Verjetnost in vzorčenje:

teoretske porazdelitve
standardne napake
ocenjevanje parametrov

as. dr. Nino RODE
prof. dr. Blaž MESEC

VERJETNOST
osnovni pojmi

- Poskus: dejanje pri katerem je izid negotov
  - met kovanca ali igralne kocke
  - pravočasen prihod na predavanja
  - psihoterapevtska obravnava
- Prostor vzorčenja: množica vseh možnih izidov
  - za pravočasen prihod na predavanja: [da, ne]
  - za met:
    - 1 kovanca: [cifra (C), grb(G)]
    - 2 kovancev: [CC, CG, GC, GG]
    - 3 kovancev: [CCC, CCG, CGC, GCC, CGG, GCG, GGC, GGG]

VERJETNOST
osnovni pojmi

- Dogodek: katerakoli podmnožica prostora vzorčenja
- Sestavljeni dogodek: dogodek, ki ga lahko razdelimo na več preprostejših dogodkov
  - pri metu 3 kovancev dobimo 1 cifro
  - trije zamudijo na predavanje
  - psihoterapija pomaga vsem klientom ki so v obravnavi v danem obdobju
- Enostavni dogodek: dogodek, ki se ga ne da več razdeliti na bolj preproste dogodke
  - pri metu kovanca dobimo cifro
  - pri metu 3 kovancev dobimo [GCG]
  - študent/ka zamudi na predavanje
  - psihoterapija pomaga klientu

VERJETNOST
osnovni pojmi

- Empirična verjetnost: delež pojavljanja danega dogodka med vsemi dogodki/izidi, ki smo jih opazili

$$P_{(dogodek)} = \frac{Št._{(dogodek)}}{Št._{(opaženi\,dogodki)}}$$

- Teoretična verjetnost: delež danega dogodka med vsemi možnimi dogodki
  - verjetnost, da bomo v 4 metih dobili 3 cifre
  - verjetnost, da študent/ka zamudi na predavanje
  - verjetnost, da s psihoterapijo pomagamo klientu
- Ugotavljamo jo najlažje, če so enostavni dogodki, iz katerih je dogodek sestavljen vsi enako verjetni

VERJETNOST
osnovni pojmi

- Pričakovana vrednost: povprečna vrednost, ki jo dobimo, ko izvedemo zelo veliko (vse možne) poskusov
  - x → vrednost vsakega od izidov
  - p → verjetnost vakega od izidiv

$$M_{(x)} = \sum x \cdot p_{(x)}$$

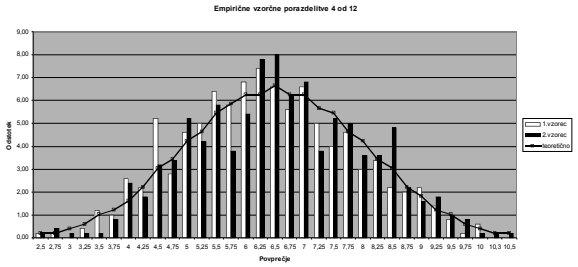
Če dobimo za C = 1€, G = 0€, je pričakovana vrednost velikega števila metov 0,5€ (pri poštenem kovancu)

- Aritmetična sredina je pričakovana vrednost podatkov, če imajo vsi enako verjetnost, da se pojavijo

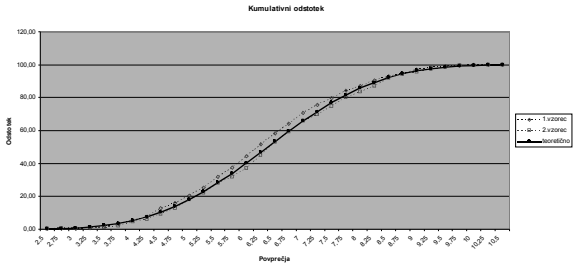
Vzorec in populacija

- Populacija: množica enot z določenimi (zanimivimi) lastnostmi
  - študentke v 3. letniku FSD
  - prebivalci Slovenije starejši od 65 let
  - državljani Republike Slovenije starejši od 65 let
  - mape za vaje pri Metodologiji s statistiko II
- Vzorec: končni del populacije, ki nam je dosegljiv in ga proučujemo, da bi pridobili informacije o lastnostih celote (populacije)
  - študentke na vajah Metodologije … II 2. 12. 2008
  - 50 ljudi starejših od 65 let, slučajno izbranih iz Registra prebivalstva RS
  - mape tretje skupine, ki so bile oddane po vajah 4. 12. 2008

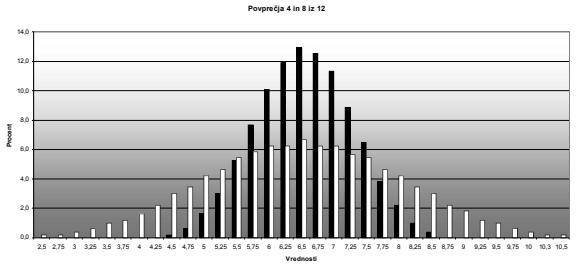
Empirična in porazdelitev teoretska porazdelitev



Empirična in porazdelitev teoretska porazdelitev



Teoretska porazdelitev različno velikih vzorcev



NAPAKA VZORČENJA

$$e = \bar{x} - \mu$$

- *e* - napaka vzorčne ocene aritmetične sredine

$\bar{x}$ - aritmetična sredina vzorca (statistika)

- μ - aritmetična sredina populacije (parameter).

NAPAKA VZORČENJA primer

- Raziskava o življenjskih razmerah starejših občanov Ljubljane (Mesec, Majcen 1982)
- 31.113 Ljubljančanov, starih 65 let ali več
- slučajnostni vzorec 730 oseb
- v populaciji je bilo 38,28 % starih 75 let in več
- v vzorcu je bilo v tej starosti 39,04 % oseb

- Napaka:

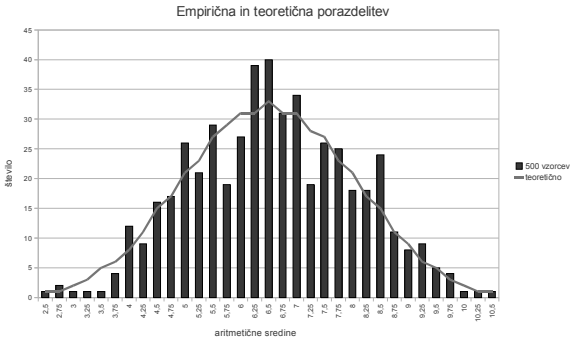
$$e_P = 39{,}04 - 38{,}28 = 0{,}76$$

PORAZDELITEV VZORČNIH OCEN

- Empirična porazdelitev: porazdelitev parametrov (npr. aritmetične sredine) iz VELIKEGA števila vzorcev
- Teoretična porazdelitev: porazdelitev parametrov izračunanih iz VSEH MOŽNIH kombinacij enot v vzorcih (vseh možnih vzorcev):

$${}^{N}_{n}K = \binom{N}{n} = \frac{N!}{n!(N!-n!)} = \frac{12!}{4!8!} = 495$$

PORAZDELITEV VZORČNIH OCEN



Standardna napaka

- Standardna napaka je standardni odklon porazdelitve statistik izračunanih iz vseh možnih vzorcev

standardni odklon populacije:

$$\sigma_{populacija} = \sqrt{\frac{30,25}{12}} = 3,452$$

standardni odklon vseh možnih vzorcev:

$$\sigma_{vzorci} = \sqrt{\frac{1066,81}{495}} = 1,468$$

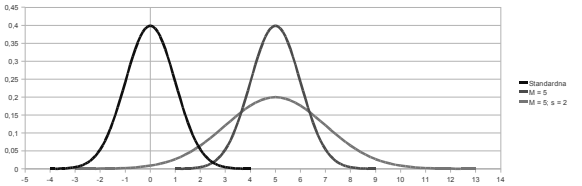
$$\sigma_{vzorcev} = \frac{\sigma_{populacije}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{3,352}{\sqrt{4}} \cdot \sqrt{\frac{12-4}{12-1}} = 1,468$$

Gausova normalna porazdelitev

- Temeljna teoretska porazdelitev (vse empirične in teoretske porazdelitve se ji približujejo)
- Veliko (recimo neskončno) enot, na vrednost vpliva veliko (recimo neskončno) dejavnikov.
- Površina pod funkcijo gostote porazdelitve je **1**.
- Je teoretična porazdelitev
- Dobro opisuje porazdelitev okoli pričakovane vrednosti, če so razlike posledica zgolj slučajnih napak.

Gausova normalna porazdelitev

- Njen položaj in oblika sta odvisna samo od dveh parametrov
  - aritmetične sredine (μ)
  - standardnega odklona (σ)



Izračun standardne napake

- Formula za izračun standardne napake aritmetične sredine

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

- Standardnega odklona populacije ne poznamo, zato uporabimo najboljšo oceno, ki nam je na voljo: oceno iz vzorca

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

- $\sqrt{\frac{N-n}{N-1}}$ je korekturni faktor pri vzorcih brez ponavljanja in manjša standardno napako.
- Smiselno ga je uporabiti, če je n > N/100

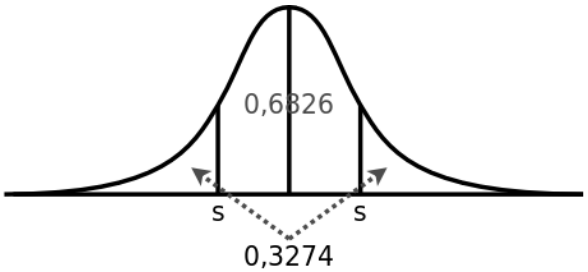
Standardne napake različnih statistik

- Aritmetična sredina $s_{\bar{x}} = \frac{s}{\sqrt{(n)}}$
- Mediana $s_{Me} = \frac{1{,}2535 \cdot s}{\sqrt{(n)}}$
- Standardni odklon $s_s = \frac{s}{\sqrt{(2 \cdot n)}}$
- Koeficient korelacije $s_x = \frac{(1-r)}{\sqrt{n-1}}$
- Razlika aritmetičnih sredin (neodvisni) $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2}$
- Razlika aritmetičnih sredin (odvisni) $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2 - 2 \cdot r \cdot s_{\bar{x}_1}^2 \cdot s_{\bar{x}_2}^2}$

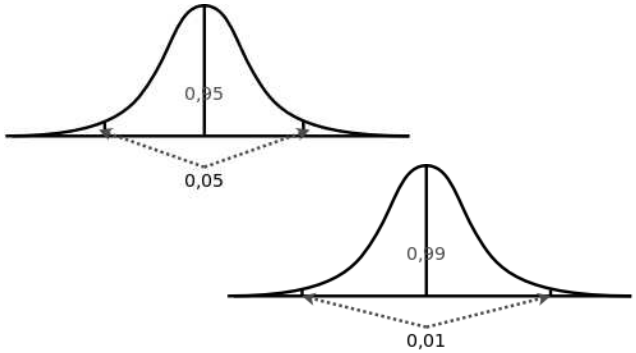
Ocenjevanje parametrov

- Ocene parametrov na podlagi vzorca praviloma:
  - Niso enake vrednosti parametra (niso točne)
  - So blizu vrednosti parametra
- Porazdeljujejo se po normalni porazdelitvi okoli parametra populacije
- Deleže pod normalno krivuljo poznamo

Ocenjevanje parametrov

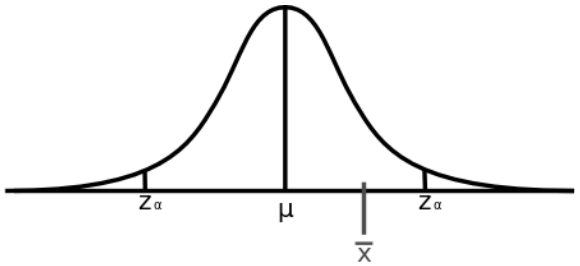


Ocenjevanje parametrov



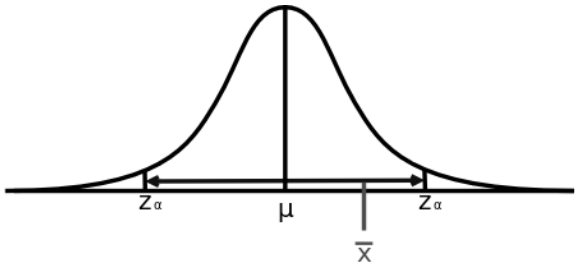
Logika intervala zaupanja

- Na podlagi statistike izračunane iz enega vzorca lahko sklepamo o vrednosti parametra



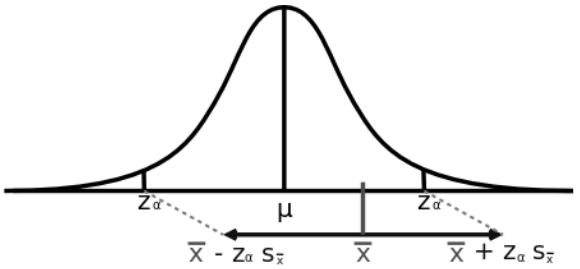
Logika intervala zaupanja

- 95% verjetnosti je, da je statistika za 1,96 standardne napake oddaljena od parametra



Logika intervala zaupanja

- V 95% primerov je torej parameter največ 1,96 standardne napake oddaljen od statistike



z-vrednosti za intervale zaupanja

% zaupanja (1 – α)	Tveganje α	z
99,9	0,001	3,29
99	0,01	2,58
95	0,05	1,96
90	0,1	1,64
80	0,2	1,28

Ocenjevanje parametrov

- Od 400 študentov VŠSD (N)
- so jih 120 (n) anketirali glede poučenosti o aidsu
- Rezultati:
  - Povprečno število točk na testu = 43 točk
  - s = 5,1 točke
- Določi meje intervala zaupanja v aritmetično sredino pri stopnji tveganja 0,05

Ocenjevanje parametrov

- $\frac{N}{n}=\frac{400}{120}=6,7$ za izračun uporabimo korekturni faktor

$$s_{\bar{x}}=\frac{s}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}=\frac{5,1}{\sqrt{120}}\cdot\sqrt{\frac{400-120}{400-1}}=0,4656\cdot 0,8377=0,39$$

$$\bar{x}-z_{0,05}\cdot s_{\bar{x}}=43-1,96\cdot 0,39=42,24$$

$$\bar{x}+z_{0,05}\cdot s_{\bar{x}}=43+1,96\cdot 0,39=43,76$$

- S 95% zaupanjem lahko trdimo, da je aritmetična sredina populacije med 42,24 in 43,76