INDUKTIVNÍ STATISTIKA
8.seminář

ODHADY PARAMETRŮ

Statistická indukce – 4. etapa

Teorie odhadů

- Odhad průměru základního souboru
- Odhad pravděpodobnosti

Testování statistických hypotéz

- Srovnání průměrů 2 základních souborů
- Srovnání pravděpodobností 2 náhodných jevů

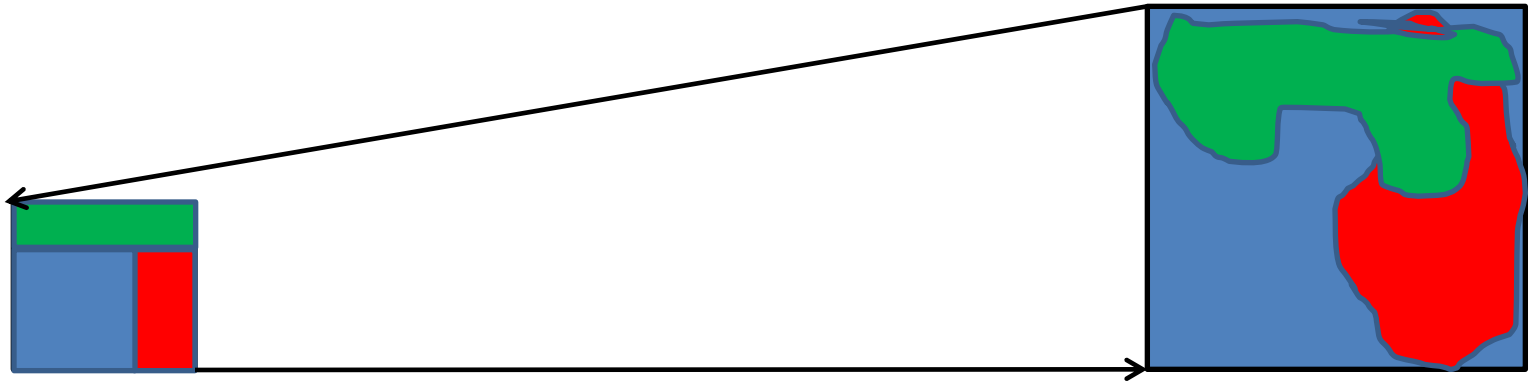
Hodnocení zavislosti

- Závislost kvantitativních veličin
- Závislost kvalitativních veličin

ZOBECNĚNÍ VÝSLEDKŮ VÝBĚROVÉHO ŠETŘENÍ NA CELÝ ZÁKLADNÍ SOUBOR

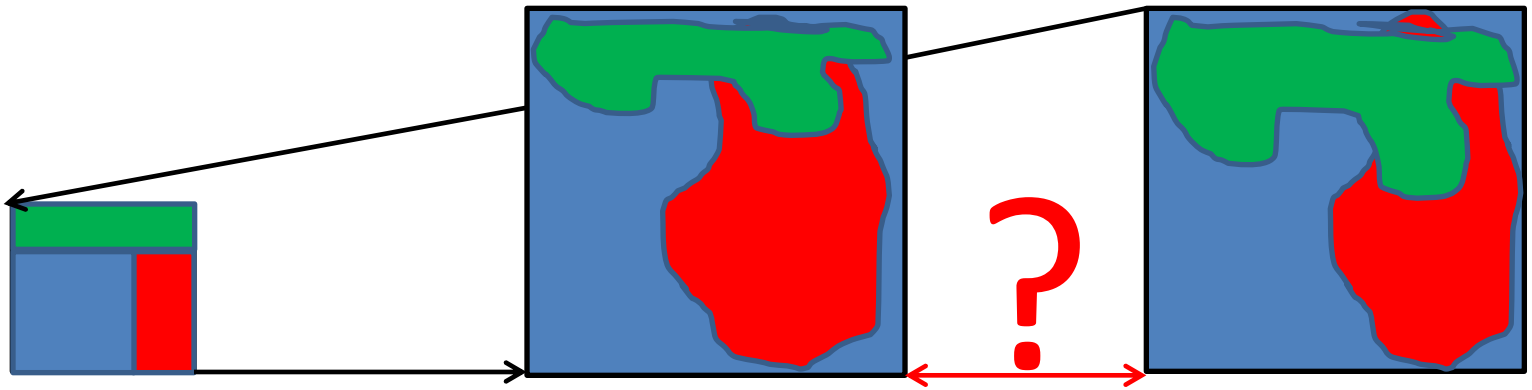
! PRAVDĚPODOBNOST, NÁHODNÁ VELIČINA A JEJÍ ROZDĚLENÍ

STATISTICKÁ INDUKCE



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

STATISTICKÁ INDUKCE



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

Pravděpodobnost náhodného jevu

- Pravděpodobnost je mírou „častosti“ výskytu tohoto jevu
- Pravděpodobnost je vlastnost náhodného jevu
- Pravděpodobnost NJ zjistíme opakováním pokusů, jejichž výsledkem může být daný jev a „měříme“ ji **relativní četností (p)** tohoto jevu v řadě opakovaných pokusů (p = k/n).

Pravděpodobnost náhodného jevu

- **Klasická definice** pravděpodobnosti – pst NJ

  je dána podílem příznivých a všech možných výsledků v experimentu, jehož možné výsledky jsou stejně pravděpodobné.
- Pravděpodobnost jevů spojených s karetními hrami, (hod kostkou, mincí ..) – dle definice
- **Nelze** dle def. vypočítat pravděpodobnost jevů v medicíně, můžeme pst jen **odhadovat pomocí relativních četností**

NÁHODNÁ VELIČINA (NV)

- **Diskrétní** (NV) - jednoznačně určena souhrnem hodnot, kt. může nabývat a pravděpodobnostmi, s nimiž těchto hodnot nabývá. Takto definujeme její rozdělení – **rozdělení pravděpodobností.**

  Graficky – tyčkový graf (rozdělení diskretní NV)

  Příklad: binomické rozdělení

  Poissonovo rozdělení vzácných jevů

  př. výsledek hodu hrací kostkou (X)

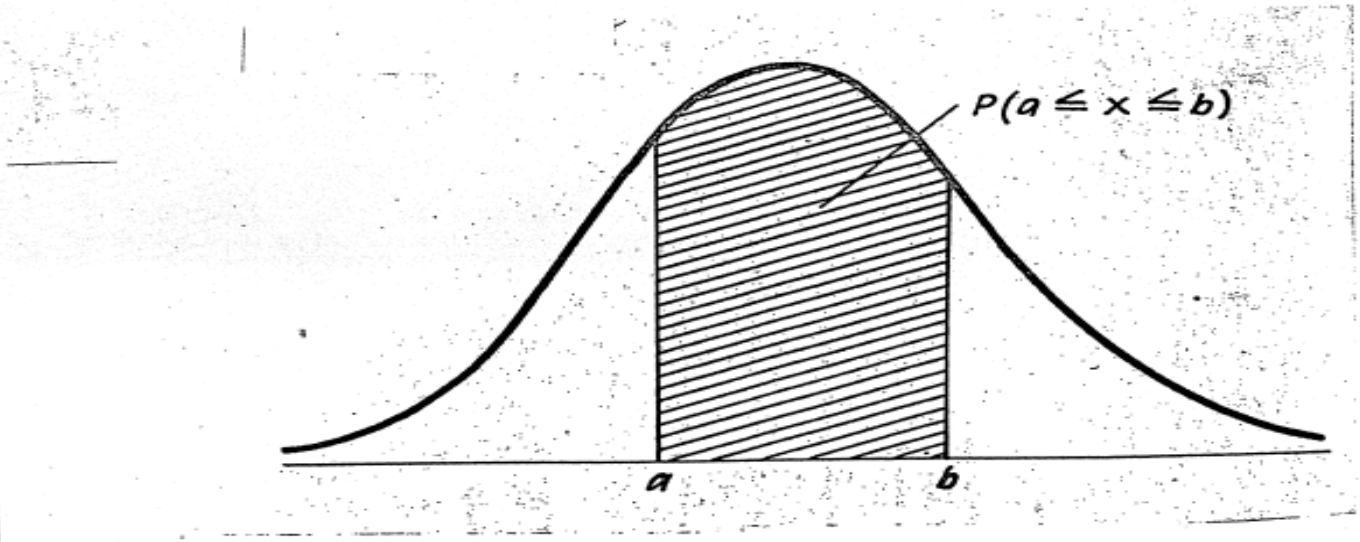
NÁHODNÁ VELIČINA (NV)

- **<u>Spojitá NV</u>** – její rozdělení je dáno <u>hustotou pravděpodobností</u> – **frekvenční křivka (funkce).**
- Matematicky – rovnice frekvenční funkce

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

- Graficky - hladká, plynulá čára

Příklad frekvenční fce spojité NV



P (a ≤ x ≤ b) - pravděpodobnost, že NV **X** nabývá hodnot z intervalu a,b

- Celá plocha pod frekvenční křivkou = 1 (100%)
- P = 0 ⇔ a = b / pravděpodobnost znázorněna úsečkou – obsah plochy = 0
- Pst malá – pokud ⟨a,b⟩ krátký

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 (nutno počítat s chybami)

ZÁKLADNÍ SOUBOR

- soubor, který nás zajímá
- teoretické rozdělení četností (matematický model)

- popis rozdělení:
  pravděpodobnostní fce

  frekvenční fce
- 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 a pravděpodobnostní rozdělení

- Každá veličina, kterou zkoumáme, je ovlivňována řadou nepatrných nahodilých vlivů, což způsobuje její **variabilitu** – tzn. veličina nabývá u různých subjektů různých hodnot.

- Měříme-li veličinu ve výběrovém souboru, pak rozložení hodnot této veličiny znázorňujeme na základě empiricky zjištěných četností (polygon četností = **výběrové rozdělení** četností).

- Každá veličina má své **pravděpodobnostní (teoretické) rozdělení**.
    - V takovém rozložení jsou **na ose x všechny hodnoty**, kterých může veličina potenciálně nabývat, a **na ose y jsou pravděpodobnosti**, se kterými se dané hodnoty vyskytují.

Empirické a pravděpodobnostní rozdělení

- **V empirickém rozdělení** (polygon četností) jsou popsány četnosti, se kterými se naměřené hodnoty vyskytovaly ve výběrovém souboru

X

- **Pravděpodobnostní rozdělení** (pravděpodobnostní křivka) vyjadřuje očekávání, jak často se budou jednotlivé hodnoty vyskytovat v nekonečně velkém souboru

-

- **Pravděpodobnostní rozdělení** náhodných veličin jsou teoretické (matematické) modely, jejichž pomocí popisujeme nejrůznější reálné situace. Navzdory různorodosti a mnohotvárnosti přírodních a společenských jevů vystačíme v praxi s malým počtem modelů(tj. typů rozdělení).

Typy pravděpodobnostních rozdělení

Diskrétní veličiny

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

Spojité veličiny

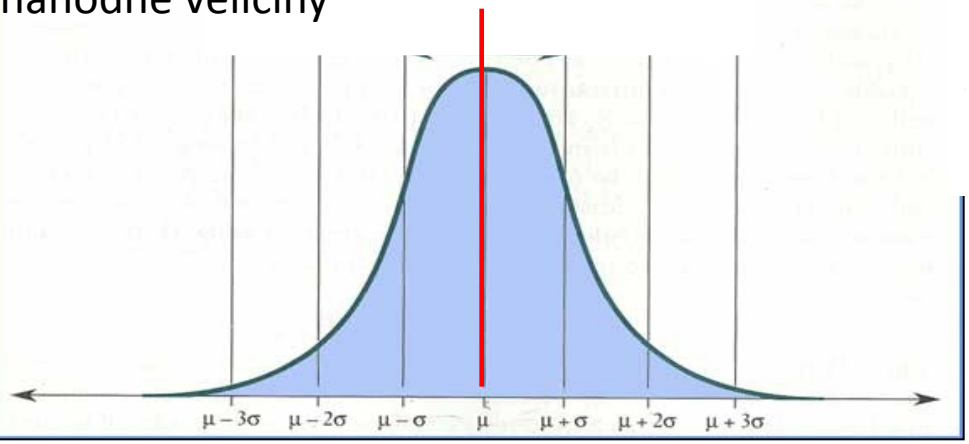
- Normální (Gaussovo) rozdělení
- Studentovo t-rozdělení
- Snedecorovo F-rozdělení
- Chí-kvadrát rozdělení

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)

Matematický model rozdělení četností spojité náhodné veličiny

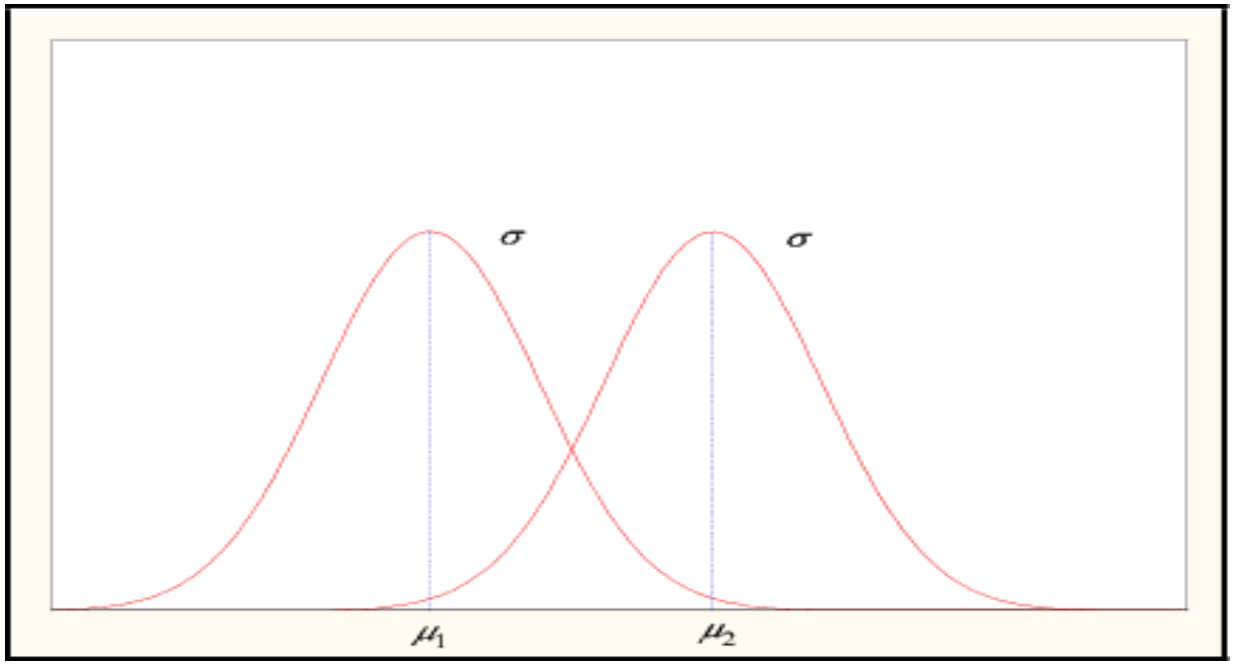


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

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

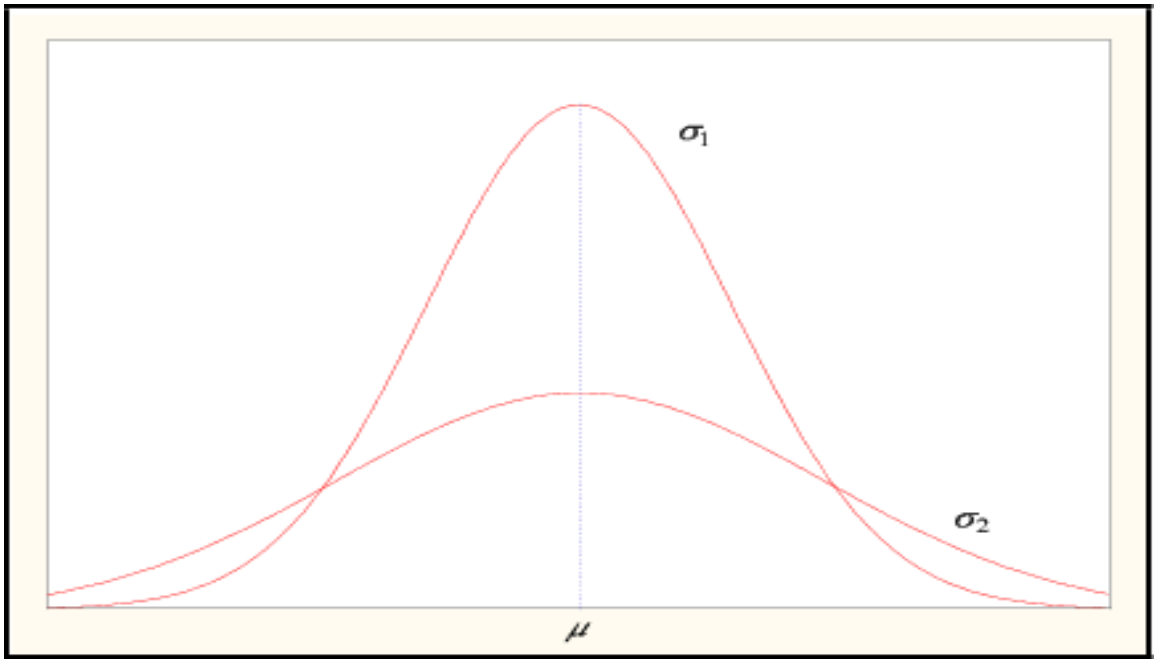
NORMÁLNÍ ROZDĚLENÍ

- Frekvenční křivky normálního rozdělení se stejnými směrodatnými odchylkami a odlišnými průměry (μ1 ≠ μ2; σ1 = σ2)

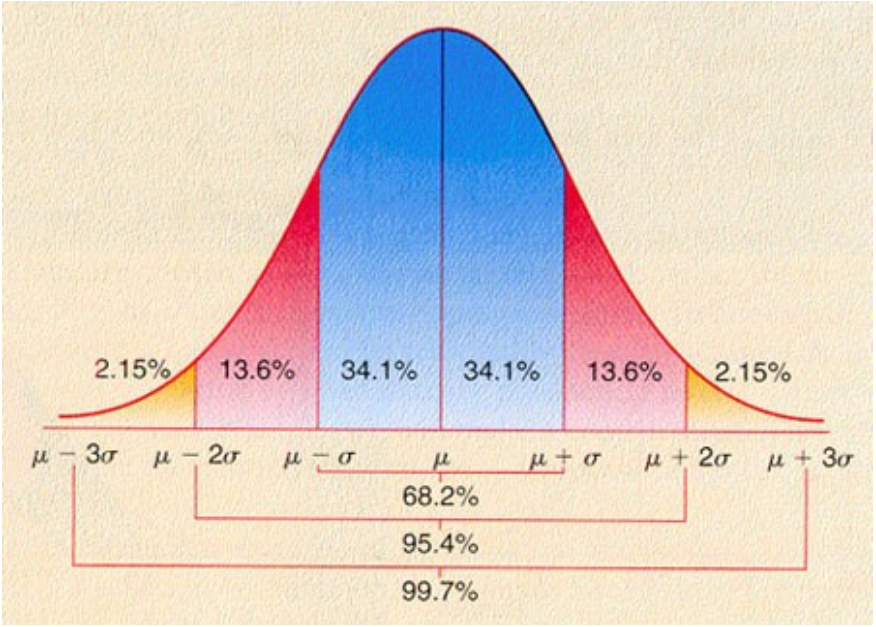


NORMÁLNÍ ROZDĚLENÍ

- Frekvenční křivky normálního rozdělení se stejnými průměry a odlišnými směrodatnými odchylkami (μ1 = μ2; σ1≠ σ2)

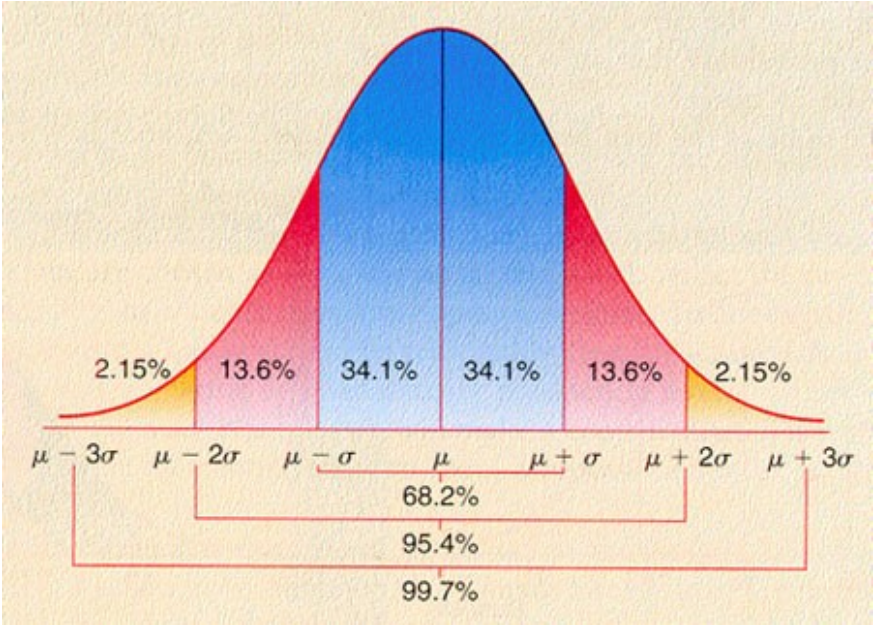


VLASTNOSTI NORMÁLNÍHO ROZDĚLENÍ



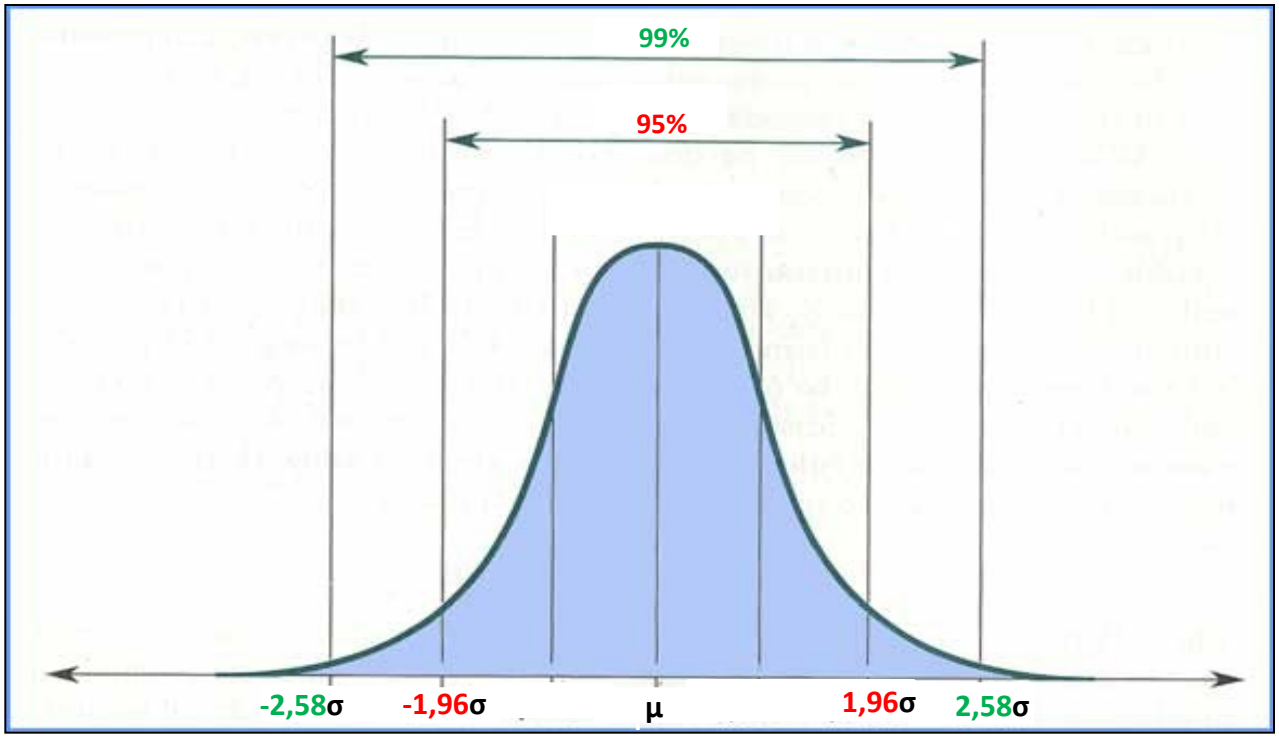
- **Frekvenční křivky normálního rozdělení mají různý tvar a polohu pro různé veličiny. Pro všechny však platí, že v intervalu:**
  - **$\mu \pm 1\,\sigma$ leží 68,2% hodnot, kterých může daná veličina nabývat**
  - **$\mu \pm 2\,\sigma$ leží 95,4% hodnot, kterých může daná veličina nabývat**
  - **$\mu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

VLASTNOSTI NORMÁLNÍHO ROZDĚLENÍ



- Častěji nás ale zajímá, v <u>jakém intervalu leží 95% (99%) hodnot sledované veličiny</u>
  - můžeme pak totiž vyjádřit tvrzení, že s pravděpodobností 95% (99%) se hodnoty sledované veličiny nacházejí právě v tomto intervalu
  - tento interval je vymezen tzv. kritickými hodnotami normálního rozdělení
- $P(\mu - 1{,}96\sigma \leq x \leq \mu + 1{,}96\sigma) = 0{,}95$
- $P(\mu - 2{,}58\sigma \leq x \leq \mu + 2{,}58\sigma) = 0{,}99$

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

Tabulky normálního rozdělení

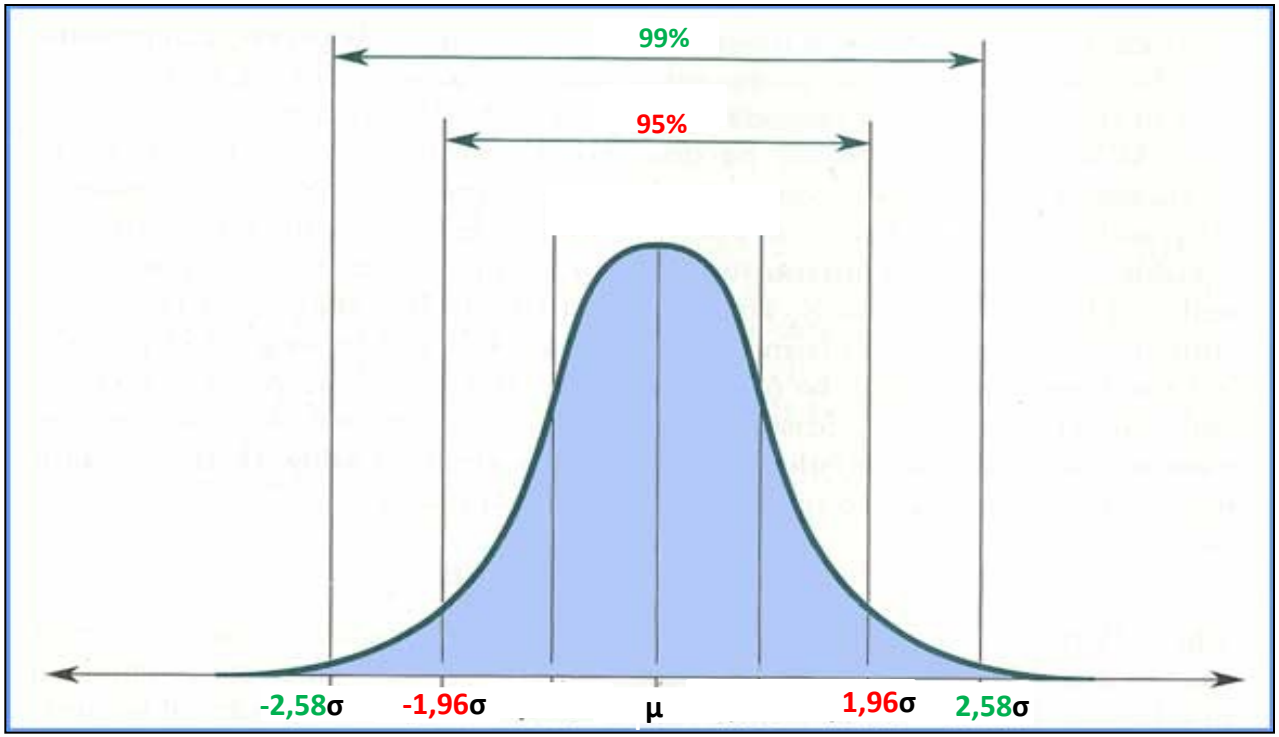
- Pro nematematiky
- Pro zvolené hranice intervalu a,b lze najít odpovídající pravděpodobnost a obráceně
- Tabelace je možná proto, že

1/ normální rozdělení je symetrické

2/ hranice a, b lze vyjádřit jako odchylky od μ v násobcích směrodatné odchylky σ

(zásada: plocha = pravděpodobnost!)

VLASTNOSTI NORMÁLNÍHO ROZDĚLENÍ



- 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
- $P (\mu - 1{,}96\sigma \leq x \leq \mu + 1{,}96\sigma) = 0{,}95$
- $P (\mu - 2{,}58\sigma \leq x \leq \mu + 2{,}58\sigma) = 0{,}99$

pravděpodobnost	násobky směrodatné odchylky
0,99	2,58
0,95	1,96
0,90	1,64
0,80	1,28
0,75	1,15
0,50	0,65

násobky směrodatné odchylky	pravděpodobnost
0,50	0,3828
1,00	0,6827
1,50	0,8664
2,00	0,9545
2,50	0,9876
3,00	0,9973

ODHADY PARAMETRŮ

1. **Bodové** odhady

2. **Intervalové** odhady

BODOVÉ ODHADY

- Neznámý parametr odhadujeme jedním číslem tj. **bodem.**
- Např. výběrový aritmetický průměr m je bodovým odhadem parametru μ

  ($\mu \approx m$, $\sigma \approx s$, $\pi \approx p$)
- Bodové odhady se „nestrefí“ přesně do odhadovaného parametru

INTERVALOVÉ ODHADY

- Neznámý parametr odhadujeme **intervalem** vytvořeným kolem tzv. nejlepšího nestranného bodového odhadu.

- **Interval spolehlivosti** (konfidenční interval)

- **Spolehlivost** si určujeme sami, obvykle buď 95% nebo 99%
  - jde o pravděpodobnost, že odhadovaný parametr se nachází v daném intervalu.

- **Hraniční hodnoty** spadají do intervalu spolehlivosti
  - (dolní hranice ≤ odhadovaný parametr ≤ horní hranice)

- zápis: **95% CI (- ; -)**
  99% CI (- ; -)

INTERVALOVÉ ODHADY

Doplněk spolehlivosti odhadu (do 100%) vyjadřuje **riziko odhadu (riziko induktivního úsudku)** – tj. riziko, že odhadovaný parametr leží mimo interval:

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

INTERVALOVÉ ODHADY

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

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

b) **Jednostranné**

– levostranné: pouze dolní hranice intervalu

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

– pravostranné: pouze horní hranice intervalu

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

U jednostranných odhadů jsou **odpovídající kritické hodnoty** pro spolehlivost: a) 95% $\pm 1{,}645$

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

ODHAD PRŮMĚRU ZÁKLADNÍHO SOUBORU (PARAMETRU μ)

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. Výběrový průměr se jako náhodná veličina vyznačuje variabilitou. Variabilita výběrových průměrů je menší než variabilita veličiny, z níž jsou průměry počítány. Díky variabilitě je každý výběrový průměr zatížen chybou – jde o tzv. **standardní chybu** průměru SE_m, kterou odhadujeme 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. 25).

Odhad průměru ZS (μ) - příklad

- Odhadněte průměrnou vitální kapacitu plic mužů ve věku 40-50 let na podkladě výběrového šetření 200 mužů s výsledky:

$m = 4,83$ $s = 0,66$ $n = 200$

Řešení

$$m \pm 1{,}96 \cdot \frac{s}{\sqrt{n-1}}$$

SE = 0,66 /√200-1 = 0,66/ 14,107= 0,04678

Pro spolehlivost 0,95

a = m - 1,96 .SE a = 4,83 – (1,96 . 0,04678)= **4,74**

b = m + 1,96.SE b = 4,83 + (1,96 . 0,04678)= **4,92**

Přesnost = 1,96.SE = 1,96.0,04678 **= 0,09**

3 formy zápisu: 1/ μ = 4,83 ± 0,09

2/ P (4,74 ≤ μ ≤ 4,92) = 0,95

3/ 95% CI (4,74 ; 4,92)

Interpretace

- Průměrná vitální kapacita plic v ZS mužů věkové kategorie 40 - 50 let se pohybuje s pravděpodobností 0,95 v rozmezí 4,83 ± 0,09, tj. od 4,74 do 4,92 litrů.

Proveďte odhad se spolehlivostí 0,99.

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

$$m \pm 1,96 \cdot \frac{s}{\sqrt{n-1}}$$

- **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)

Vlastnosti odhadu

- 95%CI (4,74 ; 4,92) přesnost ± 0,09
- 99%CI (4,71; 4,95) přesnost ± 0,12
- Porovnejme spolehlivost, délku a přesnost intervalů

Příklad: Kolik musíme / nejméně / vyšetřit osob, abychom odhadli průměrnou vitální kapacitu plic s přesností na ± 0,1 litru při 99% spolehlivosti ?

Řešení: $\text{přesnost } (0{,}1) = 2{,}58 \cdot SE_m = 2{,}58 \cdot \frac{s}{\sqrt{n-1}} = \frac{0{,}66}{\sqrt{n-1}}$

$$n - 1 = 2{,}58^2 \cdot \frac{0{,}66^2}{0{,}1^2} = 289{,}9$$

n = 291

Interpretace: Musíme vyšetřit nejméně 291 osob, abychom odhadli průměrnou vitální plic s přesností ± 0,1 při 99% spolehlivosti.

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%

Řešení

1) $SE = \frac{18,2}{\sqrt{100-1}} = 1,83$

CI 95 % (148,8 ; 156,0)

CI 99 % (147,7 ; 157,1)

2) $SE = \frac{18,2}{\sqrt{35-1}} = 3,16$

CI 95 % (146,2 ; 158,6)

CI 99 % (144,2 ; 160,6)

3) $SE = \frac{14,8}{\sqrt{100-1}} = 1,49$

CI 95 % (149,5 ; 155,3)

CI 99 % (148,6 ; 156,2)

ODHAD NEZNÁMÉ PRAVDĚPODOBNOSTI NÁHODNÉHO JEVU (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 $n \cdot p \cdot (1-p) > 9$, můžeme vycházet z normálního rozdělení.

3. Standardní chybu SE odhadujeme ze vztahu:

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

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

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

Příklad:

Odhadněte pravděpodobnost výskytu zrakové vady (se spolehlivostí 95 i 99%) u studentů LF na základě výběrového šetření u 200 studentů.

Získané výsledky interpretujte.

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

1) Platnost podmínky

$n \cdot p \cdot (1-p) > 9$

$200 \cdot 0{,}4 \cdot (1-0{,}4) > 9$

$48 > 9$

$200 \cdot 0{,}4 \cdot 0{,}6 > 9$

2) $SE = \sqrt{\frac{0{,}4 \cdot (1-0{,}4)}{200}} = 0{,}03464 \doteq 0{,}035$

$0{,}035 \times 1{,}96 = 0{,}0686 \doteq 0{,}069$

CI 95% (0,33 ; 0,47)

$p = 0{,}4 \pm 0{,}069$

S 95% pravděpodobností bude znak. vadou trpět (33-47 %)

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.**

Řešení:

1. Určete interval, ve kterém se pohybuje hemoglobin u 95% vyšetřených dětí.

$m \pm 2s$ $\quad$ 9,3 – 16,9 $\quad$ $13{,}10 \pm 2 \cdot 1{,}9 \ldots 13{,}10 \pm 3{,}8$

2. Odhadněte průměrné množství hemoglobinu v základním souboru se spolehlivostí 0,95. Jaká je přesnost tohoto odhadu?

$m \pm 1{,}96SE$ $\quad$ $SE = \frac{1{,}9}{\sqrt{99}} = 0{,}19$ $\quad$ 95% CI (12,73; 13,47)

PŘESNOST $= \pm 1{,}96\, SE_m = \pm 1{,}96 \cdot 0{,}19 = \pm 0{,}37$

3. U kolika dětí musíme provést šetření, aby přesnost odhadu průměru byla při spolehlivosti 0,95 nejméně $\pm$ 0,2.

$1{,}96SE = 0{,}2$

$1{,}96 \cdot \frac{s}{\sqrt{n-1}} = 0{,}2$

$1{,}96 \cdot \frac{1{,}90}{\sqrt{n-1}} = 0{,}2$

$n - 1 = 1{,}96^2 \cdot \frac{1{,}9^2 \cdot 1}{0{,}2^2}$

$n - 1 = 346{,}7$

$n \doteq 346{,}7 + 1$

$n \doteq 348$ (cca 350)

ODHADY PARAMETRŮ - SHRNUTÍ

- **Parametry** jsou charakteristiky základního souboru, odhadují se z výběrových charakteristik.
- **Výběrové charakteristiky** se chovají jako **náhodné veličiny,** výběr od výběru se liší.
- Pomocí výpočtu standardní chyby (SE) výběrové charakteristiky **určujeme interval,** ve kterém se hodnota odhadovaného parametru bude s určitou pravděpodobností pohybovat

μ: 95% CI $(m - 1{,}96SE\ ;\ m + 1{,}96SE)$

99% CI $(m - 2{,}58SE\ ;\ m + 2{,}58SE)$

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

- **1,96 a 2,58 jsou konstanty normálního rozložení** a vymezují plochu pod frekvenční křivkou, která představuje 95%, resp. 99% pravděpodobnost, že odhadovaný parametr bude spadat do výše uvedených intervalů

π: 95% CI $(p - 1{,}96SE\ ;\ p + 1{,}96SE)$

99% CI $(p - 2{,}58SE\ ;\ p + 2{,}58SE)$

$$SE = \sqrt{\frac{p(1-p)}{n}}$$