Normalita

Normální rozdělení a ověření normality dat

Modelová rozdělení

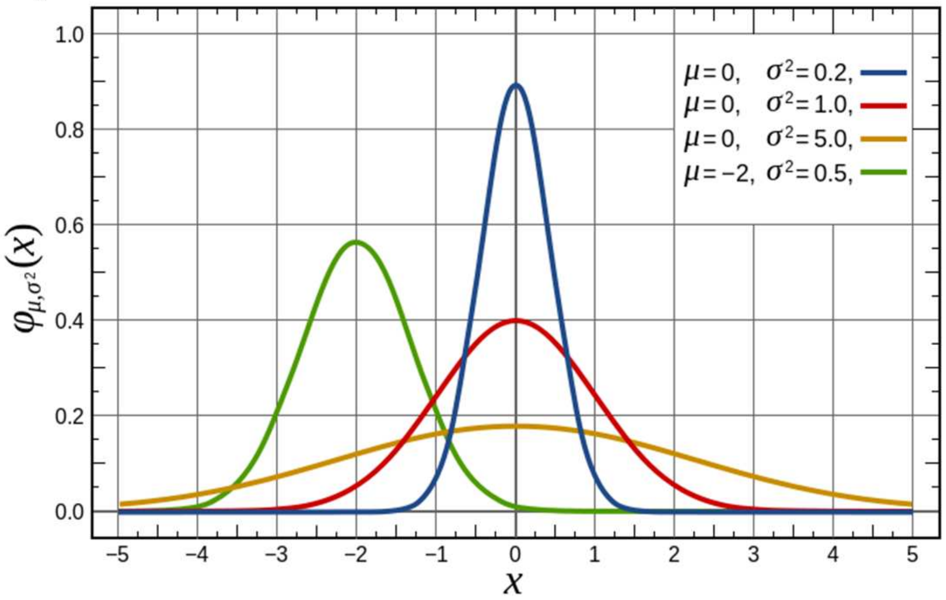
Normální rozdělení I

- 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:

μ – střední hodnota

σ^2 – rozptyl

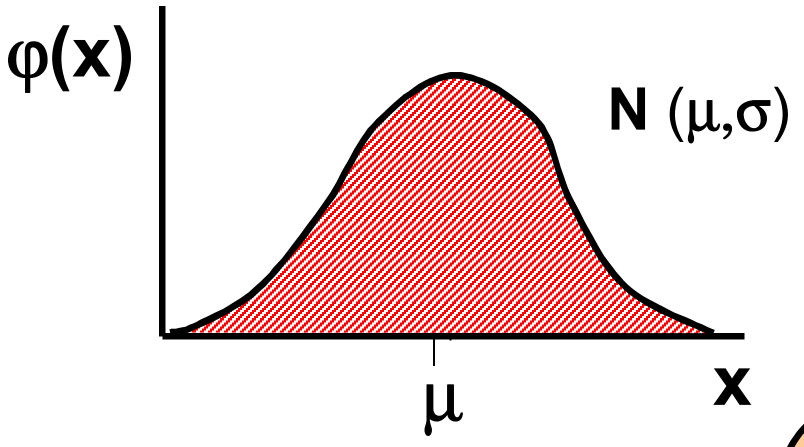
Označení: **N(μ, σ^2)**



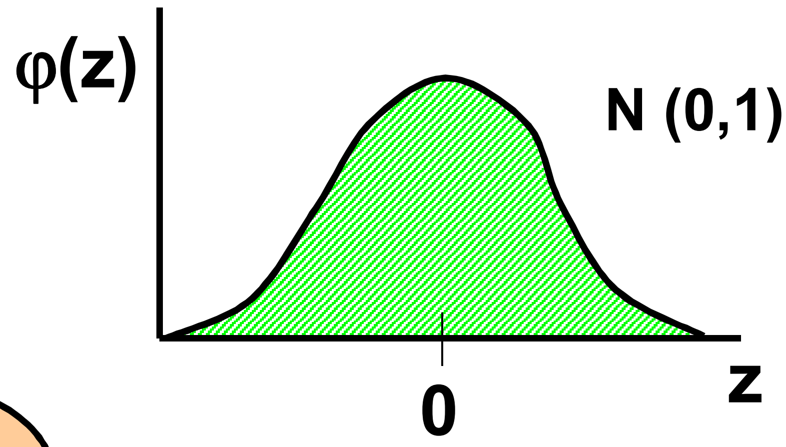
- Normalita je klíčovým předpokladem řady statistických metod
- Pro ověření normality existuje řada testů a grafických metod

Standardizované normální rozdělení

- Normální rozdělení se střední hodnotou nula a jednotkovým rozptylem



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



$$\varphi(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Vzorec pro hustotu normálního rozdělení

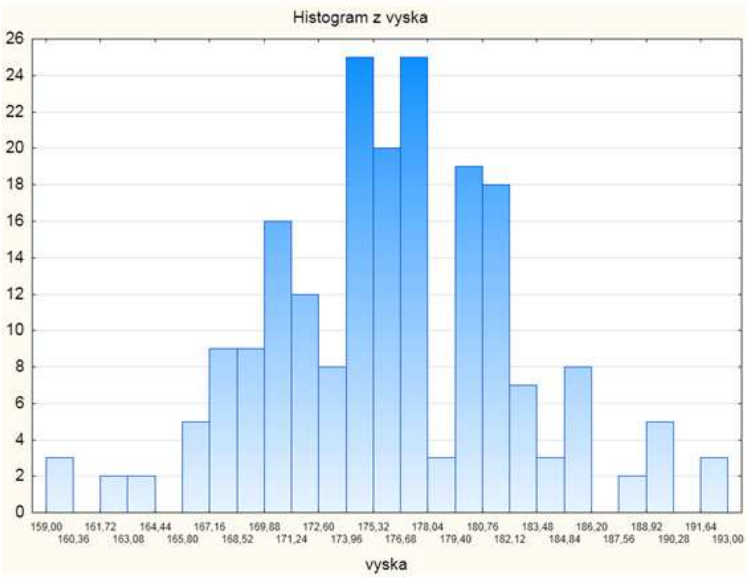
$$\varphi(z) = \frac{1}{\sqrt{2.\pi}}.e^{-\frac{z^2}{2}}$$

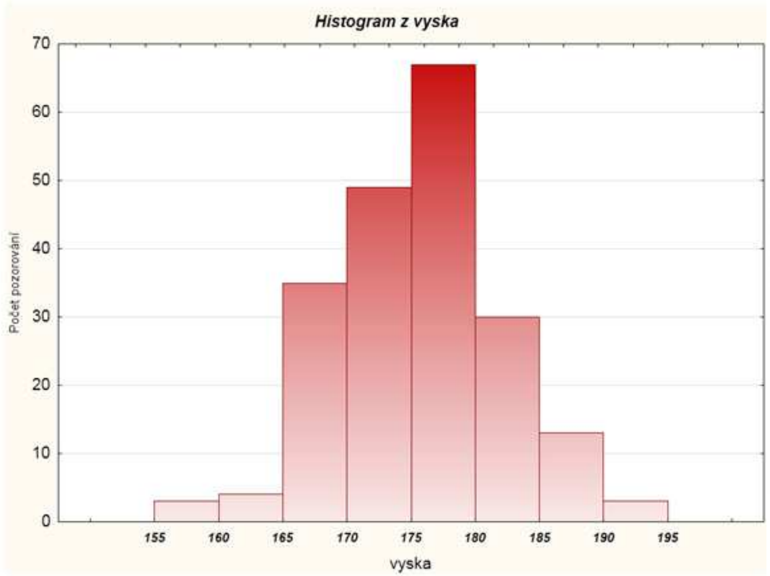
Vzorec pro hustotu standardizovaného normálního rozdělení

Tabelovaná podoba

Vizuální ověření normality

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





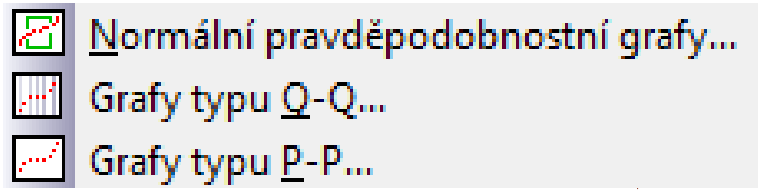
- Vhodnější jsou:
  1. Q-Q graf (kvantil-kvantilový graf)
  2. P-P graf (pravděpodobnostně-pravděpodobnostní graf)
  3. N-P graf (normální-pravděpodobnostní graf)

Opakování:
Co je kvantil?

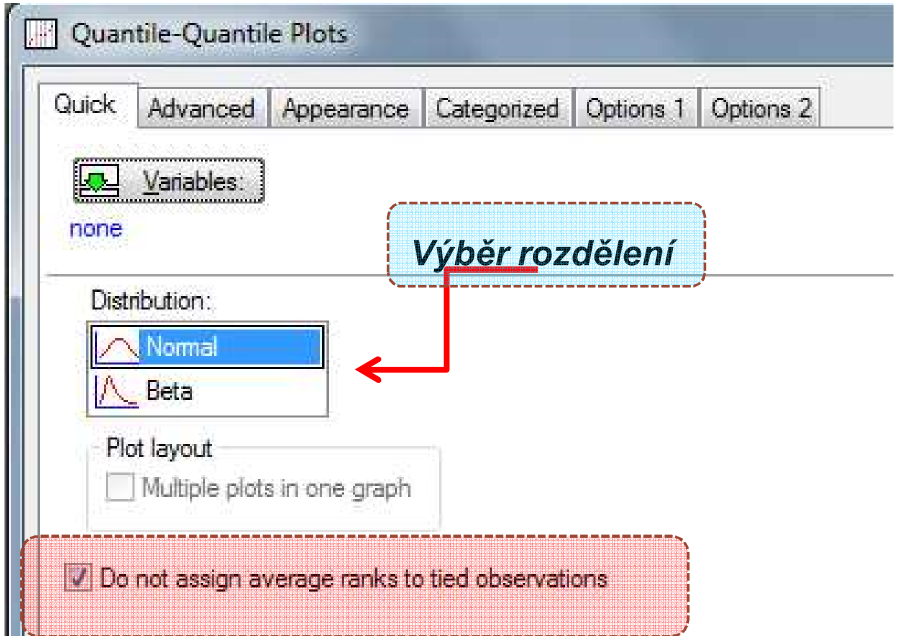
Řešení v softwaru Statistica I

- *V menu Graphs zvolíme 2D Graphs*

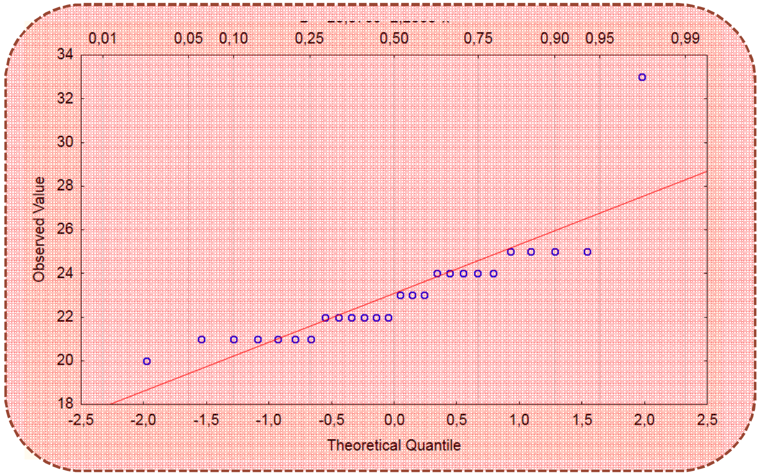
1



2

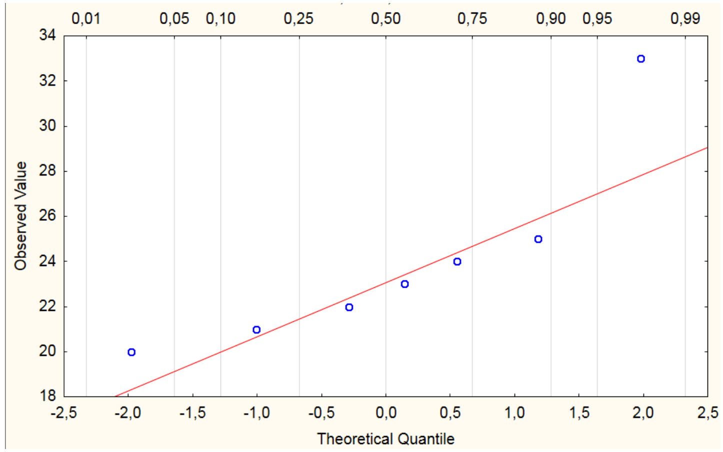


Výběr rozdělení

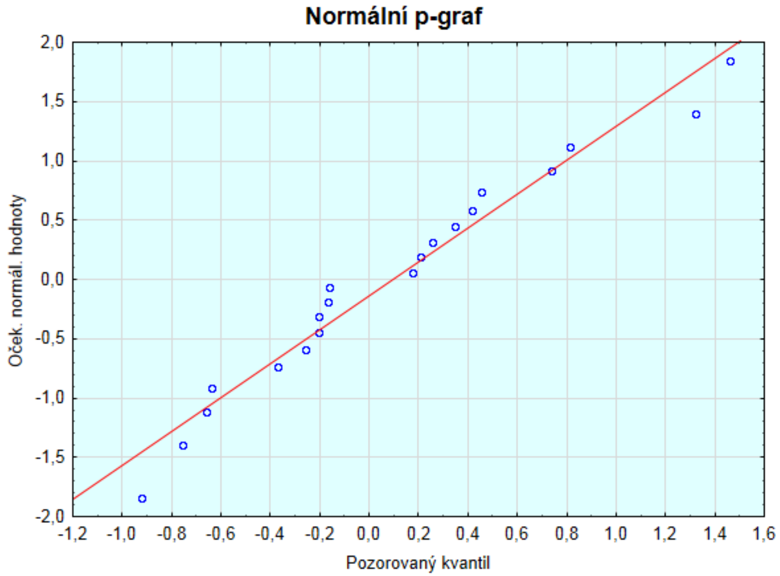


- ***V případě, že máme v datech několik stejných hodnot, je vhodné odškrtnout Neurčovat průměrnou pozici svázaných pozorování***

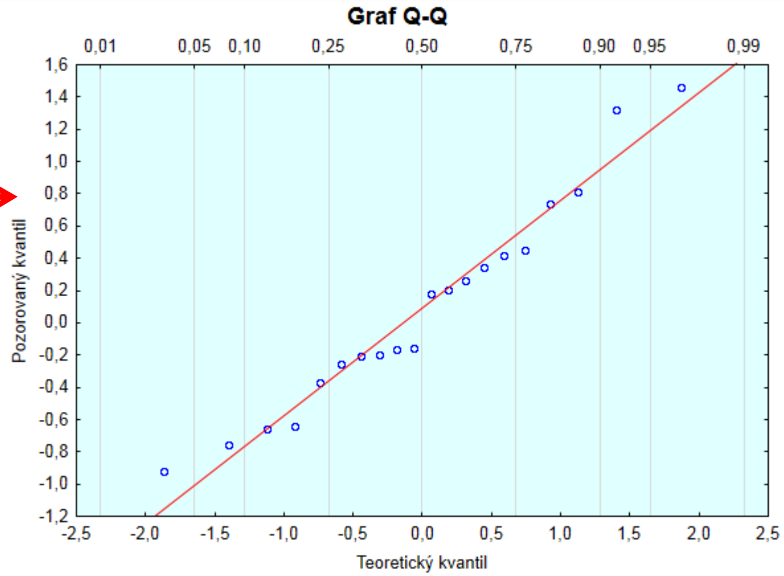
3



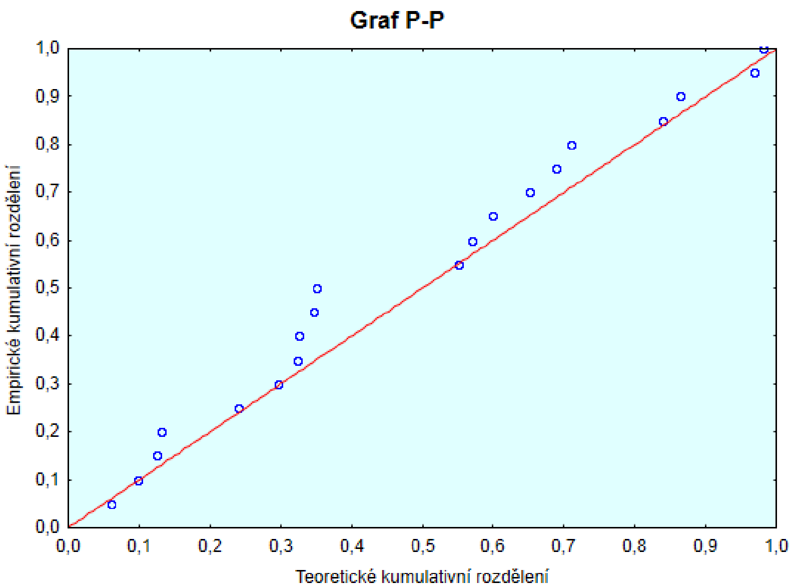
Rozdíl mezi N-P, Q-Q, P-P grafem



???



- *Pouze výměna os*
- *Znázorněn pozorovaný a teoretický kvantil*



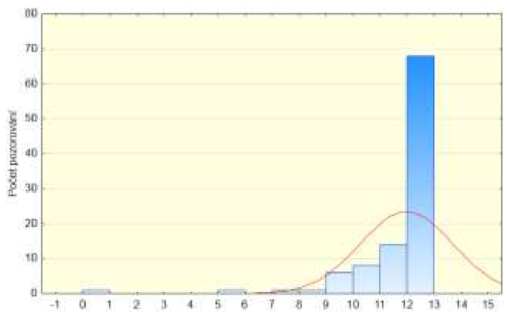
- *Vykresleno kumulativní rozdělení*

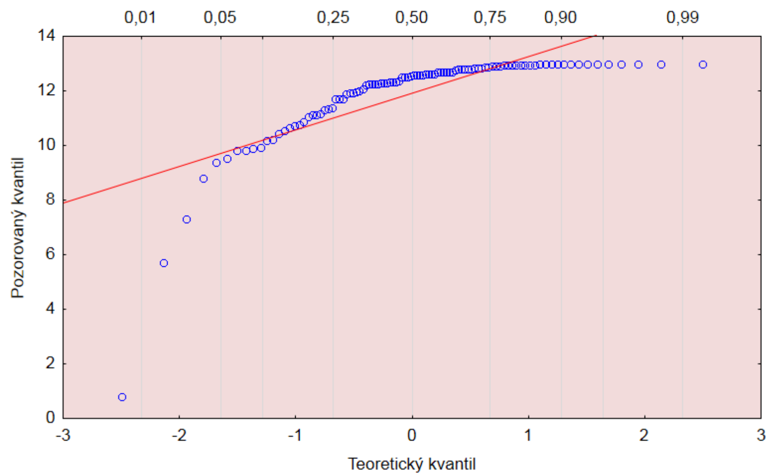
PAMATUJ:

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

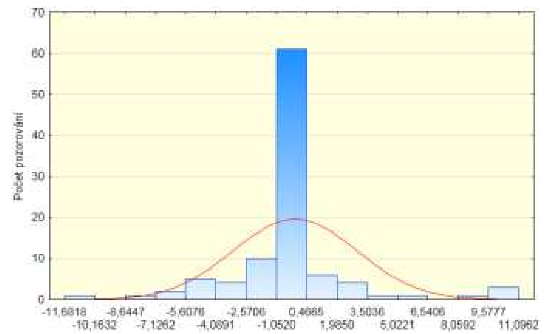
Vizualizace I.

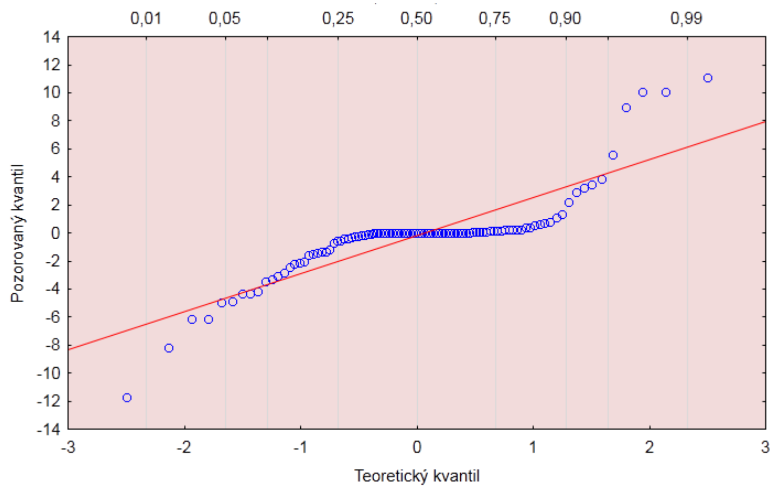
Zešikmená (nesymetrická) data





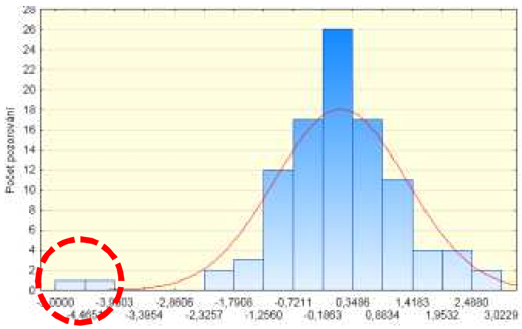
Špičatost

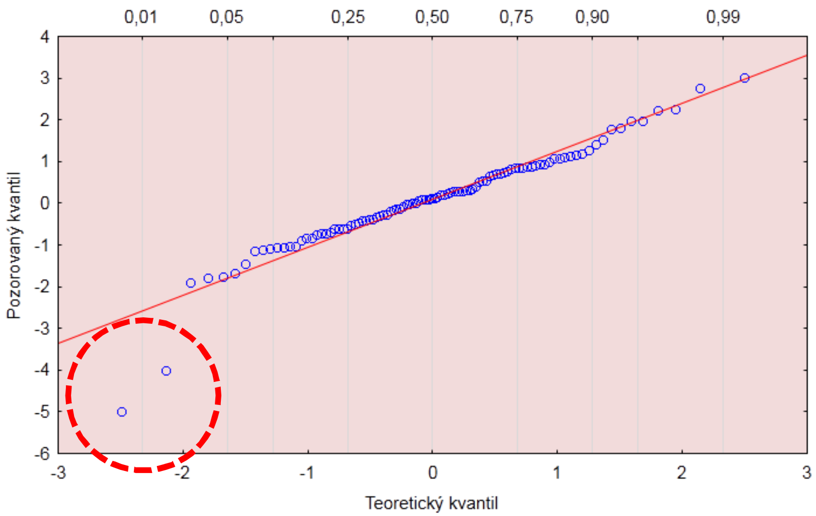




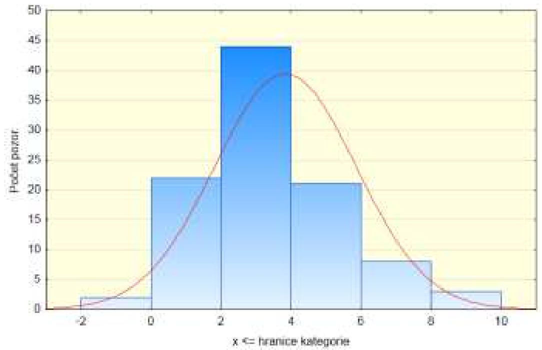
Vizualizace II.

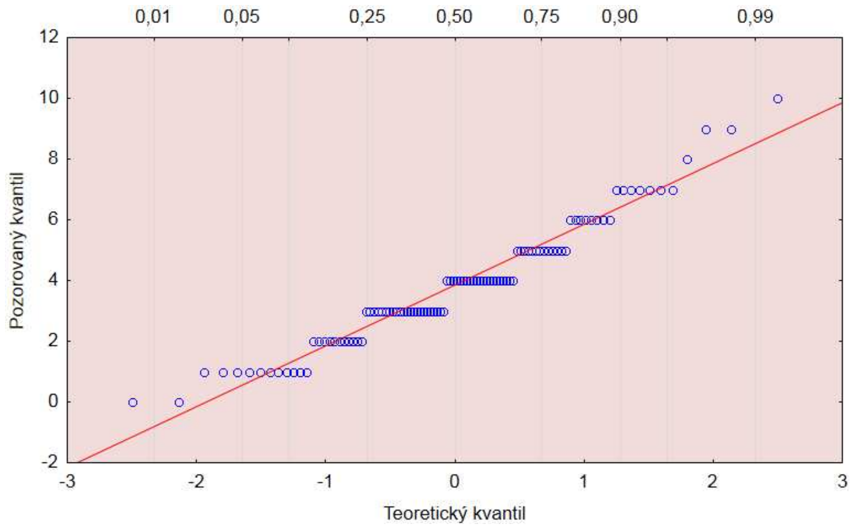
Odlehlé hodnoty





Diskrétní rozdělení





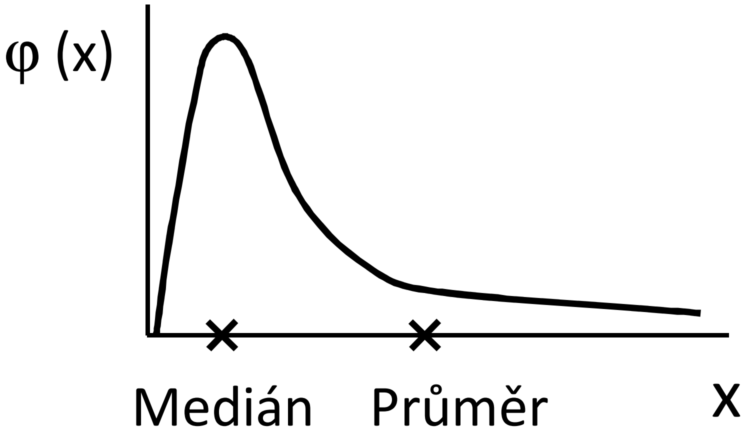
Stručný přehled modelových rozložení I.

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

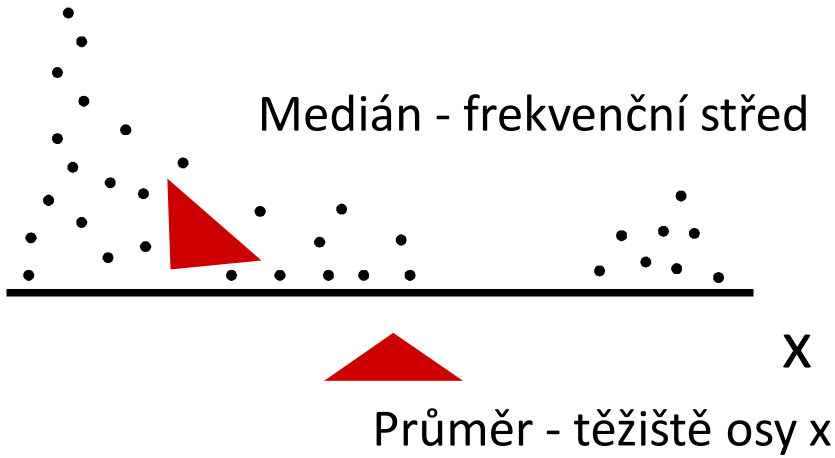
Stručný přehled modelových rozložení II.

Rozložení	Parametry	Stručný popis
Beta	Parametry distribuční funkce: α - parametr tvaru β - 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.
Studentovo	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.
Pearsonovo	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.
Fisher-Snedecorovo	Dvojí stupně volnosti - uvažuje velikost dvou vzorků	Používá se k testování hodnot průměrů - F test pro porovnání dvou výběrových rozptylů; F test, ANOVA atd.

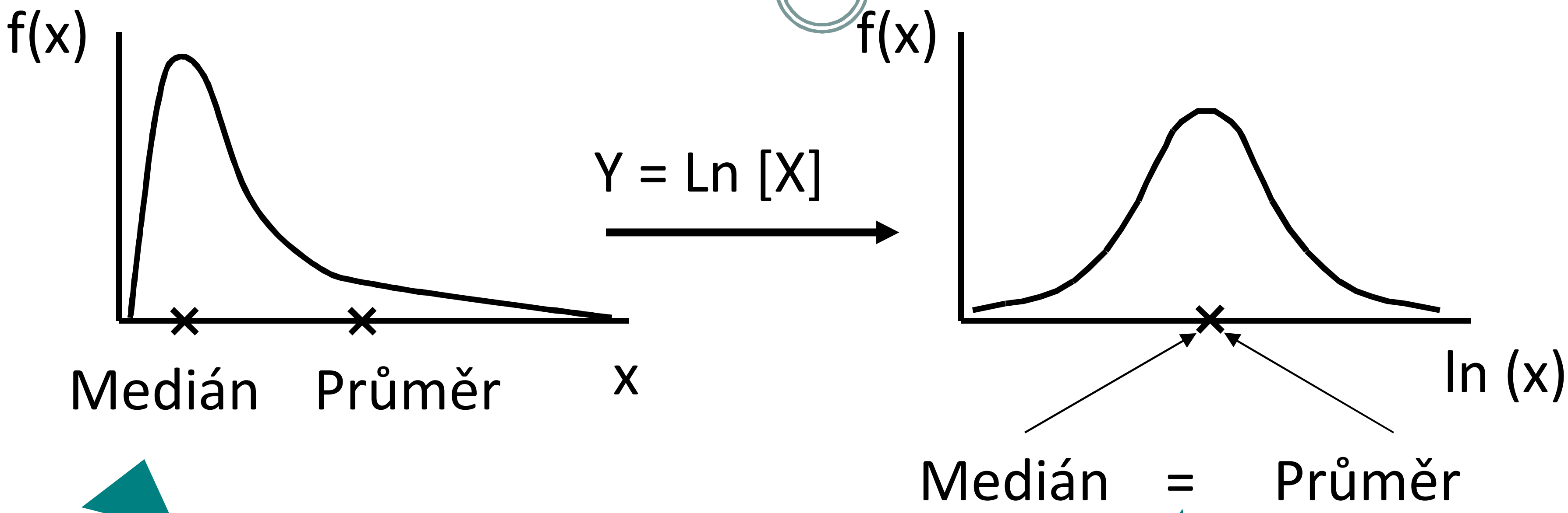
Log-normální rozložení jako častý model reálných znaků



U asymetrických rozložení je medián velmi vhodným alternativním ukazatelem středu



Log-normální rozložení lze jednoduše transformovat



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

$\overline{Y}$ ± Standardní chyba

EXP (Y) = Geometrický průměr X

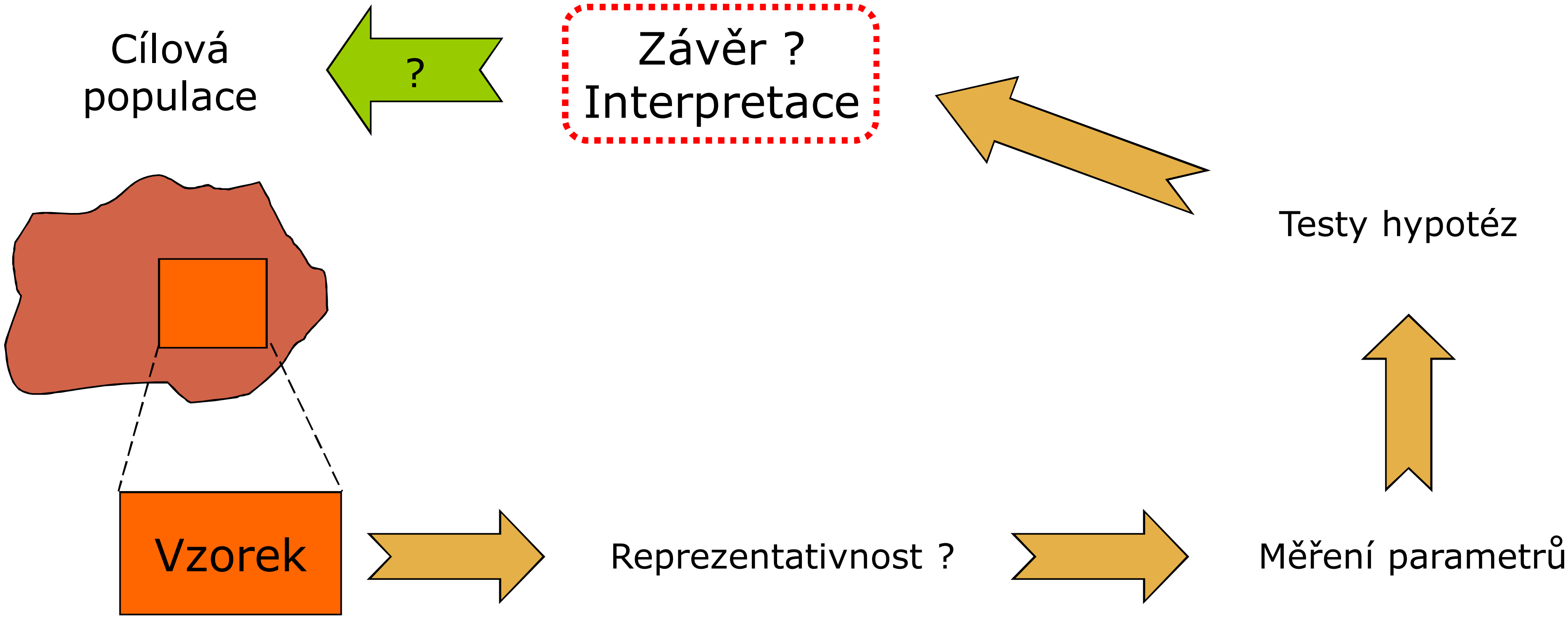
Základy testování hypotéz

Princip statistického testování hypotéz
Pojmy statistických testů
Normalita dat a její význam pro testování

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek

Princip testování hypotéz

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

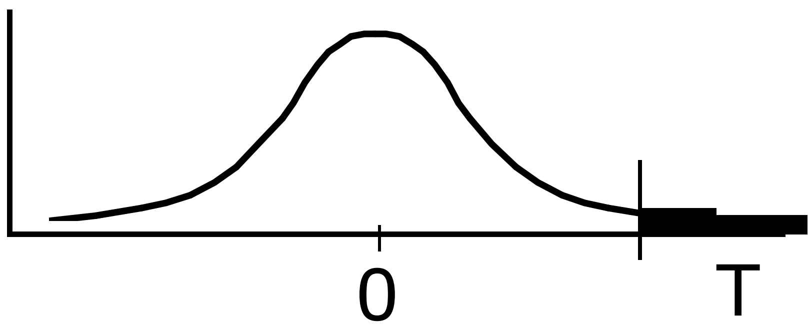


Statistické testování – základní pojmy

- Nulová hypotéza H_O — H_O: sledovaný efekt je nulový
- Alternativní hypotéza H_A — H_A: sledovaný efekt je různý mezi skupinami
- Testová statistika

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

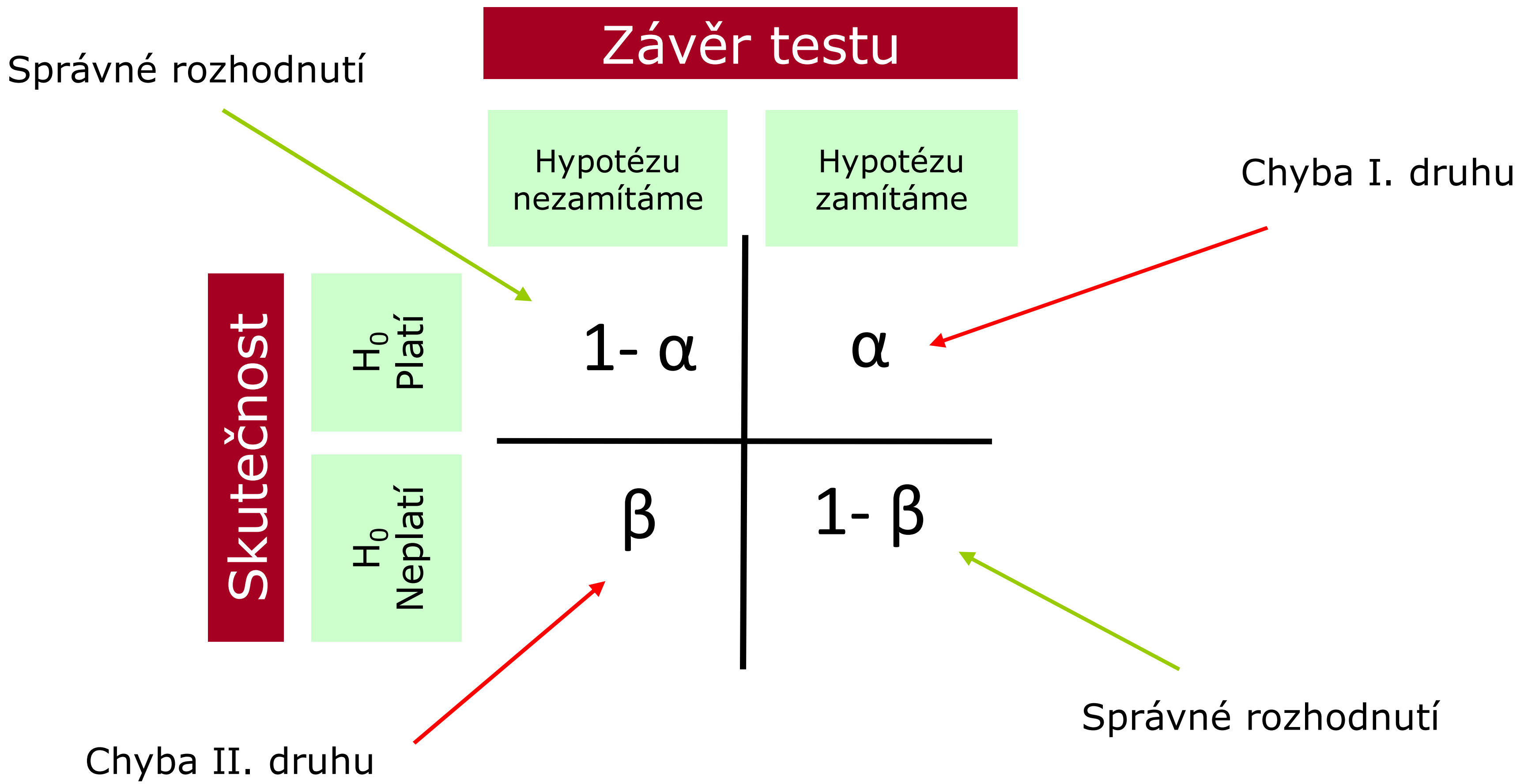
- Kritický obor testové statistiky



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

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.**



Význam chyb při testování hypotéz

Pravděpodobnost chyby 1. druhu

α → Pravděpodobnost nesprávného zamítnutí nulové hypotézy

Pravděpodobnost chyby 2. druhu

β → 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 α můžeme provést třemi různými způsoby:

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

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

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

Jedno-výběrové testy (one-sample)

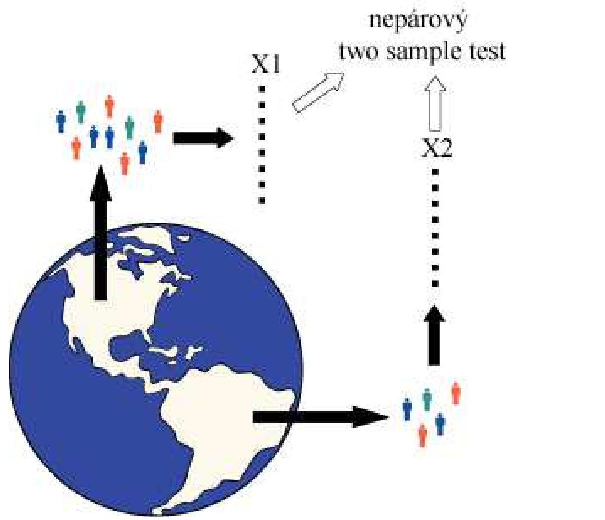
- Srovnávají jeden vzorek (one sample, jednovýběrové testy) s referenční hodnotou (popřípadě se statistickým parametrem cílové populace)
- V testu je tedy srovnáváno rozložení hodnot (vzorek) s jediným číslem (referenční hodnota, hodnota cílové populace)
- Otázka položená v testu může být vztažena k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

Dvou-výběrové 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

Nepárový vs. párový design

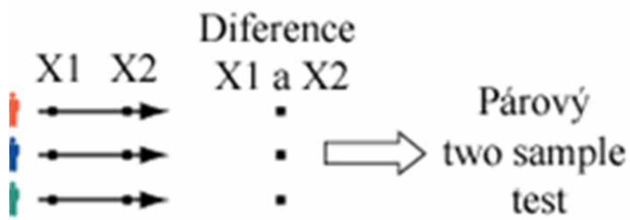
Nepárový design



- 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



Statistické testy a normalita dat

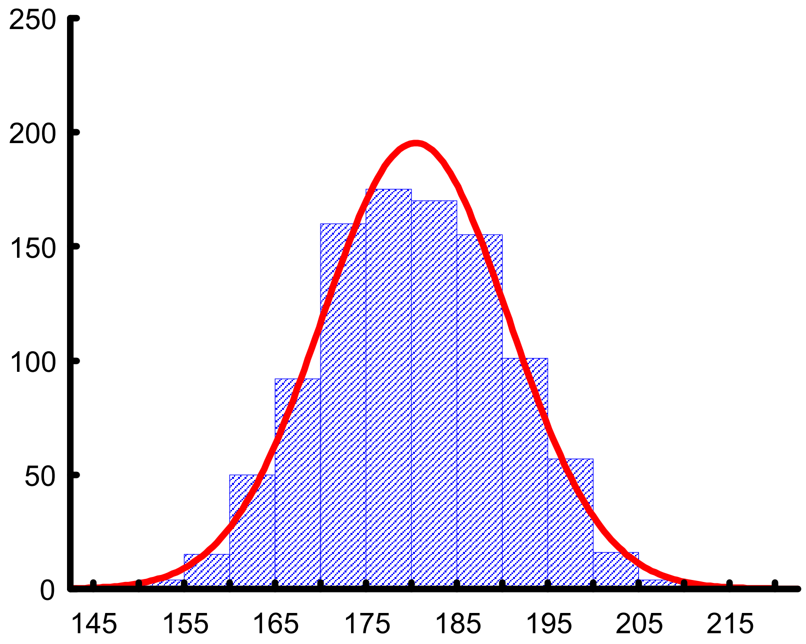
- **Normalita dat je jedním z předpokladů tzv. parametrických testů (testů založených na předpokladu nějakého rozložení) – např.** *t*-testy
- **Pokud data nejsou normální, neodpovídají ani modelovému rozložení, které je použito pro výpočet (*t*-rozložení) a test tak může lhát**

- **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 t-test	Mann Whitney test
2 skupiny dat párově:	Párový t-test	Wilcoxon test, znaménkový test
Více skupin nepárově:	ANOVA	Kruskal- Wallis 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.



•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 $\chi2$ testu dobré shody. Test dává dobré výsledky, ale je náročný na n, tedy množství dat, aby bylo možné vytvořit dostatečný počet tříd hodnot.

•Kolgomorov Smirnov 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.

•Shapiro-Wilk`s test

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