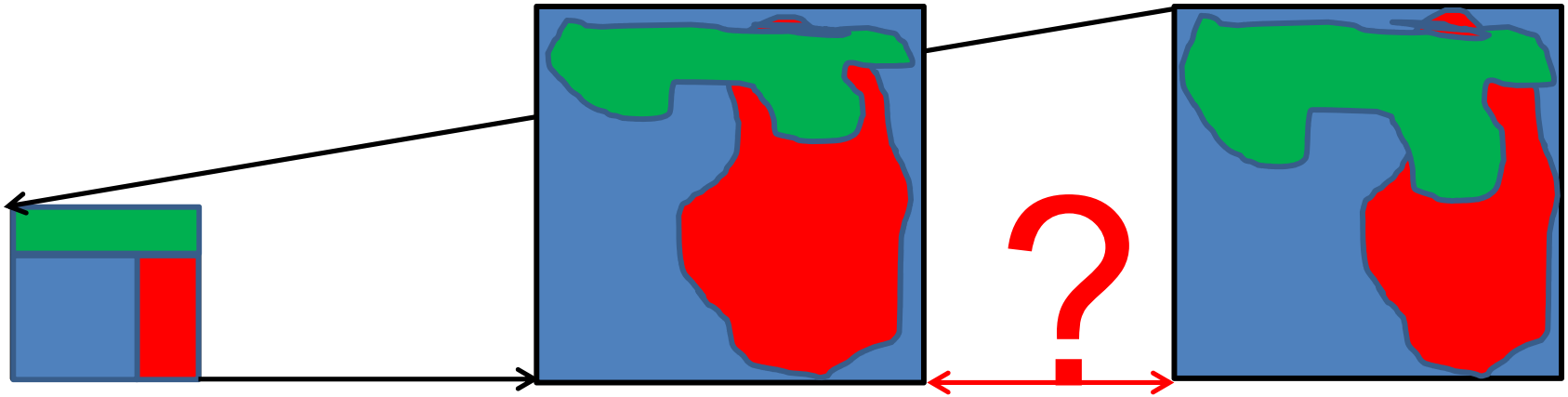
8. SEMINÁŘ

INDUKTIVNÍ STATISTIKA

1. ODHADY PARAMETRŮ

STATISTICKÁ INDUKCE

STATISTICKÁ INDUKCE



- Vlastnosti a složení výběrového souboru je přesně známé.
- Vlastnosti a složení základního souboru odhadujeme s určitou mírou nejistoty.
- Metody induktivní statistiky nejistotu neodstraňují, ale dokáží určit míru této nejistoty.

Základní a výběrový soubor

VÝBĚROVÝ SOUBOR

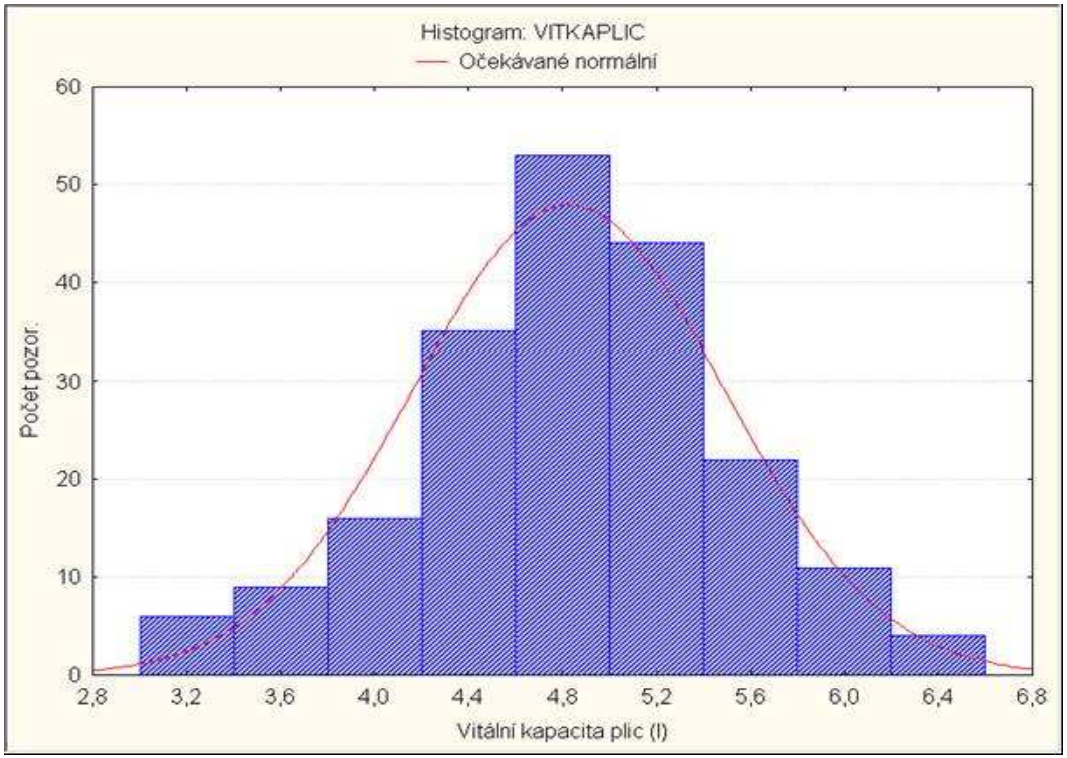
- reprezentativní náhodný výběr
- výběrové (empirické) rozdělení četností
- popis rozdělení: tabulka, graf
- stat. ukazatele = výběrové charakteristiky: **m, s, p** (ozn. latinkou)
- jsou to charakteristiky náhodných veličin a také se jako náhodné veličiny chovají, tzn. mění se výběr od výběru

ZÁKLADNÍ SOUBOR

- soubor, který nás zajímá
- teoretické rozdělení četností (matematický model)
- popis rozdělení: pravděpodobnostní rozdělení
- stat. ukazatele = parametry: **μ, σ, π** (ozn. řeckou abecedou)
- jsou to neměnné konstanty, zpravidla neznámé, pro $n \to \infty$ platí, že $m \to \mu$, $s \to \sigma$, $p \to \pi$.

Empirické rozdělení četností

- Měříme-li veličinu ve výběrovém souboru, pak rozložení hodnot této veličiny znázorňujeme na základě **empiricky** zjištěných četností (histogram).
- Jsou popsány četnosti, se kterými se naměřené hodnoty vyskytovaly ve výběrovém souboru.



Základní a výběrový soubor

VÝBĚROVÝ SOUBOR

- reprezentativní náhodný výběr
- výběrové (empirické) rozdělení četností
- popis rozdělení: tabulka, graf
- stat. ukazatele = výběrové charakteristiky: **m, s, p** (ozn. latinkou)
- jsou to charakteristiky náhodných veličin a také se jako náhodné veličiny chovají, tzn. mění se výběr od výběru

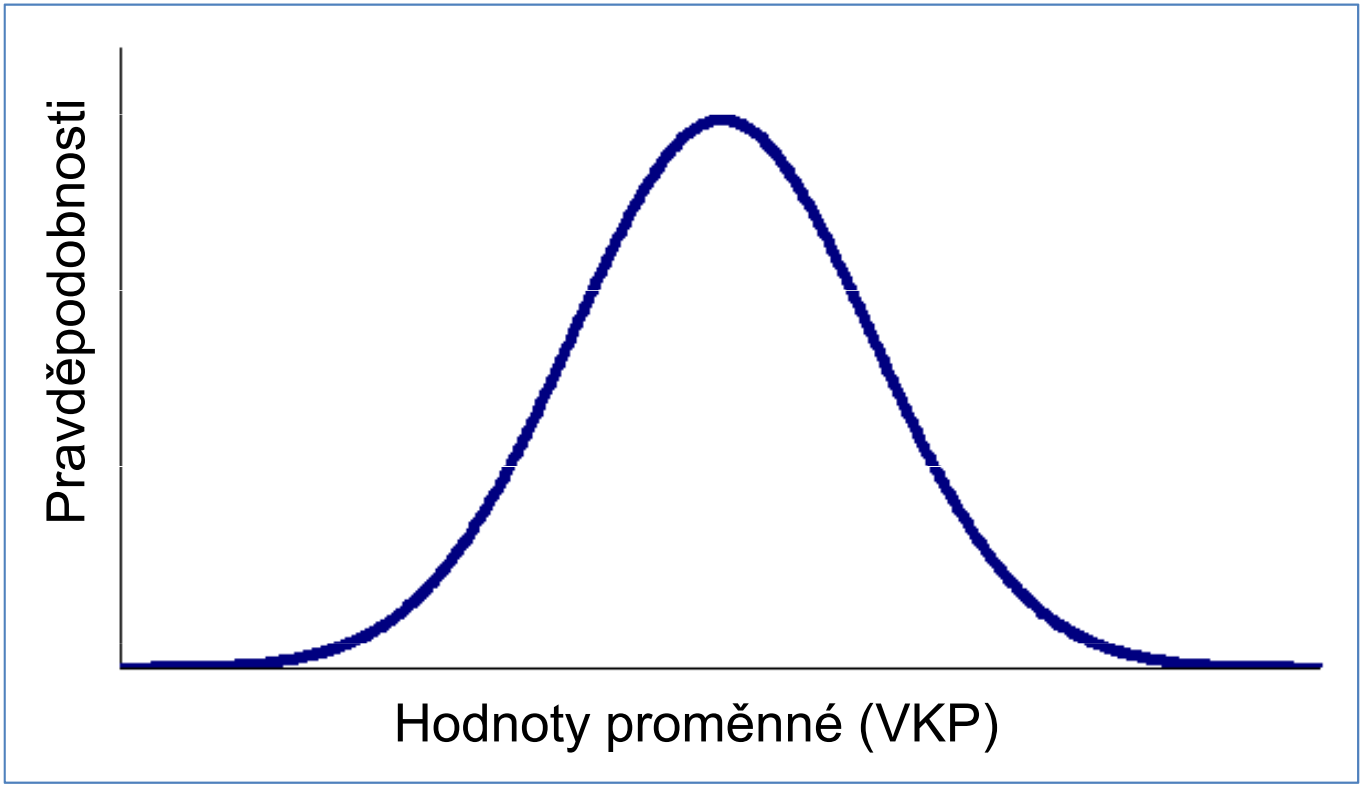
ZÁKLADNÍ SOUBOR

- soubor, který nás zajímá
- teoretické rozdělení četností (matematický model)
- popis rozdělení: pravděpodobnostní rozdělení
- stat. ukazatele = parametry: **μ, σ, π** (ozn. řeckou abecedou)
- jsou to neměnné konstanty, zpravidla neznámé, pro $n \to \infty$ platí, že $m \to \mu$, $s \to \sigma$, $p \to \pi$.

Pravděpodobnostní rozdělení

- Každá veličina má své **pravděpodobnostní (teoretické) rozdělení**.

Na ose y jsou **pravděpodobnosti** vyjadřující očekávání, jak často se budou jednotlivé hodnoty vyskytovat v nekonečně velkém souboru.



Na **ose x** jsou všechny **hodnoty**, kterých může veličina potenciálně nabývat ($\pm\infty$).

Empirické a pravděpodobnostní rozdělení

- Od empirického k pravděpodobnostnímu rozdělení (postupné vyhlazování histogramu)

- **Pravděpodobnostní rozdělení**
  - znázorňují procesy a jevy našeho každodenního života pomocí matematických modelů.
  - Tyto modely vyjadřují v matematické formě jejich zákonitosti a umožňují tak hlubší poznání zákonitostí těchto jevů.

Základní a výběrový soubor

VÝBĚROVÝ SOUBOR

- reprezentativní náhodný výběr
- výběrové (empirické) rozdělení četností
- popis rozdělení: tabulka, graf
- stat. ukazatele = výběrové charakteristiky: **m, s, p** (ozn. latinkou)
- jsou to charakteristiky náhodných veličin a také se jako náhodné veličiny chovají, tzn. mění se výběr od výběru

ZÁKLADNÍ SOUBOR

- soubor, který nás zajímá
- teoretické rozdělení četností (matematický model)
- popis rozdělení: pravděpodobnostní rozdělení
- stat. ukazatele = parametry: **μ, σ, π** (ozn. řeckou abecedou)
- jsou to neměnné konstanty, zpravidla neznámé, pro $n \to \infty$ platí, že $m \to \mu$, $s \to \sigma$, $p \to \pi$.

Typy pravděpodobnostních rozdělení

Diskrétní veličiny

- binomické rozdělení (jev – nejev)
- rovnoměrné rozdělení
- Poissonovo rozdělení (vzácné jevy)

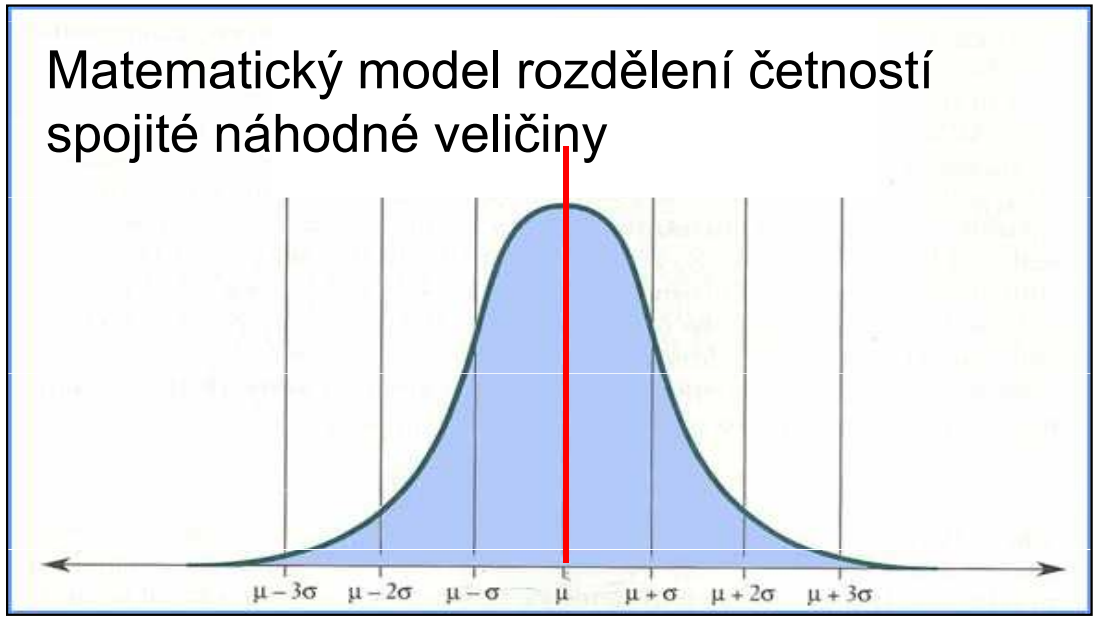
Spojité veličiny

- normální rozdělení
- Studentovo t-rozdělení
- Snedecorovo F-rozdělení
- Chí-kvadrát rozdělení

Pozn.:

- S veličinou zacházíme jako s normálně rozdělenou, pokud nemáme dostatečné důvody pro vyvrácení této domněnky.
- Rozložení většiny veličin lze převést na normální rozdělení.

NORMÁLNÍ ROZDĚLENÍ (GAUSSOVA KŘIVKA)

- 

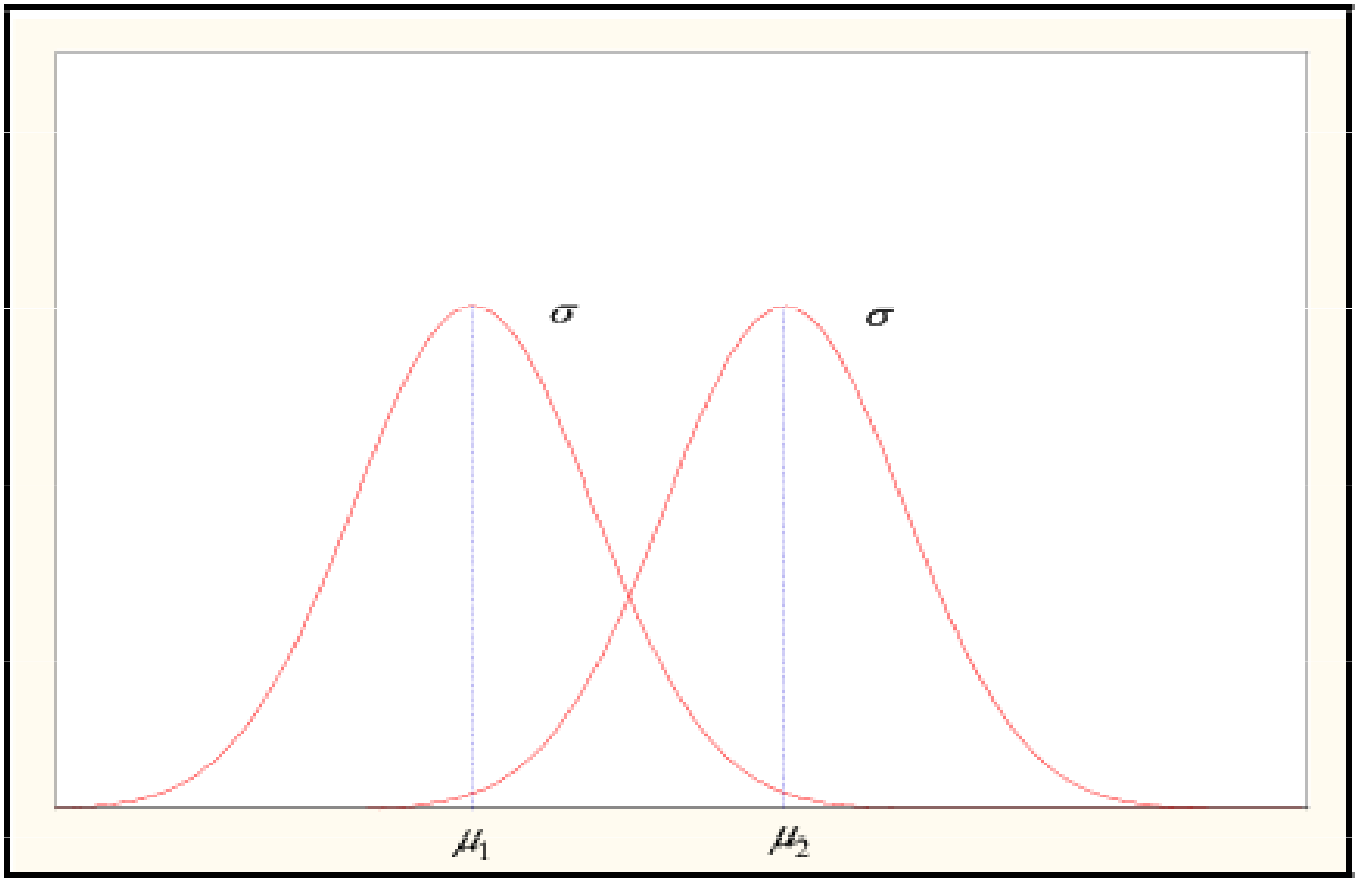
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Frekvenční křivka NR je jednoznačně určena dvěma parametry: **μ** a **σ**.
  - μ - určuje polohu křivky na ose x (analogie m)
  - σ - určuje tvar křivky (analogie s)

- Symetrické jednovrcholové rozdělení četností, parametr **μ** = průměr a zároveň nejčetnější hodnota, která půlí plochu pod křivkou na dvě stejně velké části

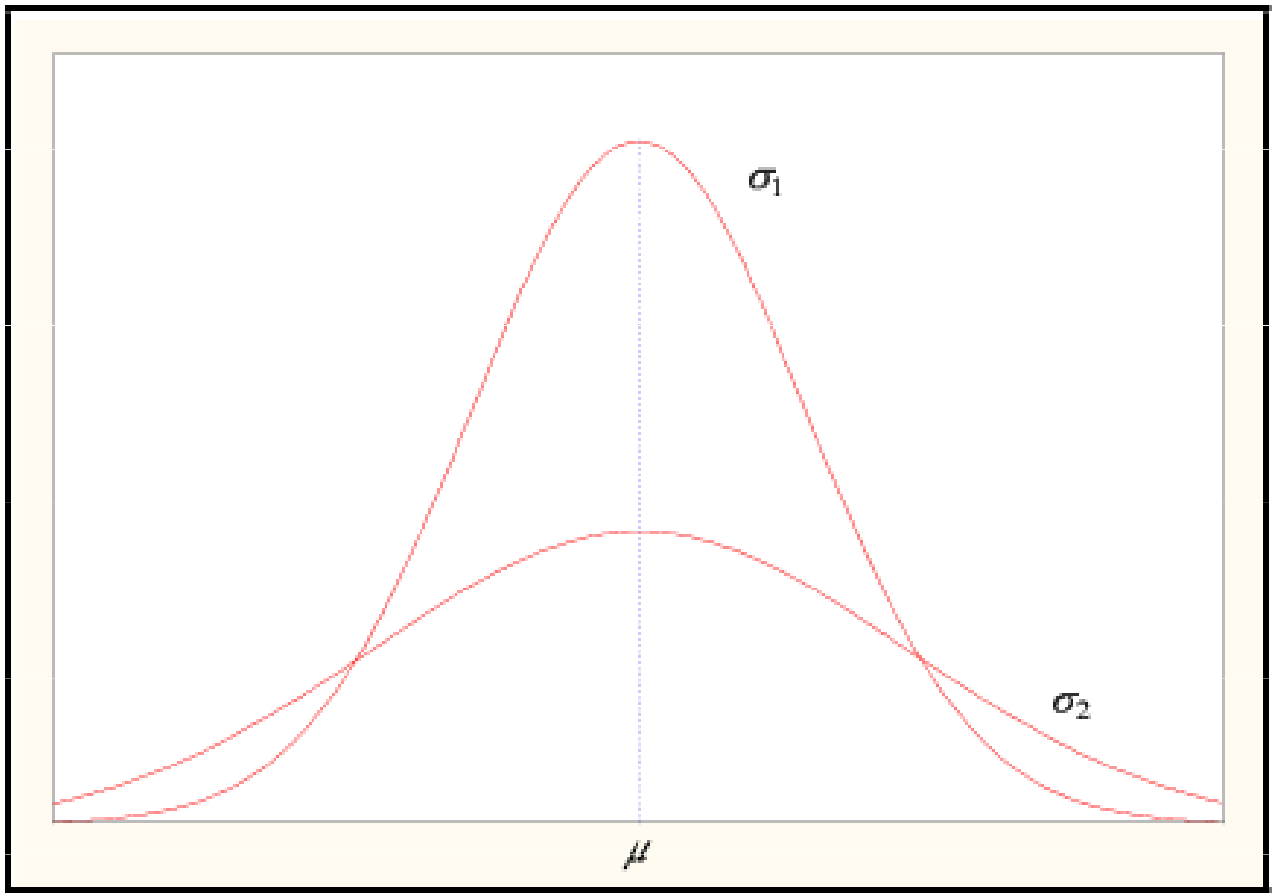
NORMÁLNÍ ROZDĚLENÍ

- μ1 □ μ2; σ1 = σ2

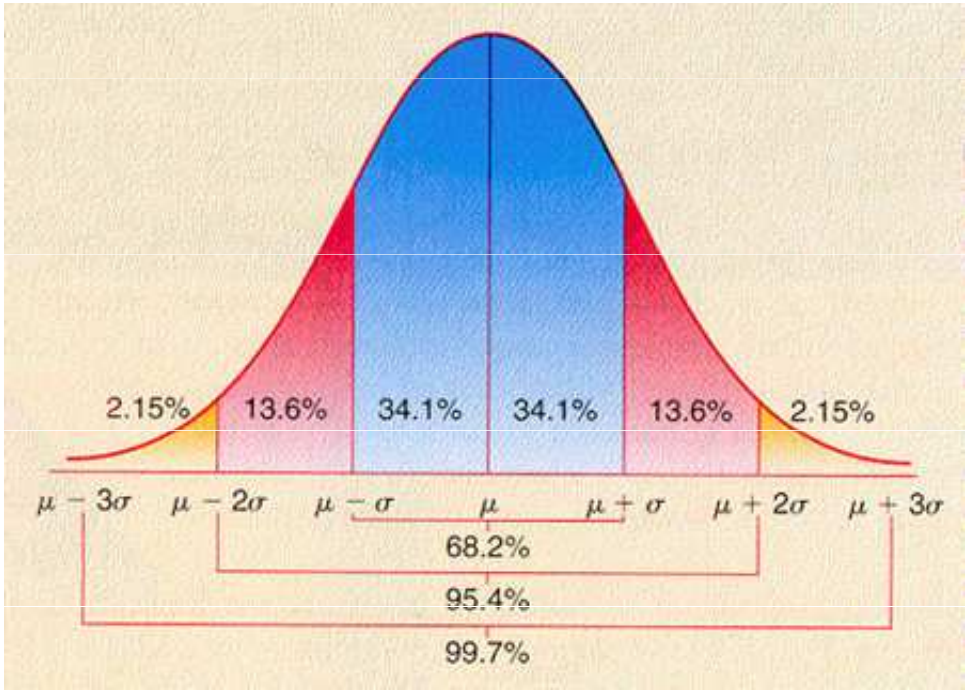


NORMÁLNÍ ROZDĚLENÍ

- μ1 = μ2; σ1 □ σ2

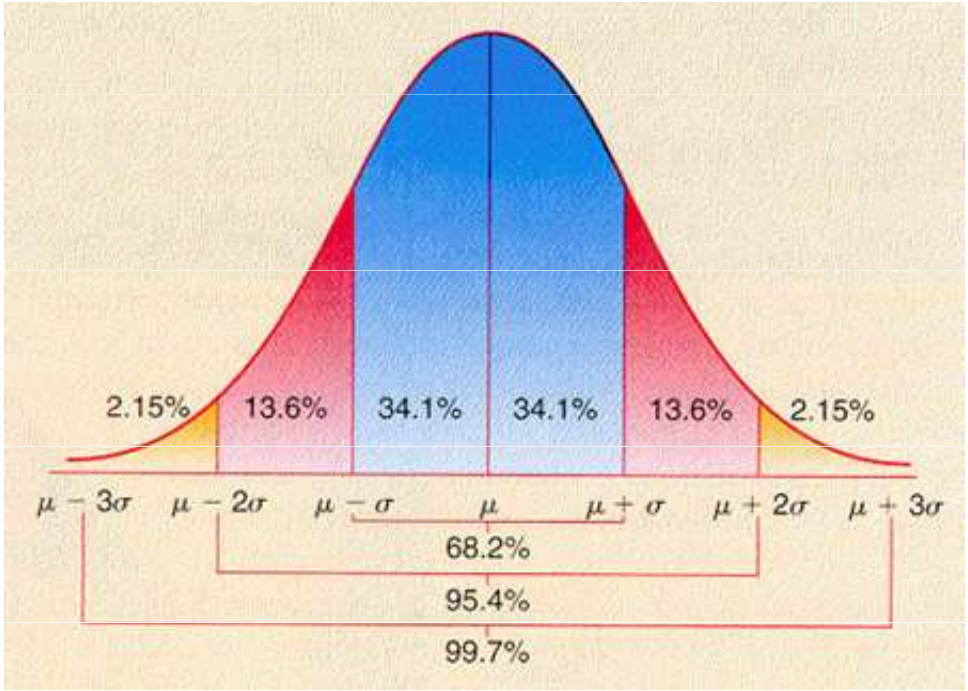


VLASTNOSTI NORMÁLNÍHO ROZDĚLENÍ



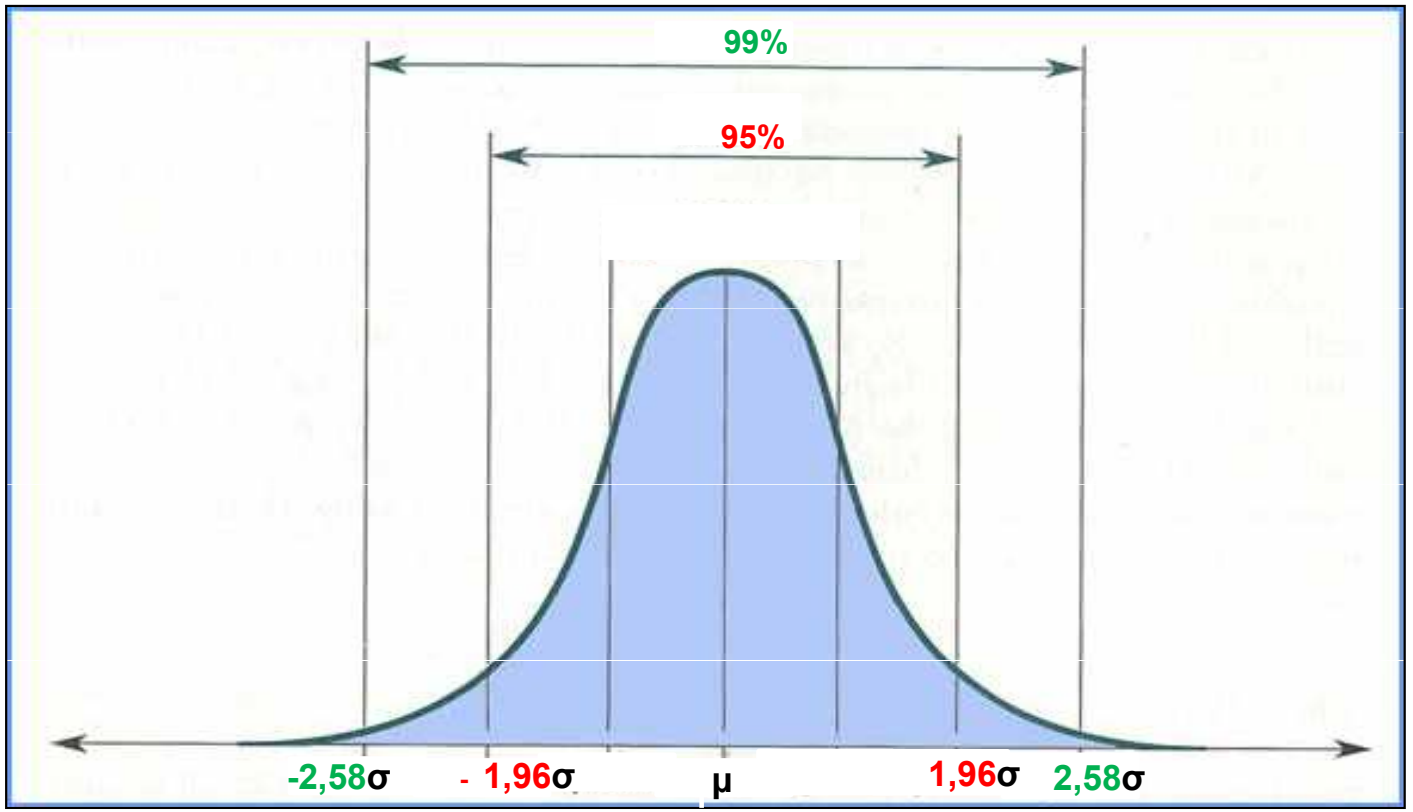
- Frekvenční křivky normálního rozdělení mají různý tvar a polohu pro různé veličiny. Pro všechny však platí, že v intervalu:
  - $\mu \pm 1\ \sigma$ leží 68,2% všech hodnot, kterých může daná veličina nabývat
  - $\mu \pm 2\ \sigma$ leží 95,4% všech hodnot, kterých může daná veličina nabývat
  - $\mu \pm 3\ \sigma$ leží 99,7% všech hodnot, kterých může daná veličina nabývat

VLASTNOSTI NORMÁLNÍHO ROZDĚLENÍ



- Častěji nás ale zajímá, v jakém intervalu leží 95% (příp. 99%) hodnot sledované veličiny
  - můžeme pak totiž vyjádřit tvrzení, že s pravděpodobností 95% (99%) se hodnoty sledované veličiny nacházejí právě v tomto intervalu, resp., že 95% hodnot, kterých sledovaná veličina nabývá, leží v tomto intervalu.
  - takový interval je vymezen tzv. **kritickými hodnotami (konstantami)** normálního rozdělení

KRITICKÉ HODNOTY NORMÁLNÍHO ROZDĚLENÍ



- **Kritické hodnoty** normálního rozložení: **1,96**σ a **2,58**σ.
- **V intervalu (μ - 1,96σ; μ + 1,96σ) se nachází 95 % všech možných hodnot sledované veličiny.**
- **V intervalu (μ – 2,58σ; μ + 2,58σ) se nachází 99% všech možných hodnot sledované veličiny.**

ODHADY PARAMETRŮ

- **Bodové odhady**
- **Intervalové odhady**

BODOVÉ ODHADY

$\mu \approx m$ $\pi \approx p$

$\sigma^2 \approx s^2$ relativní riziko v ZS ≈ RR

$\sigma \approx s$ $\rho \approx r$

- **Požadavky na bodové odhady**

(tzv. nejlepší nestranný bodový odhad)

a) **konzistence**

- s rostoucím VS se výběrová charakteristika více blíží k parametru

b) **nestrannost**

- odhady parametru provedené na základě různých VS kolísají kolem hodnoty neznámého parametru na obě strany

c) **minimální rozptyl**

- uvedené kolísání musí být co nejmenší

- **Nevýhody bodových odhadů**

- neznáme jejich **spolehlivost** a **přesnost**

INTERVALOVÉ ODHADY

- Neznámý parametr odhadujeme **intervalem** vytvořeným **kolem** tzv. nejlepšího nestranného **bodového odhadu**.
- **Interval spolehlivosti** (konfidenční interval)
- **Spolehlivost** určujeme sami, obvykle 95% nebo 99%
  - jde o **pravděpodobnost**, že odhadovaný parametr se nachází v daném intervalu.

- zápis: **95% CI (**dolní hranice **;** horní hranice**)**

  99% CI (dolní hranice **;** horní hranice**)**

INTERVALOVÉ ODHADY

Doplněk spolehlivost vyjadřuje tzv. **riziko odhadu** – tj. riziko, že odhadovaný parametr leží mimo interval:

- při spolehlivosti 95% je riziko odhadu 5%,
- při spolehlivosti 99% je riziko odhadu 1%.

ODHAD PRŮMĚRU ZÁKLADNÍHO SOUBORU (PARAMETRU μ)

CENTRÁLNÍ LIMITNÍ VĚTA

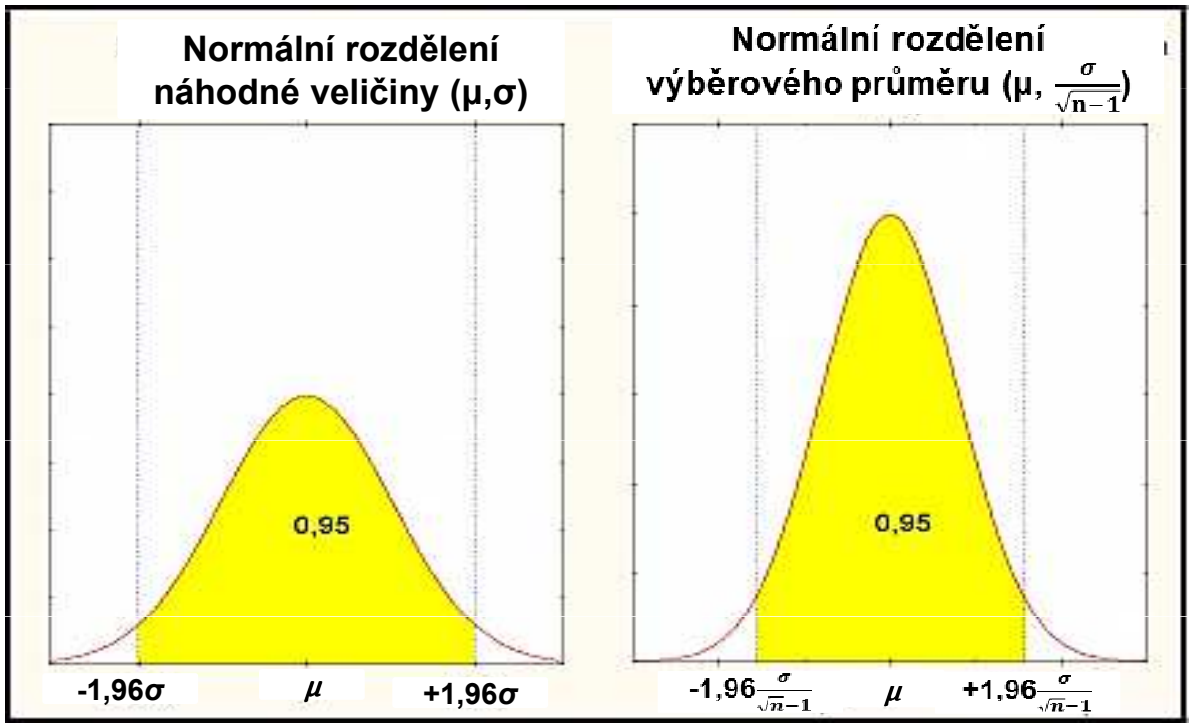
1. Nejlepší bodový odhad parametru **μ** je výběrový průměr **m.**
2. V souborech, kde **n > 30**, se výběrový průměr chová jako náhodná veličina, která má **normální rozdělení,** a to i v případě, že veličina, ze které je průměr vypočítán, normální rozdělení nemá.
3. **Standardní chyba** průměru σ_m je $\frac{\sigma}{\sqrt{n-1}}$ a odhaduje se ze vztahu:

   $SE_m = \frac{s}{\sqrt{n-1}}$

   Závěr: 95% CI $m \pm 1,96 \cdot \frac{s}{\sqrt{n-1}}$

   99% CI $m \pm 2,58 \cdot \frac{s}{\sqrt{n-1}}$

- V souborech, kde **n ≤ 30**, používáme model **Studentova rozdělení** (konstanty 1,96, příp. 2,58 se nahrazují jinými – viz skripta, str. 41). U malých souborů je odhad σ pomocí **s** málo přesný, je třeba zvětšit šířku intervalu. Toho dosáhneme použitím konstant Studentova rozdělení.



Když má náhodná veličina (např. VKP) normální rozložení s průměrem **μ** a odchylkou **σ**, pak má výběrový průměr **m** normální rozložení s průměrem **μ** a odchylkou $\sigma = SE_m = \frac{\sigma}{\sqrt{n-1}}$.

Protože hodnotu **σ** neznáme, dosazujeme za ni nejlepší bodový odhad, tj. **s**.

Když provedeme náhodný výběr o rozsahu **n > 30** a vypočítáme výběrový průměr **m**, pak jeho hodnota padne **s pravděpodobností 0,95** do vzdálenosti menší než $1,96 \cdot \frac{s}{\sqrt{n-1}}$ od parametru **μ**.

JINAK ŘEČENO:

- Parametr **μ** s pravděpodobností 0,95 leží ve vzdálenosti menší než $1,96 \cdot \frac{s}{\sqrt{n-1}}$ od **m**.

TO ZNAMENÁ, ŽE:

- Interval s krajními body $m \pm 1,96 \cdot \frac{s}{\sqrt{n-1}}$ pokrývá hodnotu parametru **μ** s pravděpodobností 0,95. Tento interval se nazývá **interval spolehlivosti pro průměr**.

Příklad na výpočet konfidenčních intervalů

Skupina A:

- Odhadněte průměrnou hladinu hemoglobinu v populaci zdravých mužů z náhodného výběru 100 jedinců s průměrnou hodnotou m = 152,4 g/l a směrodatnou odchylkou s = 18,2 g/l se spolehlivostí:
  a) 95%
  b) 99%

Skupina B:

- Odhadněte průměrnou hladinu hemoglobinu v populaci zdravých mužů z náhodného výběru 35 jedinců s průměrnou hodnotou m = 152,4 g/l a směrodatnou odchylkou s = 18,2 g/l se spolehlivostí:
  a) 95%
  b) 99%

Skupina C:

- Odhadněte průměrnou hladinu hemoglobinu v populaci zdravých mužů z náhodného výběru 100 jedinců s průměrnou hodnotou m = 152,4 g/l a směrodatnou odchylkou s = 14,8 g/l se spolehlivostí:
  a) 95%
  b) 99%

VLASTNOSTI INTERVALOVÉHO ODHADU

1) SPOLEHLIVOST

- volí se předem, jde o stanovení pravděpodobnosti, obvykle 0,95 nebo 0,99

2) PŘESNOST

- je dána délkou intervalu
- čím kratší je interval, tím je vyšší přesnost odhadu

$$m \pm \underbrace{1,96 \cdot \frac{s}{\sqrt{n-1}}}_{\text{přesnost}}$$

- **OBĚ VLASTNOSTI SPOLU SOUVISEJÍ**

- **PŘESNOST ODHADU LZE OVLIVNIT:**
  a) snížením či zvýšením **P** (spolehlivosti)
  b) snížením či zvýšením **n** (velikost souboru)
  c) snížením či zvýšením **s** (homogenita souboru)

ODHAD PRAVDĚPODOBNOSTI ZÁKLADNÍHO SOUBORU (PARAMETRU π)

1. Nejlepší bodový odhad pravděpodobnosti je relativní četnost

   $$p = \frac{k}{n} \rightarrow \pi$$

   n = počet pozorování

   k = počet pozorování, u nichž nastal sledovaný jev

2. Pro pravděpodobnosti sice platí binomické rozdělení, ale pokud platí nerovnost $\mathbf{n \cdot p \cdot (1-p) > 9}$, můžeme vycházet z normálního rozdělení.

3. Standardní chybu SE odhadujeme ze vztahu:

   $$\mathbf{SE} = \sqrt{\frac{p(1-p)}{n}} \quad \text{nebo} \quad \mathbf{SE} = \sqrt{\frac{p(100-p)}{n}} \text{ v } \%$$

Závěr: 95% CI $\mathbf{p \pm 1,96 \cdot \sqrt{\frac{p(1-p)}{n}}}$

99% CI $\mathbf{p \pm 2,58 \cdot \sqrt{\frac{p(1-p)}{n}}}$

Příklad:

Odhadněte pravděpodobnost výskytu zrakové vady u studentů LF na základě výběrového šetření u 200 studentů.

$$n = 200 \quad k = 80 \quad p = 0{,}40\ (40\%)$$

- **Než začnete počítat hranice konfidenčního intervalu pomocí kritických hodnot normálního rozložení, nezapomeňte ověřit splnění podmínky, tj. platnost nerovnosti** $n \cdot p \cdot (1 - p) > 9$.

Příklad:

Ve výběru 100 šestiměsíčních zdravých dětí náhodně vybraných z brněnské populace byl sledován hemoglobin v g%.

n = 100 m = 13,10 s = 1,9

1. Určete interval, ve kterém se pohybuje hemoglobin u 95% vyšetřených dětí.
2. Odhadněte průměrné množství hemoglobinu v základním souboru se spolehlivostí 0,95. Jaká je přesnost tohoto odhadu?
3. U kolika dětí musíme provést šetření, aby přesnost odhadu průměru byla při spolehlivosti 0,95 nejméně $\pm$ 0,2.

INTERVALOVÉ ODHADY

a) **Dvoustranné** – **určujeme** horní i dolní mez intervalu (např. u cholesterolu)

$$P(4{,}37 < \mu < 4{,}77) = 0{,}95$$

b) **Jednostranné**

– levostranné: pouze dolní hranice intervalu

$$P(\mu > 4{,}37) = 0{,}975$$

– pravostranné: pouze horní hranice intervalu

$$P(\mu < 4{,}77) = 0{,}975$$

U jednostranných odhadů jsou **odpovídající kritické hodnoty** pro spolehlivost:

a) 95% $\pm 1{,}645$

b) 99% $\pm 2{,}326$

Změna kritických hodnot u jednostranného odhadu pro spolehlivost 0,95