Statistické zjišťování

Úvod

Základní soubor
Všechny ryby v rybníce, všechny holky/kluci na škole
Cílem určit charakteristiky, pravděpodobnosti
Průměr, rozptyl, pravděpodobnost, že Maruška „kápne“ na toho pravého

Provedení vyčerpávajícího šetření – změří se všechny jednotky
Nákladné, nemožné atd.

Výběrové šetření a výběrový soubor
Vybereme pouze část ze základního souboru (100 ryb, 50 chlapců/děvčat)
Následně budeme usuzovat vlastnosti základního souboru z výběrového souboru

1)Maruška „otestuje“ 50 chlapců (na škole 5000)
2) Bude určovat charakteristiky – výběrové charakteristiky (**statistiky**)
na celé škole
Průměr, pravděpodobnost – nad 17 cm velikost PRSTŮ atd. ☺

Postup

Základní soubor má N jednotek – 5000 chlapců
Maruška vytvoří výběrový soubor o *n* jednotkách – 50 chlapců
Provede úsudek o celém základním souboru

Základní soubor
- konečně velký
-nekonečně velký

Snaha, aby výběrový soubor měl stejné vlastnosti jako základní soubor

Výběr
Prostý náhodný výběr – co se náhodně Marušce dostane pod ruku (nejčastější)
(každý chlapec má stejnou pravděpodobnost, že se dostane k Marušce)
Prostý náhodný výběr rozlišujeme na výběr s vracením/bez vracením
(s vracením – Lukáš se může dostat do Marušky výzkumu vícekrát)

Výběr s nestejnou pravděpodobností
Maruška bude hledat jednotky výběrového souboru ve školním plaveckém týmu

Bodový a intervalový odhad

Statistický soubor lze popsat pomocí charakteristik
Průměr, rozptyl, relativní četnost…

Kvůli zjednodušení zjišťujeme výběrové charakteristiky – nazýváme statistiky

Základní soubor je pevně dán (ryby v rybníce, studenti na škole)
Statistiky se však mění – mají charakter náhodné veličiny
Jedná se o náhodný výběr – průměr, rozptyl bude kolísat

Na škole je 5000 chlapců – Maruška provede náhodný výběr 50 chlapců
Nějakým způsobem zjistí velikost PRSTŮ a spočítá průměr
Průměr bude statistikou – nazveme výběrový průměr (x^-)
Získá rozdělení výběrových průměru (12,15,13,30-Uganda,….)
Díky tomu si bude moc Maruška udělat úsudek o celém základním souboru 5000

Chceme-li odhadnout hodnoty charakteristiky základního souboru
Je nutné znát pravděpodobnostní rozdělení vhodné výběrové statiky
Znát její výběrové rozdělení!!!
-velikost PRSTŮ se řídí normálním rozdělením

Odhad neznámé charakteristiky základního souboru

Buď Maruška spočítá průměr, jedno číslo 16 cm
Bodový odhad – 16cm je bodový odhad průměru základního souboru

Statistika (g) „průměr“ je odhadem charakteristiky (G) „průměr“ Z.S.

Statistiku g – nemůžeme volit jak chceme – pravidla:

Statistika nesmí vést k systematickým chybám E(g)=G
- g je **nezkresleným** (nevychýleným) odhadem G (charakteristiky ZS)

Asymptoticky nezkreslený odhad $\lim_{n\to\infty}\{E(g)-G\}=0$

Co když máme více nezkreslených statistik?
Vybereme statistiku s co nejmenším rozptylem
Vydatný odhad

Co když je odhad zkreslený? $\lim_{n\to\infty} P\{|g-G|<\varepsilon\}=1 \quad \varepsilon>0$
Požadavek **konzistentního odhadu**

Intervalový odhad

Častější použití
Odhad určité charakteristiky základního souboru (Maruška a průměr)
Pomocí intervalu – průměr bude (16,5-17,4)

Odhad základního souboru je určen intervalem (G_d,G_h)
Dolní a horní interval
Interval říká: daná charakteristika bude v něm ležet s vysokou pravděpodobností
Tato pravděpodobnost se nazývá **spolehlivost odhadu** (1-α)

Maruška bude moci třeba říci:
Průměrná velikost PRSTŮ celého základního souboru (5000 chlapců)
Se na 95% (α=5%) nachází v intervalu (16,5;17,4)

$$P(G_d < G < G_h) = 1 - \alpha$$

100.(1-α)% interval spolehlivosti – konfidenční interval

Spolehlivost vs. Přesnost

Spolehlivost je dána (1-α) – čím menší α – tím větší spolehlivost
Když α=0,01 – spolehlivost, že charakteristiky bude ležet v intervalu – 99%

ALE

S rostoucí spolehlivostí bude růst velikost intervalu – bude klesat přesnost
Maruška ví, na 95% - průměr (16,5;17,4)
Nebo Maruška ví, že na 99% - průměr (16,3;17,5)

Mezi přesností a spolehlivostí existuje nepřímá úměra !!!

Intervaly spolehlivosti

Rozlišujeme jednostranné a dvoustranné

Jednostranné – horní/dolní mez

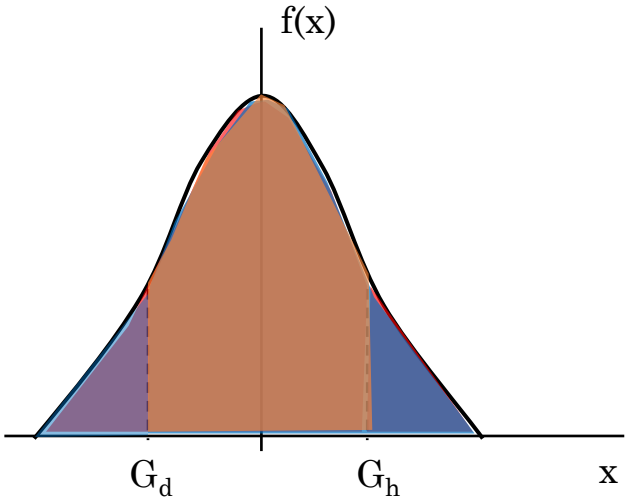
Horní mez (G_h) – pravostranný interval $P(G \geq G_h) = \alpha \quad P(G < G_h) = 1 - \alpha$

Dolní mez (G_d) – levostranný interval $P(G \leq G_d) = \alpha \quad P(G > G_d) = 1 - \alpha$

Např: Maruška může říct, že s 95% je průměrná velikost větší jak 16,8 cm
Jedná se o levostranný interval – 16,8 je doplní mez

Dvoustranný interval

$$\boldsymbol{P(G_d < G < G_h) = 1 - 2\alpha}$$



16,8

Odhad průměru základního souboru

Pro dostatečně velký rozsah výběru – pravděpodobnostní rozdělení
Výběrových průměrů přibližně normální $N[\mu;\sigma^2/n]$
Transformace na normované normální rozdělení

$$U = \frac{\bar{x} - \mu}{\sigma} . \sqrt{n}$$

Už známe, zjednodušení pro výpočet NNR v tabulkách
Když známe x-,μ,σ,n, tak vypočítáme U

$P(U < u_\alpha) = F(u_\alpha) = \alpha$

$P(U < u_{1-\alpha}) = F(u_{1-\alpha}) = 1 - \alpha$

$u_\alpha = -u_{1-\alpha}$

↓

$u_{1-0,05}$ --- $u_{0,95} = 1,645$
$u_{0,05} = -1,645$

$P(U < u_{0,05}) = F(u_{0,05}) = 0,05$

$P(U < u_{1-0,05}) = F(u_{1-0,05}) = 1 - 0,05$

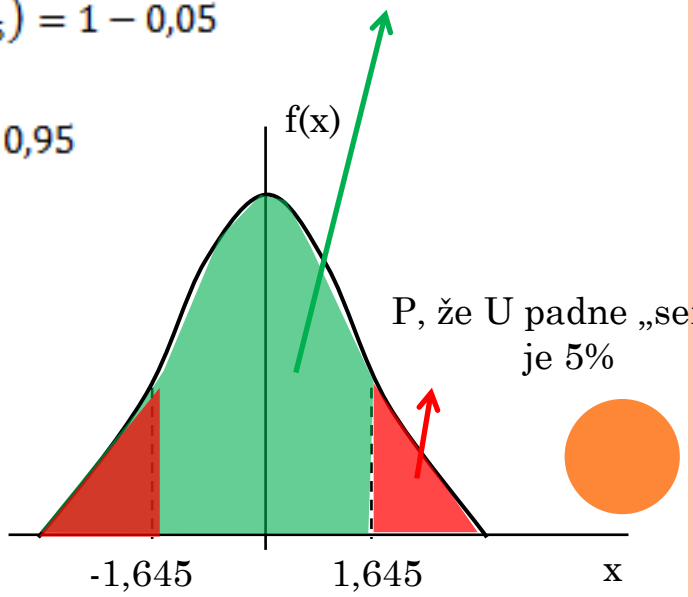
$P(U < u_{0,95}) = F(u_{0,95}) = 0,95$

↓

V tabulce najdu P=0,95
Kvantil je 1,645

$P(U > u_{0,95}) = 0,05$

P, že U padne „sem“ je 95%



Dvoustranný interval

$$P(U < u_\alpha) = F(u_\alpha) = \alpha$$

$$P(U < u_{1-\alpha}) = F(u_{1-\alpha}) = 1 - \alpha$$

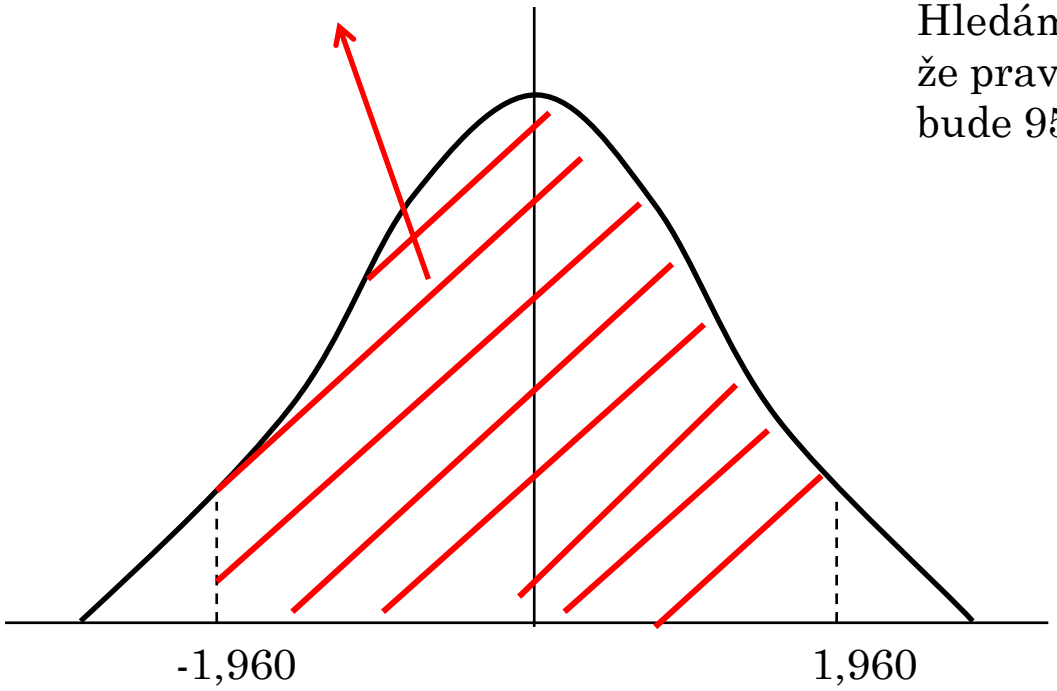
$$P(u_\alpha < U < u_{1-\alpha}) = 1 - 2\alpha$$

$$P\left(u_{\alpha/2} < U < u_{1-\alpha/2}\right) = 1 - \alpha$$

$$u_\alpha = -u_{1-\alpha}$$

$$P\left(-u_{1-\alpha/2} < U < u_{1-\alpha/2}\right) = 1 - \alpha$$

Pravděpodobnost, že U padne do „červené plochy“ je 95%



α=0,05

Hledáme takové kvantily(čísla), že pravděpodobnost, že U padne mezi ně bude 95% a tedy, že nepadne 5%

$$P\left(u_{0,025} < U < u_{0,975}\right) = 0,95$$

$u_{0,975}$ – P(0,975) = 1,960

$u_{0,025}$ = -1,960

$$P\left(-u_{1-\alpha/2} < U < u_{1-\alpha/2}\right) = 1-\alpha$$

$$U = \frac{\bar{x}-\mu}{\sigma}.\sqrt{n}$$

$$P\left(-u_{1-\alpha/2} < \frac{\bar{x}-\mu}{\sigma}.\sqrt{n} < u_{1-\alpha/2}\right) = 1-\alpha$$

V tomto případě známe rozptyl základního souboru – proto σ
Známe x^- - průměr spočítaný z výběrového souboru, Marušky 50 chlapců
Neznáme μ – což je průměr základního souboru

Jednoduchá matematická operace

$$-u_{1-\alpha/2} < \frac{\bar{x}-\mu}{\sigma}.\sqrt{n} \rightarrow -\bar{x}-u_{1-\alpha/2}.\frac{\sigma}{\sqrt{n}} < -\mu \rightarrow \bar{x}+u_{1-\alpha/2}.\frac{\sigma}{\sqrt{n}} > \mu$$

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

$$P\left(\bar{x}-u_{1-\alpha/2}.\frac{\sigma}{\sqrt{n}} < \mu < \bar{x}+u_{1-\alpha/2}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Pravděpodobnost, že střední hodnota základního souboru leží v daném intervalu, je 1-α

Příklad:

Maruška naměří 50 hodnot
Sečte je a vydělí množstvím (50) a získá výběrový průměr ($\bar{x}$) – 16,5
Budeme znát rozptyl velikosti prstů u mužů
Jedná se o rozptyl základního souboru – důležité!!!! Známe jej!!!
Rozptyl (σ^2) =4

Maruška chce stanovit 95% interval spolehlivosti
Pro průměr základního souboru (μ) – stanoví interval
Kde se průměr bude nacházet s 95% pravděpodobností

$$P\left(\bar{x} - u_{1-\alpha/2}.\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_{1-\alpha/2}.\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

α=0,05 – n=50 – σ=2 – $\bar{x}$=16,5
$u_{0,975}$ = 1,960 / $-u_{0,975}$=-1,960

$$16{,}5 - 1{,}960.\frac{2}{\sqrt{50}} < \mu < 16{,}5 + 1{,}960.\frac{2}{\sqrt{50}}$$

$$15{,}945 < \mu < 17{,}054$$

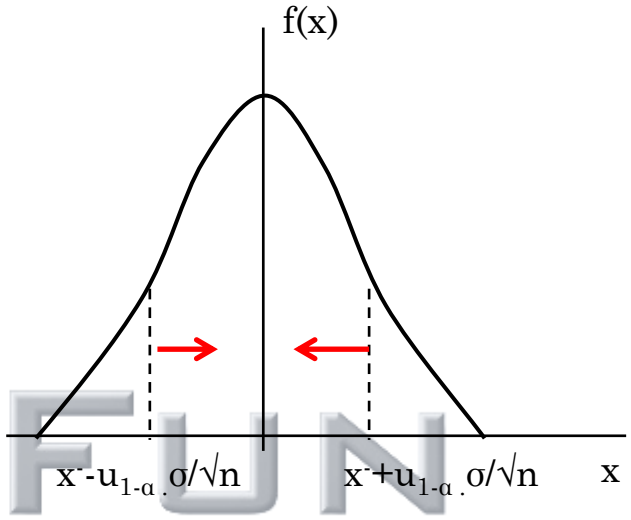
Maruška zjistila, že na 95% je průměr základního souboru mezi 15,945cm a 17,054cm

Určení jednostranných intervalů

Levostranný interval

$$P\left(\bar{x} - u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} < \mu\right) = 1 - \alpha$$

Pozor už není α/2!!!
Pravděpodobnost, že střední hodnota
Bude větší než: $x^- - u_{1-\alpha} . \sigma/\sqrt{n}$



Pravostranný interval

$$P\left(\mu < \bar{x} + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Pravděpodobnost, že střední hodnota
Bude menší než: $x^- + u_{1-\alpha} . \sigma/\sqrt{n}$

Maruška nebude znát rozptyl základního souboru – σ^2

Dejte si pozor – je v tom rozdíl!!!
Zde musí Maruška vypočítat výběrový rozptyl ($s_x'^2$)
Vezme jednotlivé údaje (x_i) a počítá ☺

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

Znovu když budeme předpokládat velký rozsah výběru n>30
Můžeme dané pravděpodobnostní rozdělení nahradit NR
Stejné jako minule pouze nahradíme σ - (s_x')

$$P\left(\bar{x} - u_{1-\alpha/2}.\frac{s_x'}{\sqrt{n}} < \mu < \bar{x} + u_{1-\alpha/2}.\frac{s_x'}{\sqrt{n}}\right) = 1-\alpha$$

Levostranný

$$P\left(\bar{x} - u_{1-\alpha}.\frac{s_x'}{\sqrt{n}} < \mu\right) = 1-\alpha$$

Pravostranný

$$P\left(\mu < \bar{x} + u_{1-\alpha}.\frac{s_x'}{\sqrt{n}}\right) = 1-\alpha$$

Pouhé dosazení do vzorce, ale pozor na:

1) n>30
2) Jestli známe rozptyl Z.S. nebo ne!!!
3) Oboustranný interval $u_{1-\alpha/2}$
4) Jednostranný $u_{1-\alpha}$

Maruška nebude tak aktivní
a získá výběrový soubor pouze o 20 chlapcích

Použije Studentovo t-rozdělení s **(n-1)** stupni volnosti
Statistika má tvar:

$$t = \frac{\bar{x} - \mu}{s_x'} . \sqrt{n}$$

Pozor budou jiné kvantily – $\mathbf{t_{1-\alpha}}$

$$P\left(\bar{x} - t_{1-\alpha/2} \cdot \frac{s_x'}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/2} \cdot \frac{s_x'}{\sqrt{n}}\right) = 1 - \alpha$$

Pro jednostranné intervaly stejný postup
Pouze kvantily NNR nahradíme kvantily t-rozdělení
Dáme si pozor na ***počet stupňů volnosti*** a vyhledáme v tabulkách

Odhad rozptylu základního souboru

Maruška bude dělat závěry nad výpočtem rozptylu z výběrového souboru
Setkali jsme se s bodovým odhadem rozptylu v základním souboru σ^2
Tento odhad je výběrový rozptyl ($s_x'^2$) –nezkreslený, konzistentní

Je třeba rozlišovat, zda-li známe, tentokrát průměr ZS (μ)
A nebo jej neznáme – zde se bude řešit pouze příklad, že známe

Cílem je konstrukce intervalu spolehlivosti pro rozptyl
(interval ve kterém se bude hodnota rozptylu Z.S. nacházet s 1-α P)

Využijeme χ^2 (chí kvadrát) rozdělení s **v=n-1** stupni volnosti
Nezapomínat na stupně volnosti – podle nich hledáme v tabulkách

Statistika má tvar: $\frac{(n-1)s_x'^2}{\sigma^2}$

$$\frac{(n-1)s_x'^2}{\sigma^2}$$

Interval spolehlivosti budeme odvozovat z:

$$P(\chi^2_{\alpha/2} < \frac{(n-1)s_x'^2}{\sigma^2} < \chi^2_{1-\alpha/2}) = 1-\alpha$$

Hodnota statistiky leží v intervalu daných kvantily χ^2 s P=1-α
Našim cílem je zjistit σ^2 – provedeme úpravy na osamostatnění σ^2
100(1-α)% interval spolehlivosti pro rozptyl základního souboru:

$$\frac{(n-1)s_x'^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s_x'^2}{\chi^2_{\alpha/2}}$$

Zde žádné $\chi^2_{1-\alpha} = \chi^2_{1-\alpha}$ – toto rozdělení není symetrické

Ale pro n>30 aproximujeme pomocí kvantilů NNR

Pravostranný interval spolehlivosti

$$\sigma^2 < \frac{(n-1)s_x'^2}{\chi^2_{\alpha}}$$

Levostranný interval spolehlivosti

$$\frac{(n-1)s_x'^2}{\chi^2_{1-\alpha}} < \sigma^2$$