POLB139

2. 11. 2020

# STATISTICKÁ INFERENCE, POPULACE, VZOREK, INTERVAL SPOLEHLIVOSTI

# JAKÉ TYPY DAT MŮŽEME POUŽÍT?

- Census
  - Naše statistiky se vztahují ke všem případům v populaci
- Výběrový soubor (vzorek)
  - Jak moc mě zajímají statistiky vztahující se k vzorku?
  - Chci generalizovat na celou populaci – statistická inference

Populace vs. Vzorek

# JAK FUNGUJE VZOREK?

- Když obvolám 1000 studentů MUNI a zeptám se, koho by volili, kdyby zítra byly volby, jako moc dobrý dostanu odhad toho, jak by ty volby fakt dopadly?????

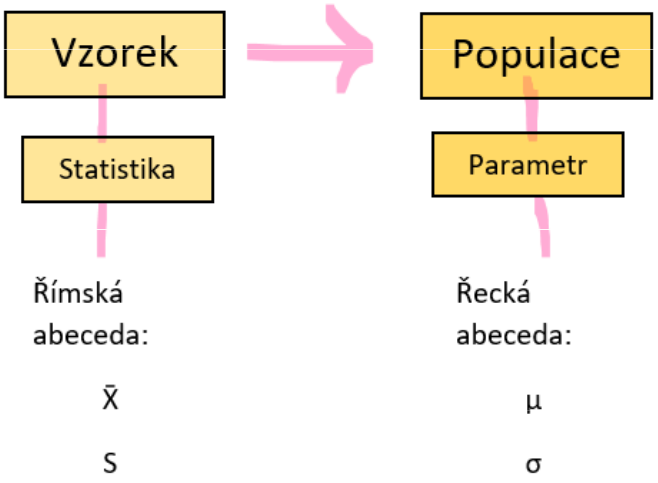
- Klíč: výběr vzorku, sampling

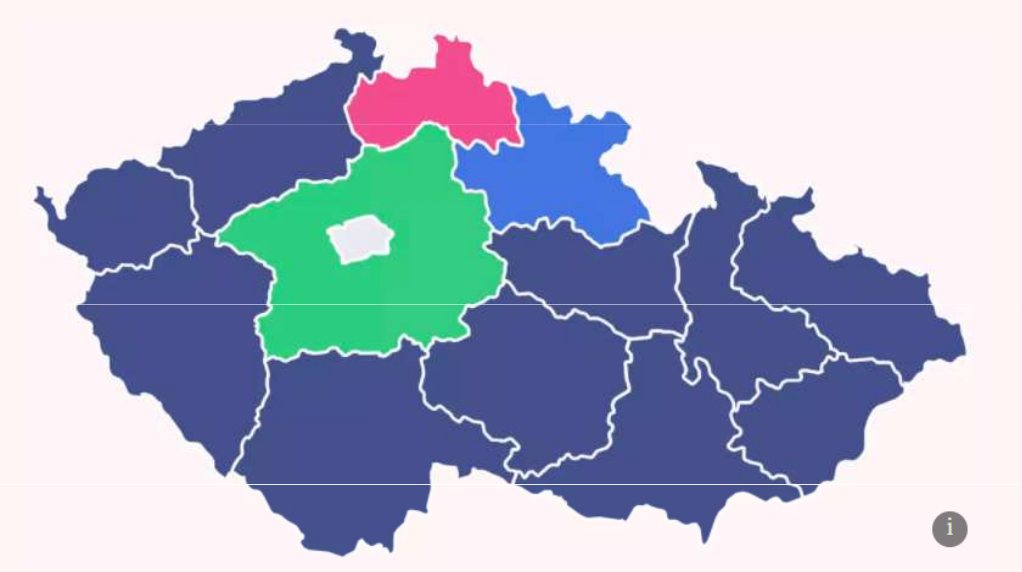
# N = 1000 OBČANŮ ČR

- Deskriptivní statistika, jednorozměrná analýza
  - Míry centrální tendence
  - Relativní četnosti
  - Rozpětí, IGR, rozptyl, směrodatná odchylka atd.
- Dvourozměrná analýza
  - Korelace, regrese

# N = 1000 OBČANŮ ČR

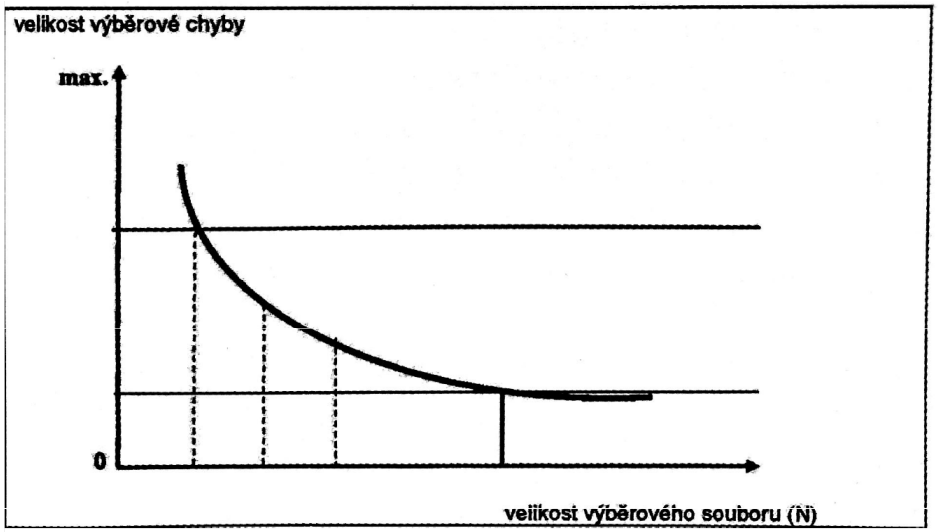
- Cíl je dělat závěry o celé populaci občanů ČR
- Inferenční statistika:



# VZORKY

- Reprezentativní
- Náhodné
- Čím větší, tím lepší (skoro vždy)
- Velikost vzorku ale nenahradí mizernou kvalitu
- Garabage in, garbage out
- Od určité míry se přesnost parametru nezlepšuje
- Úvaha: kolik mám skupin, jak přesný potřebuji odhad, jaká je velikost efektu, neexistuje fixní pravidlo

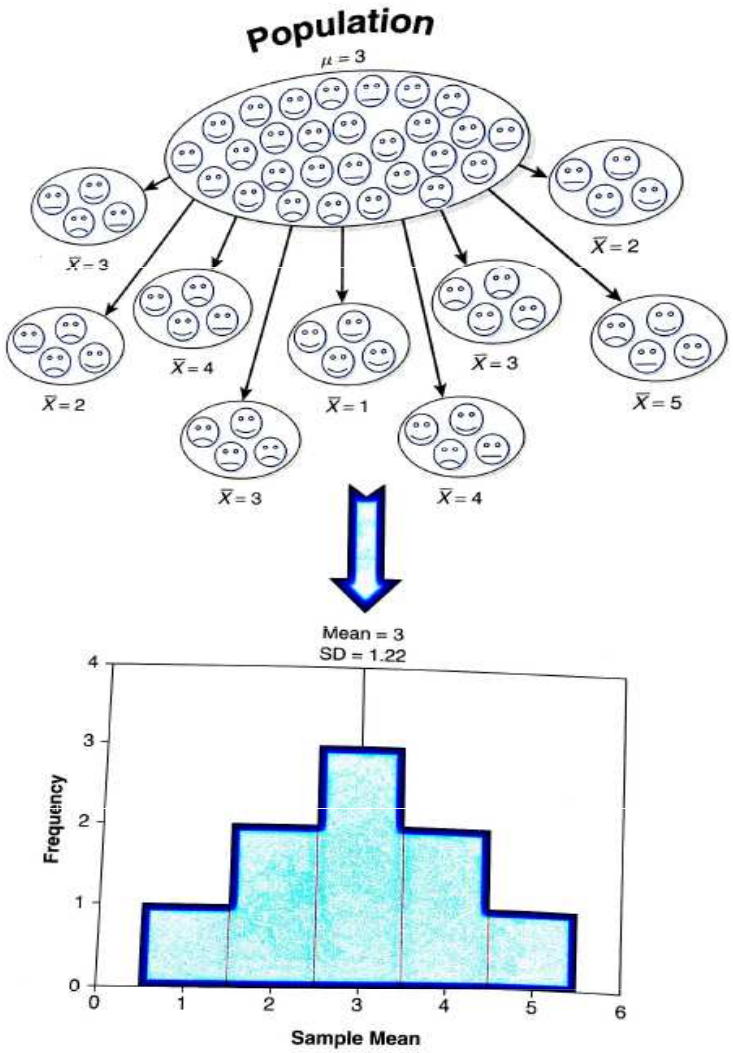


# SAMPLING DISTRIBUTION

- Vzorkovací rozdělení, rozdělení/distribuce vzorků, **vzorková distribuce**
- Z celé populace vyberu náhodně vzorek (např. 100 případů) a zjistím, že jejich průměrná sympatie k premiérovi je 5,1.
- Z populace vyberu náhodně druhý vzorek (jiných 100 případů), zjistím že $\bar{X} = 4,1$.
- Z populace vyberu náhodně další vzorek (jiných 100 případů), zjistím, že $\bar{X} = 7,3$.
- Z populace vyberu náhodně další vzorek (jiných 100 případů), zjistím $\bar{X} = 2,6$.
- Atd.. Opakuji mnohokrát
- Ve skutečnosti mám jen jeden vzorek, ale snažím si představit, že je to jen jeden možný vzorek z mnoha, má určitou míru výběrové chyby, velikost výběrové chyby se liší napříč vzorky. **$\bar{X} = \mu$???????**

# SAMPLING DISTRIBUTION

- Distribuce průměrů pro velmi mnoho (nekonečně mnoho) náhodných vzorků
- Dostanu mnoho různých hodnot pro průměr jednotlivých vzorků
- Čím více mám vzorků, tím více se distribuce průměrů bude připomínat normální rozdělení.
- Nekonečně mnoho vzorků z populace = dokonalé normální rozdělení průměrů těch jednotlivých distribucí a průměrná hodnota průměrů pro jednotlivé vzorky bude rovna skutečnému populačnímu průměru.
- Bude-li populační průměr sympatií k premiérovi 5,2, čím více budu mít vzorků, tím více se jejich průměr bude blížit hodnotě 5,2.


Population
$\mu = 3$
$\bar{X} = 3$
$\bar{X} = 2$
$\bar{X} = 2$
$\bar{X} = 4$
$\bar{X} = 1$
$\bar{X} = 3$
$\bar{X} = 5$
$\bar{X} = 3$
$\bar{X} = 4$
Mean = 3
SD = 1.22
Frequency
0
1
2
3
4
0
1
2
3
4
5
6
Sample Mean

# SAMPLING DISTRIBUTION

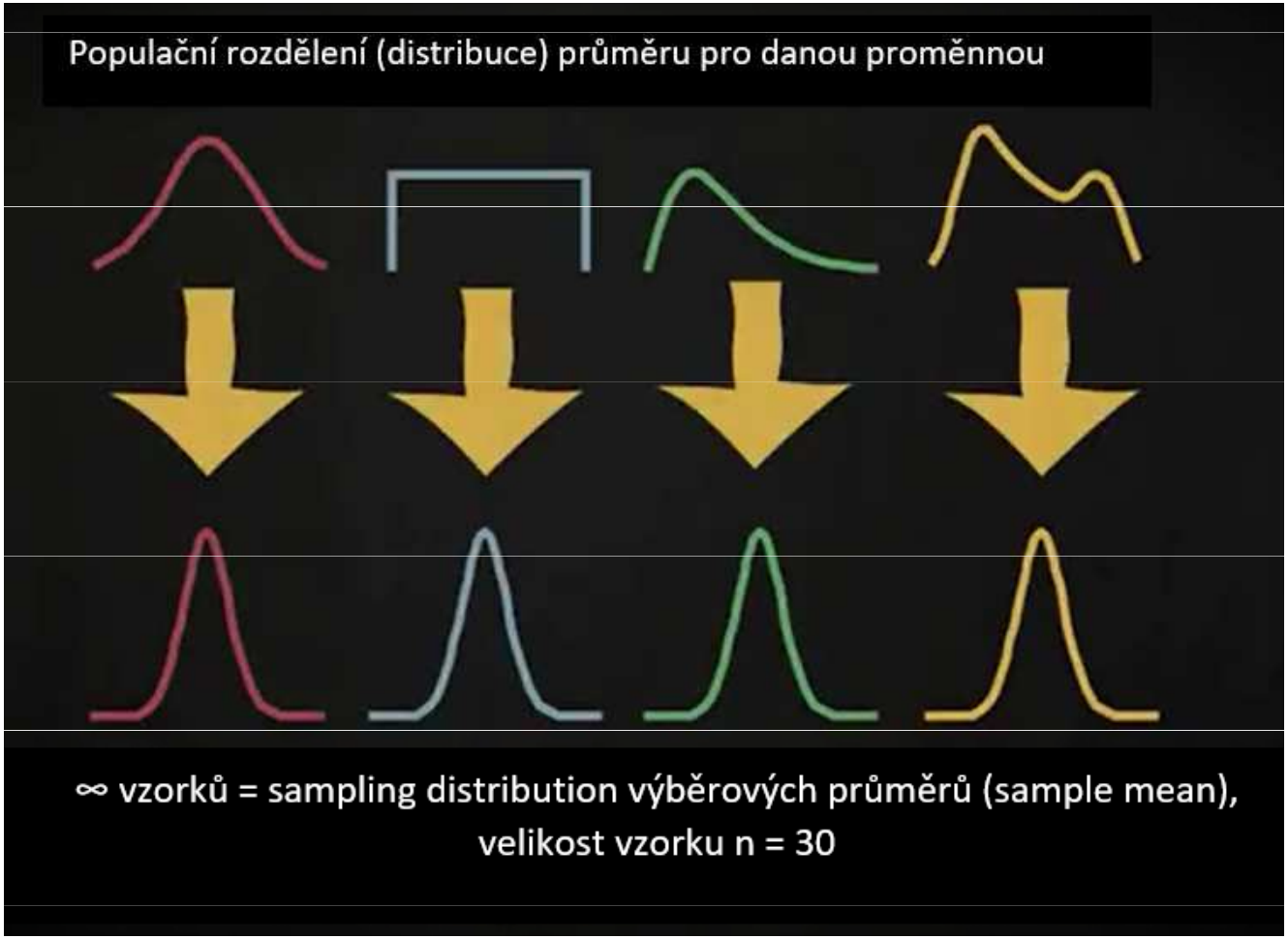
- https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm

- Není to distribuce v populaci (population distribution)
- Není to ani distribuce v jednom vzorku (sample distribution)
- Je to distribuce průměrů pro nekonečně mnoho individuálních vzorků (sampling distribution)

- Kdybychom byly schopni to udělat, tak zjistíme, že průměr této distribuce odpovídá reálnému průměru v populaci
- Jedná se o hypotetický konstrukt! (Nejsme schopni v reálnu udělat velmi mnoho vzorků na té stejné populaci. Navíc neznáme většinou ani populační parametr, abychom si ověřili, že se sampling distribution blíží).

# CENTRAL LIMIT THEOREM

- Centrální limitní teorém
- Provedeme-li velmi mnoho výběrů, bude se distribuce výběrových průměrů (sampling distribution) blížit normální distribuci. Celkový průměr těchto průměrů se bude blížit skutečné hodnotě průměru v dané populaci.
- Platí to i v případě, že distribuce v populaci není normální!!
- Sampling distribution (výběrových) průměrů  bude normální.
- Za předpokladu, že **máme dostatečně velký vzorek**.


Populační rozdělení (distribuce) průměru pro danou proměnnou
∞ vzorků = sampling distribution výběrových průměrů (sample mean),
velikost vzorku n = 30

# POZOR NA ROZDÍLY

- Population mean (průměr všech hodnot v dané populaci, $\mu$)
- Sample mean (průměr v individuálním vzorku, $\bar{x}$)
- Sampling distribution of the sample mean (distribuce výběrových průměrů/průměrů jednotlivých vzorků $\mu_{\bar{x}} = \mu$)

# SAMPLING DISTRIBUTION (VÝBĚROVÁ DISTRIBUCE)

- Má i směrodatnou odchylku
- Směrodatná odchylka od průměru průměrů
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Problém může být v tom, že neznáme populační s.d.
- Pokud neznáme populační s.d., tak ji odhadujeme ze směrodatné odchylky vzorku
- Tento odhad nazýváme směrodatnou chybou (s.e)
- $se = \frac{s}{\sqrt{n}}$

# SMĚRODATNÁ CHYBA

- Liší se od směrodatné odchylky
- S.D.: rozptýlenost naměřených hodnot kolem průměru vzorku. Čím větší, tím méně jsou si hodnoty podobné. Popisná statistika.
- S.E.: rozptýlenost (nikoliv hodnot) průměrů velkého množství vzorků téže populace, naznačuje, zda jde naměřený průměr dobře zobecnit na populaci. Pokud je S.E. velká, mám velkou variabilitu průměrů a můj průměr nemusí dobře odhadovat populační průměr. Inferenční statistika.

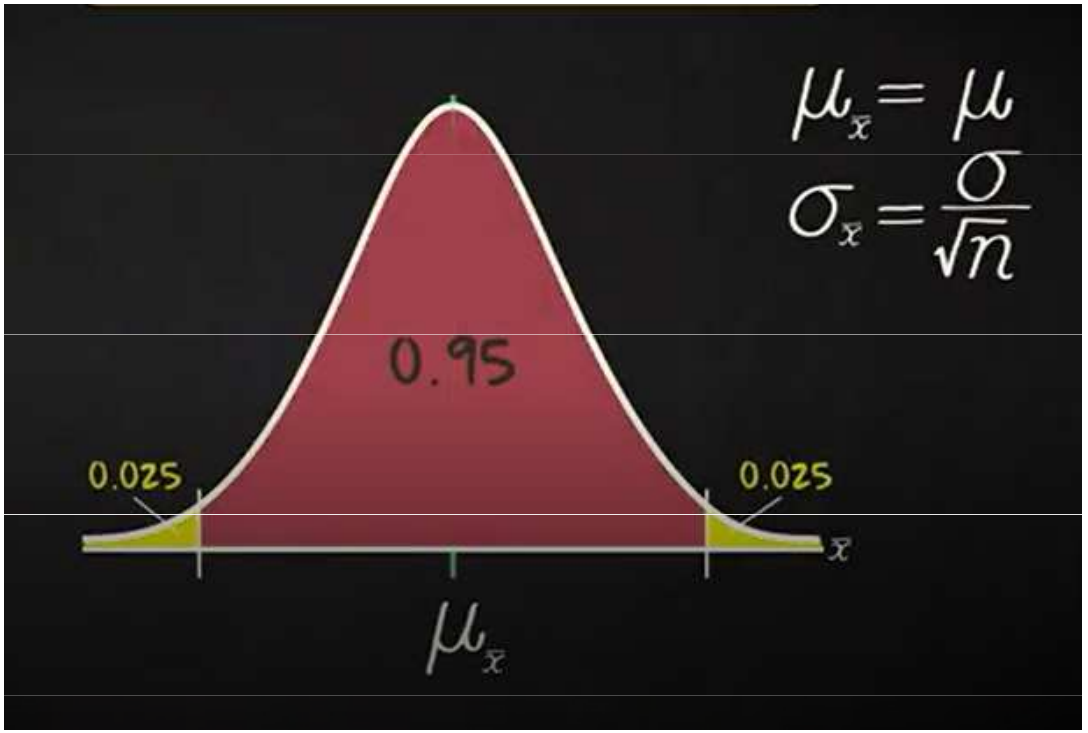
# STATISTICKÁ INFERENCE

- Odhadujeme populační parametry, nebo testujeme hypotézy
- Odhady:
  - Bodové, odhaduje konkrétní hodnotu (např. průměr)
  - Intervalové. Odhaduje interval hodnot, do kterého populační parametr pravděpodobně spadá.
- Zajímá nás i ta pravděpodobnost, se kterou parametr do intervalu spadá.
- P je blízko 1, většinou 95% (0,95)
- 95% interval spolehlivosti

# INTERVAL SPOLEHLIVOSTI

- Vycházím ze sampling distribution (normální distribuce, $\mu$, $\sigma/n-1$)

# PLUS MÍNUS DVĚ SMĚRODATNÉ ODCHYLKY OD PRŮMĚRU LH2

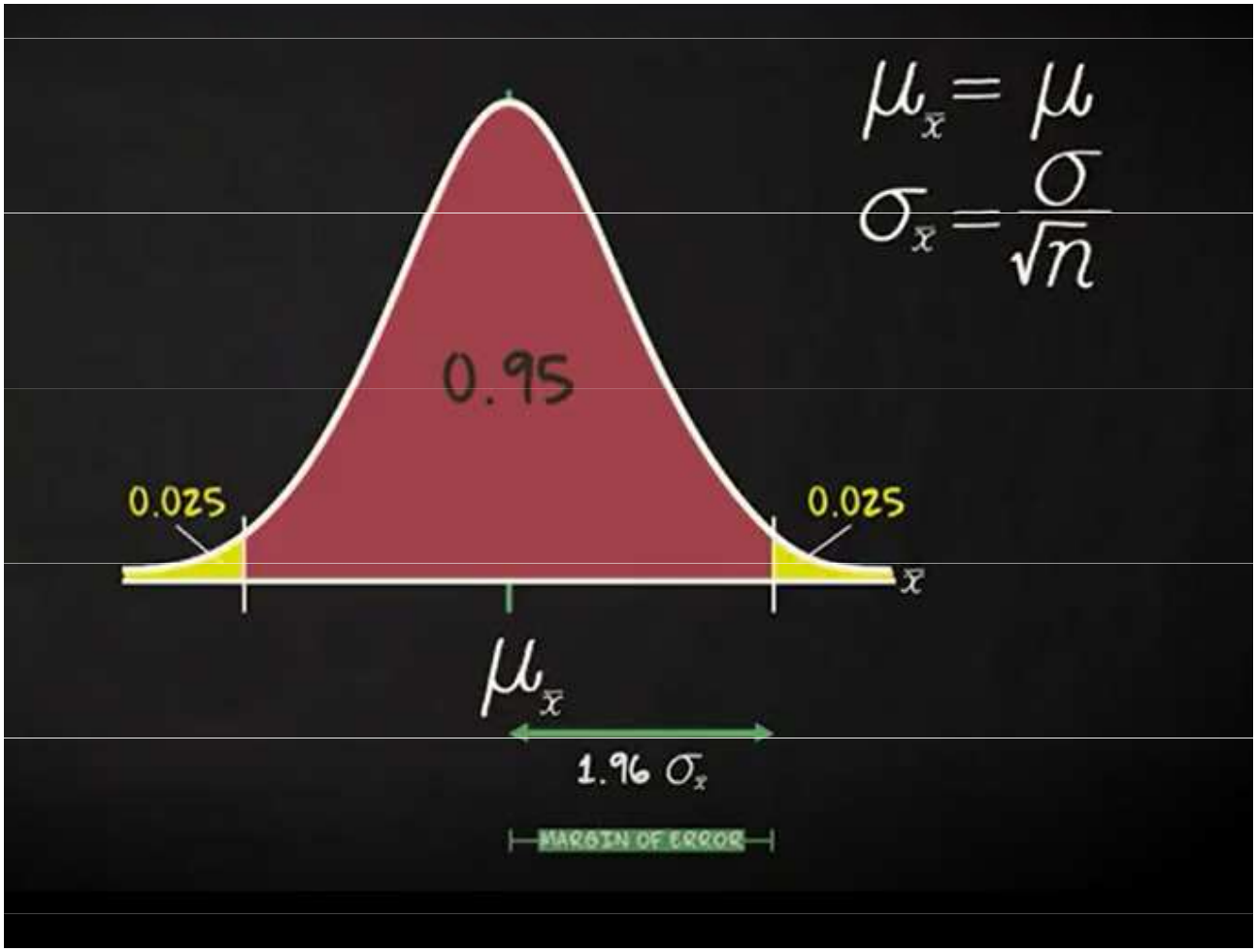


## Snímek 18

**LH2** Pravděpodobnost, že najdeme průměr menší než 2 směrodatné odchylky pod průměrem je 0,025. Pravděpodobnost, že naměřený průměr bude větší něž dvě směrodatné odchylky nad průměrem je taky 0,025.

Lenka Hrbková, 11/2/2020

LH3
LH4
LH6

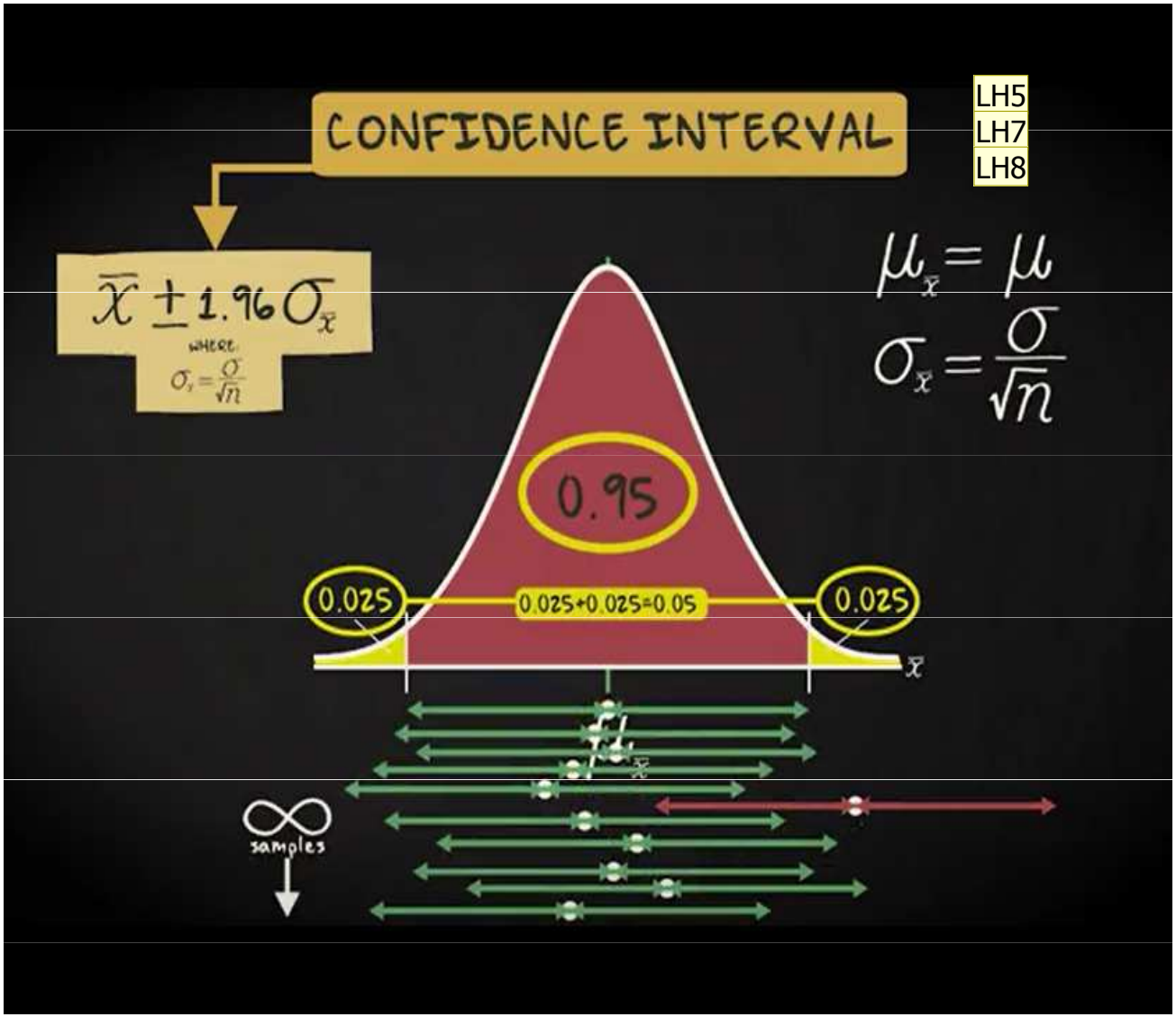

$\mu_{\bar{x}} = \mu$
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
0.95
0.025
0.025
$\bar{x}$
$\mu_{\bar{x}}$
1.96 $\sigma_{\bar{x}}$
MARGIN OF ERROR

## Snímek 19

**LH3** Když si najdu hodnotu pro pravděpodobnst 0,025 v Z-tabulce, zjistím, že je to -1,96. Pro hodnotu 0,975 (1-0,025) je to 1,96.
Lenka Hrbková, 11/2/2020

**LH4** Takže je 95% pravděpodobnost, že můj průměr bude spadat do intevalu 1,96 směrodatných odchylek (plus a mínus) od průměru mju.
Lenka Hrbková, 11/2/2020

**LH6** Bude-li můj sample mean (výběrový průměr, který jsme naměřila na vzorku) spadat někam do té červené plochy(plus mínus 1,96 směrodatných odchylek od populačního průměru, tak bude jeho interval spolehlivosti obsahovat i ten populační průměr.
Lenka Hrbková, 11/2/2020


CONFIDENCE INTERVAL
LH5
LH7
LH8
$\bar{x} \pm 1.96\sigma_{\bar{x}}$
WHERE
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
$\mu_{\bar{x}} = \mu$
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
0.95
0.025
0.025+0.025=0.05
0.025
$\bar{x}$
samples

## Snímek 20

**LH5** Interval spolehlivosti prů můj sample mean je v podstatě můj naměřený průměr vzorku (sample mean) plus mínus 95% úroveň spolehlivosti krát směrodatná odchylka.
Lenka Hrbková, 11/2/2020

**LH7** $\bar{x} \pm 1.96\sigma_{\bar{x}}$
Lenka Hrbková, 11/2/2020

**LH8** Z náhodného vzorku naší populace spočítáme průměr a k němu přičteme a odečteme výběrovou chybu pomocí uvedeného vzorce. Tím dostaneme dvě hodnoty, které ohraničují interval spolehlivosti na úrovni 95%. Pokud bychom tuto operaci opakovali mnohokrát, tak bude pravý populační průměr mjů spadat do 95% spočítaných intervalů spolehlivosti pro jednoltivé průměry vzroků. (Na obrázku nespadá mjů  do konfidenčního intervalu toho průměru, jehož interval je vyznačen červeně).
Lenka Hrbková, 11/2/2020

# DŮVĚRA VLÁDĚ (0-10)

- $\bar{x} = 5{,}2;\ \sigma = 2{,}1;\ n = 100$
- $\overline{x} \pm 1.96\sigma_{\overline{x}}$
- 1) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 2{,}1/10 = 0{,}21$
- 2) 1,96*0,21 = 0,41
- 3) 5,2 + 0,41 = 5,61
- 4) 5,2 – 0,41 = 4,79
- 5) **95% CI (4.79, 5.61) - populační průměr s 95% pravděpodobností spadá do tohoto intervalu.**

# CO KDYŽ NEZNÁM POPULAČNÍ SMĚRODATNOU ODCHYLKU?

LH9

$$\bar{x} \pm 1.96\sigma_{\bar{x}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

**Snímek 22**

**LH9** Tu potřebuju k vypočítání směrodatné odchylky sampling distribution.
Lenka Hrbková, 11/2/2020

# ODHADNEME POPULAČNÍ SMĚRODATNOU ODCHYLKU

- Odhad pomocí výběrové směrodatné odchylky (směrodatné odchylky vzorku).
- Odhad se nazývá směrodatná chyba! **Standard error**, SE

$$\bar{x} + z_{95\%} se$$

$$\bar{x} - z_{95\%} se$$

$$\bar{x} + 1{,}96 se$$

$$\bar{x} - 1{,}96 se$$

VÝBĚROVÁ SD

$$se = \frac{s}{\sqrt{n}}.$$

# $N = 200,\ \bar{X} = 4,3,\ S = 1,6$

- Spočítejte mi 95% CI
- 1) spočítejte **SE**:
  - $se = \frac{s}{\sqrt{n}}$ .
  - Se = 1,6/14,14 = 0,11
- 2) **Horní hranice** CI:
  - 4,3 + 1.96*0, 11 = 4,52
- 3) **Spodní hranice** CI:
  - 4,3 – 1.96*0, 11 = 4,08
- 4) 95% CI (4.08, 4.52)

# CO KDYŽ CHCEME JINÝ CI?

- Můžeme chtít jinou hladinu spolehlivosti, ne nutně 95%
- Někdy se používá 99%
- Někdy 90%

LH13
LH14

- Chci 99% CI: interval -2,58 až +2,58
- Pro náš příklad:
  - 4,3 + 2,58*0,11 = 4,58,
  - 4,3 - 2,58*0,11 = 4,02

**99% CI (4.02 – 4.58)**

## Snímek 25

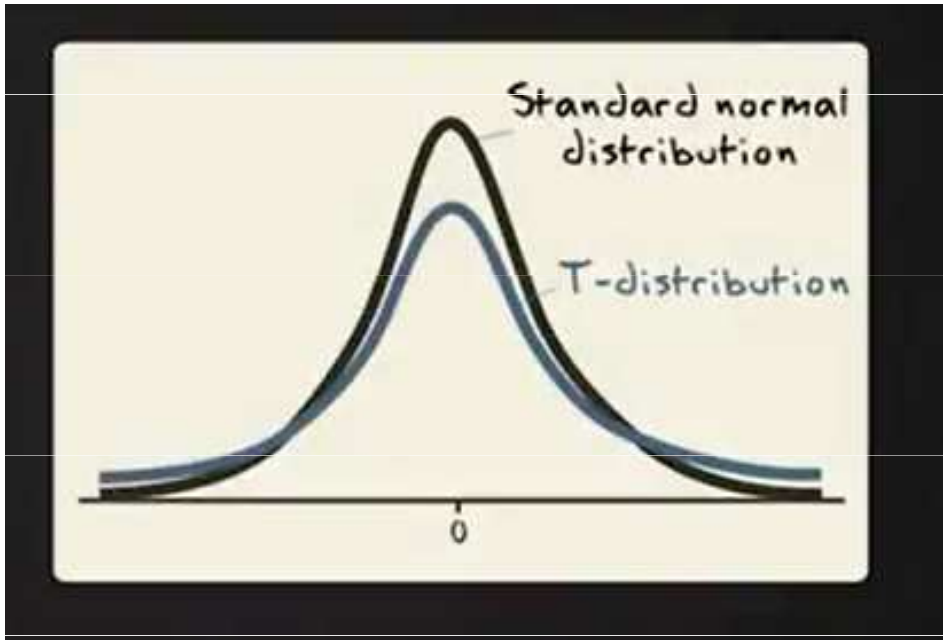
**LH13** Pro 99% spolehlivost je ta hodnota 2,58. Pro 90% je to 1,64. Nejčastěji se ale používá 1,96 (pracujeme s 95% spolehlivostí).
Lenka Hrbková, 11/2/2020

**LH14** 99% CI = je o něco větší než u 95%. Znamená to, že máme větší jistotu, že populační parametr spadá do tohoto intervalu (99% pravděpodobnost místo 95%) Ale na druhou stanu je ten interval větší, tzn. že náš odhad je míň přesný.
Lenka Hrbková, 11/2/2020

# CI PRO PRŮMĚR NA MALÉM VZORKU?

LH10

- Nelze pracovat s z-distribucí (tzv. standardní normální distribuce)
- Ale s tzv. t-distribucí (t-rozdělení)
- T-rozdělení – mění tvar podle velikosti vzorku
- Čím větší vzorek, tím víc jako normální
- Má tlustší konce a větší SD než normální
- Tvar závisí **na stupních volnosti**
- Degrees of freedom, df = **n – 1**



**Snímek 26**

**LH10** Protože aleé vzorky nemají normálně rozdělení. Pracuju s t-rozdělením.
Lenka Hrbková, 11/2/2020

# N = 30, X = 4,3, S = 1,6

LH11
LH12

- Jaký je 95%CI???
  - Horní hranice CI = $\bar{x} + (t_{df} * SE)$
  - Dolní hranice CI = CI = $\bar{x} - (t_{df} * SE)$
- Kritické hodnoty najdeme v tabulce: https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf
  - t = hodnota t v tabulce pro df29 a 95% je 2,052
  - SE = 0,29
  - $\bar{x} + (2{,}052*0{,}29) = 4{,}9$
  - $\bar{x} – (2{,}052*0´{,}29) = 3{,}75$

## Snímek 27

**LH11** Tady nepracuju se z-tabulou ale t-tabulkou, kde hledám kritickou hodnotu pro danou úroveň spolehlivosti v závislosti na počtu stupňů volnosti.
Lenka Hrbková, 11/2/2020

**LH12** Pro 95% a 29 stupňů volnosti (30-1) je to hodnota 2,052. TUto hodnotu vynásobím standardní chybou. Pak je postup stejný jako před tím.
Lenka Hrbková, 11/2/2020

# CI PRO I PRO PROCENTA VÝSKYTU

- Relativní četnosti
- Pokud mám nějakou proporci ve vzorku (např voliči Pirátů), zají má mě, CI pro populační proporci. LH15
- **Horní hranice CI: p + 1,96 x** $\sqrt{p(1-p)/n}$
- **Dolní hranice CI: p - 1,96 x** $\sqrt{p(1-p)/n}$
- p = pozorovaná **relativní četnost v našem vzorku** (%)
- n = velikost vzorku

**Snímek 28**

**LH15** Například u tech průzkuů volebích. Zajímá mě interval, do kterého spadá populační proporce voličů Pirátů. Odhaduju ho na základě proporce (%) voličů Pirátů z mého vzorku.
Lenka Hrbková, 11/2/2020

# Ú K O L :

- Volební model
- N = 917
- Voliči Pirátů p = 16%
- Spočítejte 95% CI pro toto procento voličů.

# ZÁVĚR

- S každým vzorkem se pojí výběrová chyba
- Tím, že vyberu vzorek, tak se více či méně odchýlím od přesných parametrů v populaci
- Odhaduju-li populační parametr ze vzorku, je dobré nepoužívat bodový odhad (protože je tam vždycky ta chyba)
- Použiju intervalový odhad, s určitou mírou jistoty mi řekne, že parametr (např. průměr) leží v daném intervalu.
- Pracuju se standardním normálním rozdělením (95% pravděpodobnost, že statistika mého vzorku (např výběrový průměr) leží plus mínus **1,96** směrodatných odchylek od populačního průměru).
- Když pracuju s populačními daty, tak to ani nemusím řešit, protože nemusím nic odhadovat. Prostě si to změřím a vím to s jistotou.