Úvod do teorie odhadu.

Bodové a intervalové odhady

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

doc. Ing. Petr Seďa, Ph.D.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co se dnes dozvíte?

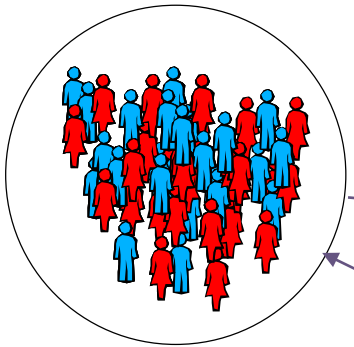
- Základní a výběrový soubor, náhodný výběr.
- Teoretické a výběrové statistiky.
- Bodové odhady a jejich vlastnosti.
- Intervalový odhad střední hodnoty.
- Studentovo rozdělení.
- Intervalový odhad podílu výskytu.

Teorie odhadu

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Základní a výběrový soubor:

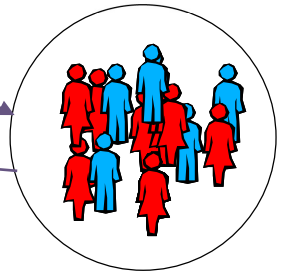
základní soubor



výběrový soubor (vzorek)
- obsahuje vybrané statistické jednotky
- parametry vzorku lze spočítat

výběr

odhad



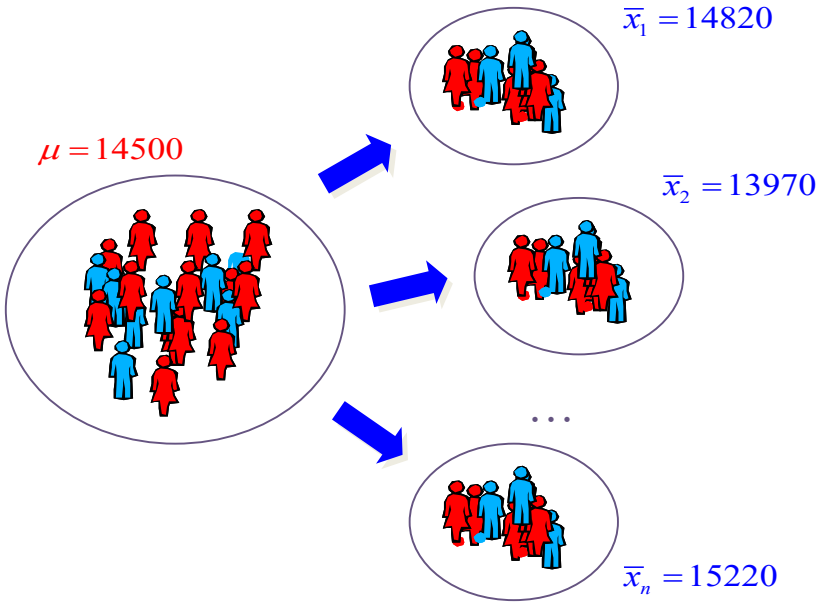
výběrový soubor

základní soubor (populace)
- obsahuje všechny možné statistické jednotky
- parametry populace neznáme, ale chceme určit

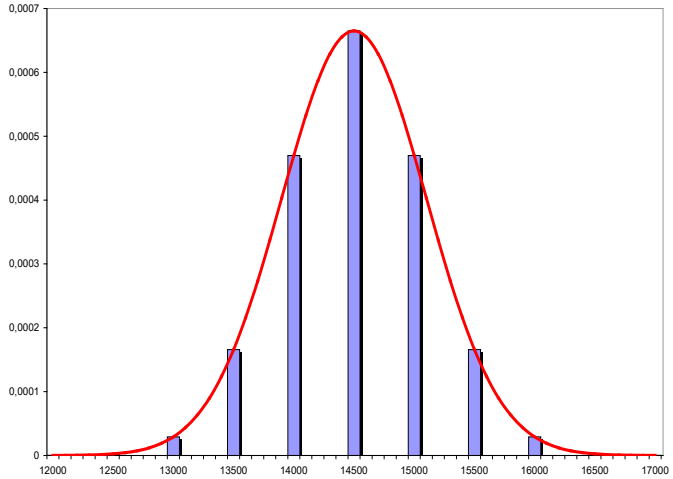
SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Parametry výběru jsou náhodné veličiny:

populace o velikosti $N \rightarrow n = C_k(N)$ různých k-prvkových výběrů



rozdělení výběrového průměru



SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Statistické šetření:

předmět šetření - statistický soubor, na němž se zkoumání provádí

obsah šetření – statistické znaky (proměnné), které se v daném souboru sledují

metody šetření – způsoby, kterými se šetření (zjišťování) provádí, a metody následné analýzy

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Výběrové plány (metody výběru):

náhodný (pravděpodobnostní) výběr - všechny jednotky populace mají stejnou šanci, že budou zařazeny do výběru

záměrný výběr - jednotky jsou do výběrového souboru vybírány záměrně, o pořadí výběru rozhodují různá kritéria

- v praxi se snažíme o náhodný výběr, ale každý výběr je v podstatě svým způsobem záměrný (organizátor výběru vnáší do volby subjektivní prvky)

Teorie odhadu

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Výběrové metody (metody šetření):

odhady parametrů - na základě znalosti výběrového souboru se provádí co nejlepší odhad parametrů znaků základního souboru

bodový odhad - *neznámý parametr se odhaduje jedinou nejpravděpodobnější hodnotou*

intervalový odhad – *neznámý parametr se odhaduje pomocí intervalu hodnot, které jsou s danou pravděpodobností možnými hodnotami parametru*

testování hypotéz – na základě znalosti výběrového souboru se potvrzuje nebo vyvrací dané tvrzení o parametrech znaků základního souboru

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vlastnosti dobrého odhadu:

nezkreslenost (nevychýlenost, nestrannost) - střední hodnota nezkresleného odhadu se rovná odhadovanému parametru $E(u_n) = Q$

konsistence - s rostoucím počtem prvků výběrového souboru se konsistentní odhad zpřesňuje (jeho variabilita se snižuje) $\lim_{n \to \infty} \sigma(u_n) = 0$

vydatnost – vydatný odhad má nejmenší variabilitu ze všech možných

Bodový odhad

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Odhad střední hodnoty μ:

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

nejlepším odhadem střední hodnoty je výběrový průměr

parametry výběrového průměru:

$$E(\bar{x}) = \mu_{\bar{x}} = \mu$$
$$SE(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

SE – standard error
střední chyba odhadu

Odhad střední hodnoty σ^2:

$$E\left(\sigma^2\right) = s^2$$

nejlepším odhadem rozptylu populace je výběrový rozptyl

výpočet výběrového rozptylu:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$

$n-1$ stupeň volnosti

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Rozptyl populace a výběrový rozptyl:

rozptyl populace (***N*** – velikost populace)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

výběrový rozptyl (***n*** – velikost vzorku)

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}{n-1}$$

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Centrální limitní věta:

centrální limitní věta – rozdělení průměru vzájemně nezávislých náhodných veličin s konečnou střední hodnotou $\boldsymbol{\mu}$ a konečným rozptylem $\boldsymbol{\sigma^2}$ konverguje k normálnímu rozdělení $\boldsymbol{N(\mu\,;\sigma^2)}$

důsledek: výběrový průměr má normální rozdělení

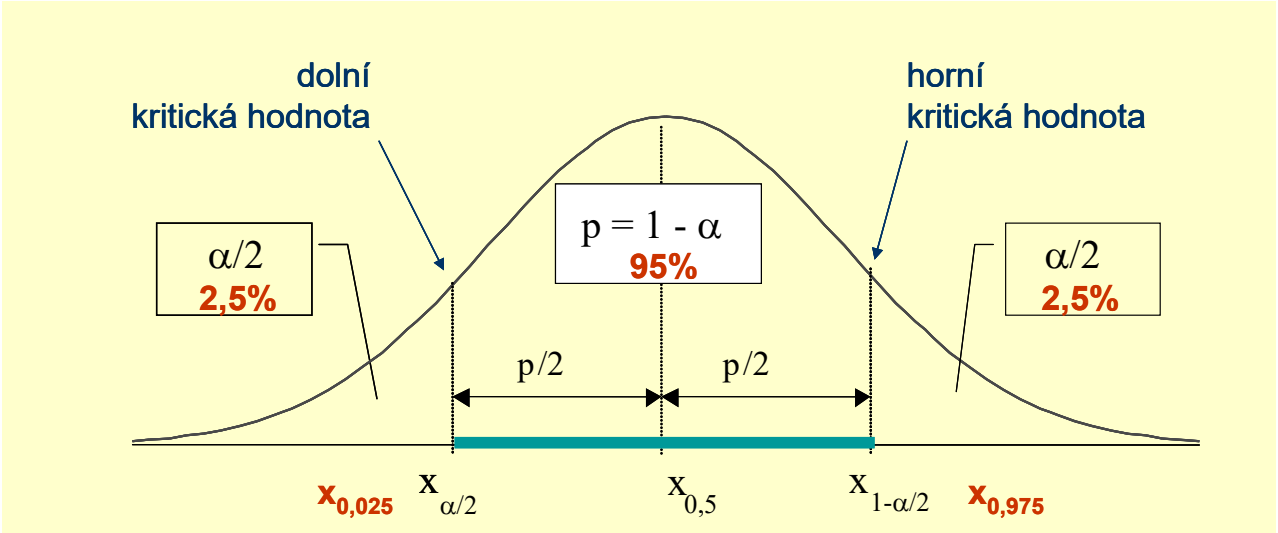
SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Intervalový odhad střední hodnoty:

p - interval spolehlivosti – konfidenční interval

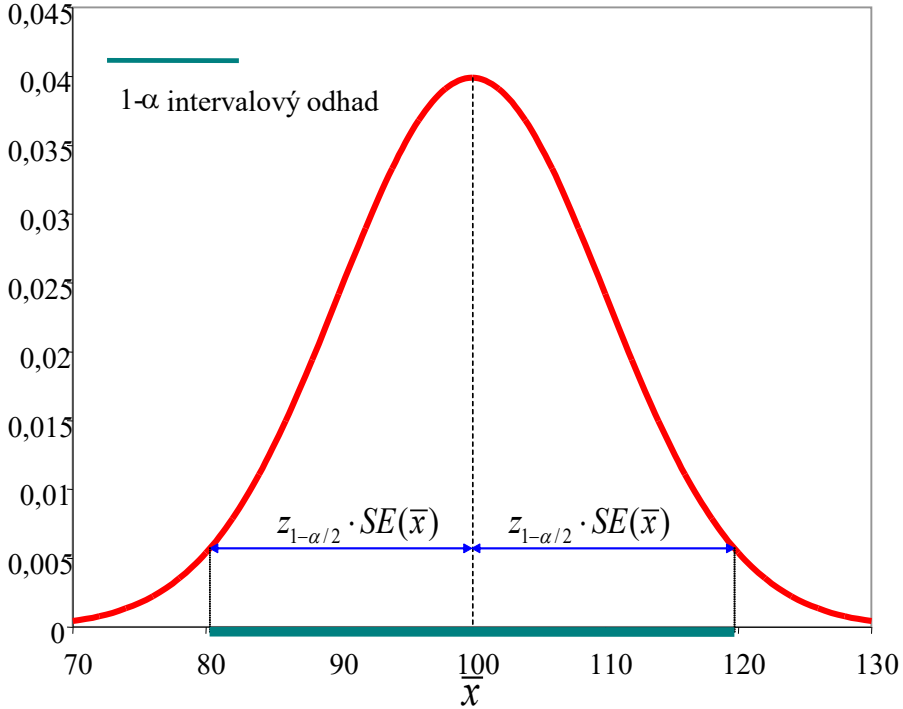
$p = 1 - \alpha$ spolehlivost odhadu (obvykle 95%)

α hladina významnosti = riziko chyby



SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Interval spolehlivosti a střední chyba:



SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Interval spolehlivosti střední hodnoty:

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$\alpha/2$ – kvantil rozdělení výběrového průměru

$1-\alpha/2$ – kvantil rozdělení výběrového průměru

jak určit střední chybu odhadu, neznáme-li σ ?

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

je-li vzorek dostatečně velký ($n > 30$), lze variabilitu populace odhadnout pomocí výběrové variability

Příklad 1 – normální rozdělení

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Při průzkumu vytíženosti praktického lékaře byla zjištěna u 50 náhodně vybraných pacientů průměrná doba na vyšetření 14,3 minuty se směrodatnou odchylkou 4,6 minuty. Odhadněte průměrnou délku pobytu jednoho pacienta u lékaře pomocí 95% intervalu spolehlivosti.

$$14{,}3 - 1{,}96 \cdot \frac{4{,}6}{\sqrt{50}} < \mu < 14{,}3 + 1{,}96 \cdot \frac{4{,}6}{\sqrt{50}}$$

$$14{,}3 - 1{,}3 < \mu < 14{,}3 + 1{,}3$$

$$13{,}0 < \mu < 15{,}6$$

S 95% pravděpodobností se bude průměrná délka pobytu u lékaře pohybovat mezi 13,0 a 15,6 minutami.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co když je vzorek malý?

$n < 30$

odhad s podhodnocuje skutečnou variabilitu σ

normální rozdělení → Studentovo rozdělení

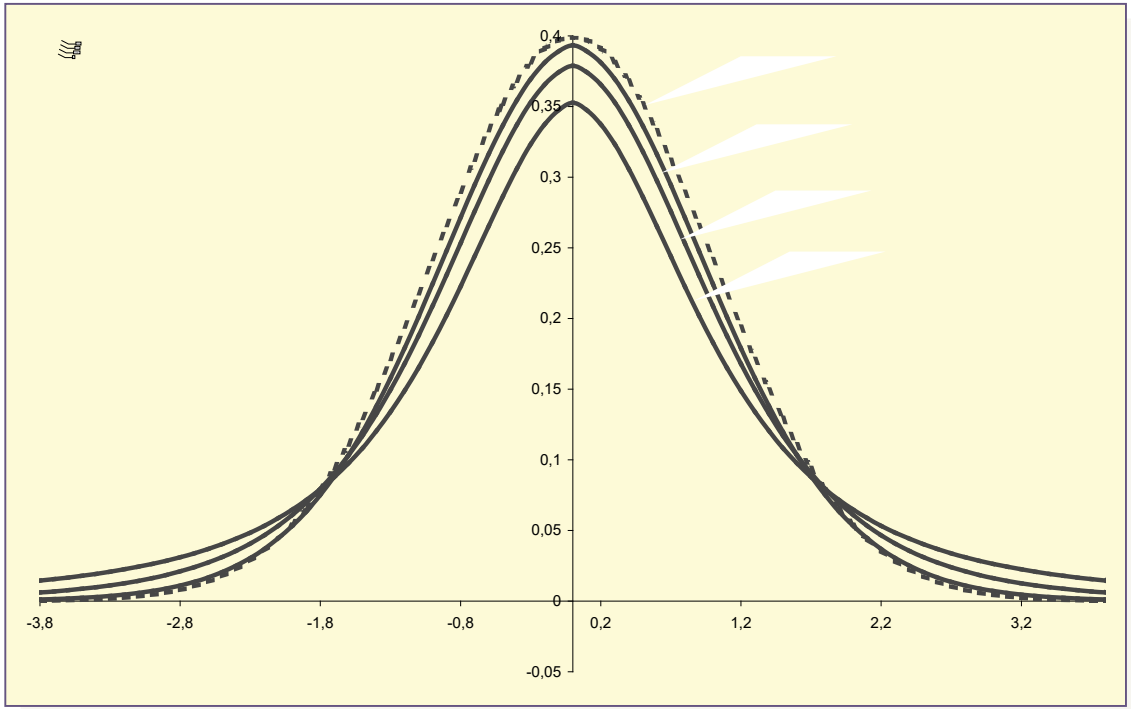
$$\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

$t(v)$ – Studentovo rozdělení s $v = n - 1$ stupni volnosti

Intervalové odhady

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Studentovo *t* rozdělení:



SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vlastnosti Studentova t rozdělení:

***t* - rozdělení** má podobný průběh a tvar jako normální

***t* - rozdělení** má stejnou střední hodnotu, ale větší rozptyl než normované normální rozdělení

***t* - rozdělení** se s rostoucí hodnotou parametru v blíží normovanému normálnímu rozdělení

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jak počítat interval spolehlivosti?

$\boldsymbol{n > 30}$

$$\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

$\boldsymbol{n < 30}$

$$\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

$t(v)$ – Studentovo rozdělení s $v = n - 1$ stupni volnosti

Příklad 2 – Studentovo rozdělení

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Prodejna chce zjistit průměrný počet zákazníků v páteční odpolední směně. Po dobu 2 měsíců tedy sleduje počet zákazníků, kteří prošli pokladnami prodejny, s tímto výsledkem: 527, 418, 495, 554, 392, 548, 449, 511. Určete 95% intervalový odhad pro průměrný počet zákazníků obsloužených v jedné směně.

$$486,75 - 2,365 \cdot \frac{60,57}{\sqrt{8}} < \mu < 486,75 + 2,365 \cdot \frac{60,57}{\sqrt{8}}$$

$$436,1 < \mu < 537,4$$

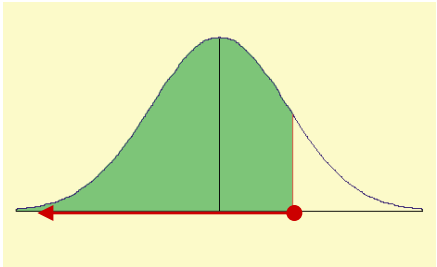
Průměrný počet zákazníků obsloužených v páteční odpolední směně se tedy bude dlouhodobě pohybovat mezi 437 a 537.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jednostranné intervaly:

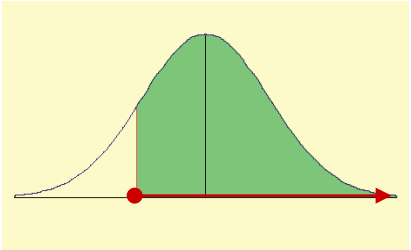
pravostranný interval:

$$\mu < \bar{x} + t_{1-\alpha}(n-1) \cdot \frac{s}{\sqrt{n}}$$



levostranný interval:

$$\mu > \bar{x} - t_{1-\alpha}(n-1) \cdot \frac{s}{\sqrt{n}}$$



SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Odhad podílu výskytu vlastnosti π:

$$E(\pi) = p = \frac{m}{n}$$

pokud platí podmínky Moivre-Laplaceovy věty:
- rozsah výběrového souboru $n > 30$
- rozptyl binomického rozdělení $n \cdot p \cdot (1-p) > 9$

$$p - z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$$

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jednostranné intervaly pro odhad podílu:

Jak budou vypadat jednostranné intervaly pro odhad podílu?

pravostranný interval: $\pi < p + z_{1-\alpha} \cdot \sqrt{\frac{p(1-p)}{n}}$

levostranný interval: $\pi > p - z_{1-\alpha} \cdot \sqrt{\frac{p(1-p)}{n}}$

Příklad 3 – odhad podílu

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

V předvolební kampani do Senátu si jeden z kandidátů X.Y. chce ověřit svoje možnosti na zvolení, a proto si nechá udělat průzkum preferencí. Namátkovým dotazováním u 200 potenciálních voličů bylo přitom zjištěno, že kandidáta X.Y. by volilo 106 z nich. Zaručuje mu tento výsledek vítězství již v prvním kole? Určete na základě jednostranného intervalového odhadu se spolehlivostí 95%.

$$p = \frac{106}{200} = 0,53$$

podmínky Moivre-Laplaceovy věty:

$n = 200 > 30 \quad n \cdot p \cdot (1 - p) = 49,82 > 9$

Příklad 3 – odhad podílu

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

V předvolební kampani do Senátu si jeden z kandidátů X.Y. chce ověřit svoje možnosti na zvolení, a proto si nechá udělat průzkum preferencí. Namátkovým dotazováním u 200 potenciálních voličů bylo přitom zjištěno, že kandidáta X.Y. by volilo 106 z nich. Zaručuje mu tento výsledek vítězství již v prvním kole? Určete na základě jednostranného intervalového odhadu se spolehlivostí 95%.

zadání vede na levostranný 95% interval:

$$\pi > p - z_{1-\alpha} \cdot \sqrt{\frac{p(1-p)}{n}}$$

po zadání konkrétních hodnot:

$$\pi > 0,53 - 1,645 \cdot \sqrt{\frac{0,53 \cdot 0,47}{200}} \doteq \boxed{0,47}$$

Se spolehlivostí 95% nemá kandidát záruku nadpoloviční většiny, která je nutná k vítězství v prvním kole.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jak určit rozsah výběru?

Δ – přípustná chyba odhadu (polovina intervalu spolehlivosti)

$$n \geq \left(\frac{z_{1-\alpha/2} \cdot s}{\Delta} \right)^2 \qquad n \geq \frac{z_{1-\alpha/2}^2 \cdot p(1-p)}{\Delta^2}$$

variabilitu $\boldsymbol{s}$ nebo podíl $\boldsymbol{p}$ odhadujeme předvýběrem vzorku o velikosti $\boldsymbol{n_1 > 30}$

poté provedeme doplňkový výběr o velikosti $\boldsymbol{n_2 = n - n_1}$

Příklad 4 – rozsah výběru

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Senátor X.Y. není spokojen s dosaženými předvolebními odhady a chce, aby tyto odhady měly maximální chybu $\pm 2\%$ se spolehlivostí 95%. Jak velkého počtu potenciálních voličů se má jeho volební štáb dotázat?

předvýběrem vyšlo $p = 0{,}53$

$$n \geq \frac{1{,}96^2 \cdot 0{,}53 \cdot 0{,}47}{0{,}02^2} \doteq \boxed{2400}$$

Vzhledem k tomu, že dosud bylo dotázáno pouze 200 voličů, je třeba provést průzkum ještě u dalších 2200 voličů. Je ovšem otázka, zda je tak rozsáhlý průzkum v silách volebního štábu.

Literatura

1. Janáček J. *Statistika jednoduše*. Praha: Grada, 2022. **(kapitola 5).**

2. Ramík J. a Čemerková Š. *Statistika A.* Opava, Karviná: SLU, 2000. **(kapitola 6 a 7).**

Děkuji za pozornost.