud je nemáte načtené ze společného cvičení). **Přidejte** za proměnnou *váha* novou **proměnnou BMI** (body mass index – index tělesné hmotnosti), kterou vypočítáte z proměnné *výška* a *váha*.

*Poznámka:* V případě, že jste ze samostatného cvičení nepřepsali odlehlou hodnotu proměnné *výška*, učiňte tak nyní (hodnotu 100 přepište na hodnotu 144).

2. Vypište zvlášť pro muže a ženy (proměnná *pohlaví*) **základní popisné statistiky** následujících proměnných: *váha*, *výška*, *BMI* (počet hodnot, průměr, medián, směrodatnou odchylku, minimum a maximum). Výsledek znázorněte v jedné tabulce (nápopověda: změňte nastavení formy výstupů v sekci *By Group*).

3. Vykreslete kategorizované histogramy proměnných *výška*, *váha* a *BMI* pro muže a ženy zvlášť. Zkuste si proložit histogramy postupně **normálním rozdělením** a dalšími rozděleními ze záložky *Advanced* → *Fit types*.

# Samostatné cvičení – ověřování normality dat



4. Pro proměnné *výška*, *váha* a *BMI* (opět pro muže a ženy zvlášť) vykreslete **Q-Q graf**, **N-P graf** a **P-P graf**. Které proměnné dle těchto diagnostických grafů podle vás mají normální rozložení? Zapište svůj odhad do připravené tabulky.
5. Otestujte normalitu dat proměnných *výška*, *váha* a *BMI* pro muže a ženy zvlášť pomocí **Shapirova-Wilkova testu**. Zapište výsledek (p-hodnotu) do připravené tabulky. Srovnajte své odhady z diagnostických grafů s výsledky testů.
6. V případě, že se dle diagnostických grafů nebo S-W testu data řídí normálním rozdělením, jaký je v uvedených případech odhad **parametrů tohoto rozdělení** (střední hodnoty a rozptylu)? Hodnoty zaznamenejte do tabulky.



# Samostatné cvičení – ověřování normality dat



**Tabulka:** Vizuální a testové ověření normality.

Proměnná	Normalita dle Q-Q / N-P / P-P grafu (ano/ne)	p-hodnota Shapirova-Wilkova testu	Odhad střední hodnoty	Odhad rozptylu
Výška				
Muži				
Ženy				
Váha				
Muži				
Ženy				
BMI				
Muži				
Ženy				

# Samostatné cvičení – ověřování normality dat



## Výsledky:

**Tabulka:** Vizuální a testové ověření normality.

Proměnná	Normalita dle Q-Q / N-P / P-P grafu (ano/ne)	p-hodnota Shapirova-Wilkova testu	Odhad střední hodnoty	Odhad rozptylu
<b>Výška</b>				
<b>Muži</b>	Ne/ne/ne	0.037		
<b>Ženy</b>	Ano/ano/ano	0.539	161.2	17.3
<b>Váha</b>				
<b>Muži</b>	Ne/ne/ne	0.004		
<b>Ženy</b>	Ano/ano/ano	0.784	65.9	25.1
<b>BMI</b>				
<b>Muži</b>	Ano/ano/ano	0.529	25.3	3.6
<b>Ženy</b>	Ano/ano/ano	0.200	25.4	4.3

# Samostatné cvičení – ověřování normality dat



## Poznámky k nejčastějším chybám:

1. Parametry normálního rozdělení jsou: **střední hodnota** a **rozptyl**. Nejlepším nestranným odhadem střední hodnoty u normálního rozdělení je **průměr** (nikoliv medián, ale měl by v případě normálního rozdělení stejný nebo podobný jako průměr), nejlepším nestranným odhadem rozptylu jako parametru je **výběrový rozptyl**.
2. Nepleťte si rozptyl a směrodatnou odchylku. **Směrodatná odchylka je odmocnina z rozptylu**. Na rozdíl od rozptylu je ve stejných jednotkách jako hodnocený parametr.

## Další chyby:

1. Přehozené skupiny pohlaví (záměna žen a mužů).
2. Odhad střední hodnoty a rozptylu měl být vyplněn pouze tam, kde jste pomocí testu nezamítli nulovou hypotézu o normalitě dat.
3. Správná interpretace např. výšky může být: „Pomocí Shapirova-Wilkova testu můžeme předpokládat, že se výška u žen v našem hodnoceném souboru řídí normálním rozdělením. U mužů jsme však nulovou hypotézu zamítli, tedy test prokázal, že výška u mužů nemá normální rozdělení.“