6. Modelová rozdělení pravděpodobnosti, popisné statistiky

Rozdělení pravděpodobnosti

Normální rozdělení jako statistický model

Přehled a aplikace modelových rozdělení

Popisné statistiky

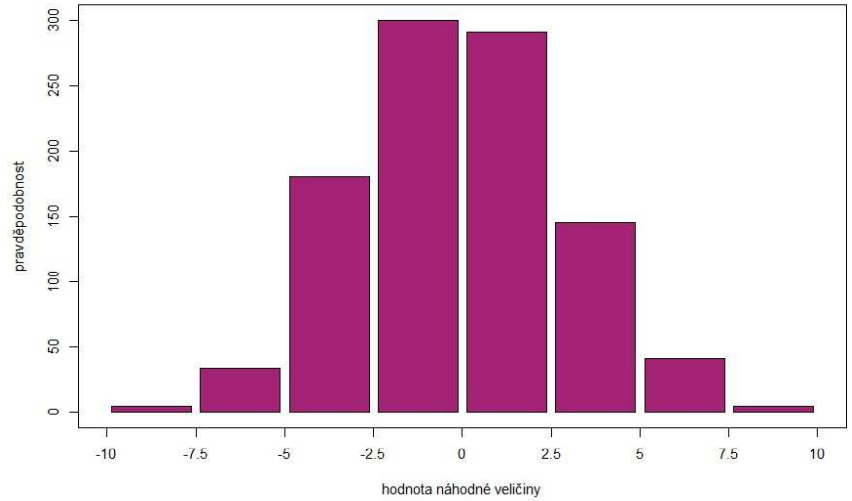
Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek

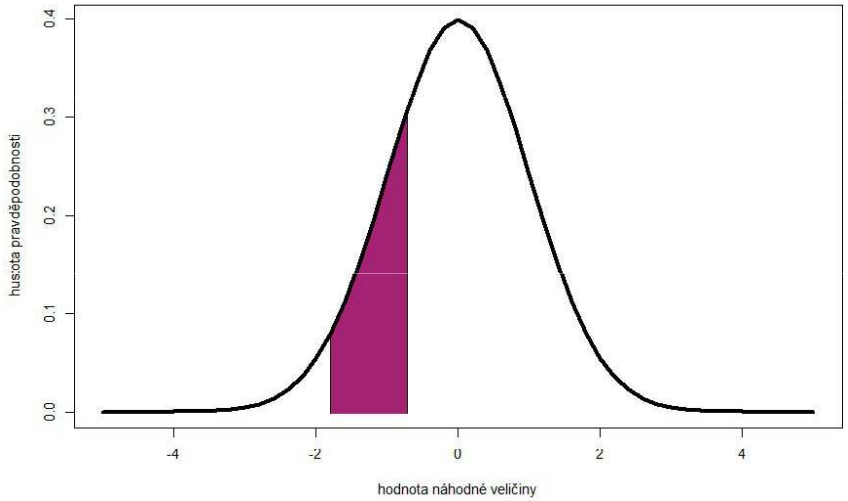
Anotace

- Klasickým postupem statistické analýzy je na základě vzorku cílové populace identifikovat typ a charakteristiky modelového rozdělení dat, využít jeho matematického modelu k popisu reality a získané výsledky zobecnit na hodnocenou cílovou populaci.
- Využití tohoto přístupu je možné pouze v případě shody reálných dat s modelovým rozdělením, v opačném případě hrozí získání zavádějících výsledků (neparametrické statistiky).
- Nejklasičtějším modelovým rozdělením, od něhož je odvozena celá řada statistických analýz je tzv. normální rozdělení, známé též jako Gaussova křivka.

Rozdělení (rozložení, distribuce) pravděpodobnosti (dat)

- Funkce přiřazující intervalu hodnot náhodné veličiny pravděpodobnost (obecně), resp. přiřazující hodnotě náhodné veličiny určitou hustotu pravděpodobnosti (derivace pravděpodobnosti podle hodnoty).
- V případě diskrétní náhodné veličiny lze ztotožnit intervaly s konkrétními hodnotami a tvrdit, že rozdělení pravděpodobnosti přiřazuje hodnotám pravděpodobnost.

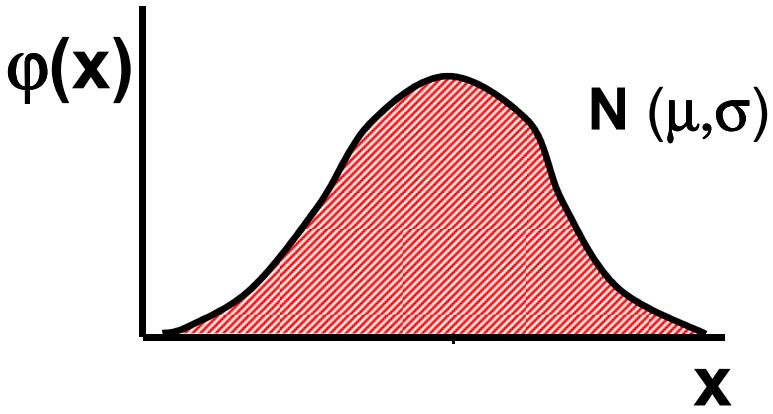




Rozdělení (rozdělení, distribuce) pravděpodobnosti (dat)

- Rozdělení pravděpodobnosti pro spojité a diskrétní náhodné veličiny se liší (páry podobných rozdělení).
- Každá náhodná veličina má určité rozdělení, které může a nemusí být známé.
- Rozdělení je určeno charakteristickými parametry. Jejich typ a počet se liší na základě komplexity rozdělení:
  - průměr,
  - rozptyl,
  - špičatost,
  - šikmost aj.
- Při analýze určujeme výběrové parametry, které nejsou totožné s reálnými parametry rozdělení.

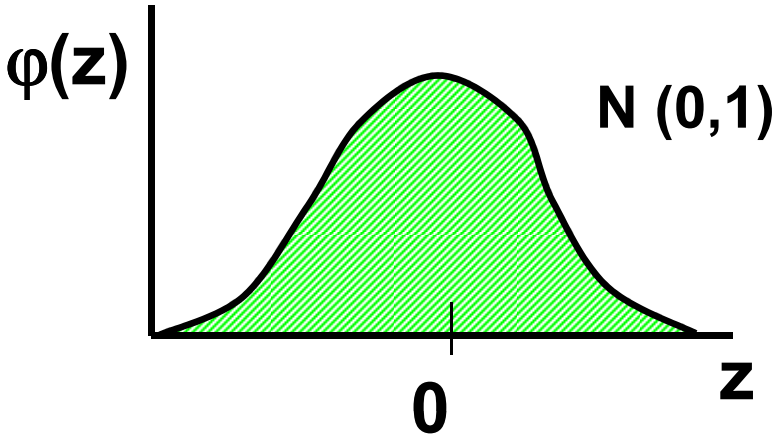
rozdělení hodnot jako model: Normální rozdělení



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

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

Standardizovaná forma



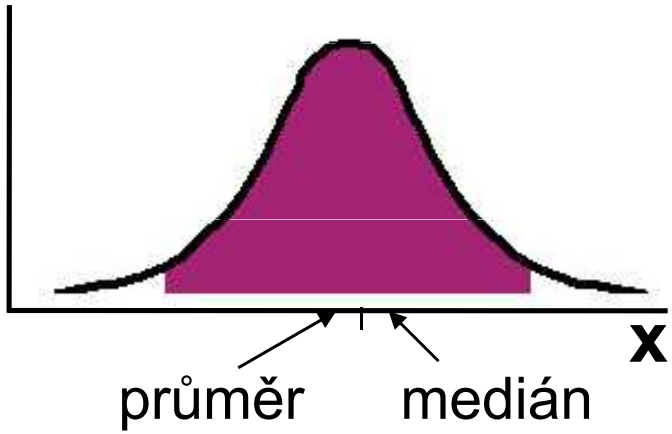
$$\varphi(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

Parametry charakterizující normální rozdělení a jejich význam

$E(x) \sim \bar{x} \sim \mu$

$D(x) \sim s^2 \sim \sigma^2$



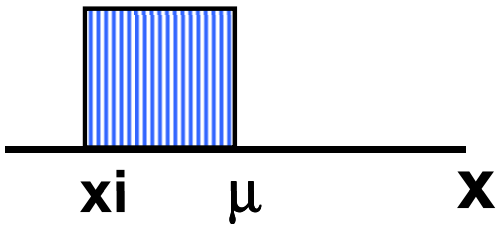
a) $\mu \sim \bar{x}$

průměr - ukazatel středu

b) $\sigma^2 \sim s^2$

rozptyl

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$



c) $\sigma \sim s$

směrodatná odchylka

$$s = \sqrt{s^2}$$

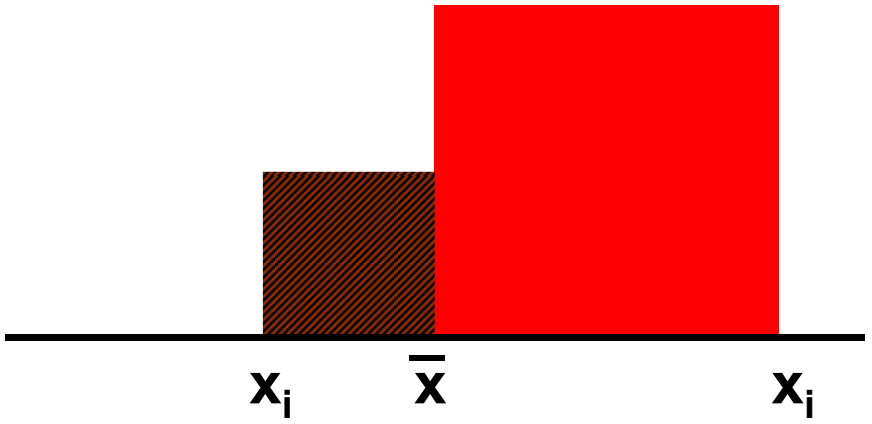
Pravidlo ± 3s

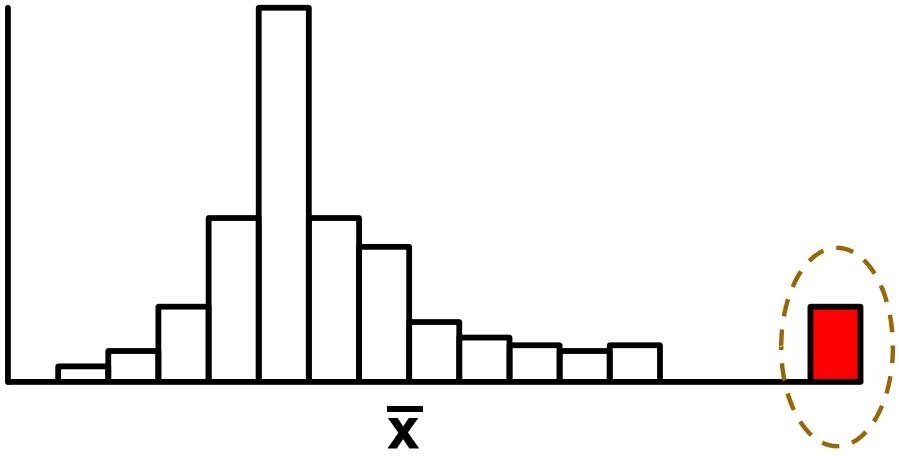
d) **koeficient variance**

$$c = s / \bar{x}$$

Rozptyl není univerzálním ukazatelem variability

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$





$\Rightarrow$ neúměrně zvýší s^2

- Rozptyl a směrodatná odchylka jsou citlivé na odlehlé hodnoty (jiné než normální rozdělení).

Normální rozdělení jako model

I. Použitelnost modelu

A) X: spojitý znak - hmotnost jedince (myši)

1,2; 1,4; 1,6; 1,8; 2,0; 2,4; 3.8

n = 7 opakování

medián = 1,8

průměr = $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}\sum_{i=1}^{7} x_i = \frac{1}{7}(1{,}2+1{,}4+1{,}6+1{,}8+2{,}0+2{,}4+3{,}8) = \frac{1}{7}14{,}2 = 2{,}03$

rozptyl (s^2) = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{7}(x_i-2{,}03)^2}{6} = 0{,}766$

sm. odchylka (s) = $\sqrt{s^2} = \sqrt{0{,}766} = 0{,}875$

? **Je předpoklad normálního rozdělení oprávněný ?**
Jaký předpokládáte možný rozsah hodnot tohoto znaku ? **?**

Normální rozdělení jako model

I. Použitelnost modelu

B) X: spojitý znak - hmotnost jedince (myši)

1,2; 1,4; 1,6; 1,8; 2,0; 2,2; 2,4; 3,8; 8,9

n = 9 opakování

medián = 2

průměr = $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{1}{9}(1{,}2+1{,}4+1{,}6+1{,}8+2{,}0+2{,}2+2{,}4+3{,}8+8{,}9) = \frac{1}{9}25{,}3 = 2{,}81$

rozptyl (s^2) = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{9}(x_i-2{,}81)^2}{8} = 5{,}79$

sm. odchylka (s) = $\sqrt{s^2} = \sqrt{5{,}79} = 2{,}269$

Jak hodnotíte model u těchto dat ?

Stochastické rozdělení jako model

1. **Předpoklad: Znak x je rozložen podle daného modelu** ✓

2. **Znak x je naměřen o n hodnotách s modelovými parametry: $\bar{x}$ a s** → **Platnost modelu ?**

3. **Znak x je převeden na formu odpovídající tabulkovému standardu:** →

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

4. **Využije se tabelované (modelové) distribuční funkce pro testy o rozdělení hodnot x**

Normální rozdělení jako model - příklad

Tabulky distribuční funkce

- **<u>Data z průzkumu jsou publikována jako:</u>**

Kosti prehistorického zvířete:
n = 2000
průměrná délka = 60 cm
sm. odchylka (s) = 10 cm

✓ **Předpokládáme, že je oprávněný model normálního rozdělení**

? Jaká je pravděpodobnost, že by velikost dané kosti překročila velikost 66 cm: P (x > 66) ? $Z = \frac{x - \mu}{\sigma}$

$P(x > 66) = 1 - P(x \leq 66)$ a platí, že $P(X \leq x) = F(X)$

tedy $P(x > 66) = 1 - P(x \leq 66) = 1 - P(\frac{x - m}{s} \leq \frac{66 - 60}{10}) = 1 - F(0,6) = 0,27425$

? Kolik kostí mělo zřejmě délku větší než 66 cm ? $P(x > 66) * n = 0,27425 * 2000 = 548$

? Jaký podíl kostí ležel svou délkou v rozsahu x od 60 cm do 66 cm ?

$P(60 < x < 66) = P\left(\frac{60 - 60}{10} < Z < \frac{66 - 60}{10}\right) = F(0,6) - F(0) = 0,22575$ ➡ 22,6% kostí leží v rozsahu 60-66cm

Stručný přehled modelových rozdělení I.

rozdělení	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 rozdělení.
Weibullovo	**α - parametr tvaru** **β - parametr rozsahu hodnot**	Změnou parametru a lze modelovat distribuci doby přežití, např. stresovaného organismu. rozdělení využívané i jako model k odhadu 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 rozdělení.
Triangulární	**$f(x) = [b - ABS(x - a)] / b^2$** **$a - b < x < a + b$**	Pravděpodobnostní funkce pro typ rozdělení, kdy jsou střední hodnoty výrazně pravděpodobnější než hodnoty okrajové.
Gama **(Exponenciální)**	**Parametry distribuční funkce:** **α - parametr tvaru** **β - parametr rozsahu hodnot**	Umožňuje flexibilně modelování distribučních funkcí nejrůznějších tvarů. Např. χ^2 rozdělení je rozdělení typu Gama. Gama rozdělení s a = 1 je známo jako exponenciální rozdělení.

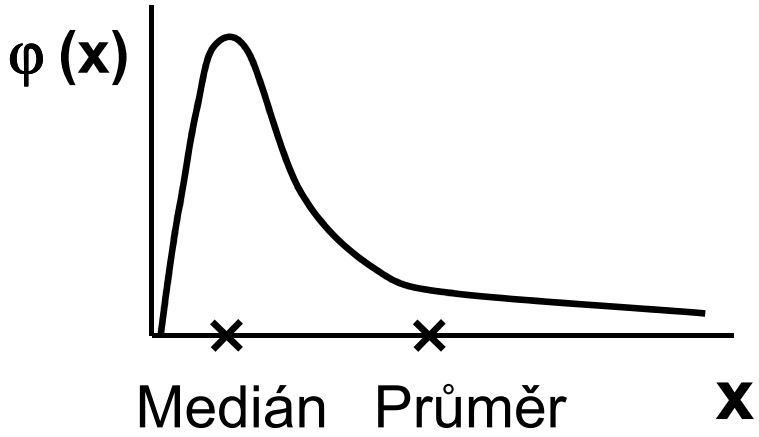
Stručný přehled modelových rozdělení II.

rozdělení	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 rozdělení pro menší vzorky čísel. Pro větší soubory (n > 100) se limitně blíží k normálnímu rozdělení.
Pearsonovo	**Stupně volnosti - uvažuje velikost vzorku**	Slouží především k porovnání četností jevů ve dvou a více kategoriích. Používá se k modelování rozdělení 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.

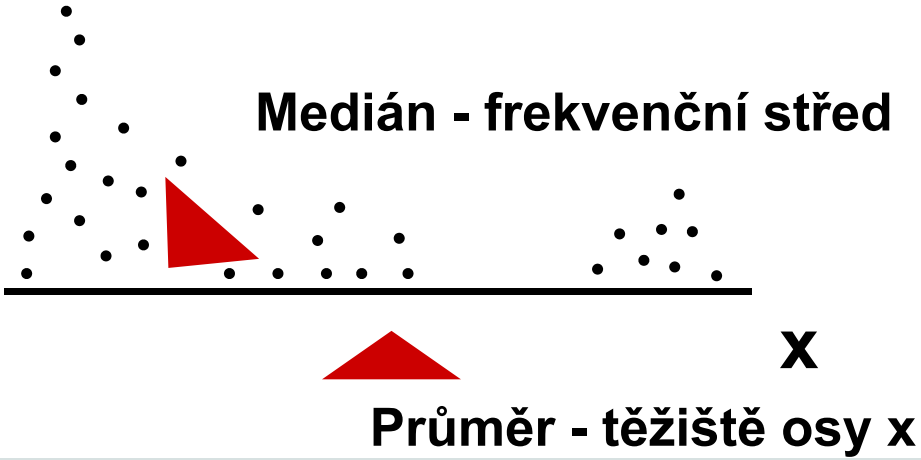
Stručný přehled modelových rozdělení II.

rozdělení	Parametry	Stručný popis
Binomické	**Průměr (μ)** **Rozptyl (σ^2)**	Diskrétní obdoba normálního rozdělení - symetrická funkce popisující intervalovou četnost výskytu jevu v nezávislých pokusech; nejpravděpodobnější jsou průměrné hodnoty znaku.
Poissonovo	**Lambda**	Rozdělení řídkých (málo pravděpodobných) jevů. Pro n > 30 se používá k aproximaci binomického rozdělení (jednoduchá matematický forma funkce).
Geometrické	**Lambda**	Diskrétní podoba exponenciálního rozdělení. Udává počet opakování experimentu do prvního úspěchu při konstantní pravděpodobnosti úspěchu.
Bernoulliho	**Pravděpodobnost jevu p**	Binární rozdělení pravděpodobnosti, kdy jev nastane s pravděpodobností p a nenastane s pravděpodobností 1-p.

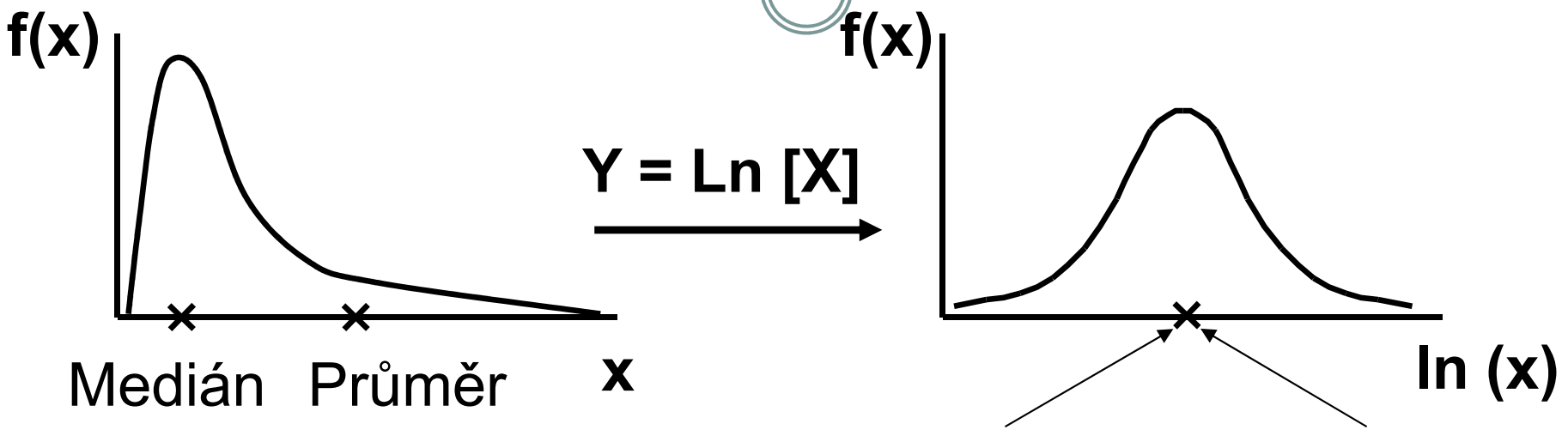
Log-normální rozdělení jako častý model reálných znaků



U asymetrických rozdělení je medián velmi vhodným alternativním ukazatelem středu



Log-normální rozdělení lze jednoduše transformovat



EXP (Y) = Geometrický průměr X

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

$\overline{Y}$ ± **Standardní chyba**

Transformace dat - legitimní úprava rozdělení

✓ **Základní typy transformací vedou k normalitě rozdělení nebo k homogenitě rozptylu**

Logaritmická transformace

Logaritmická transformace je velmi vhodná pro data s odlehlými hodnotami na horní hranici rozsahu. Při porovnání průměrů u více souborů dat je pro tuto transformaci indikující situace, kdy se s rostoucím průměrem mění proporcionálně i směrodatná odchylka, a tedy jednotlivé proměnné mají stejný koeficient variance, ačkoli mají různý průměr.

Za takovéto situace přináší logaritmická transformace nejen zeslabení asymetrie původního rozdělení, ale také vyšší homogenitu rozptylu proměnných. Pro transformaci se nejčastěji používá přirozený logaritmus a pokud jsou v původním souboru dat nulové hodnoty, je vhodné použít operaci **Y = ln (X+1)**.

Je-li průměr logaritmovaných dat (tedy průměrný logaritmus) zpětně transformován do původních hodnot, výsledkem není aritmetický, ale geometrický průměr původních dat.

Transformace dat - legitimní úprava rozdělení

✓ **Základní typy transformací vedou k normalitě rozdělení nebo k homogenitě rozptylu**

Odmocninová transformace

Transformace je vhodná pro proměnné mající Poissonovo rozdělení, tedy proměnné vyjadřující celkový počet nastání určitého jevu (spíše vzácného) v **n** nezávisle opakovaných pokusech. Obecněji lze tento typ transformace doporučit v případě normalizace dat typu počtu jedinců (buněk, apod.). Jde o transformaci:

$$Y=\sqrt{x} \quad \textbf{nebo} \quad Y=\sqrt{x+1} \quad \textbf{nebo} \quad Y=\sqrt{x}+\sqrt{x+1}$$

Transformace s přičtenou hodnotou 1 jsou efektivní, pokud **X** nabývá velmi malých nebo nulových hodnot. Situace indikující vhodnost odmocninové transformace je také proporcionalita výběrového rozptylu a průměru, tedy obecně jestliže $\mathbf{s^2_x = k}$ (výběrový průměr).

Transformace dat - legitimní úprava rozdělení

Arcsin transformace

Tzv. **úhlová transformace** - velmi vhodná pro data typu podílů výskytu určitého jevu (znaku) mezi **n** hodnocenými jedinci - tedy pro data mající binomické rozdělení. Pokud se určitý znak vyskytuje r-krát mezi **n** možnostmi (jedinci, opakováními), pak lze vyjádřit relativní četnost jeho výskytu jako **p = r/n** s variabilitou **p.(1-p)/n**. Arcsin transformace odstraní ze souborů dat podíly blízké 0 nebo 1, a tak efektivně sníží variabilitu odhadů středu. Transformace však není schopná odstranit variabilitu vyvolanou rozdílným počtem opakování v jednotlivých variantách - v takovém případě lze doporučit provedení vážených transformací dat. Velmi častou formou této transformace je:

$$Y = \arcsin \sqrt{p}$$

- tedy transformace podílů do hodnot, jejichž sinus je roven druhé odmocnině původních hodnot. Pokud celkový počet jedinců (opakování), mezi kterými je výskyt znaku monitorován, je n < 50, pak lze doporučit velmi efektivní empirická opatření pro transformaci podílů blízkých 0 nebo 1. Pro tento případ lze nahrazovat nulové podíly hodnotou 1/4n a 100 % podíly hodnotou (n-1/4)/n. Pokud se mezi hodnotami vyskytuje větší množství krajních hodnot (menší než 0,2 a větší než 0,8), lze doporučit transformaci:

$$Y = \frac{1}{2}\left[\arcsin \sqrt{\frac{x}{n+1}} + \arcsin \sqrt{\frac{x+1}{n+1}}\right]$$

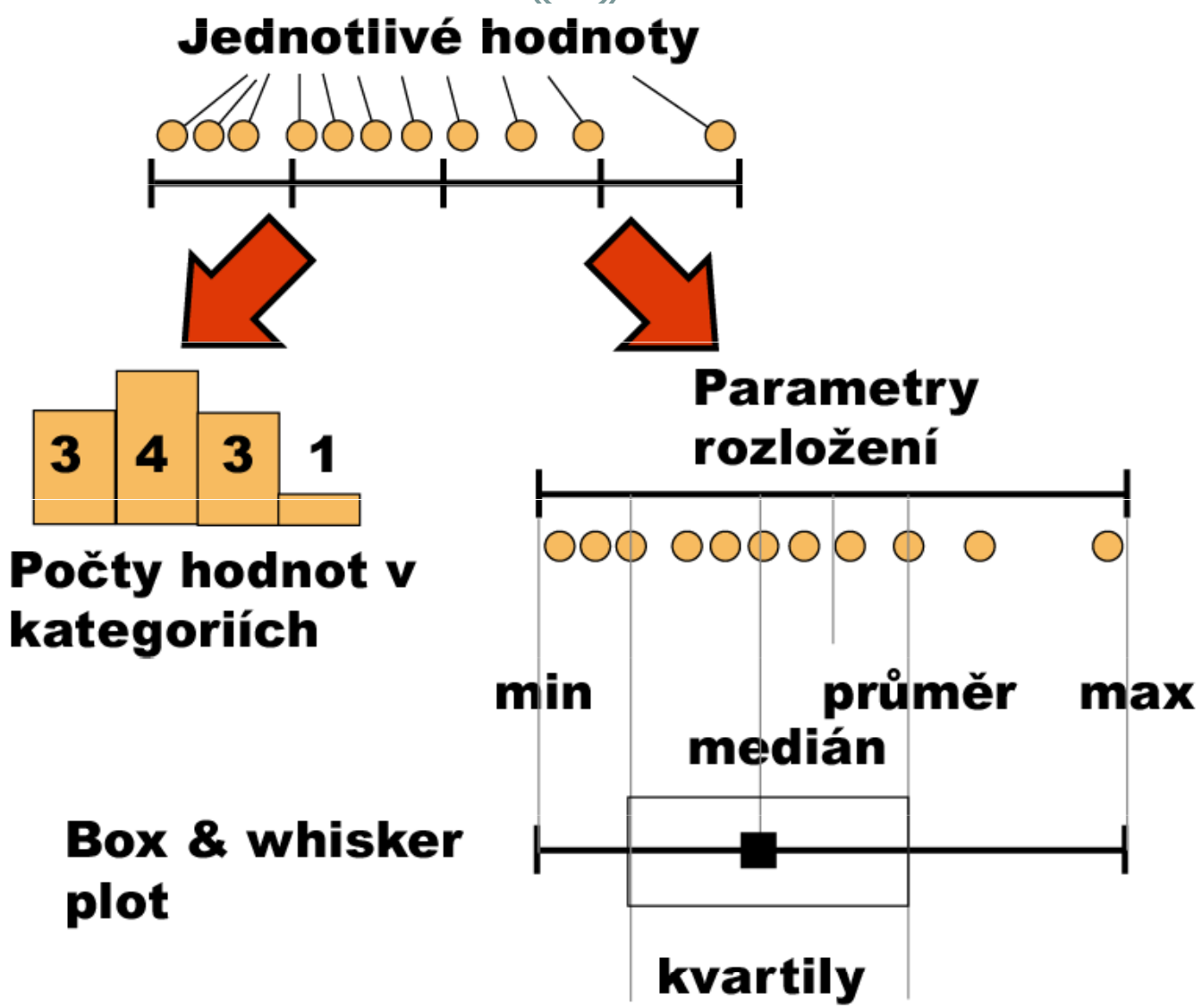
Popisná statistika

- Popisná analýza dat je po vizualizaci dat dalším krokem v procesu statistického hodnocení. Poskytuje představu o rozsazích hodnocených dat a umožňuje vyhodnotit, srovnáním s literárními údaji nebo dosavadní zkušeností, jejich realističnost.
- Již při výběru vhodné popisné statistiky se uplatňuje znalost rozdělení dat. Některé popisné statistiky, odvozené od modelových rozdělení, je možné využít pouze v případě, že data mají dané modelové rozdělení. Typickým příkladem je průměr a směrodatná odchylka, jejichž předpokladem je přítomnost symetrického, resp. normálního rozdělení.

Typy proměnných

- **Kvalitativní/kategorická**
  - binární - ano/ne
  - nominální - A,B,C … několik kategorií
  - ordinální- 1<2<3 …několik kategorií a můžeme se ptát, která je větší
- **Kvantitativní**
  - nespojitá – čísla, která však nemohou nabývat všech hodnot (např. počet porodů)
  - spojitá – teoreticky jsou možné všechny hodnoty (např. krevní tlak)

Řada dat a její vlastnosti

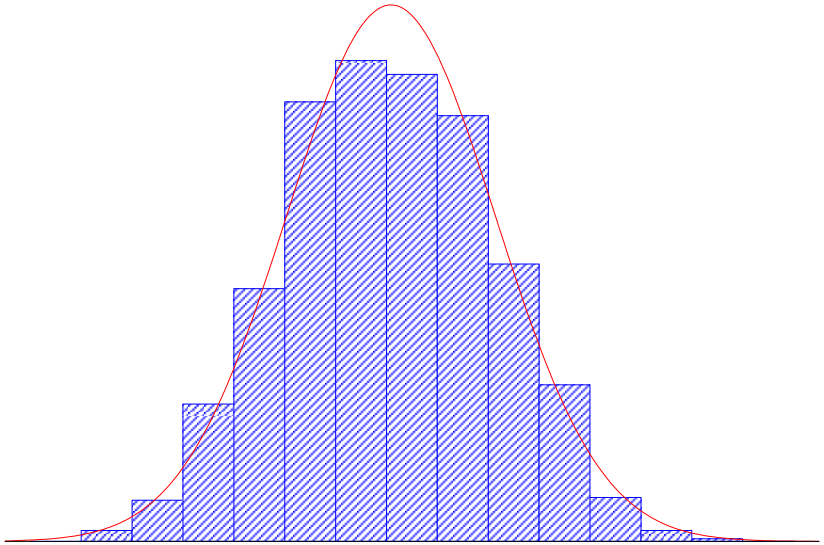


Frekvenční rozdělení

Kategorie	Četnost
B	5
C	8
D	1

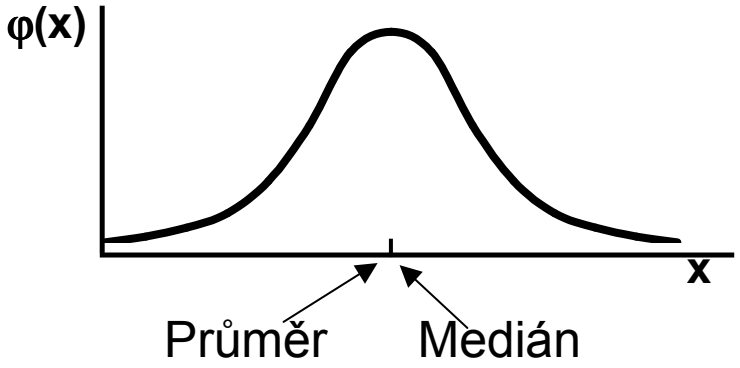
Kvalitativní data
Tabulka s četností jednotlivých kategorií.

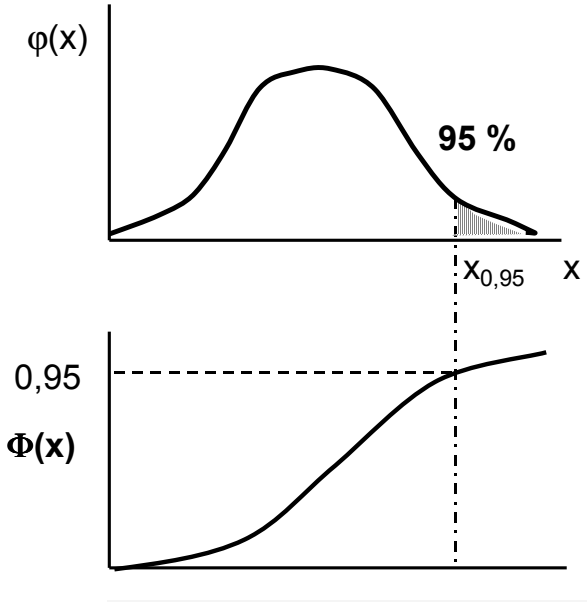
Kvantitativní data
Četnost hodnot rozdělení v jednotlivých intervalech.



Parametry rozdělení

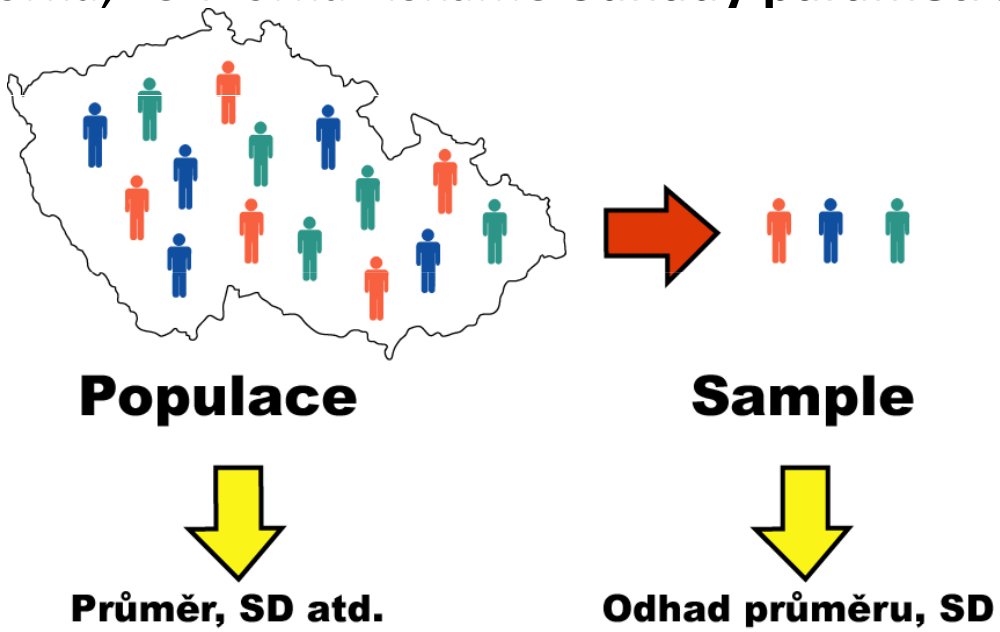
- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozdělení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
  - Středu (medián, průměr, geometrický průměr)
  - Šířky rozdělení (rozsah hodnot, rozptyl, směrodatná odchylka)
  - Tvaru rozdělení (skewness, kurtosis)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





Populace a vzorek

- Populace představuje veškeré možné objekty vzorkování, např. veškeré obyvatelstvo ČR při sledování na úrovni ČR, z populace získáme reálné parametry rozdělení
- Z populace je prováděno vzorkování za účelem získání reprezentativního vzorku (**sample**) populace, toto vzorkování by mělo být náhodné, důležitá je také velikost vzorku, ze vzorku získáme **odhady parametrů rozdělení**



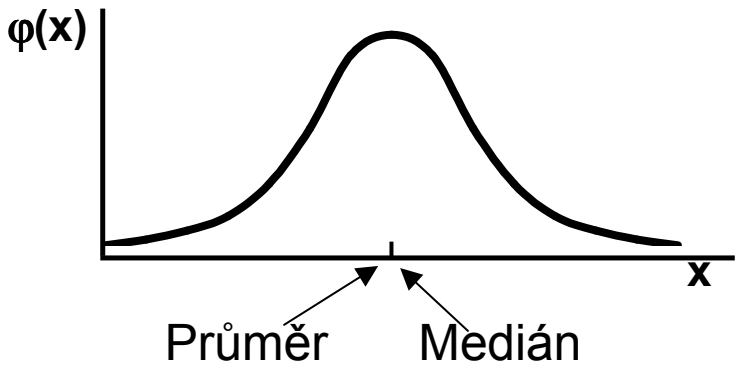
Ukazatele středu rozdělení I

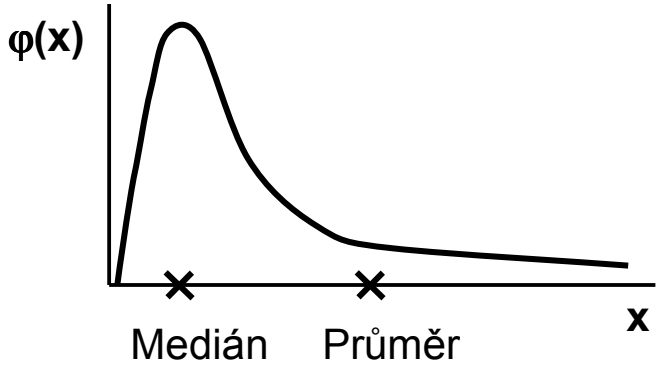
- **Průměr** – vhodný ukazatel středu u normálního/symetrického rozdělení, kde $\mathbf{x_i}$ jsou jednotlivé hodnoty a **n** jejich počet

$$E(x) = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

- **Medián** – jde vlastně o 50% kvantil, tj. polovina hodnot leží nad a polovina pod mediánem

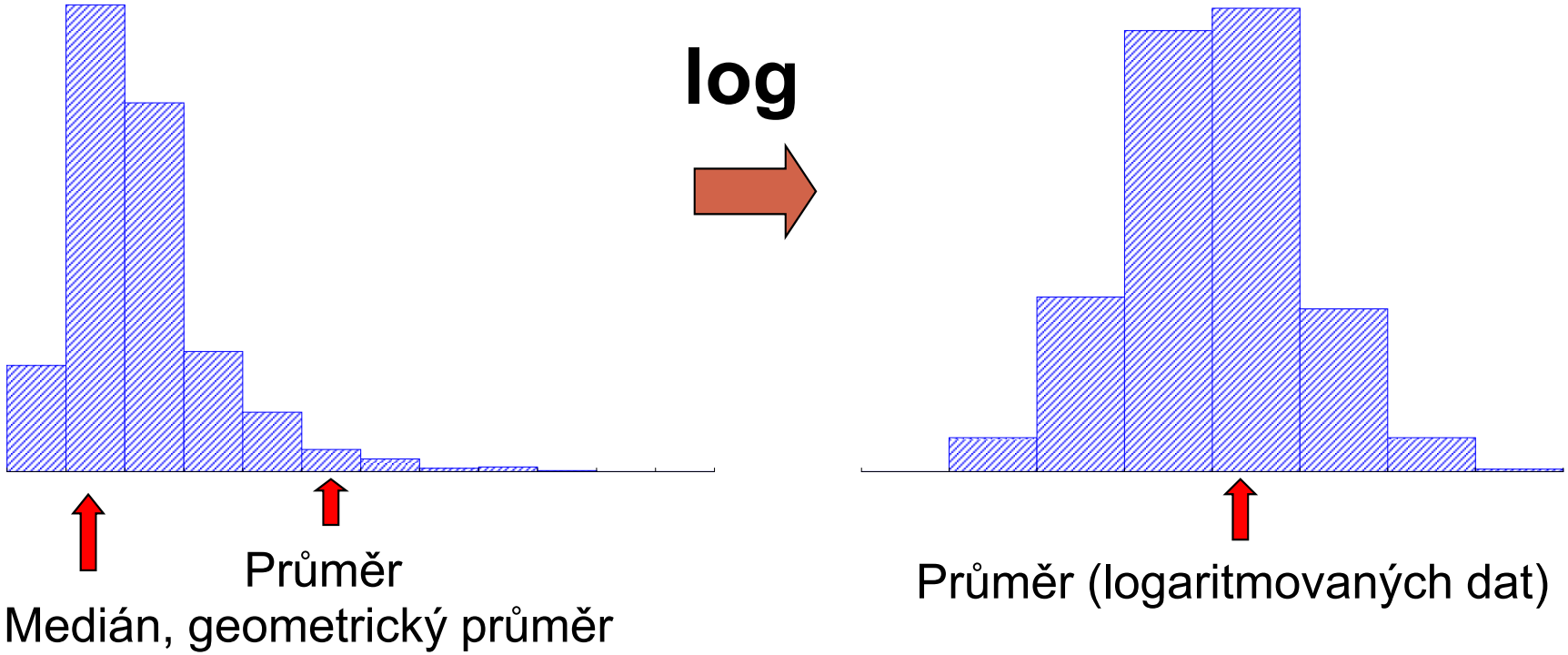
- V případě symetrického rozdělení jsou jejich hodnoty v podstatě shodné





Ukazatele středu rozdělení II.

- Geometrický průměr – antilogaritmus průměru logaritmovaných dat, je vhodný pro doleva asymetrická data (lognormální rozdělení), která jsou v biologii velmi častá, jeho hodnota v podstatě odpovídá mediánu
- Takto asymetrická data je možné převést logaritmickou transformací na normální rozdělení

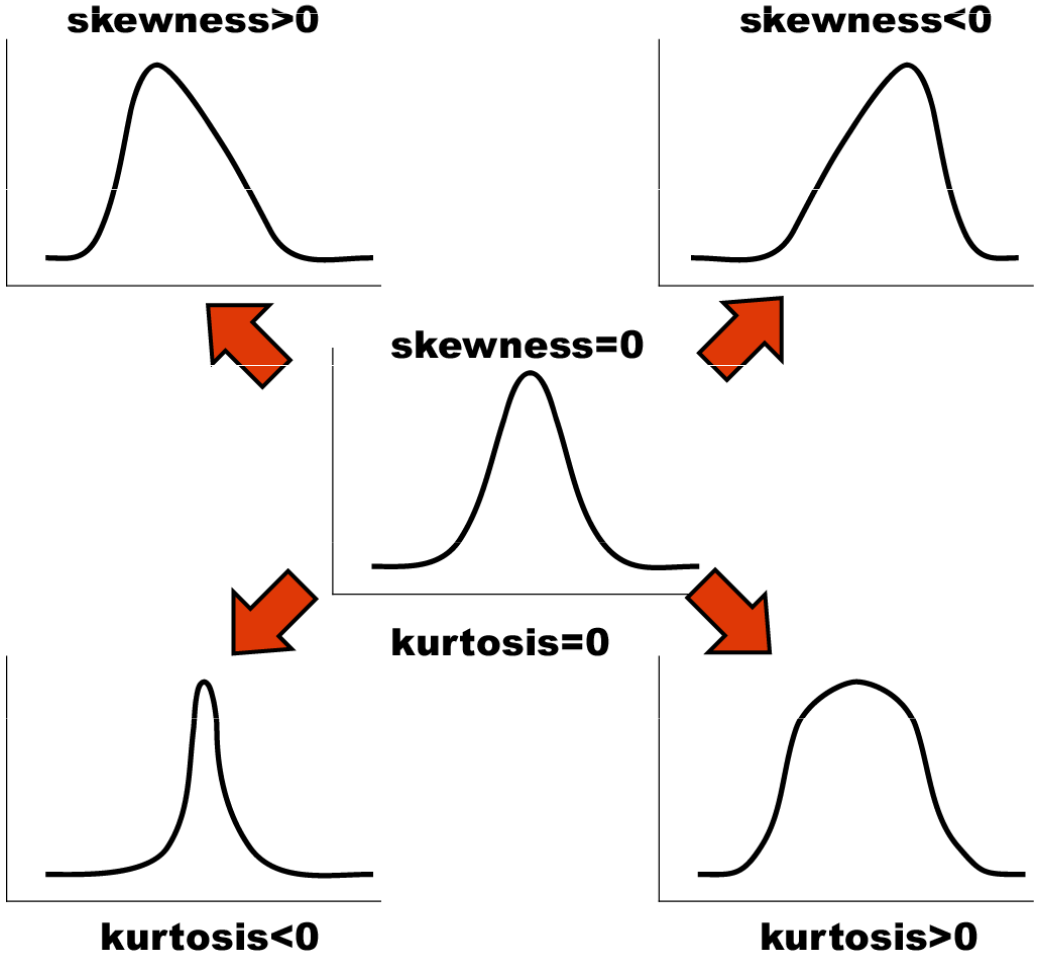


Ukazatele šířky rozdělení

- **Rozptyl** je ukazatelem šířky rozdělení získaný na základě odchylky jednotlivých hodnot od průměru. $s^2 = \frac{\sum (x_i - x)^2}{n - 1}$
- Obdobně jako u průměru je jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozdělení
- **Směrodatná odchylka** je druhá odmocnina z rozptylu
- **Koeficient variance** - podíl SD ku průměru (u normálního rozdělení by se 95% hodnot mělo vejít do průměr ±3 SD), pokud je SD větší než 1/3 průměru jsou teoreticky pravděpodobné záporné hodnoty v rozdělení – ukazatel problémů s normalitou dat

Ukazatele tvaru rozdělení

- **Skewness** – ukazatel „šikmosti“ rozdělení, asymetrie rozdělení
- **Kurtosis** – ukazatel „špičatosti/plochosti“ rozdělení



Další parametry rozdělení

- **Počet hodnot –** důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Střední chyba odhadu průměru** - je založena na směrodatné odchylce rozdělení a **počtu hodnot**, vlastně jde o směrodatnou odchylku rozdělení průměru. Říká jak přesný je náš výpočet průměru. Čím větší počet hodnot rozdělení, tím je náš odhad skutečného průměru přesnější.
- **Suma hodnot**
- **Modus** – nejčastější hodnota, vhodný např. při kategoriálních datech
- **Minimum, maximum**
- **Rozsah hodnot**
- **Harmonický průměr** - převrácená hodnota průměru převrácených hodnot (vždy platí harmonický průměr < geometrický průměr < aritmetický průměr)

6a. Párové testy

Párový t-test

Wilcoxonův test

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

Levenův test

- Test sloužící k rozhodnutí, zda mají dva nebo více vzorků stejný rozptyl.
- H0: rozptyl je stejný.
  HA: rozptyl se liší.
- Testová statistika:

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i (Z_i - Z)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i} N_i \left(Z_{i,j} - Z_i\right)^2}$$

N je celkový počet hodnot

N_i je počet hodnot v i-té skupině

k je počet skupin

$\bar{x}_i$ je průměr hodnot i-té skupiny (resp. medián)

$Z_{i,j} = |x_{i,j} - \bar{x}_i|$

Z_i je průměr $Z_{i,j}$

Z je průměr všech $Z_{i,j}$

Párové dvouvýběrové testy – předpoklady

- Skupiny dat jsou spojeny přes objekt měření, příkladem může být měření parametrů pacienta před léčbou a po léčbě (nemusí jít přímo o stejný objekt, dalším příkladem mohou být např. krysy ze stejné linie).
- Oba soubory musí mít shodný počet hodnot, protože všechna měření v jednom souboru musí být spárována s měřením v druhém souboru. Při vlastním výpočtu se potom počítá se změnou hodnot (diferencí) subjektů v obou souborech.
- Před párovým testem je vhodné ověřit si zda existuje vazba mezi oběma skupinami – vynesení do grafu, korelace.

Existuje několik možných designů experimentu, stručně lze sumarizovat:

1. pokus je párový a jako párový se projeví
2. párové provedení pokusu – párově se neprojeví
   - možná párovost není
   - špatně provedený pokus – malé n, velká variabilita, špatný výběr jedinců
3. čekali jsme nezávislé a jsou
4. čekali jsem nezávislé a nejsou
   - vazba
   - náhoda

Párový dvouvýběrový t-test

- Tento test nemá žádné předpoklady o rozložení vstupních dat, protože je počítán až na základě jejich diferencí.
- Tyto diference by měly být normálně rozloženy a otázkou v párovém t-testu je, zda se průměrná hodnota diferencí rovná nějakému číslu, typicky jde o srovnání s nulou jako důkaz neexistence změny mezi oběma spárovanými skupinami.
- V podstatě jde o one sample t-test, kde místo rozdílu průměru vzorku a cílové populace je uveden průměr diferencí a srovnávané číslo (0 v případě otázky, zda není rozdíl mezi vzorky).
- Pro srovnání s 0 (testovou statistikou je t rozložení): $t = \frac{\bar{D}}{s}\sqrt{n} \qquad \upsilon = n - 1$
- Někdy je obtížné rozhodnout, zda jde nebo nejde o párové uspořádání, párový test by měl být použit pouze v případě, že můžeme potvrdit vazbu (korelace, vynesení do grafu), jedním z důvodů proč toto ověřovat je fakt, že v případě párového t-testu není nutné brát ohled na variabilitu původních dvou souborů, tento předpoklad však platí pouze v případě vazby mezi proměnnými. Výpočet obou typů testů se vlastně liší v použité s, jednou jde o s diferencí, v druhém případě o složený odhad rozptylu obou souborů.
- Zda je párové uspořádání efektivnější lze určit na základě:
  - Síly vazby
  - Je-li s_D výrazně menší než s_{x1-x2}
- Závislost je možné rozepsat pomocí vzorce: $s_D^2 \cong \sigma_{x_1}^2 + \sigma_{x_2}^2 - 2Cov(x_1; x_2)$
- v případě Cov=0, tedy v případě neexistence vazby pak s_D^2 odpovídá součtu původních rozptylů, tedy přibližně S_{x1-x2}.

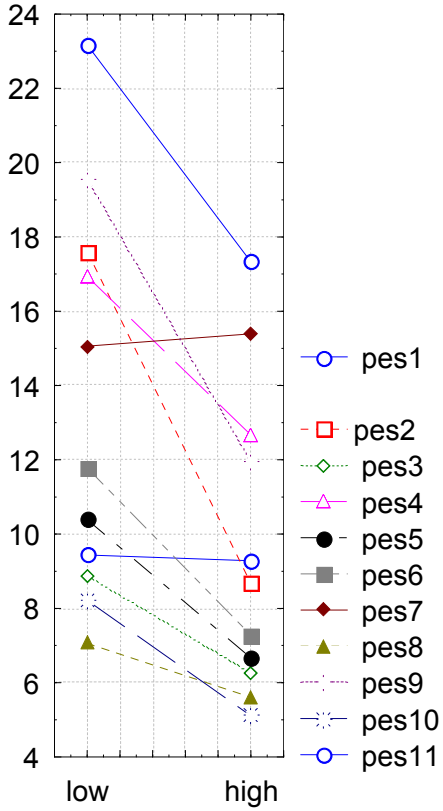
Párový dvouvýběrový t-test – příklad

Byl prováděn pokus s dietou 11 diabetických psů, každý pes byl vystaven dvěma dietám s odlišným typem sacharidů (snadno vstřebatelné X pozvolna se rozkládající na glukózu), hodnoty krevní glukózy v průběhu jednotlivých diet mají být srovnány pro zjištění vlivu diety na hladinu krevní glukózy. Protože každý pes absolvoval obě diety, jde o párové uspořádání, kdy výsledky hodnoty v obou pokusech jsou spojeny přes pokusné zvíře.

1. ***Nulová hypotéza zní, že skutečný průměrný rozdíl mezi oběma dietami je 0, alternativní hypotéza zní, že to není 0.***
2. ***Pro každého psa je spočítán rozdíl mezi jeho hladinou glukózy při obou dietách a měly by být ověřeny předpoklady pro one sample t-test – tedy alespoň přibližně normální rozložení.***
3. ***Je spočítána testová charakteristika, výpočet vlastně probíhá jako one-sample t-test, kde je zjišťována významnost průměru diferencí obou souborů jako rozdíl mezi touto hodnotou a nulou (nula je hodnota, kterou by průměrná diference měla nabývat, pokud platí nulová hypotéza). T=4.37 s 10 stupni volnosti, skutečná hodnota p=0,0014 a tedy na hladině p=0,05 můžeme nulovou hypotézu zamítnou***

$$t = \frac{rozdíl\_průměru\_vzorku\_a\_populace}{SE(průměru)} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

4. ***Závěrem můžeme říci, že nulová hypotéza neexistence rozdílu mezi oběma dietami byla zamítnuta, což znamená, že high-fibre dieta má významný vliv na snížení hladiny krevní glukózy.***



Neparametrická obdoba párového t-testu

Wilcoxon test

- Jsou vytvořeny diference mezi soubory, je vytvořeno jejich pořadí bez ohledu na znaménko a poté je sečteno pořadí kladných a pořadí záporných rozdílů. Menší z těchto dvou hodnot je srovnána s kritickou hodnotou testu a pokud je menší než kritická hodnota testu, pak zamítáme hypotézu shody obou souborů hodnot. Pro test existuje aproximace na normální rozložení, ale pouze pro velká n>25.

$$t = \frac{Menší\_suma\_diferencí - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Před zásahem	Po zásahu	Změna	Absolutní pořadí
6	2	4	10
2,5	3	-0,5	1,5
6,3	5	1,3	6
8,1	9	-0,9	5
1,5	2	-0,5	1,5
3,4	4	-0,6	3
2,5	1	1,5	8
1,11	2	0,89	4
2,6	4	-1,4	7
1	3	-2	9

Wilcoxonův test – příklad I

člověk	A	B	diference	pořadí
1	142	138	4	4,5
2	140	136	4	4,5
3	144	147	-3	3
4	144	139	5	7
5	142	143	-1	1
6	146	141	5	7
7	149	143	6	9,5
8	150	145	5	7
9	142	136	6	9,5
10	148	146	2	2

A…….parametr krve před podáním léku

B…….parametr krve po podání léku

W_+ …… Σpořadí kladných rozdílů = 51

W_- …… = 4

$$W = \min(W_+; W_-) = 4$$

počet párů = n = 10

Pokud je W menší než kritická hodnota testu, pak zamítáme hypotézu shody distribučních funkcí obou skupin.

Wilcoxonův test – příklad II

Byla testována nová dieta pro laboratorní krysy, při pokusu byl zjišťován její vliv na různých liniích krys, bylo proto zvoleno párové uspořádání kdy krysy v obou dietách jsou spojeny přes svoji linii, tj. na začátku byly dvojice krys stejné linie, jedna z nich byla náhodně přiřazena k dietě, druhá z dvojice pak do druhé diety.

1. nulová hypotéza je, že váha krys není ovlivněna použitou dietou, alternativní, že ovlivnění dietou existuje
2. spočítáme diference – tyto diference jsou nenormální a proto je vhodné využít neparametrický test
3. Spočítáme sumu pořadí kladných a záporných diferencí, zde je menší suma záporných diferencí – 31
4. výsledkem výpočtu je $p>0,05$ a tedy nemáme dostatečné důkazy pro zamítnutí nulové hypotézy, nelze říci, že by nová dieta byla efektivnější než stará
5. pro doplnění výsledků je vhodné zjistit také skutečnou velikost rozdílu hmotností ve skupinách, např. ve formě mediánu