IV. Statistická analýza dat - úvod

IV.a Teoretické pozadí statistické analýzy

IV.b Základní typy dat

IV.c Modelová rozložení

IV.d Popisná statistika dat

IV.e Provádění odhadů

IV.f Základy testování hypotéz

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

IV.a Teoretické pozadí statistické analýzy

Jak vznikají informace

Rozložení dat

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

Anotace

- Základním principem statistiky je pravděpodobnost výskytu nějaké události. Prostřednictvím vzorkování se snažíme odhadnout skutečnou pravděpodobnost událostí.Klíčovou otázkou je velikost vzorku, čím větší vzorek, tím větší šance na projevení se skutečné pravděpodobnosti výskytu jevu.

JAK vznikají informace ? základní pojmy

Skutečnost

Náhoda
(vybere jednu z možností pokusu)

Jev

podmnožina všech možných výsledků pokusu/děje, o které lze říct, zda nastala nebo ne

Pozorovatel

Rozliší, co nastalo
a) ***podle možností***
b) ***podle toho, jak potřebuje***

Jevové pole

třída všech jevů, které jsme se rozhodli nebo jsme schopni sledovat

Skutečnost + Jevové pole = Měřitelný prostor

Experimentální jednotka - *objekt, na kterém se provádí šetření*

Populace - *soubor experimentálních jednotek* **Znak -** *vlastnost sledovaná na objektu*
Sledovaná veličina - *číselná hodnota vyjadřující výsledek náhodného experimentu*

Znak se stává náhodnou veličinou, pokud se jeho hodnota zjišťuje vylosováním objektu ze základního souboru

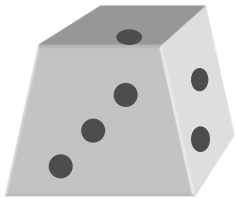
Výběr - výběrová populace - cílová populace

Náhodný výběr **Reprezentativnost**

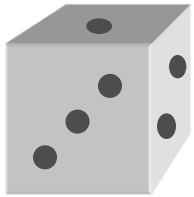
JAK vznikají informace ?

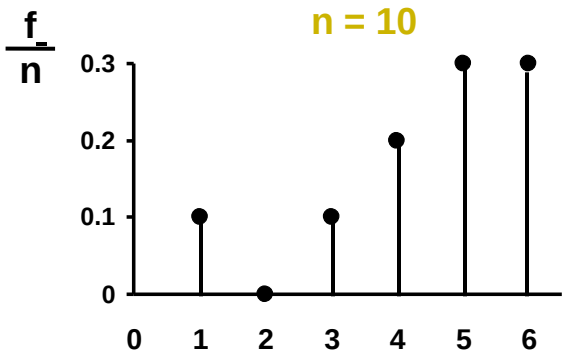
„Empirical approach“

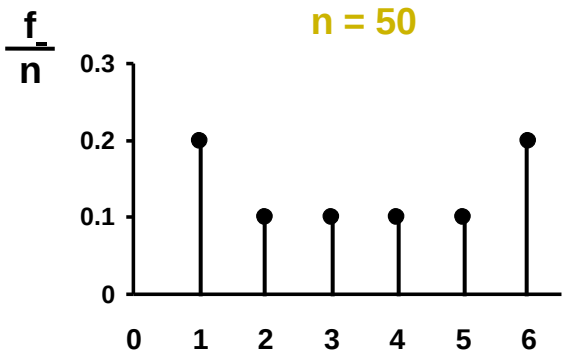
„Classical approach“

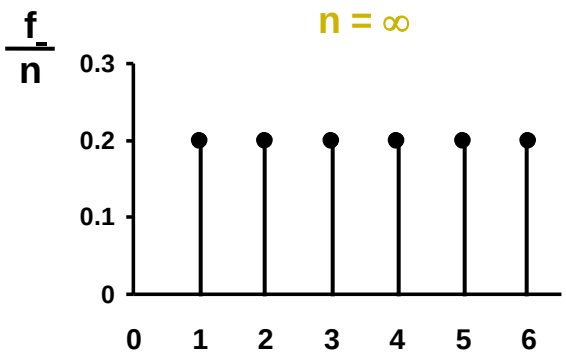


Empirický postup









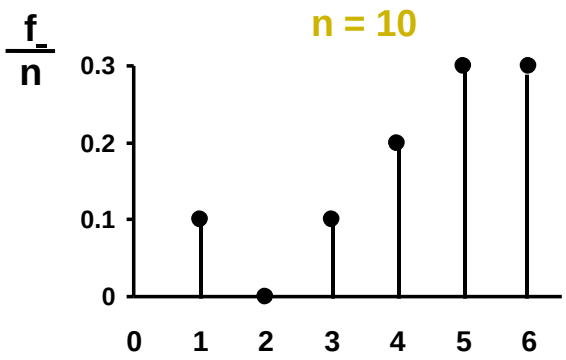
možné jevy: čísla 1 – 6

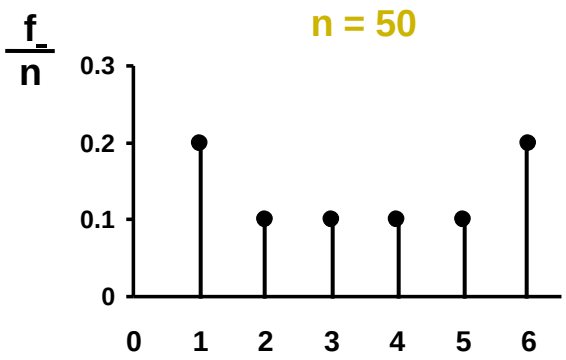
n – počet hodů (opakování)

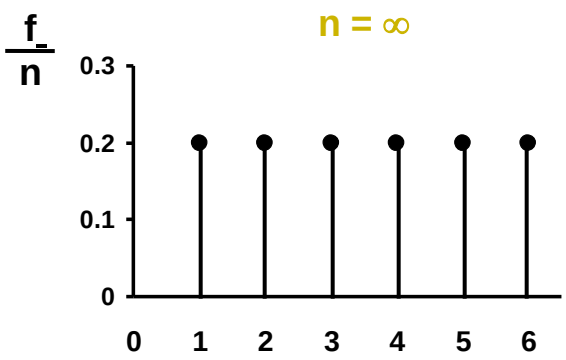
U složitých stochastických systémů se pravda získá až po odvedení značného množství experimentální práce: musíme dát systému šanci se projevit

JAK vznikají informace ?

Empirický postup







možné jevy: čísla 1 – 6

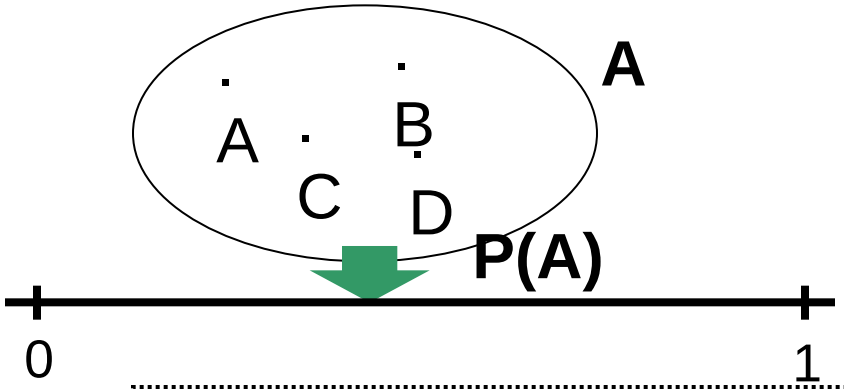
n – počet hodů (opakování)

Při realizaci náhodného experimentu roste se zvyšujícím se počtem opakování pravdivá znalost systému (výsledky se stávají stabilnější) …. diskutabilní je ale ovšem míra zobecnění konkrétního experimentu

Empirický zákon velkých čísel

Při opětovné nezávislé realizaci téhož náhodného experimentu se podíl výskytů sledovaného jevu mezi všemi dosud provedenými realizacemi zpravidla ustaluje kolem konstanty.

Pravděpodobnost je libovolná reálná funkce definovaná na jevovém poli A, která každému jevu A přiřadí nezáporné reálné číslo P(A) z intervalu 0 - 1.

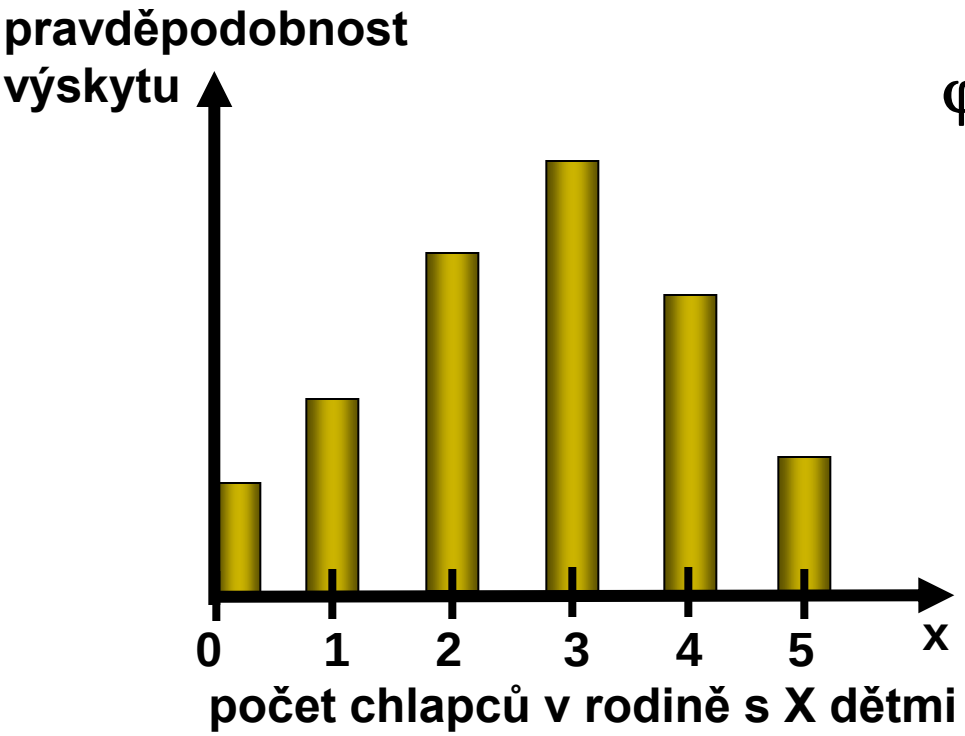


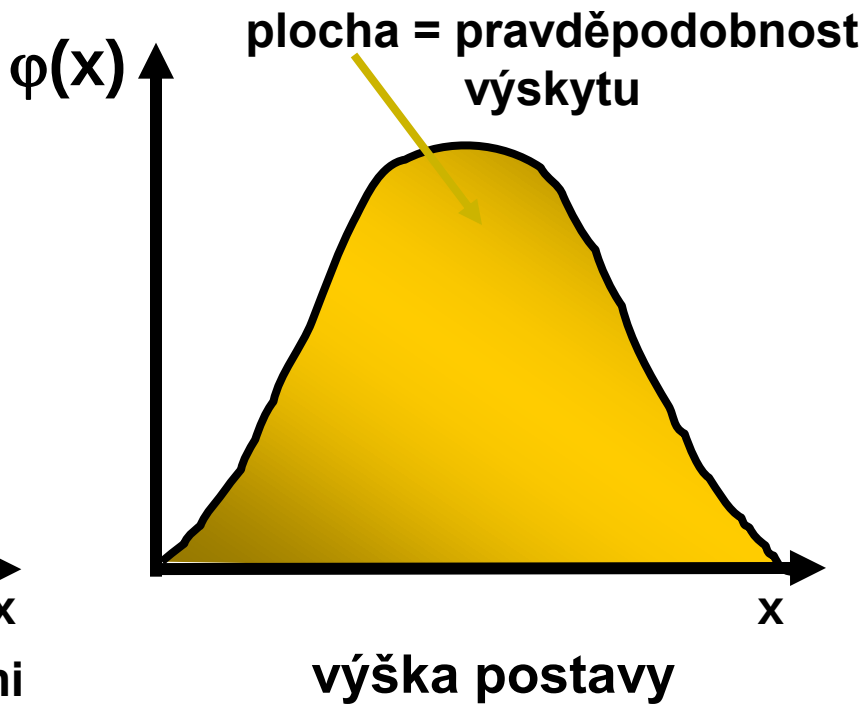
Z praktického hlediska je pravděpodobnost
idealizovaná relativní četnost

P (A) = 1 ..... jev jistý
P (A) = 0 ..... jev nemožný
P (A ∩ B) = P (A) . P (B)..... nezávislé jevy
P (A ∩ B) = P (A) . P (B/A) .....závislé jevy
P (A / B) = P (A ∩ B) / P (B) .....podmíněná pravděpodobnost

Pravděpodobnost výskytu jevu – rozložení dat

- existuje pravděpodobnost výskytu jevů (nedeterministické závěry)
- „vše je možné“: pouze jev s pravděpodobností 0 nikdy nenastane
- pravděpodobnost lze zkoumat retrospektivně i prospektivně





IV.b Základní typy dat

Spojitá a kategoriální data

Základní popisné statistiky

Grafický popis dat

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

Anotace

- Realitu můžeme popisovat různými typy dat, každý z nich se specifickými vlastnostmi, výhodami, nevýhodami a vlastní sadou využitelných statistických metod - od binárních přes kategoriální, ordinální až po spojitá data roste míra informace v nich obsažené.
- Základním přístupem k popisné analýze dat je tvorba frekvenčních tabulek a jejich grafických reprezentací – histogramů.

Jak vznikají informace ?
– různé typy dat znamenají různou informaci

Samotná znalost typu dat ale na dosažení informace nestačí .....

Jak vznikají informace ?
– různé typy dat znamenají různou informaci

Statistika středu

Data poměrová

PRŮMĚR

Data intervalová

Data ordinální

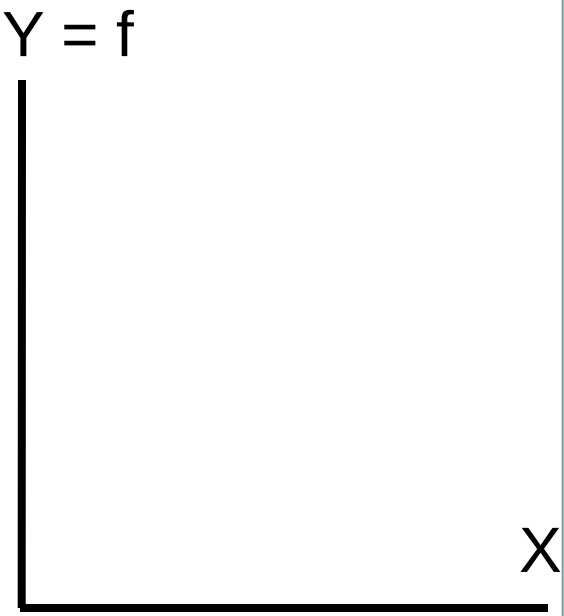
MEDIÁN

Data nominální

MODUS

Spojitá data

Diskrétní data

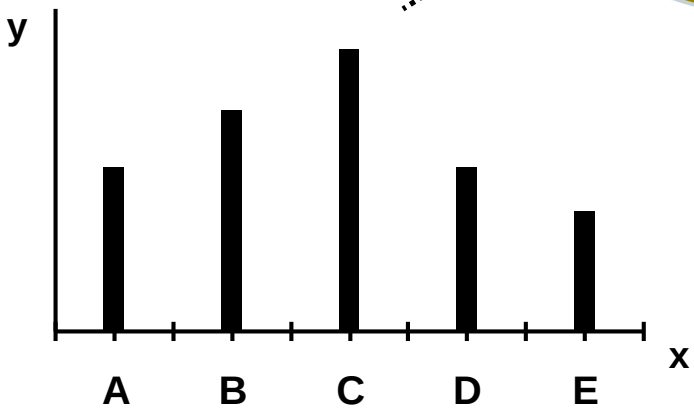


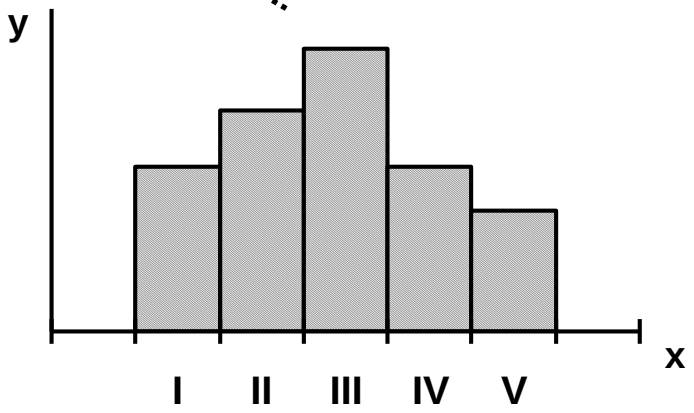
JAK vznikají informace ?
- opakovaná měření informují rozložením hodnot

Y: frekvence

- absolutní / relativní

KOLIK se naměřilo





CO se naměřilo

X: měřený znak

Diskrétní data ← → **Spojitá data**

Jak vznikají informace ?
- frekvenční tabulka jako základní nástroj popisu

DISKRÉTNÍ DATA

Primární data

Počty epizod pro n = 100 hemofiliků

0
0
1
2
1
1
3
1
1
2
.
.
.
.
.
.
n = 100

Frekvenční sumarizace

N: 100 dětí (hemofiliků)
x: znak: počet krvácivých epizod za měsíc

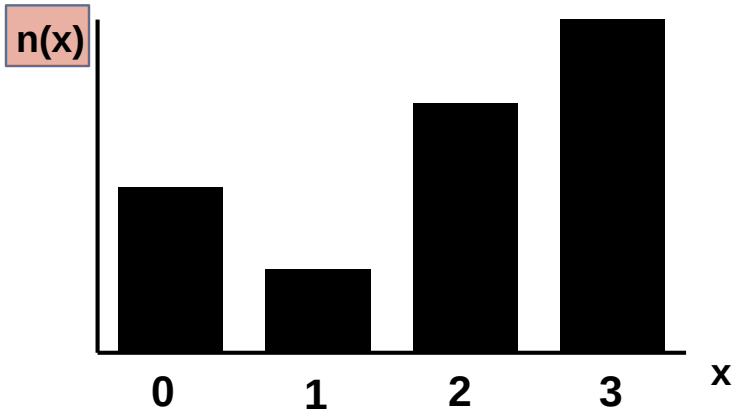
x	n(x)	N(x)	p(x)	F(x)
0	20	20	0,2	0,2
1	10	30	0,1	0,3
2	30	60	0,3	0,6
3	40	100	0,4	1,0

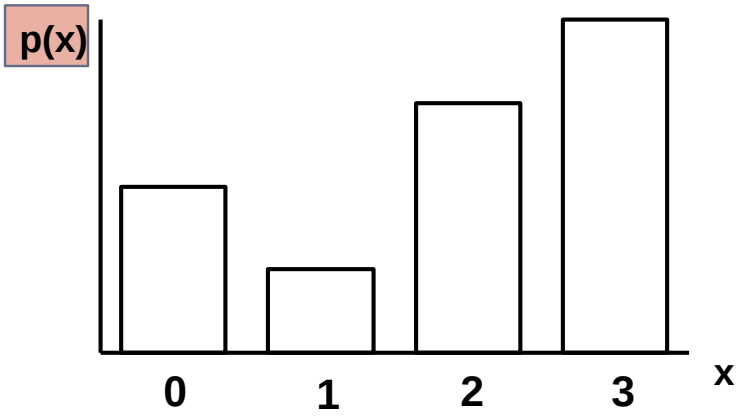
n(x) – absolutní četnost x
N(x) – kumulativní četnost hodnot nepřevyšujících x;

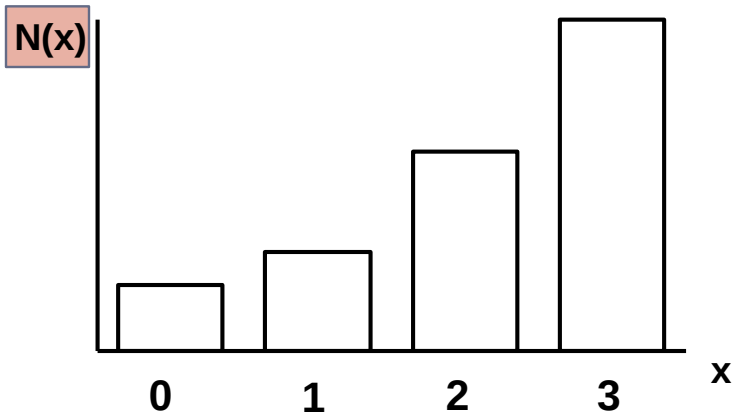
$$N(x) = \sum_{t \le x} n(t)$$

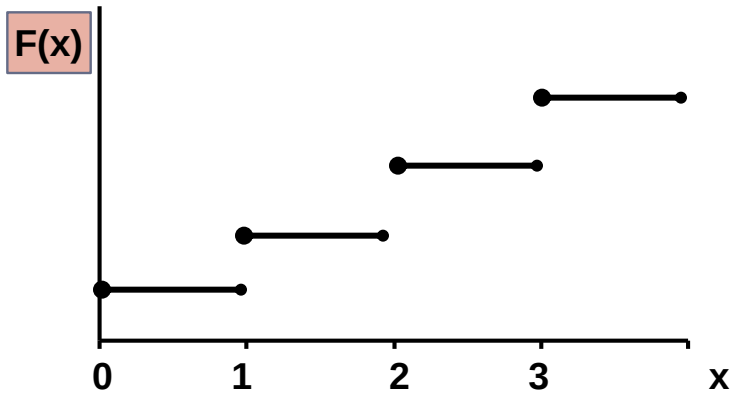
p(x) – relativní četnost; p(x) = n(x) / n
F(x) – kumulativní relativní četnost hodnot nepřevyšujících x; F(x) = N(x) / n

Jak vznikají informace ?
Grafické výstupy z frekvenční tabulky









Jak vznikají informace ?
- frekvenční tabulka jako základní nástroj popisu

SPOJITÁ DATA

Příklad: **x: koncentrace látky v krvi n = 100 pacientů**

Primární data

Hodnoty pro n = 100 osob

1,21
1,48
1,56
0,31
1,21
1,33
0,33
.
.
.
n = 100

Frekvenční sumarizace

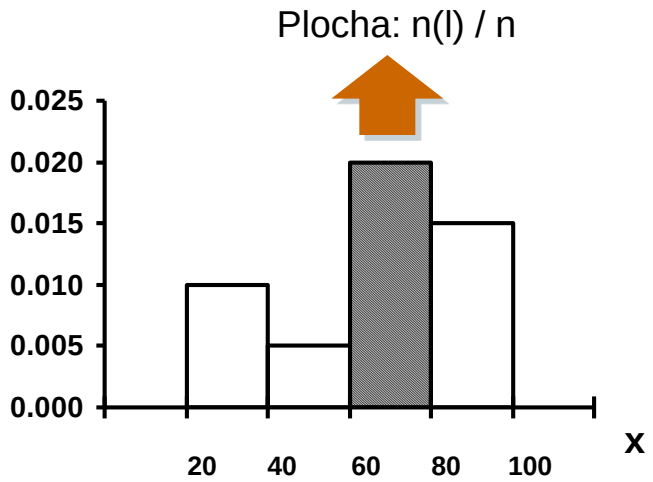
n = 100 opakovaných měření (100 pacientů)
x: koncentrace sledované látky v krvi (20 – 100 jednotek)

interv	d(l)	n(l)	n(l)/n	N(x'')	F(x'')
<20, 40)	20	20	0,2	20	0,2
<40, 60)	20	10	0,1	30	0,3
<60, 80)	20	40	0,4	70	0,7
<80, 100)	20	30	0,3	100	1,0

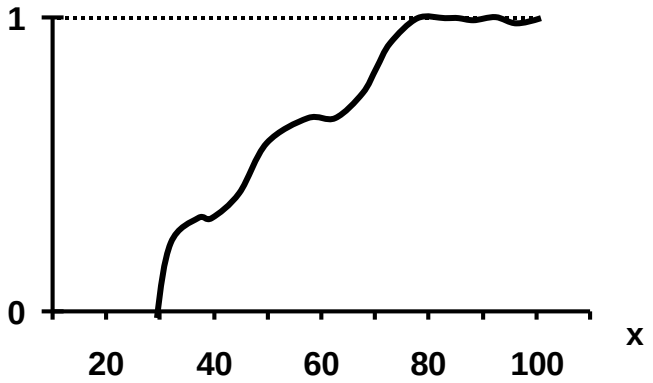
d(l) – šířka intervalu
n(l) – absolutní četnost
n(l) / n – intervalová relativní četnost
N(x'') – intervalová kumulativní četnost do horní hranice X''
F(x'') – intervalová relativní kumulativní četnost do horní hranice X''

Jak vznikají informace ?
- frekvenční sumarizace spojitých dat

Histogram



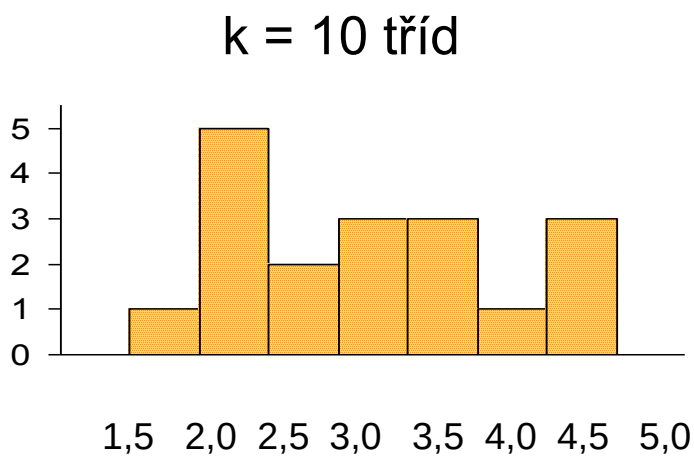
Výběrová distribuční funkce

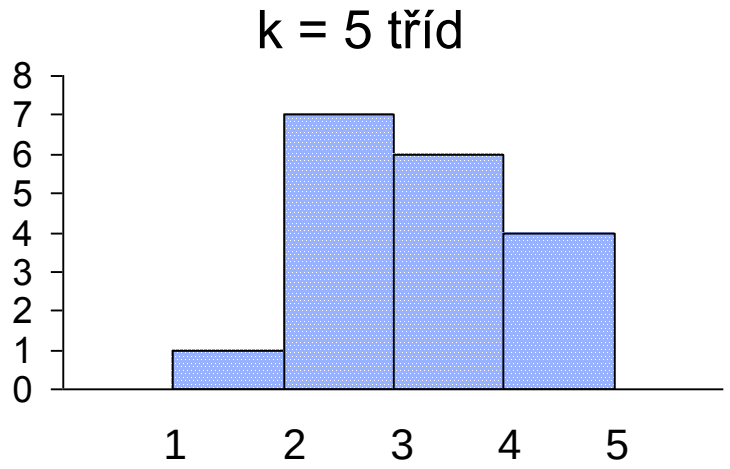


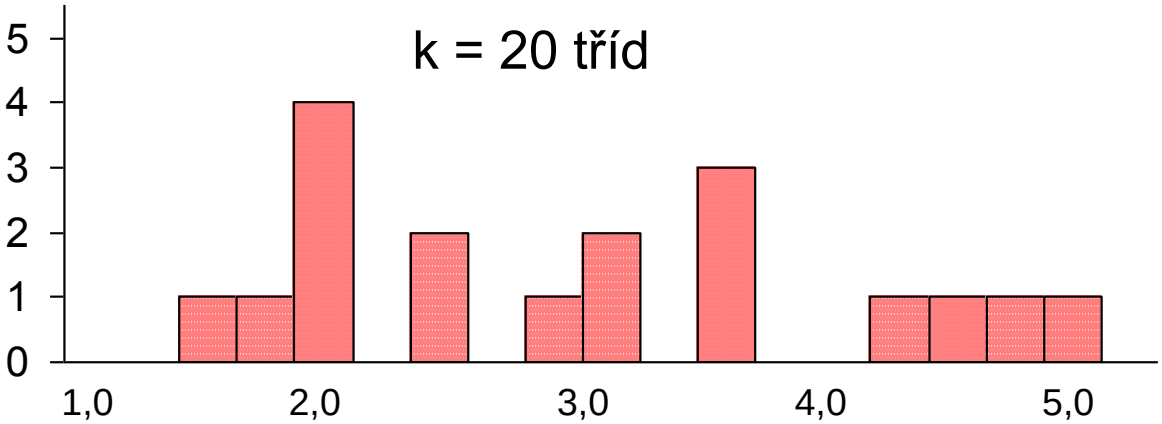
$$f(x)=\frac{n(l)\,/\,n}{d(l)} \Rightarrow$$ **Intervalová hustota četnosti**

F(x) $\Rightarrow$ **Intervalová relativní kumulativní četnost**

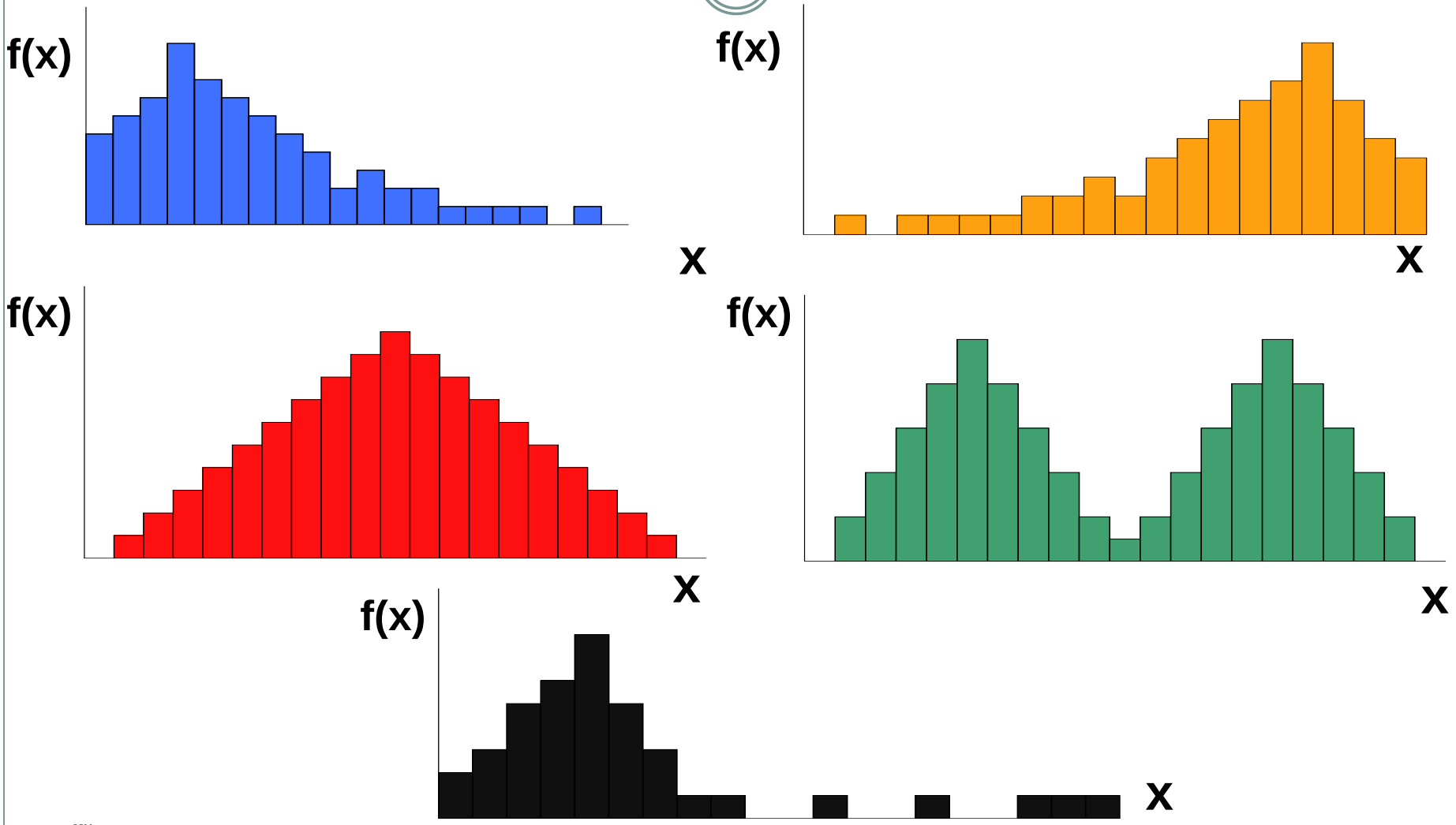
Počet zvolených tříd a velikost souboru určují kvalitu výstupu



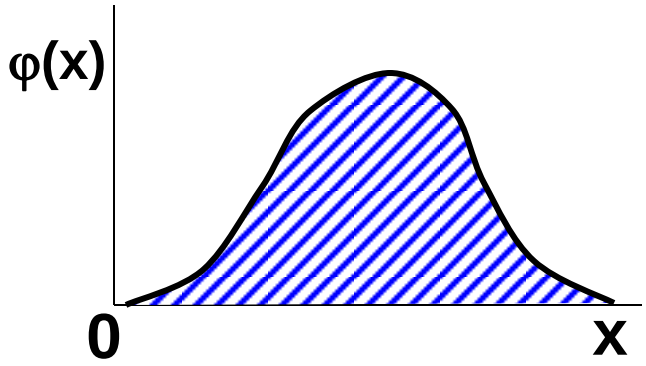




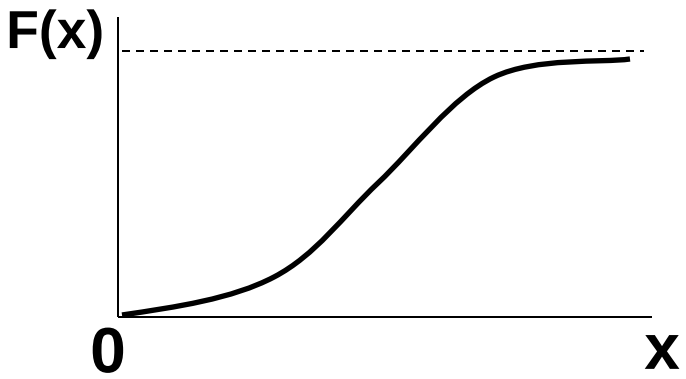
Histogram vyjadřuje tvar výběrového rozložení



Pojem ROZLOŽENÍ - příklad spojitých dat



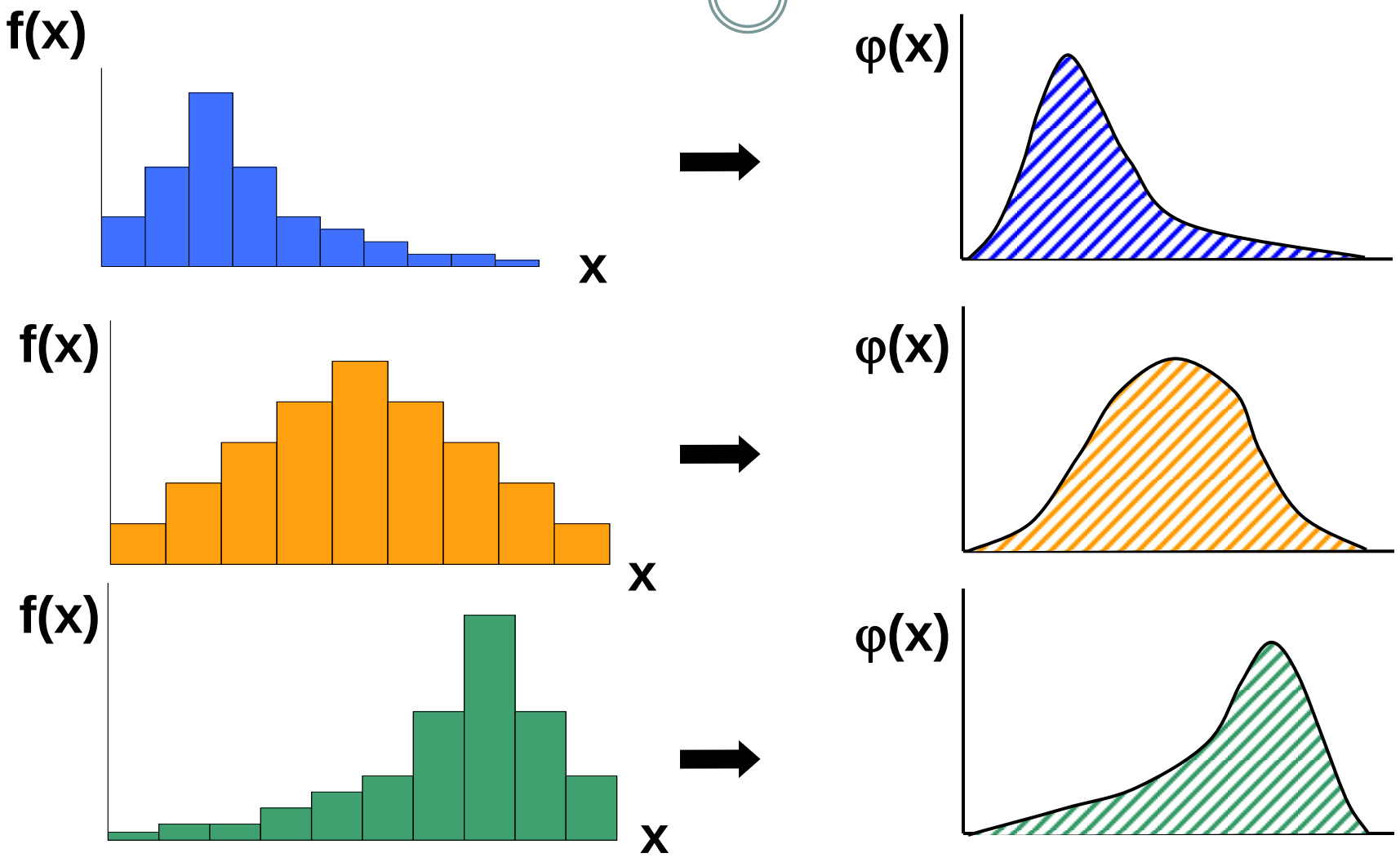
Rozložení



Distribuční funkce

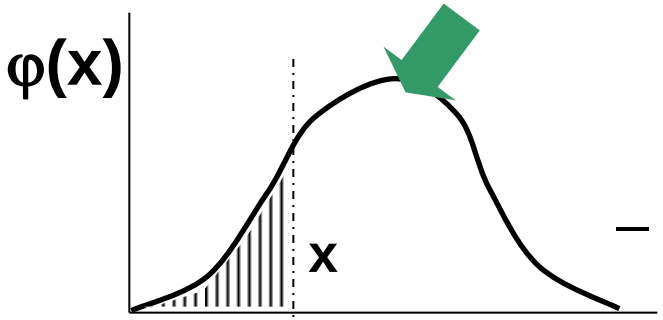
Je - li dána distribuční funkce, je dáno rozložení

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



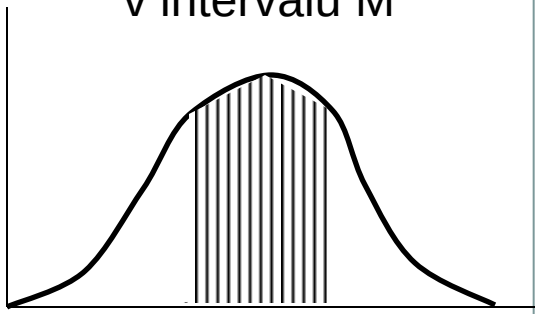
Distribuční funkce jako užitečný nástroj pro práci s rozložením

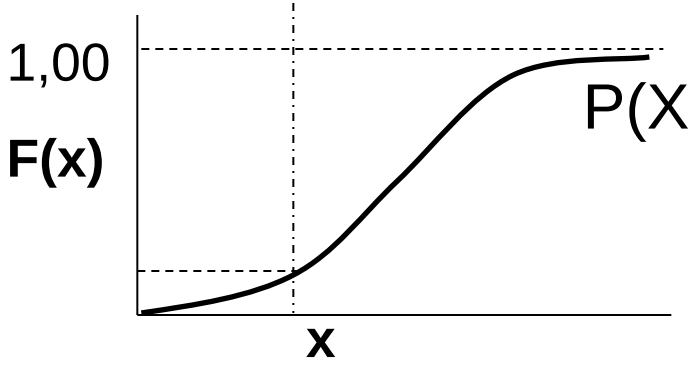
Plocha = relativní četnost



$$\int_{-\infty}^{\infty} \varphi(x)\ d(x) = 1$$

F(x):
Pravděpodobnost, že se X vyskytuje v intervalu M





$$P(X \leq x) = \Phi(x) = F(x'')$$

$\Phi(x)$ … distribuční funkce

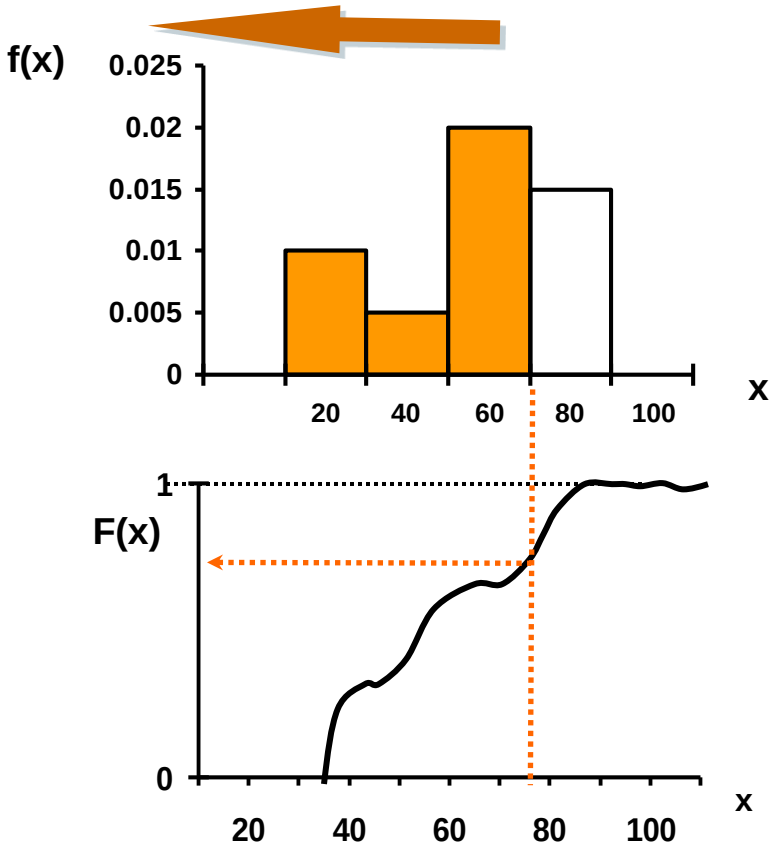
$$P(X \leq x) = \int_{M} \varphi(x)\ d(x)$$

Známe-li distribuční funkci, pak známe rozložení sledované veličiny.

Pro jakoukoli množinu hodnot (**M**) lze určit P, že X do této množiny patří.

Jak vznikají informace ?
- frekvenční sumarizace spojitých dat

Grafické výstupy z frekvenční tabulky – spojitá data



Uspořádání čísel podle velikosti a konstrukce rozložení umožňuje pravděpodobnostní zařazení každé jednotlivé hodnoty

KVANTIL

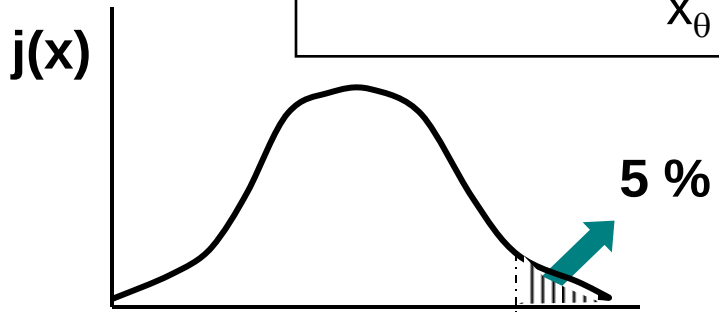
$X_{0.1}$; $X_{0.9}$; $X_{0.5}$; X_{θ}

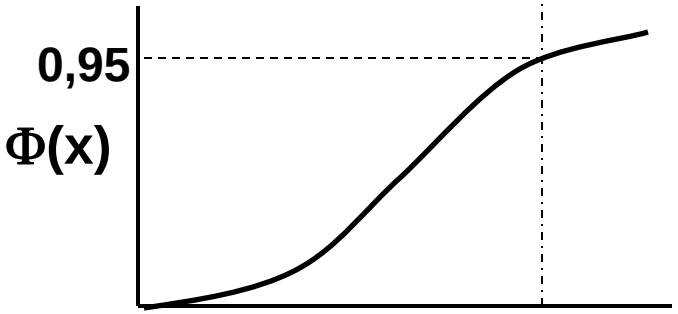
Otázka: Jak velké musí být X, aby 5 % všech hodnot bylo nad ním?

θ = 0,95 … Pravděpodobnost

Hledáme: $P(X \leq x_\theta) = 0{,}95 = \theta$

$x_\theta = (X_{0,95}) = ?$





F (xθ) = θ

Kvantil je číslo, jehož hodnota distribuční funkce je rovna P, pro kterou je kvantil definován

Jakékoliv číslo na ose x je kvantilem

IV.c Modelová rozložení

Normální rozložení jako statistický model

Aplikace modelových rozložení

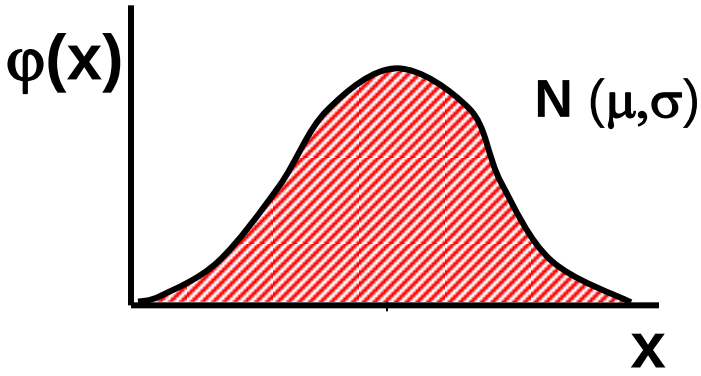
Přehled modelových rozložení

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 rozložení 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 rozložením, v opačném případě hrozí získání zavádějících výsledků.
- Nejklasičtějším modelovým rozložením, od něhož je odvozena celá řada statistických analýz je tzv. normální rozložení, známé též jako Gaussova křivka.

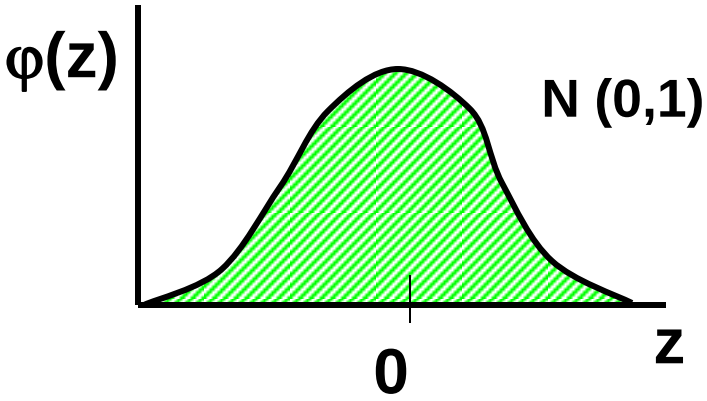
Rozložení hodnot jako model: Normální rozložení



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

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

Standardizovaná forma



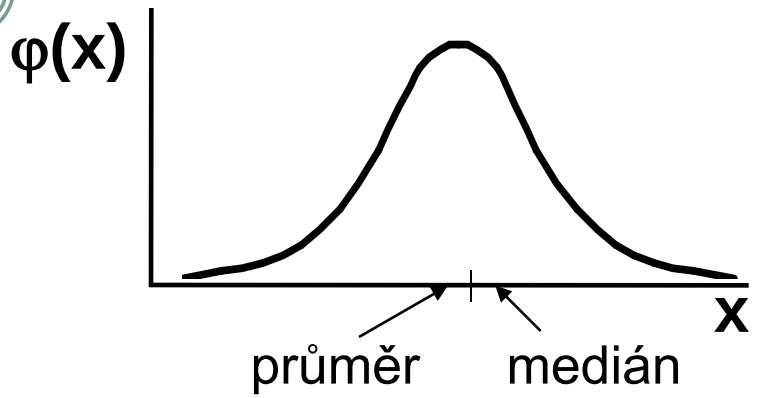
$$\varphi(z) = \frac{1}{\sqrt{2.\pi}}.e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

Parametry charakterizující normální rozložení a jejich význam

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

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



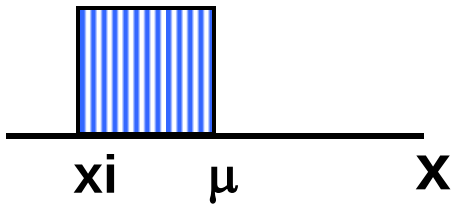
a) $\mu \sim \bar{x}$

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

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

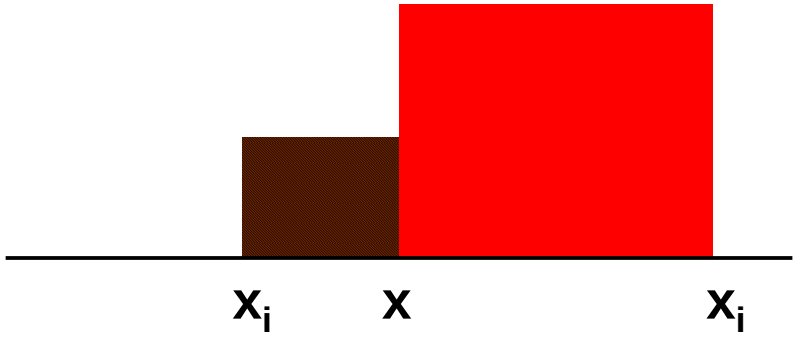
Pravidlo ± 3s

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

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

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

?

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

?

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

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

Jak hodnotíte model u těchto dat ?

Stochastické rozložení 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 rozložení hodnot x**

Normální rozložení jako model - příklad

Tabulky distribuční funkce

- **Data z průzkumu jsou publikována jako:**

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

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

? 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

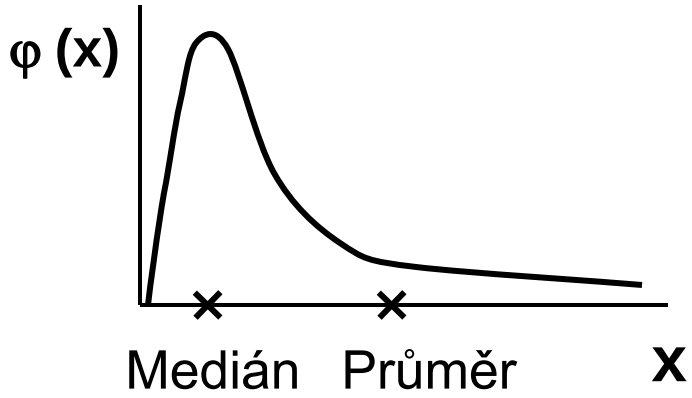
Stručný přehled modelových rozložení I.

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.

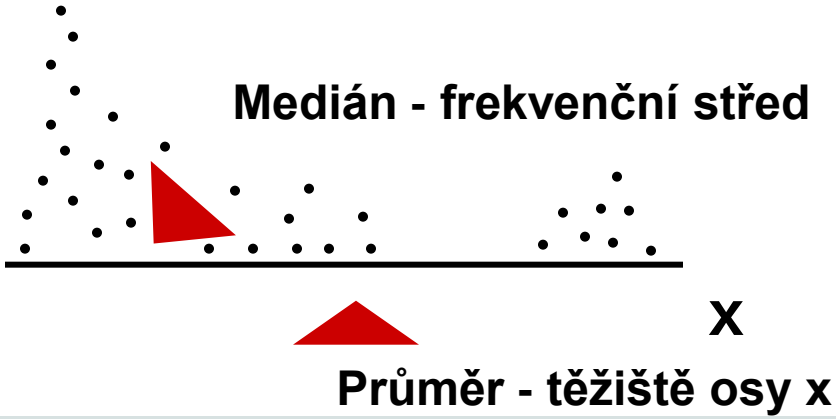
Stručný přehled modelových rozložení II.

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.

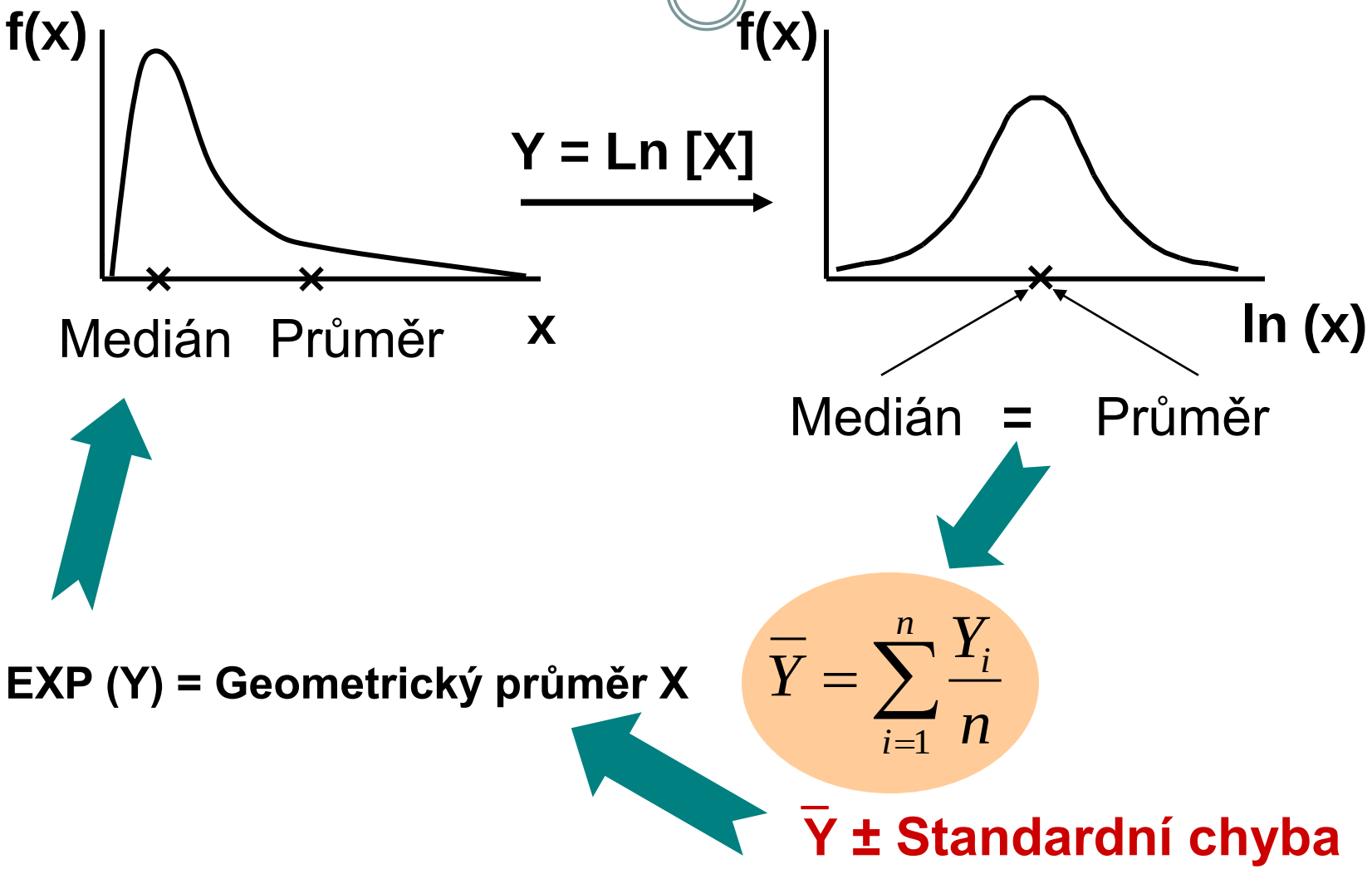
Log-normální rozložení jako častý model reálných znaků



U asymetrických rozložení je medián velmi vhodným alternativním ukazatelem středu



Log-normální rozložení lze jednoduše transformovat



Transformace dat - legitimní úprava rozložení

✓ **Základní typy transformací vedou k normalitě rozložení 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 rozložení, 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 rozložení

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

Odmocninová transformace

Transformace je vhodná pro proměnné mající Poissonovo rozložení, 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 rozložení

Arcsin transformace

Tzv. **úhlová transformace** - velmi vhodná pro data typu podílů výskytu určitého jevu (znaku) mezi **n** hodnocenými jedinci - tedy pro data mající binomické rozložení. 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:

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

IV.d Popisná statistika dat

Popisné statistiky dat

Vizualizace dat

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

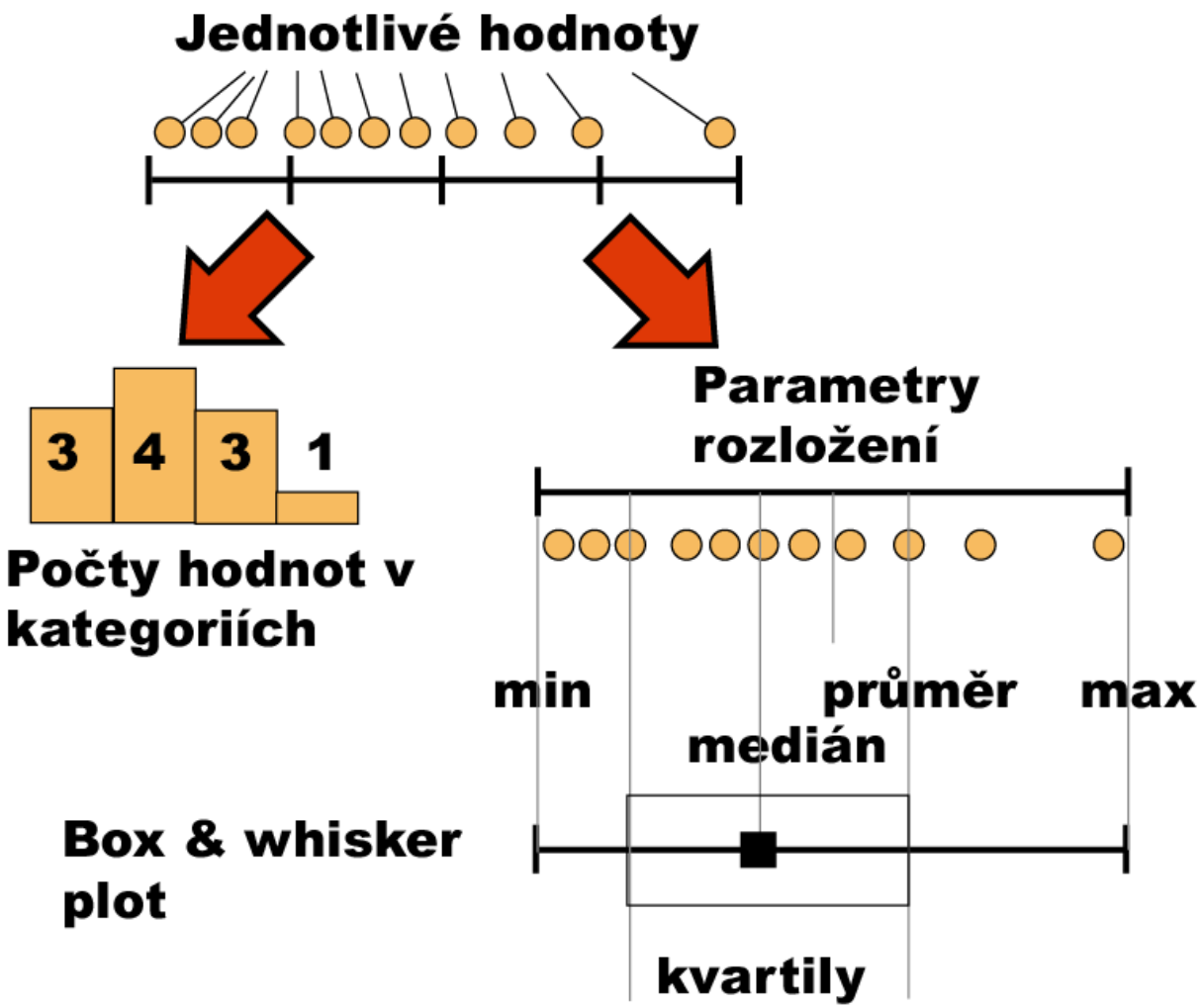
Anotace

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

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

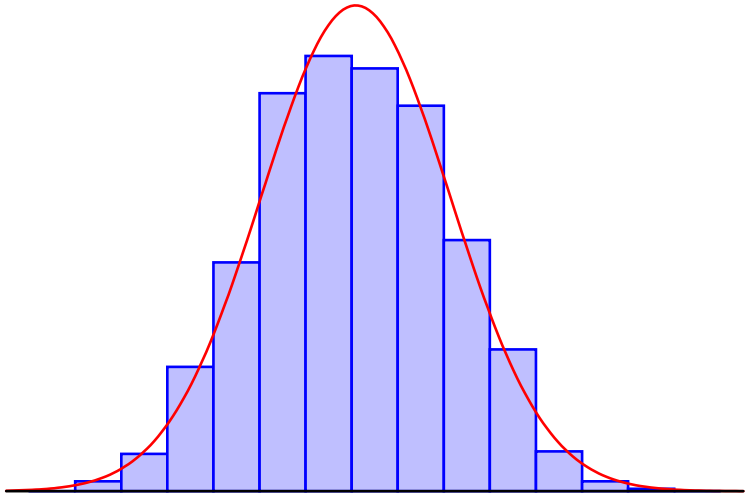


Frekvenční rozložení

Kategorie	Četnost
B	5
C	8
D	1

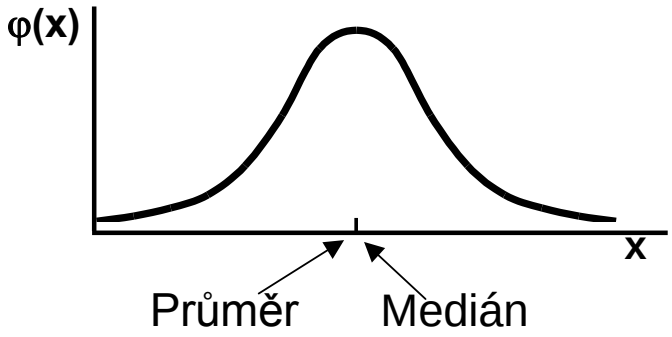
Kvalitativní data
Tabulka s četností jednotlivých kategorií.

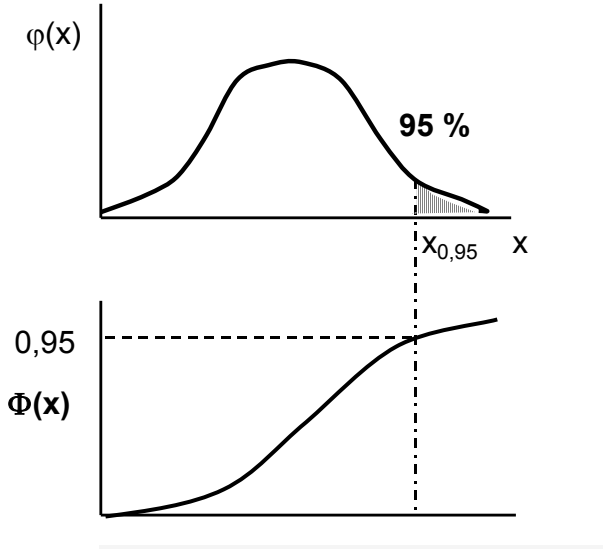
Kvantitativní data
Četnost hodnot rozložení v jednotlivých intervalech.



Parametry rozložení

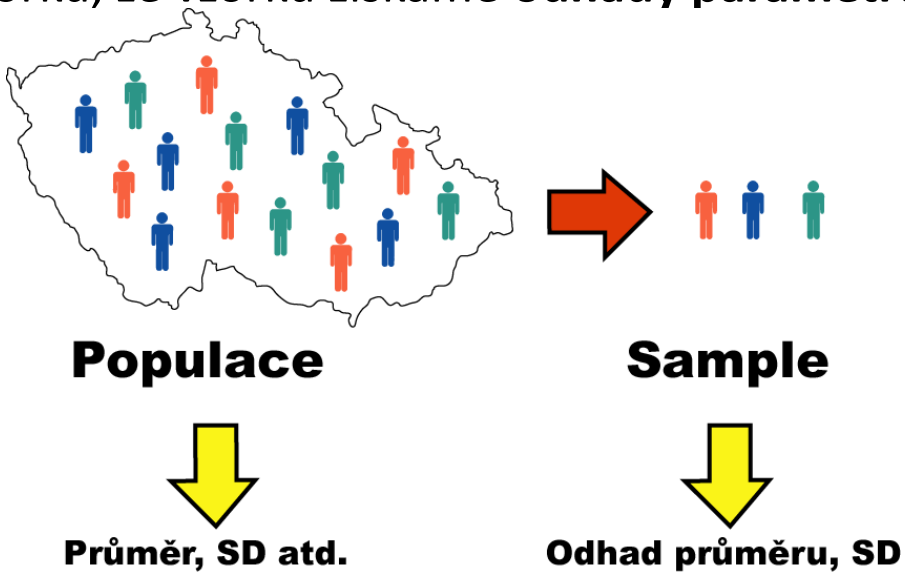
- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
  - Středu (medián, průměr, geometrický průměr)
  - Šířky rozložení (rozsah hodnot, rozptyl, směrodatná odchylka)
  - Tvaru rozložení (skewness, kurtosis)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





Populace a vzorek

- Populace představuje veškeré možné objekty vzorkování, např. veškeré obyvatelstvo ČR při sledování na úrovni ČR, z populace získáme reálné parametry rozložení
- 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 rozložení**



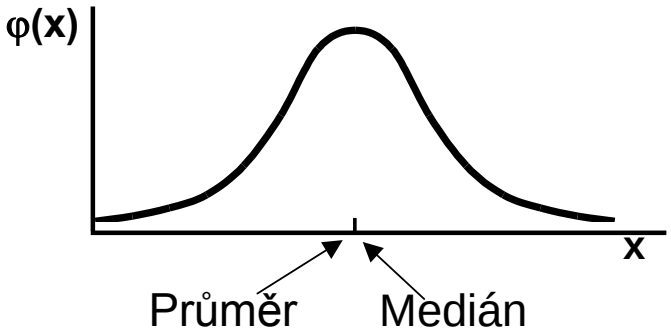
Ukazatele středu rozložení I

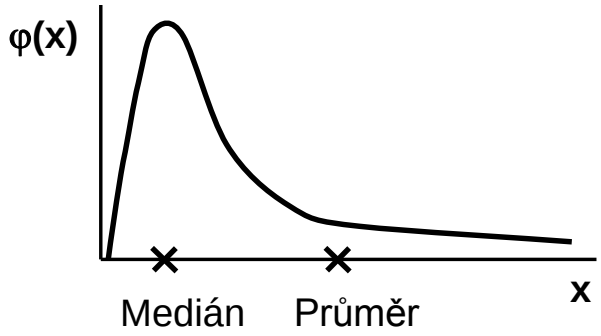
- **Průměr** – vhodný ukazatel středu u normálního/symetrického rozložení, kde $\mathbf{x_i}$ jsou jednotlivé hodnoty a **n** jejich počet

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

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

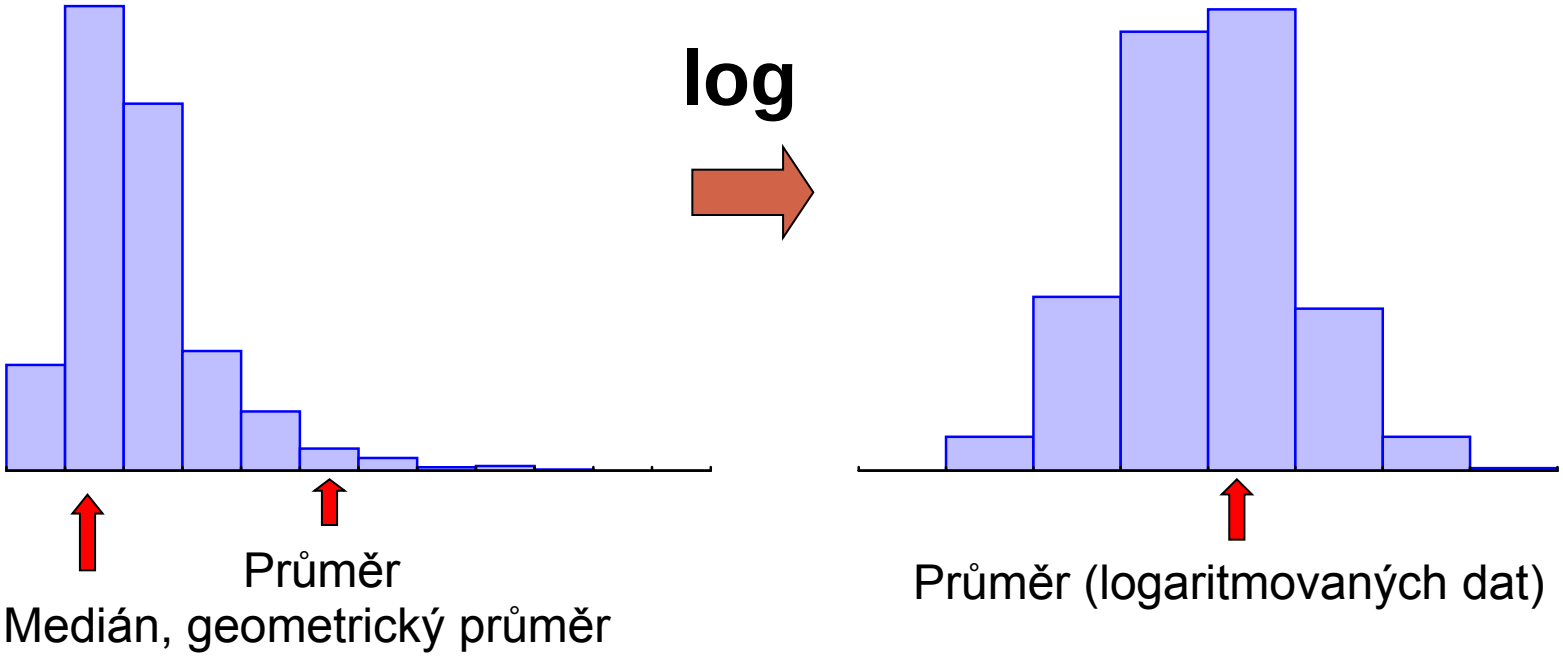
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné





Ukazatele středu rozložení II.

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

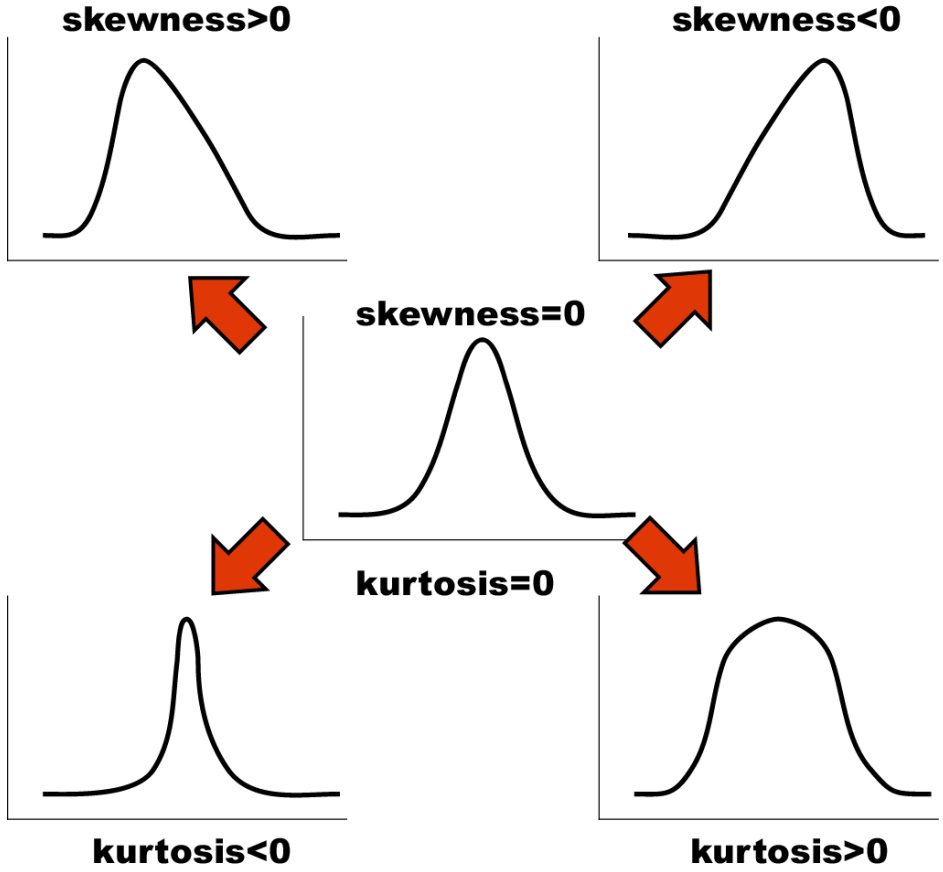


Ukazatele šířky rozložení

- **Rozptyl** je ukazatelem šířky rozložení 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 rozložení
- **Směrodatná odchylka** je druhá odmocnina z rozptylu
- **Koeficient variance** - podíl SD ku průměru (u normálního rozložení by se 95% hodnot mělo vejít do průměr ⌖3 SD), pokud je SD větší než 1/3 průměru jsou teoreticky pravděpodobné záporné hodnoty v rozložení – ukazatel problémů s normalitou dat

Ukazatele tvaru rozložení

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



Další parametry rozložení

- **Počet hodnot –** důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Střední chyba odhadu průměru** - je založena na směrodatné odchylce rozložení a **počtu hodnot**, vlastně jde o směrodatnou odchylku rozložení průměru. Říká jak přesný je náš výpočet průměru. Čím větší počet hodnot rozložení, 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)

IV.e Provádění odhadů

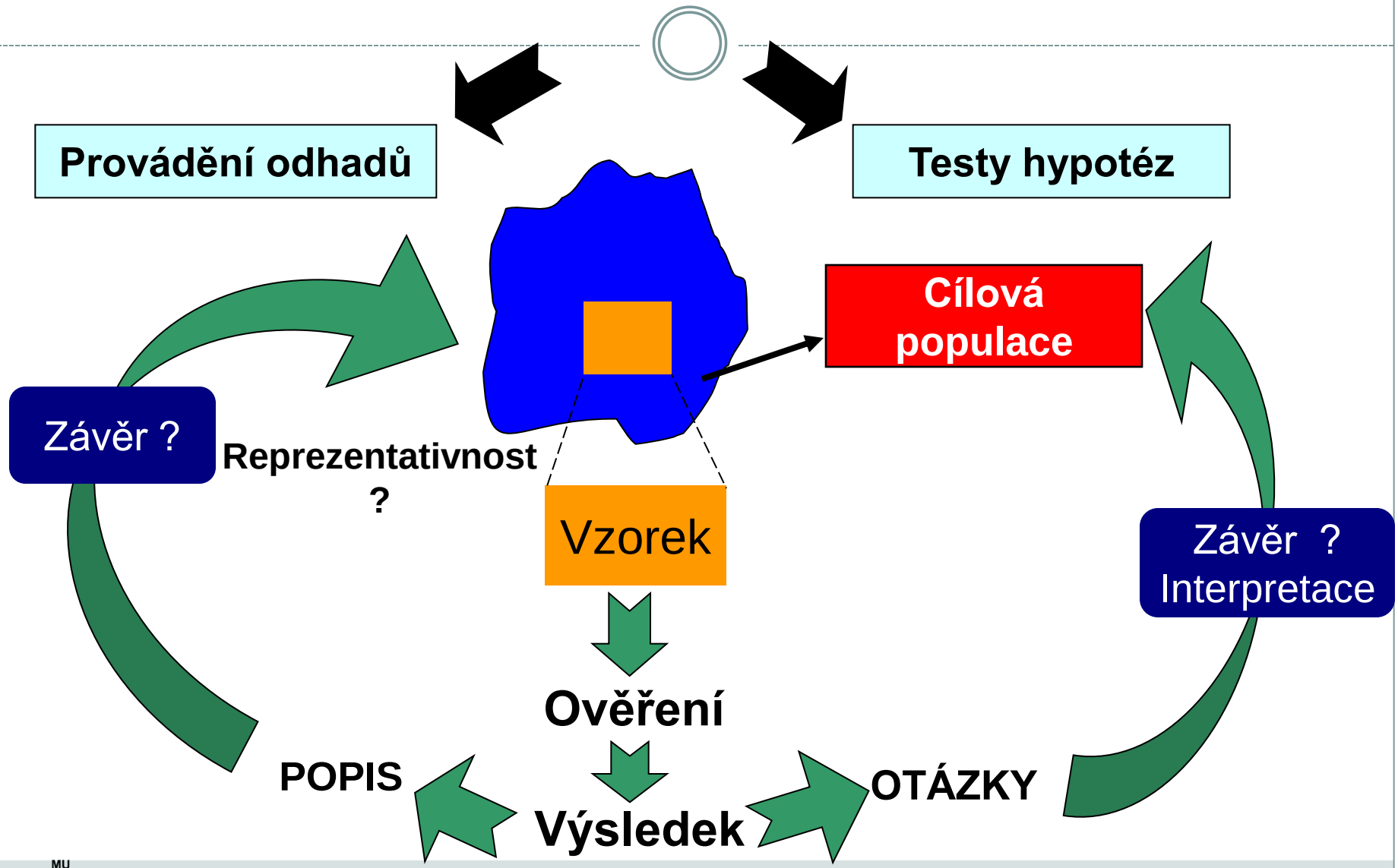
Bodové a intervalové odhady
Význam intervalu spolehlivosti

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

Anotace

- Dva základní přístupy statistického hodnocení jsou popis dat a testování hypotéz. Při popisu dat je třeba si uvědomit, že popisné statistiky získané ze vzorku nejsou skutečnou hodnotou v cílové populaci, ale pouze jejím odhadem. Přesnost odhadu závisí jednak na variabilitě dat, jednak na velikosti vzorku, při navzorkování celé cílové populace by výsledná popisná statistika již byla přesnou hodnotou, nikoliv odhadem.
- Odhady a s nimy související intervaly spolehlivosti jsou univerzálním statistickým postupem a je možné je dopočítat k libovolné popisné statistice.

Statistika v průzkumném studiu



INTERVAL SPOLEHLIVOSTI
velmi užitečná míra věrohodnosti odhadů

ODHADY

- Bodové
  - Číslo (chyba)
  - (Odhad parametru)
- Intervalové
  - **Interval pravděpodobných hodnot**
  - Spolehlivost
  - (Pravděpodobnostní interpretace)

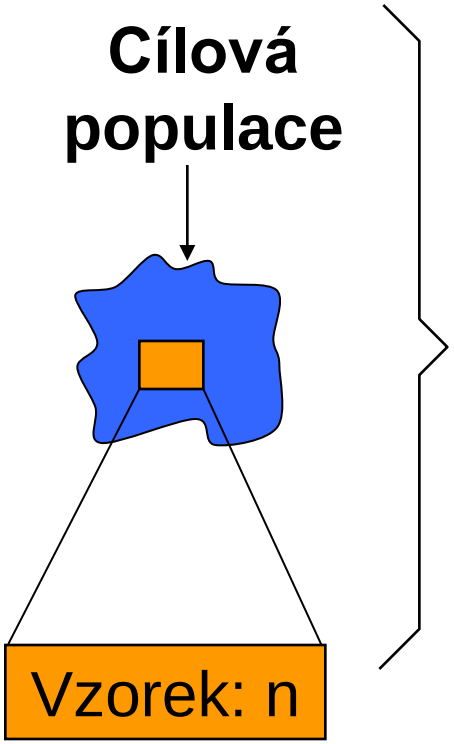
Obecný tvar:

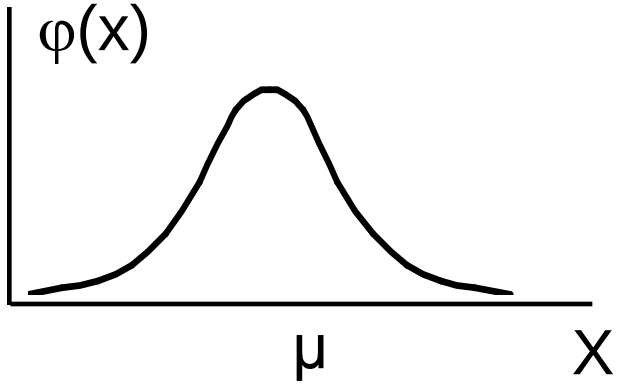
$$P\ (L_1 < \text{Odhad} < L_2) \geq 1 - \alpha/2$$

Odhadovaný parametr ± Kvantil modelového rozložení × SE (odhadu)

K_V pro $(1 - \alpha/2)$

NORMÁLNÍ ROZLOŽENÍ: model pro odhad průměru





Prezentace

n; $\bar{x}$; s

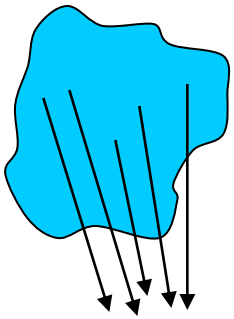
n; $\bar{x}$; $\frac{s}{\sqrt{n}}$

n; $\bar{x}$; c

n; $\bar{x}$; Interval spolehlivost i pro odhad průměru

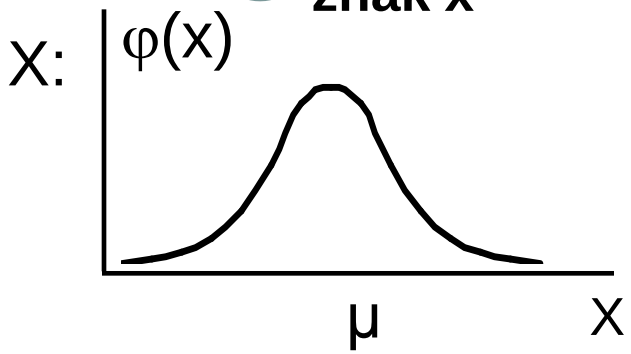
$\bar{x}$ ..... odhad průměru

NORMÁLNÍ ROZLOŽENÍ: odhad průměru je rovněž normálně rozložen



Cílová populace

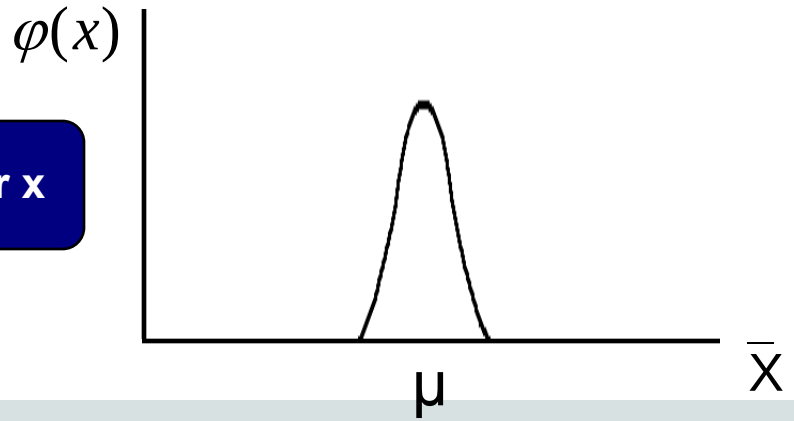
znak x



x: μ ± 3s

Náhodné výběry o n = 100

$\bar{x}1$ $\bar{x}2$ $\bar{x}3$ $\bar{x}4$ .... $\bar{x}i$



průměr x

$$\mu \pm 3 . \frac{s}{\sqrt{n}}$$

$\frac{s}{\sqrt{n}}$ ~ Standardní chyba odhadu průměru

ODHAD PRŮMĚRU: Vztahy

Bodový

$$\bar{x}; \left(\frac{s}{\sqrt{n}} \right)$$

Intervalový

$$\bar{x} - t_{1-\alpha/2}^{(\nu=n-1)} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}^{(\nu=n-1)} \cdot \frac{s}{\sqrt{n}}$$

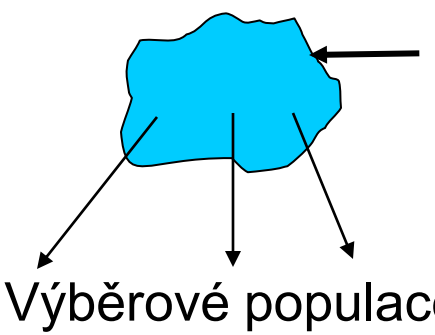
$$\mu : \bar{x} \pm t_{1-\alpha/2}^{(\nu=n-1)} \cdot \frac{s}{\sqrt{n}}$$

$$\mu : \bar{x} \pm t_{1-\alpha/2}^{(\nu=n-1)} \cdot s_{\bar{x}}$$

t ... příslušný kvantil Studentova rozložení
1 - α ... spolehlivost hodnoceného intervalu

Interval spolehlivosti odhadu průměru je pouze informací o přesnosti tohoto odhadu

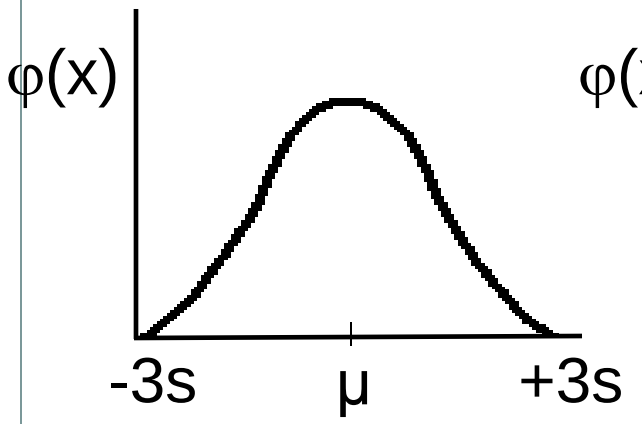
Interval spolehlivosti je hodnocen pro (1 - α) procentní spolehlivost



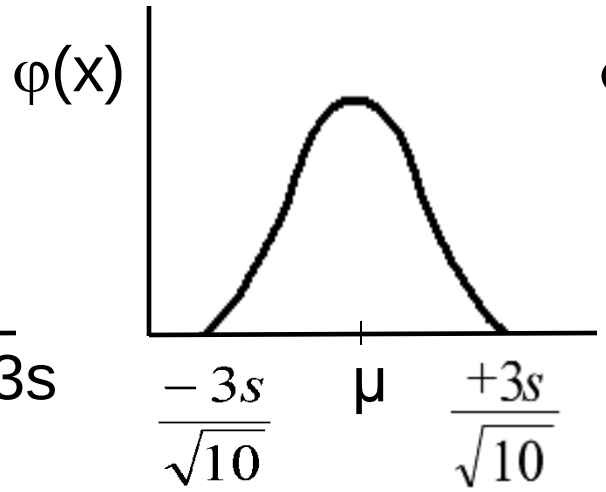
Šířku intervalu určuje:

a) velikost vzorku
b) rozptyl (variabilita) vzorku
c) požadovaná spolehlivost

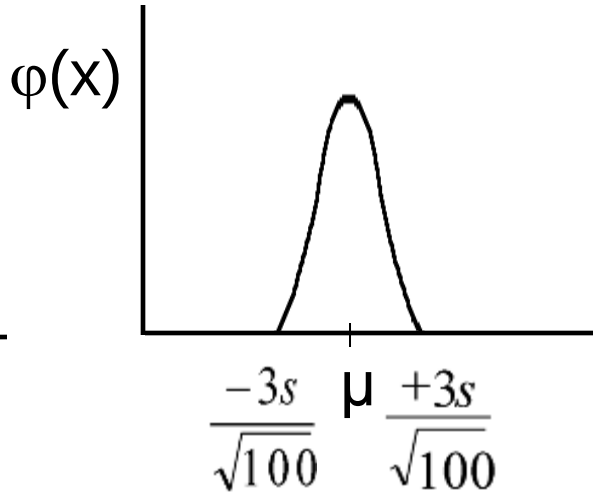
Původní proměnná x



Výběr n=10 pro odhad průměru



Výběr n=100 pro odhad průměru



ODHAD PRŮMĚRU: Příklad

X: Cena výrobku v n = 21 obchodech

Data:

$$n = 21; \bar{x} = 3{,}58; s^2 = 0{,}12$$

$$s_{\bar{x}} = \sqrt{0{,}12/21} = 0{,}075$$

95% Interval spolehlivosti:

$$t_{1-\alpha/2}^{(u = n-1)} = t_{0{,}975}^{(20)} = 2{,}086$$

$$\mu : \bar{x} \pm 2{,}086.s_{\bar{x}}$$

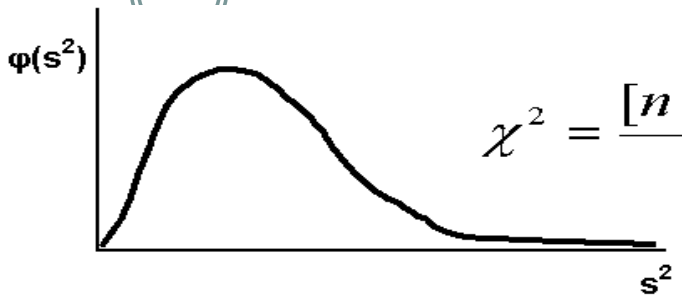
$$3{,}58 - 2{,}086.0{,}075 \le \mu \le 3{,}58 + 2{,}086.0{,}075$$

$$3{,}423 \le \mu \le 3{,}737$$

$$P\ (3{,}423 \le \mu \le 3{,}737) \ge 0{,}95$$

Interval spolehlivosti pro odhad rozptylu

$s^2 \sim \sigma^2$ pro velká n



$$\chi^2 = \frac{[n-1]s^2}{\sigma^2}$$

Interval spolehlivosti

a) pro σ^2 : $\frac{(n-1)s^2}{\chi^2_{\alpha/2}{}^{(n-1)}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}{}^{(n-1)}}$

b) pro σ : $\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}{}^{(n-1)}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}{}^{(n-1)}}}$

c) pro $\sigma/\sqrt{\mathrm{n}}$: $\sqrt{\frac{(n-1)s^2}{n\chi^2_{\alpha/2}{}^{(n-1)}}} \leq \frac{\sigma}{\sqrt{n}} \leq \sqrt{\frac{(n-1)s^2}{n\chi^2_{(1-\alpha/2)}{}^{(n-1)}}}$

$\sigma/\sqrt{n}$

-směrodatná odchylka odhadu průměru (S.E.)

Interval spolehlivosti pro odhad rozptylu: příklad

Příklad: měření produkce metabolitu (x) u buněk dvou nádorových linií

Linie 1
n = 50
$s^2(x) = 10$ (mg/ml)^2
$s(x) = 3{,}16$ mg/ml
$\bar{x} = 2$ mg/ml
$\bar{s}_x = 0{,}447$ mg/ml

95% IS

$$\frac{49 * 10}{77{,}22} \leq \sigma^2 \leq \frac{49 * 10}{31{,}56}$$

$$6{,}98 \leq \sigma^2 \leq 15{,}53$$

c = 1,58

Linie 1
n = 100
$s^2(x) = 16$ (mg/ml)^2
$s(x) = 4$ mg/ml
$\bar{x} = 2{,}8$ mg/ml
$\bar{s}_x = 0{,}4$ mg/ml

95% IS

$$\frac{99 * 16}{128{,}42} \leq \sigma^2 \leq \frac{99 * 16}{73{,}36}$$

$$12{,}33 \leq \sigma^2 \leq 13{,}49$$

c = 1,43

Výpočet mediánu z frekvenčních dat a jeho odhady

a) Určete medián tohoto souboru dat: 1,3,4,5,7,8 [4,5]

b) Určete medián tohoto souboru dat: 5,1,8,3,4 [4]

c) Tento příklad je ukázkou výpočtu mediánu u velkého souboru dat. V následující tabulce je uveden rozbor rozložení souboru dat od 179 krav, kde sledovanou veličinou byl počet dní od narození telete do znovuobnovení menstruačního cyklu. Uvedená data jsou velmi zjednodušena a jsou zde uvedena pouze pro ilustraci:

Class limits (days)	0,5-20,5	20,5-40,5	40,5-60,5	60,5-80,5	80,5-100,5	100,5-120,5	120,5-140,5	140,5-160,5	160,5-180,5	180,5-200,5	200,5-220,5
Frequency	8	33	50	32	15	20	11	6	2	1	1
Cumulative frequency	8	41	91	123	138	158	169	175	177	178	179

Frekvence zastoupení dosahuje nejvyšší hodnoty u třídy od 40,5 – 60,5 dnů. Druhý (menší) frekvenční pík lze pozorovat u intervalu od 100,5 do 120,5 dní. Existence dvou maxim (bimodální data) je důkazem nenormality tohoto konkrétního souboru.

Výpočet mediánu z frekvenčních dat a jeho odhady

Jelikož n =179, pak je medián devadesátá hodnota od počátku souboru, a dále je zřejmé, že bude velmi blízko horní hranici třídy 40,5 – 60,5 dní. Za předpokladu, že 50 hodnot této třídy je v ní rovnoměrně rozmístěno lze použít následující vzorec:

$$M = X_L + \frac{gl}{f}, kde$$

X_L = hodnota X (sledované veličiny) na spodní hranici třídy obsahující medián: zde 40,5 dní

g = pořadová hodnota mediánu minus kumulativní frekvence do horní hranice předchozí třídy, tj. 90 - 41= 49

l = třídní interval: 20 dní

f = frekvence ve třídě obsahující medián

Dosadíme-li do uvedeného vzorce, získáme odhad mediánu jako 60 dní. Průměr tohoto datového souboru je 69,9, což je významně odlišná hodnota, a potvrzuje znovu nenormální charakter dat.

U velkých vzorků z normálních populací je výběrový odhad mediánu normálně rozložen kolem populační hodnoty se směrodatnou odchylkou $1{,}253\sigma/\sqrt{n}$. U normálního rozložení, kde medián i průměr představují odhad stejné hodnoty, je medián méně přesný než průměr. Proto hlavní význam mediánu spočívá u nesymetrických distribucí.

Existuje velmi jednoduchá metoda pro výpočet intervalu spolehlivosti pro odhad mediánu a jako horní a spodní hranice slouží pořadová čísla vypočítaná podle následujícího vztahu:

$$\frac{(n+1)}{2} \pm \frac{z\sqrt{n}}{2}, kde$$

n představuje velikost datového souboru, **z** je kvantil standardizovaného normálního rozložení pro příslušnou pravděpodobnost. U našeho příkladu je n = 179 a pro 95% interval spolehlivosti je z přibližně rovno 2. Horní a spodní limit pro odhad mediánu tedy je $90 \pm \sqrt{179} = 77$ a 103. 95% interval spolehlivosti je tedy tvořen počty dní, které mají pořadí 77 a 103:

77: Počet dní = 40,5+(36)(20)/50 = 55 dní

103: Počet dní = 60,5+(12)(20)/32 = 68 dní

Medián cílové populace byl tedy odhadnut 95% intervalem spolehlivosti jako hodnota ležící mezi 55 a 68 dny. Interpretujte tento výsledek.

IV.f Základy testování hypotéz

Princip statistického testování hypotéz

Pojmy statistických testů

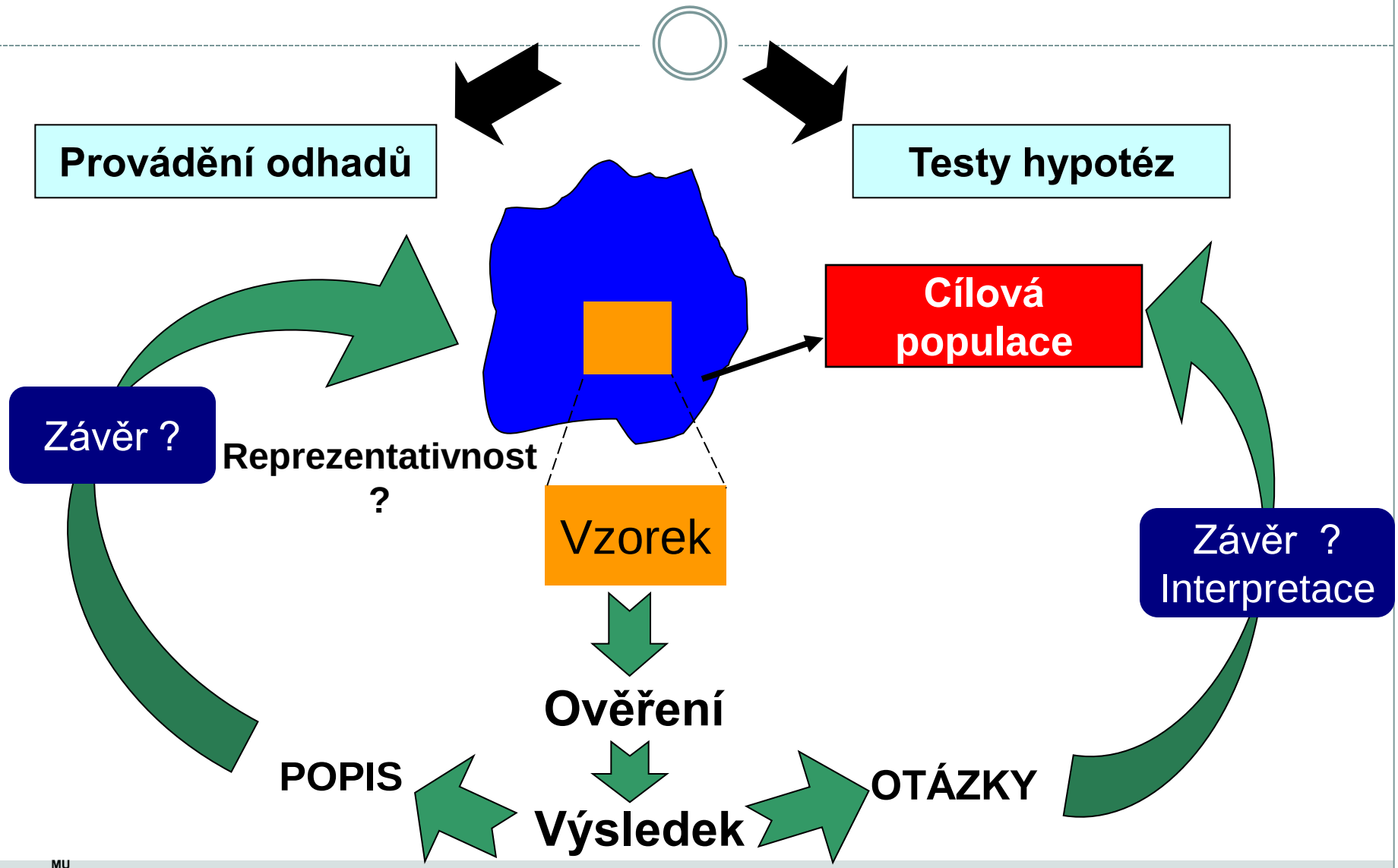
Normalita dat a její význam pro testování

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

Anotace

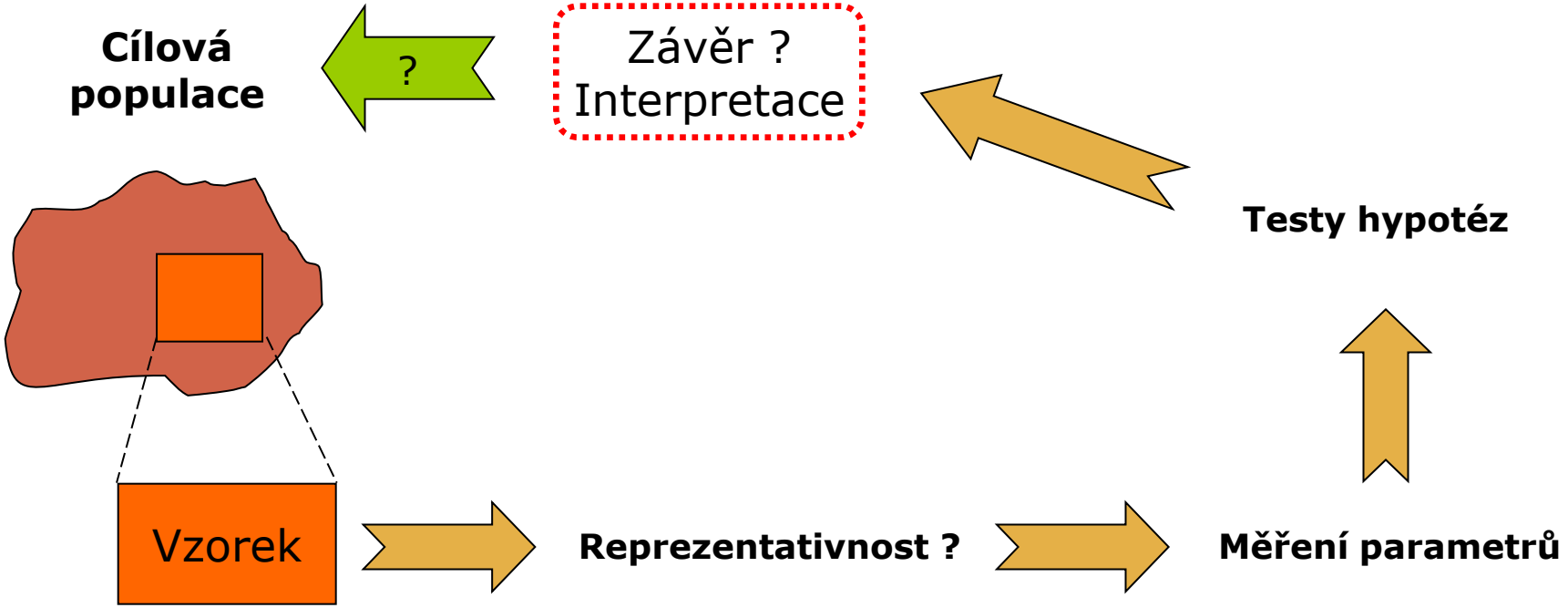
- Testování hypotéz je po popisné statistice druhým hlavním směrem statistických analýz. Při testování pokládáme hypotézy, které se snažíme s určitou pravděpodobností potvrdit nebo vyvrátit.
- Tzv. nulovou hypotézu lze nejlépe popsat jako situaci, kdy předpokládáme vliv náhody (rozdíl mezi skupinami je pouhá náhoda, vztah dvou proměnných je pouhá náhoda apod.), alternativní hypotéza předpokládá vliv nenáhodného faktoru.
- Výsledkem statistického testu je v zásadě pravděpodobnost nakolik je hodnocený jev náhodný nebo ne, při překročení určité hranice (nejčastěji méně než 5% pravděpodobnost, že jev je pouhá náhoda) deklarujeme, že pravděpodobnost náhody je pro nás dostatečně nízká abychom jev prohlásili za nenáhodný
- Statistická významnost je ovlivnitelná velikostí vzorku a tak je pouze indicií k prohlášení např. rozdílu dvou skupin pacientů za skutečně významný. V ideální situaci je nezbytné aby rozdíl byl významný nejenom statisticky (=nenáhodný), ale i prakticky (=nejde pouze o artefakt velikosti vzorku).

Statistika v průzkumném studiu



Princip testování hypotéz

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

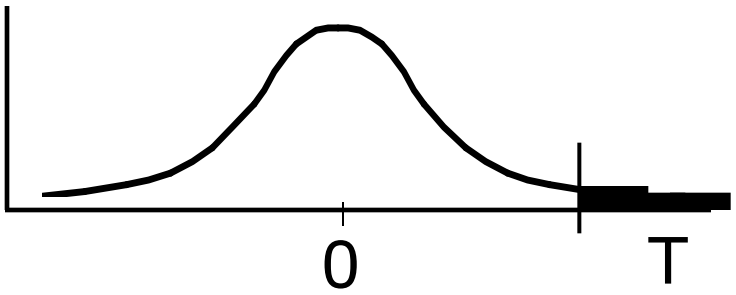


Statistické testování – základní pojmy

- **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.

Co znamená náhodný rozdíl?

Léčba

Placebo

X_1

X_2

Rozdíl ?

Je tu **rozdíl**?

Jak by vypadal rozdíl, kdyby byl **náhodný**?

Nasimulujme si ho !!! ☺

Mnoho-krát . . . .

X_1

X_2

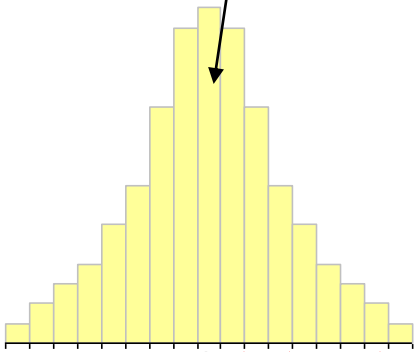
Rozdíl

X_1

X_2

Rozdíl

Rozložení možných náhodných rozdílů



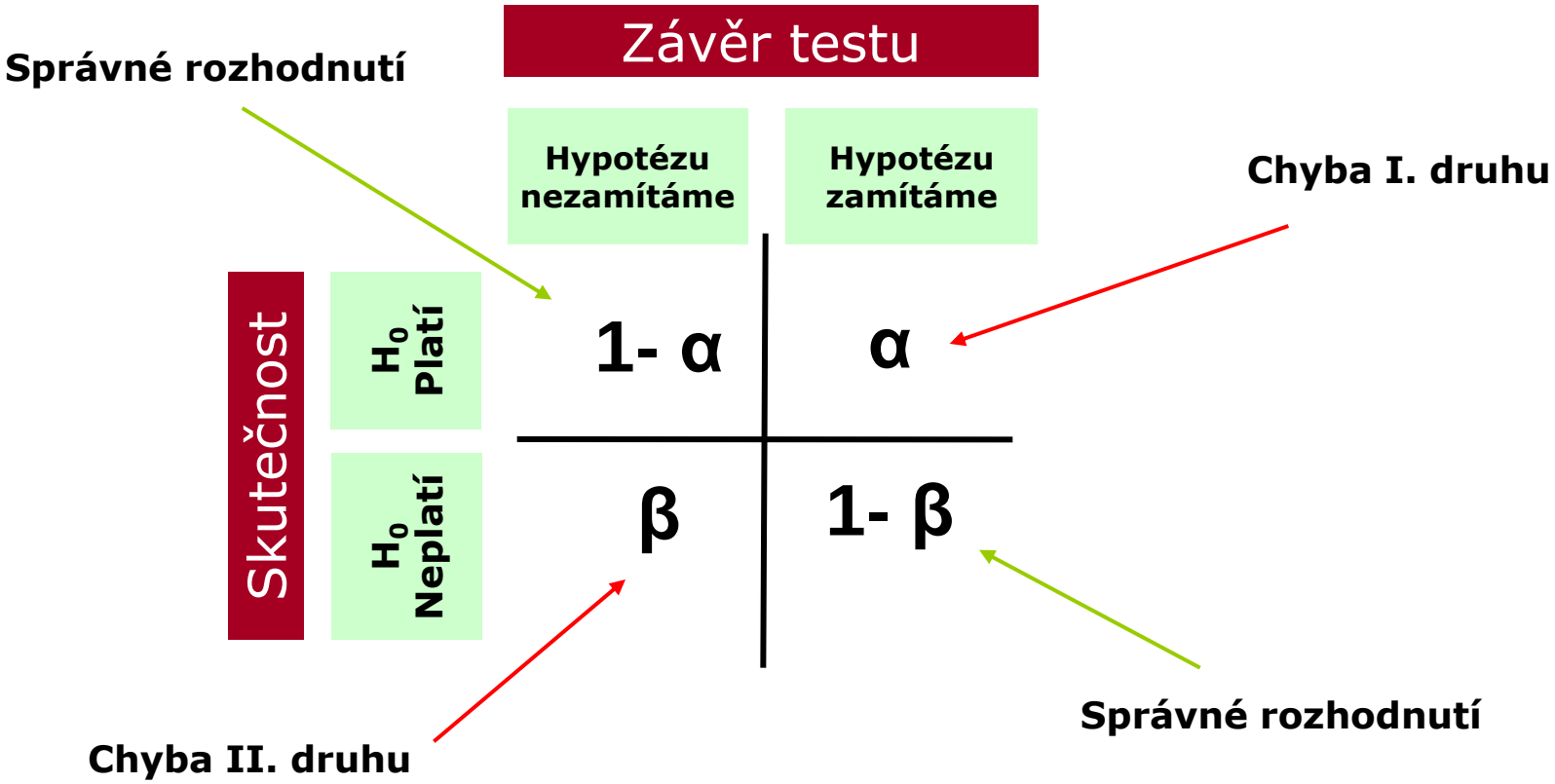
Rozdíl ?

Kde leží **skutečný rozdíl**?

Jak moc je **pravděpodobné**, že je **náhodný**?

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**

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

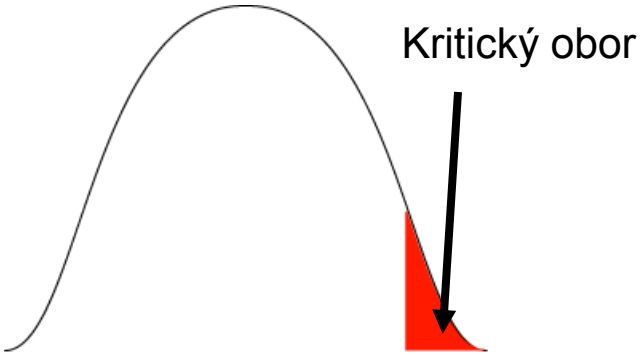
Dvou-výběrové testy (two-sample)

- 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

One-tailed vs. Two-tailed testy

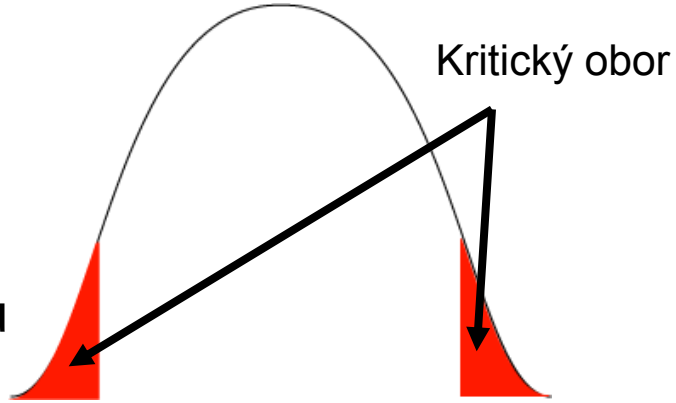
One – tailed testy

- 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



Two – tailed testy

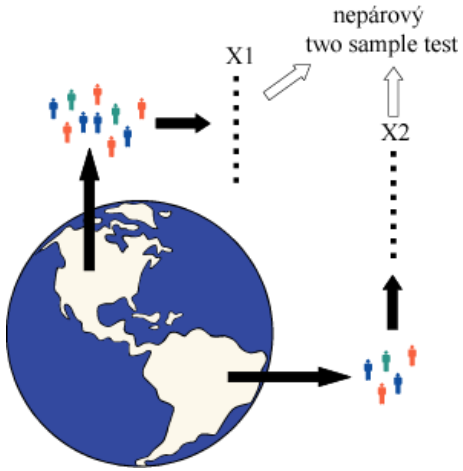
- Hypotéza testu se ptá na otázku **rovná se/nerovná se**
- Test může mít trojí výstup – **menší - rovná se – větší než**
- Situace **nerovná se** je tedy souhrnem dvou možných výstupů testu (**menší+větší**)



Nepárový vs. párový design

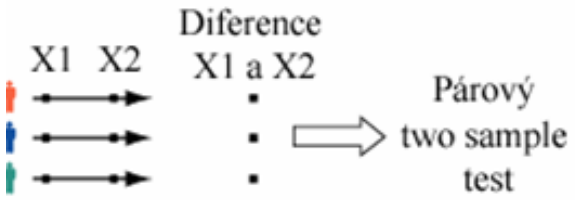
Nepárový design

- 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



Statistické testy a normalita dat

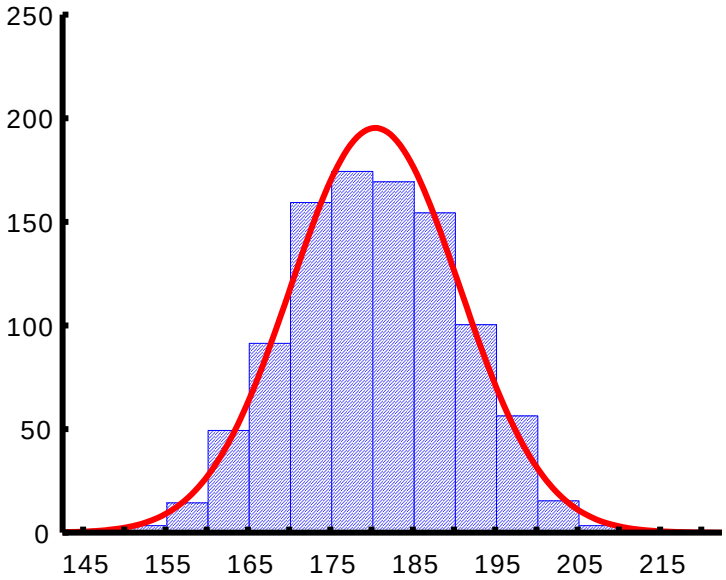
- **Normalita dat je jedním z předpokladů tzv. parametrických testů (testů založených na předpokladu nějakého rozložení) – např.** *t*-testy
- **Pokud data nejsou normální, neodpovídají ani modelovému rozložení, které je použito pro výpočet (*t*-rozložení) a test tak může lhát**

- **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 $\chi 2$ testu dobré shody. Test dává dobré výsledky, ale je náročný na n, tedy množství dat, aby bylo možné vytvořit dostatečný počet tříd hodnot.

•Kolgomorov Smirnov test

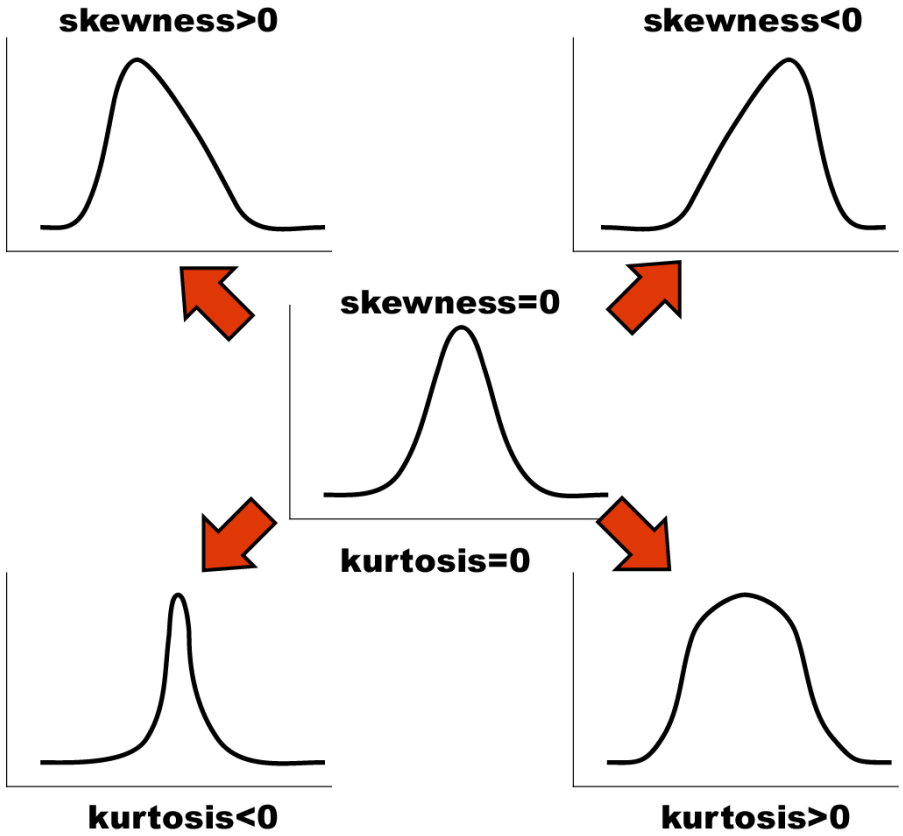
Tento test je často používán, dokáže dobře najít odlehlé hodnoty, ale počítá spíše se symetrií hodnot než přímo s normalitou. Jde o neparametrický test pro srovnání rozdílu dvou rozložení. Je založen na zjištění rozdílu mezi reálným kumulativním rozložením (vzorek) a teoretickým kumulativním rozložením. 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.

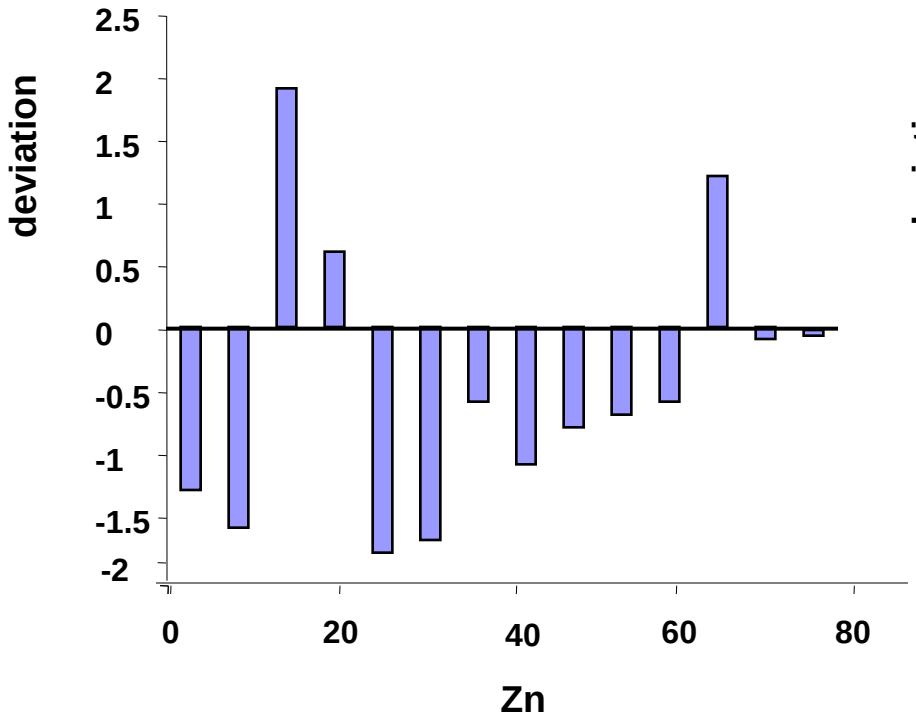
Šikmost a špičatost jako testy normality

- Parametry normálního rozložení, skewness a kurtosis mohou být využity pro testování normality, ale pouze pro velké vzorky (šikmost – 100, špičatost – 500).

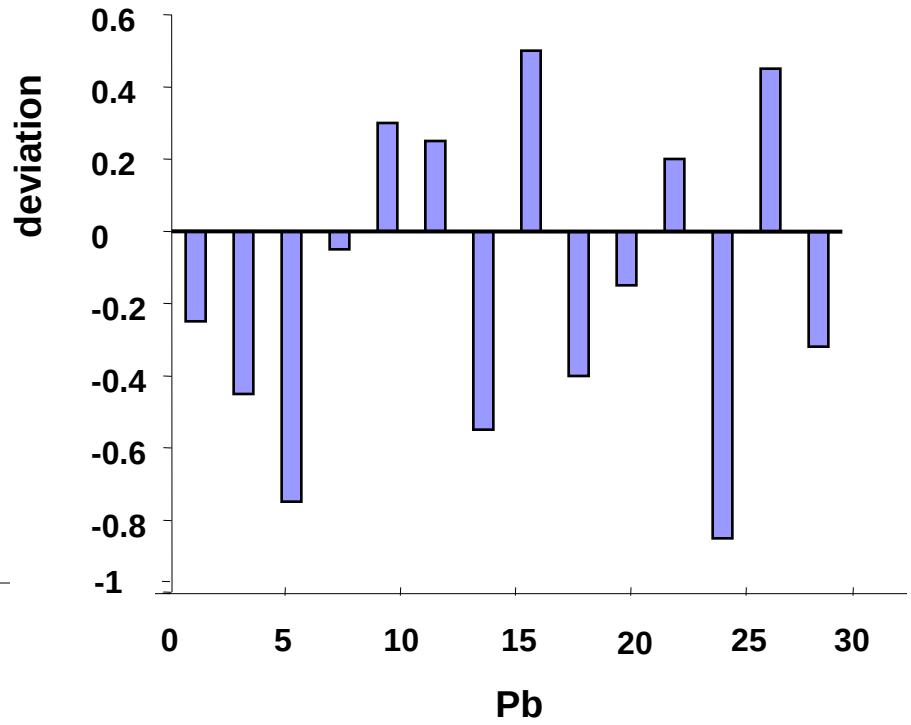


Grafická diagnostika normality

Rootgram

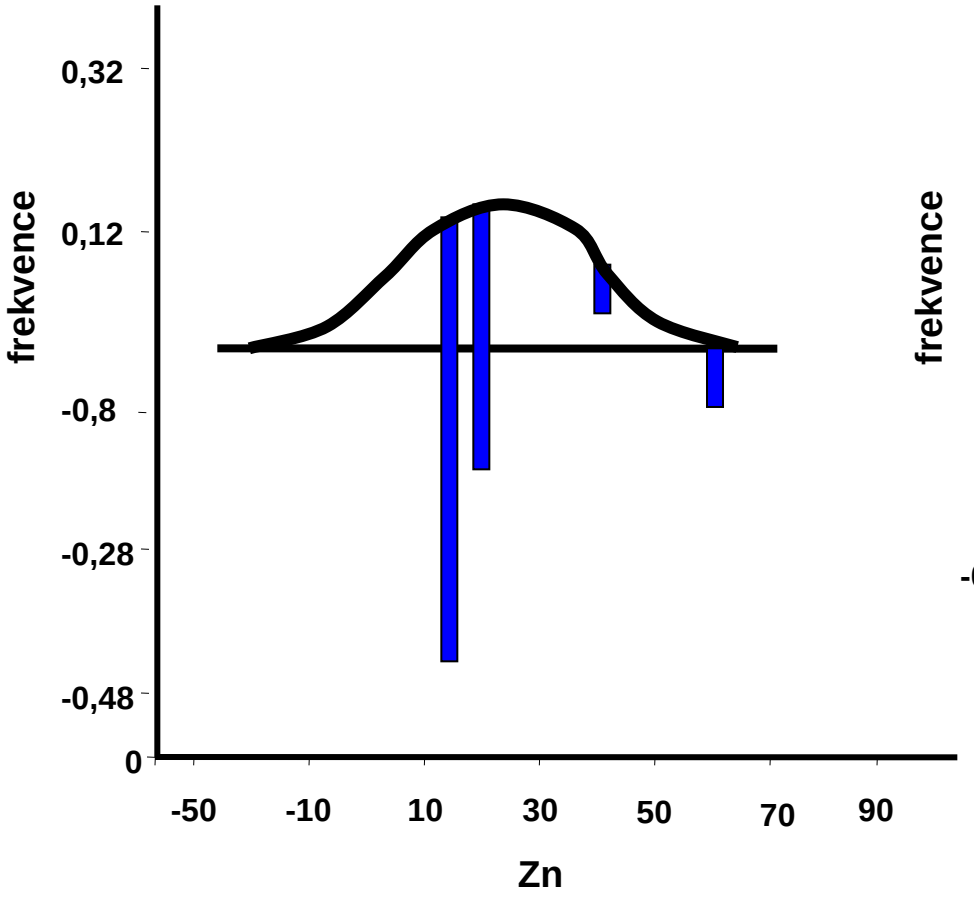


Rootgram

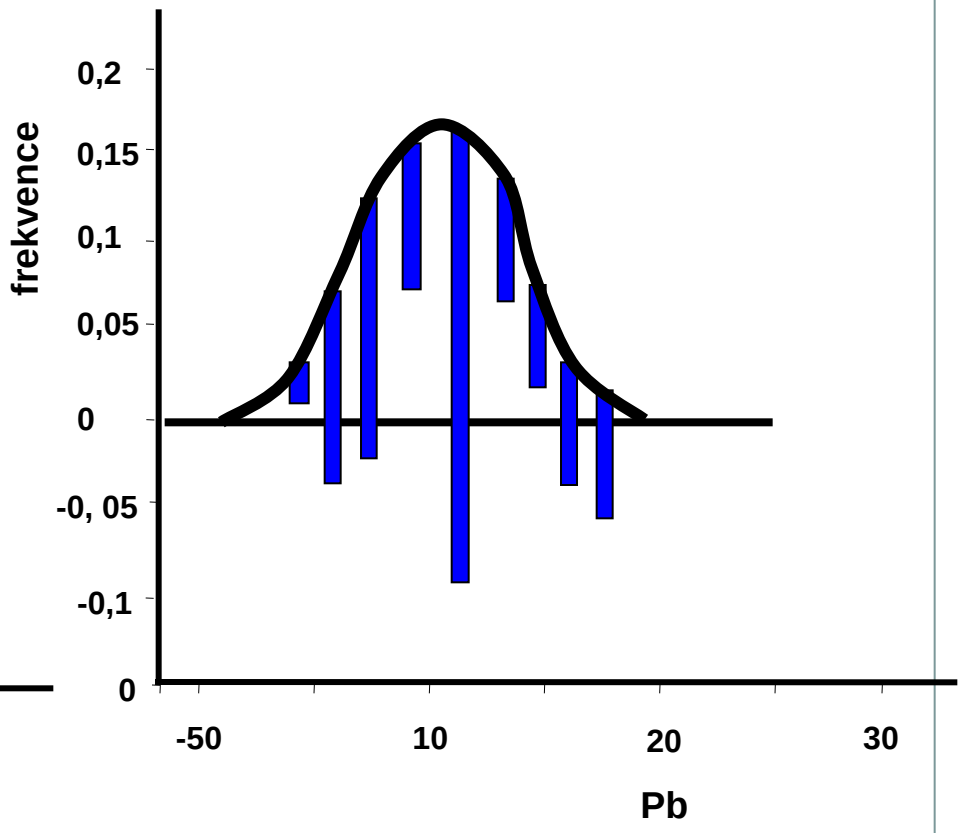


Grafická diagnostika normality

Hanging Histobars.

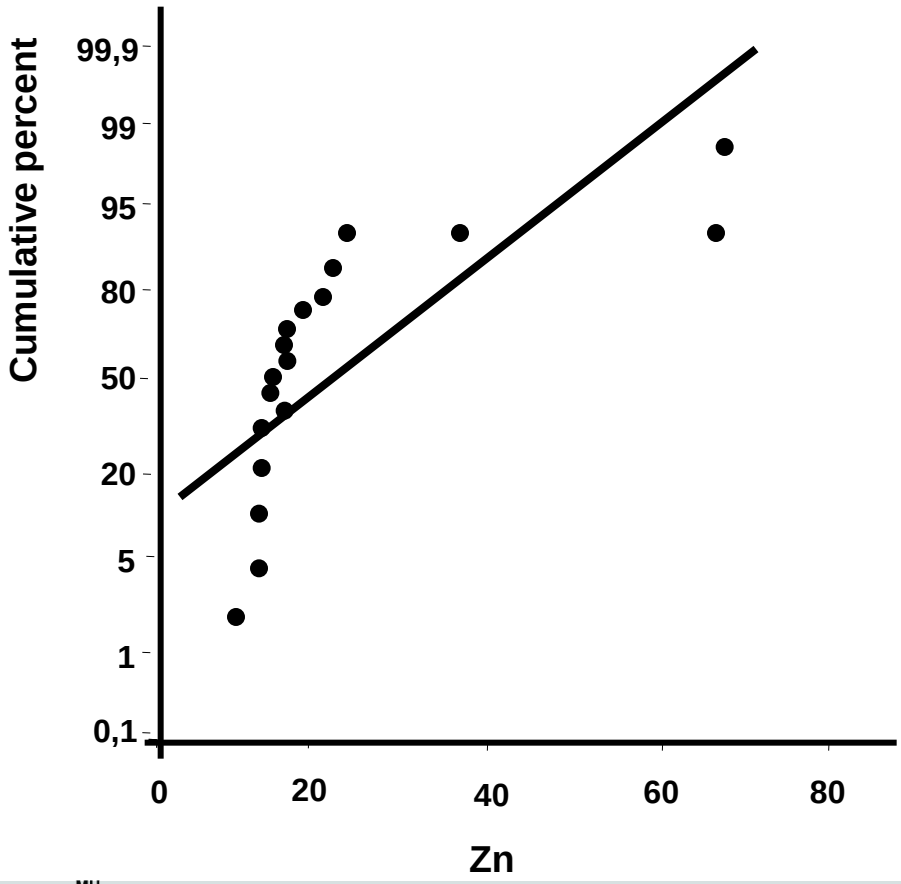


Hanging Histobars.

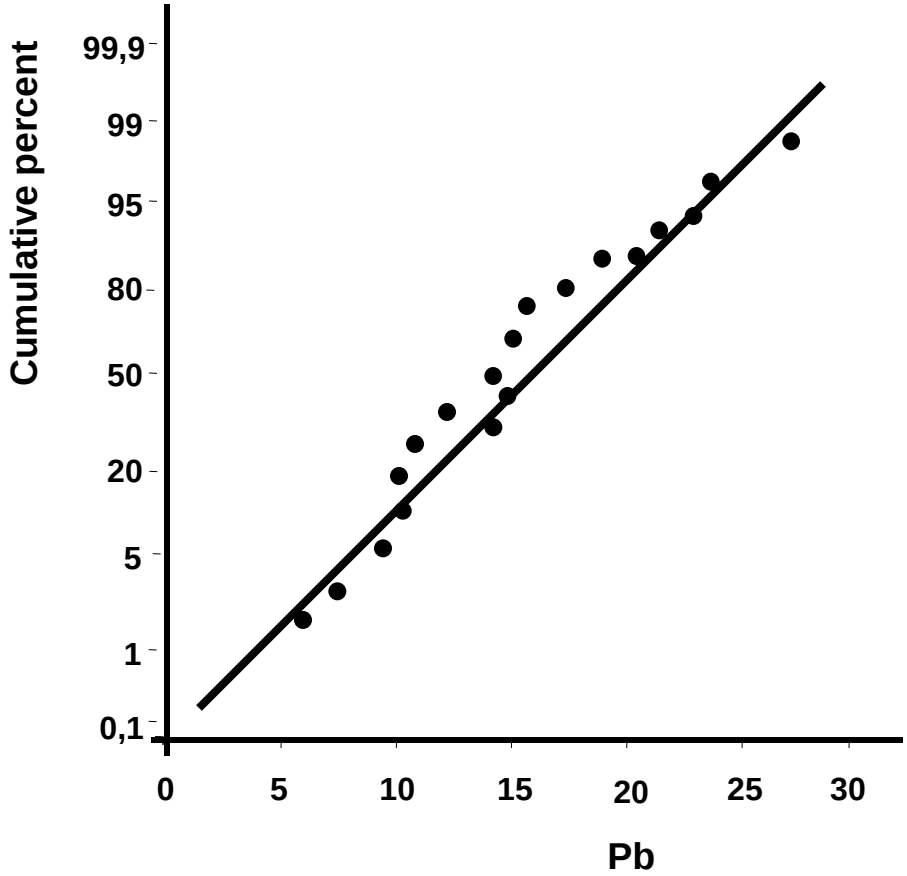


Grafická diagnostika normality

Normal Probability Plot



Normal Probability Plot



Grafická diagnostika normality