Induktivní statistika

8. seminář

Statistické šetření

Etapy statistického šetření

1. Plán šetření
2. Sběr dat
3. Popis a technické zpracování
4. **Rozbory a závěry**

Statistická indukce – 4. etapa

Teorie odhadů

- Odhad průměru základního souboru
- Odhad pravděpodobnosti

Testování statistických hypotéz

- Srovnání dvou průměrů
- Srovnání pravděpodobností

Hodnocení zavislosti

- Závislost kvantitativních veličin
- Závislost kvalitativních veličin

ZOBECNĚNÍ VÝSLEDKŮ VÝBĚROVÉHO ŠETŘENÍ NA CELÝ ZÁKLADNÍ SOUBOR

! PRAVDĚPODOBNOST, NÁHODNÁ VELIČINA A JEJÍ ROZDĚLENÍ

Pravděpodobnost (skripta kapitola 4)

Všechny stat. výroky – pravděpodobnostní charakter a jejich věrohodnost vyjadřuje:

- spolehlivost - pravděpodobnost, že tento výrok **platí**

- r**iziko -** pravděpodobnost, že **neplatí** daný výrok

Pravděpodobnost náhodného jevu je kvantitativní charakteristika, která je mírou častosti jeho výskytu.

Pravděpodobnost je vlastnost náhodného jevu stejně jako např. délka nebo hmotnost jsou vlastnosti určitého předmětu.

Klasická definice **P (A) = m/n** (např. hod kostkou, pravděpodobnost 1/6, že padne dané číslo)

m = počet příznivých výsledků v experimentu

n = počet všech možných výsledků

Příklady: kostka, mince, karty, sportka, narození chlapce/dívky

Podmínkou je komplex podmínek – souhrn předpisů, za nichž se experiment provádí (např. homog. kostky)

Pravděpodobnost reakce pacienta na určitou léčbu?

Pouze **odhad** pomocí **relativních četností**

n pacientů

x vyléčeno

Podíl **x/n** odhaduje pravděpodobnost vyléčení **P**

Vlastnosti pravděpodobnosti

1. $0 < \text{nebo} = P(A) < \text{nebo} = 1$
2. $P(A) = 0$ pro jevy nemožné
3. $P(A) = 1$ pro jevy jisté

Pravidla pro počítání

a) Pravidlo pro sčítání

A, B disjunktní – **vzájemně se vylučují** (pravděpodobnost, že nastane jeden jev nebo druhý).

P (A **nebo** B) = **P (A) + (B)** kostka(strany) 5, 6 je rovno 1/6 + 1/6 = 1/3

A, B nejsou disjunktní – **vzájemně se nevylučují** (pravděpodobnost, že nastanou oba jevy)

P (A **nebo** B) = P (A) + P (B) – P (A i B)

př. A = diabetes (D)

B = hypertenze (H)

P (D nebo H) = P (D) + P (H) – P (D i H)

Vlastnosti pravděpodobnosti

b) **Pravidlo pro násobení**

A, B **nezávislé jevy** (výskyt jednoho jevu není ovlivněn výskytem druhého jevu)

P (A **i** B) = **P (A) . P (B)**

A, B **závislé jevy** (jeden jev podmiňuje druhý jev)

P (A **i** B) = P (A) . **P (B/A)** – podmíněná pravd. jevu B, že nastal jev A.

= P (B) . P (A/B) př. dvě nemoci u člověka

V praxi – máme určit podmíněn. pravd. P (A/B) ev. P (B/A), kdy P (A i B) je známa: P (A/B), P (B/A) jsou tzv. **podmíněné pravděpodobnosti.**

Platí: P (A/B) = P (A i B)/ P (B)

P (B/A) = P (A i B)/P (A)

Pravidlo pro sčítání i násobení se dá rozšířit na více jevů.

Podmíněné pravděpodobnosti jsou užitečné pro hodnocení **rizika** nemoci v populaci

Vlastnosti pravděpodobnosti

Např. Ze srovnání pravděpodobností

P (Ca), P (Ca/K), P(Ca/N)

Lze usuzovat na riziko kouření (K,N) na výskyt karcinomu plic (Ca)

Např. **P (Ca/K) / P (Ca)** - udává, kolikrát je větší pravděpodobnost výskytu karcinomu plic u kuřáků než v celé populaci.

P (Ca/K) / P (Ca/N) - udává, kolikrát je větší pravděpodobnost výskytu karcinomu plic u kuřáků než u nekuřáků.

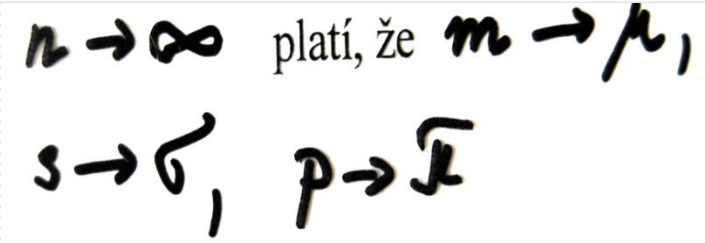
Výběrový/základní soubor

Výběrový soubor

- reprez. náhodný výběr
- **výběrové** (empirické) **rozdělení četností**
- popis rozdělení: tabulka, graf
- stat. ukazatele = **výběrové charakteristiky**: **m, s, p** (ozn. latinkou)
- jsou to charakteristiky náhodných veličin, tzn. **mění se výběr od výběru** + je nutné počítat s chybami (výběrové, náhodné)

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 **konstanty**, zpravidla neznámé, pro



Statistická indukce = *usuzování z vlastností výběru na vlastnosti základního souboru*

Empirické a pravděpodobnostní rozdělení

- každá veličina, kterou zkoumáme, je ovlivněna řadou nepatrných **náhodných 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
- 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 zaneseny pravděpodobnosti, se kterými se dané hodnoty vyskytují.
- 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

Typy pravděpodobnostních rozdělení

Diskrétní veličiny

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

Spojité veličiny

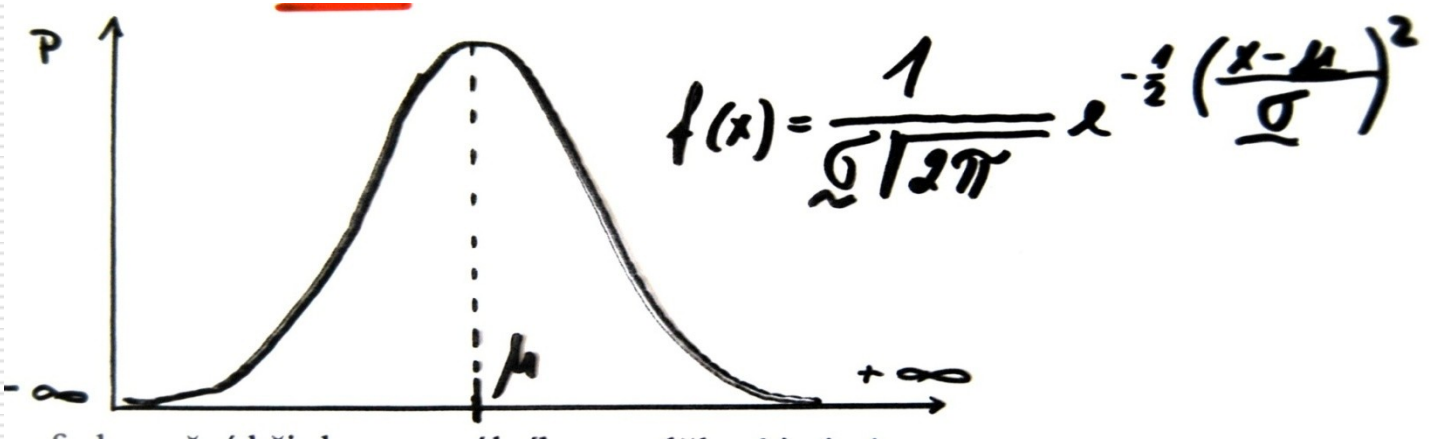
- **normální rozdělení**
- Studentovo t-rozdělení
- Snedecorovo F-rozdělení (Fisherovo – Snedecorovo rozdělení)
- Chí-kvadrát rozdělení

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

- matematický model rozdělení četností **náhodné veličiny**



frekvenční křivka normálního rozdělení je jednoznačně určena dvěma parametry: **μ, σ**

μ určuje polohu křivky (analogie **m**) **mí**

σ určuje tvar křivky (analogie **s**) **sigma**

x –

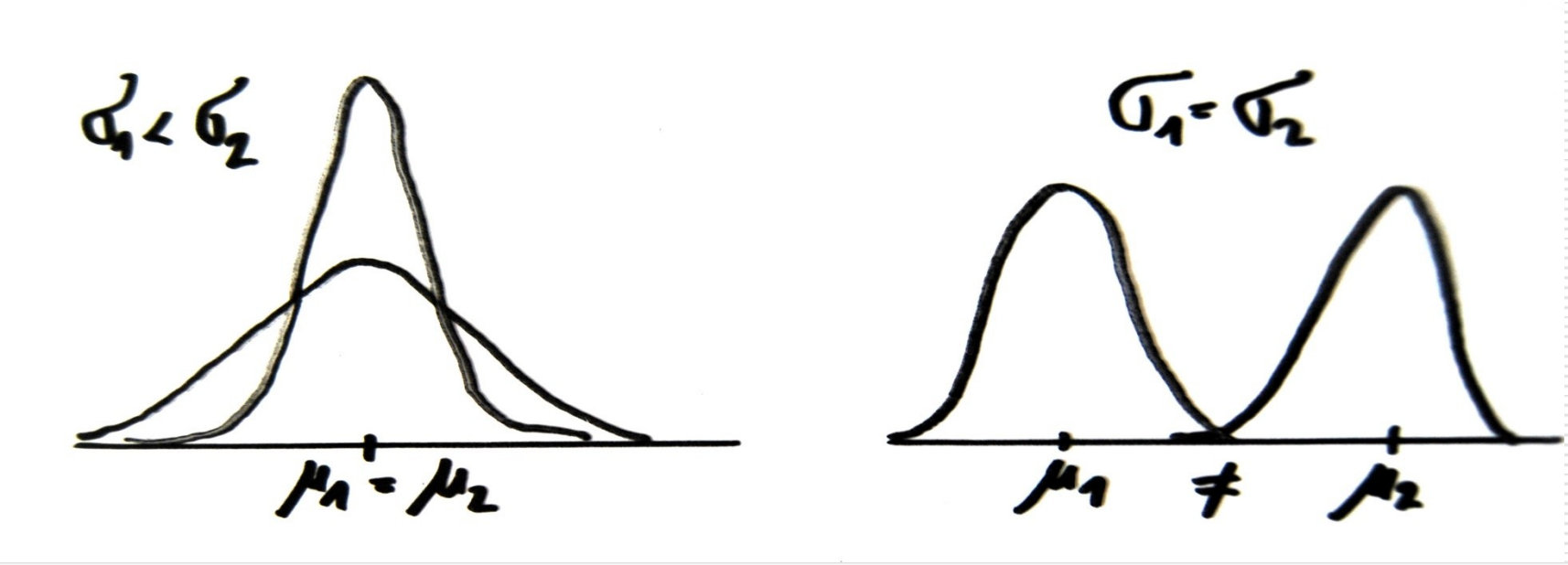
e – základ přirozených logaritmů = 2,72

π – Ludolfovovo číslo = 3,14

Normální rozdělení

μ určuje polohu křivky (analogie **m -** VS) **mí -** ZS

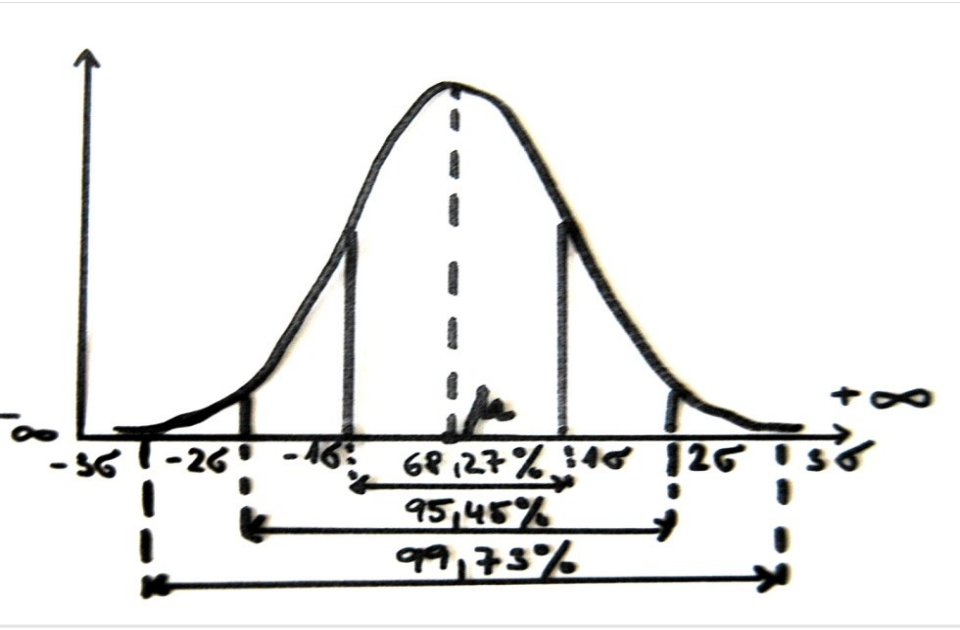
σ určuje tvar křivky (analogie **s -** VS) **sigma -** ZS

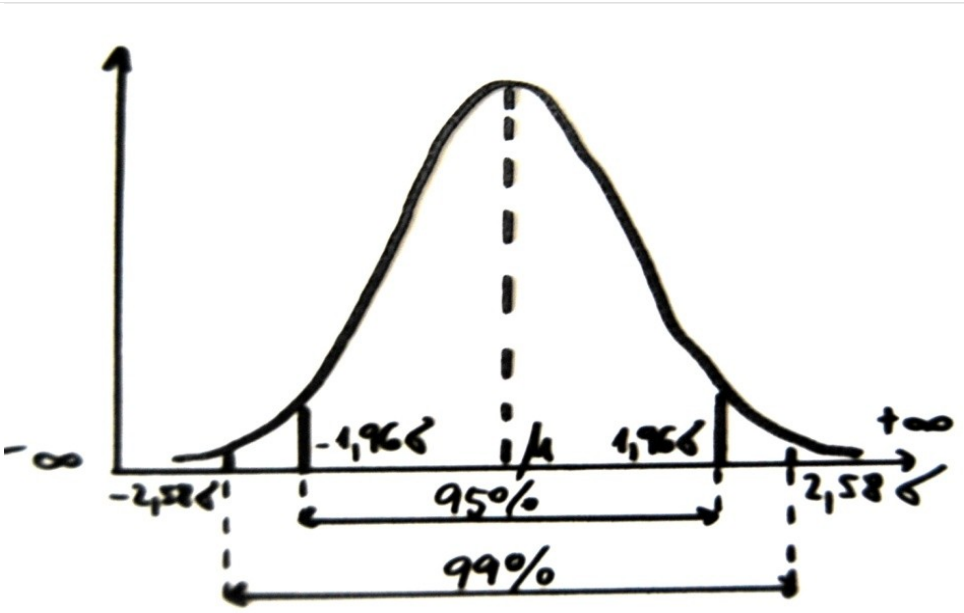


Vlastnosti normálního rozložení

- Frekvenční křivky normálního rozložení mají pro různé veličiny různý tvar (**σ**) a polohu (**μ**)

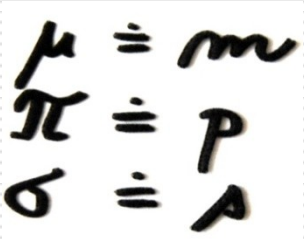
- Pro všechny ale platí, že intervaly, ve kterých se odhadovaná proměnná nachází s pravděpodobností 95% nebo 99%, lze vyjádřit jako odchylky od **μ** v násobcích **σ** :





Odhady parametrů – bodové, intervalové

1)Bodové odhady = odhad jedním číslem



průměr
relativní četnost
směrod. odchylka

Požadavky na bodové odhady:

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** = charakteristika s minimálním rozptylem.
- **Interval spolehlivosti** (konfidenční interval - **CI**)
- **Spolehlivost** si určujeme sami (na začátku výzkumu) – buď 95% nebo 99%

  jde o pravděpodobnost, že odhadovaný parametr se nachází v daném intervalu

  95% CI (-;-)

  99% CI (-;-)
- doplněk spolehlivosti vyjadřuje **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) [mí]

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
3. V souborech, kde **n < 30**, používáme model Studentova rozdělení (konstanty 1,96, příp. 2,58 se nahrazují jinými – viz. skripta statistiky str. 25 - tabulka)
4. Každý výběrový průměr je zatížen chybou – jde o tzv. **standardní chybu** průměru SE_m, kterou odhadujeme ze vztahu: (střední chyba)

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

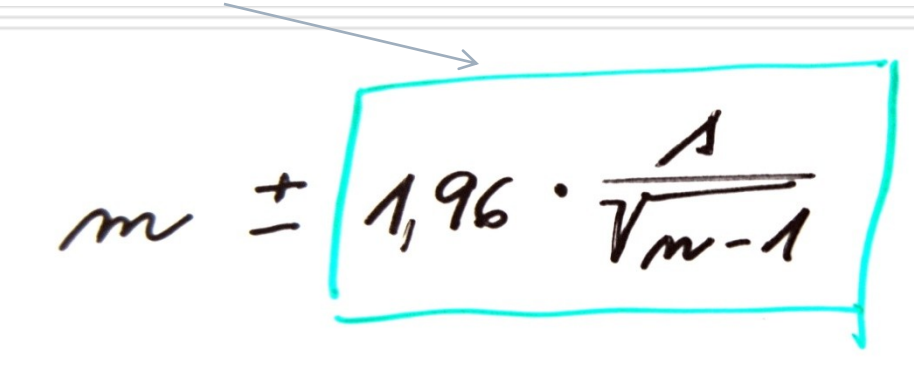
Závěr:

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

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

Vlastnosti 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



Obě vlastnosti spolu souvisí

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 ZS (parametru π) [pí]

1. Nejlepší bodový odhad 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 chceme pracovat s normálním rozdělením, platí

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

můžeme vycházet z normálního rozdělení

3. Standardní chybu **SE** odhadujeme ze vztahu:

$$SE = \sqrt{\frac{p(1-p)}{n}} \qquad SE = \sqrt{\frac{p(100-p)}{n}} \ \%$$

čísla

Pro 95 % CI $\pi = p \pm 1{,}96\sqrt{\frac{p(1-p)}{n}}$

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

Příklad 1:

Příklad:
Odhadněte pravděpodobnost výskytu zrakové vady u studentů LF na základě výběrového šetření u 200 studentů
n = 200 k = 80 p = 0,40 (40%)

!Ověřit platnost podmínky!

Řešení 1:

1) Platnost podmínky

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

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

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

$48 > 9$

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

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

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

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

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

Příklad 2:

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 2:

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

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

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

2) $SE = \frac{18,2}{\sqrt{35-1}} = 3,12$

CI 95 % (146,3; 158,5)

CI 99 % (144,4; 160,5)

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

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

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

Děkuji za pozornost