Tjedan 8: Uzorkovanje, procjena i intervali pouzdanosti

Kako iz dijela saznati nešto o cjelini

2025-04-19

Table of contents

```
library(tidyverse)
```

Ishodi učenja

Nakon ovog predavanja moći ćete

1. Objasniti razliku između populacije i uzorka te između parametra i statistike.
2. Opisati kako veličina uzorka utječe na preciznost procjene populacijskog parametra.
3. Objasniti što je distribucija uzorkovanja (sampling distribution) i zašto je važna.
4. Opisati centralni granični teorem i demonstrirati ga simulacijom.
5. Izračunati standardnu pogrešku prosjeka i objasniti njezino značenje.
6. Konstruirati i interpretirati interval pouzdanosti za prosjek.
7. Prepoznati uobičajene pristranosti u uzorkovanju i objasniti zašto su online ankete problematične.
8. Kritički ocijeniti marginu pogreške u medijskim izvještajima o anketama.

1 Temeljni problem statistike

Statistika rješava jedan temeljni problem. Želimo znati nešto o cijeloj populaciji, ali nemamo pristup cijeloj populaciji. Želimo znati koliki je prosječni dnevni medijski ekran-time svih odraslih Hrvata, ali ne možemo pitati svaku od 3.5 milijuna odraslih osoba. Želimo znati preferiraju li čitatelji kratke ili dugačke članke, ali ne možemo testirati svaki članak na

svakom čitatelju. Želimo znati koliki je CTR novog oglasa, ali ne možemo ga pokazati svim korisnicima interneta.

Umjesto toga, uzimamo **uzorak** (manji dio populacije), mjerimo što nas zanima u uzorku i na temelju toga donosimo zaključak o cijeloj populaciji. Ovo zvuči jednostavno, ali otvara niz pitanja. Koliko veliki uzorak trebamo? Koliko možemo vjerovati procjeni iz uzorka? Kako znamo da uzorak nije pristran?

Ovo predavanje daje odgovore na ta pitanja. Koncepti koje ćemo naučiti (distribucija uzorkovanja, centralni granični teorem, standardna pogreška, interval pouzdanosti) su temelj za sve statističke testove koji dolaze u nastavku kolegija.

---

2 Naši podaci: populacija i uzorci

Za ovo predavanje imamo luksuz koji u stvarnom životu nikad nemamo — poznajemo cijelu populaciju. Dataset sadrži 50 000 odraslih osoba iz fiktivnog hrvatskog grada, s podacima o dobi, spolu, obrazovanju, primarnom izvoru vijesti, povjerenju u medije i dnevnoj medijskoj konzumaciji.

```
pop <- read_csv("../resources/datasets/media_population.csv")
glimpse(pop)
```

```
Rows: 50,000
Columns: 8
$ person_id           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
$ age                 <dbl> 42, 24, 28, 68, 22, 42, 58, 19, 69, 35, 22, 69, 56~
$ gender              <chr> "ženski", "muški", "muški", "ženski", "muški", "mu~
$ education           <chr> "viša/prvostupnik", "srednja", "srednja", "osnovna~
$ primary_news_source <chr> "portal", "društvene mreže", "društvene mreže", "T~
$ media_trust         <dbl> 4, 2, 4, 9, 1, 1, 6, 4, 6, 3, 6, 3, 2, 2, 2, 7, 7,~
$ daily_media_min     <dbl> 178, 130, 120, 224, 127, 198, 248, 153, 293, 174, ~
$ willing_to_pay      <dbl> 0, 14, 0, 45, 0, 11, 0, 32, 42, 0, 0, 48, 0, 26, 0~
```

Zato što poznajemo cijelu populaciju, možemo izračunati prave populacijske parametre i onda vidjeti koliko dobro ih procjenjuju uzorci različitih veličina. Ovo je pedagoški trik jer u stvarnom istraživanju nikad ne znate populacijske parametre (da ih znate, ne biste trebali statistiku). Ali ovdje ih znamo pa možemo procijeniti kvalitetu naših procjena.

```r
# PRAVI populacijski parametri (u praksi ih NIKAD ne znamo)
pop_params <- pop |>
  summarise(
    mu_trust = round(mean(media_trust), 2),
    sigma_trust = round(sd(media_trust), 2),
    mu_media_min = round(mean(daily_media_min), 1),
    sigma_media_min = round(sd(daily_media_min), 1),
    udio_portal = round(mean(primary_news_source == "portal"), 3)
  )

pop_params
```

```
# A tibble: 1 x 5
  mu_trust sigma_trust mu_media_min sigma_media_min udio_portal
     <dbl>       <dbl>        <dbl>           <dbl>       <dbl>
1     4.87        1.98         174.            65.5       0.304
```

Zapamtite ove brojeve. To su istine o populaciji. Sve što radimo dalje pokušava se približiti ovim vrijednostima iz uzoraka.

---

3 Populacija vs uzorak: terminologija

Statistika strogo razlikuje populaciju i uzorak te njihove mjere.

Populacija je cjelokupni skup jedinica o kojima želimo donijeti zaključak. Mjere populacije zovemo **parametri** i označavamo ih grčkim slovima poput μ (prosjek), σ (standardna devijacija), ili p (proporcija).

Uzorak je podskup populacije koji zaista mjerimo. Mjere uzorka zovemo **statistike** i označavamo ih latinskim slovima poput $\bar{x}$ (prosjek uzorka), s (standardna devijacija uzorka), $\hat{p}$ (proporcija uzorka).

Statistike su procjene parametara. Procjena nikad nije savršena jer uzorak nije cijela populacija, ali dobra procjena može biti dovoljno blizu za praktične svrhe.

```r
set.seed(42)

# Uzmimo jedan uzorak od 100 osoba
uzorak_100 <- pop |> slice_sample(n = 100)

# Usporedba parametra i statistike
tibble(
```

```
  mjera = c("Prosjek povjerenja", "SD povjerenja", "Udio portal"),
  populacija = c(
    round(mean(pop$media_trust), 2),
    round(sd(pop$media_trust), 2),
    round(mean(pop$primary_news_source == "portal"), 3)
  ),
  uzorak_100 = c(
    round(mean(uzorak_100$media_trust), 2),
    round(sd(uzorak_100$media_trust), 2),
    round(mean(uzorak_100$primary_news_source == "portal"), 3)
  )
)
```

```
# A tibble: 3 x 3
  mjera              populacija uzorak_100
  <chr>                   <dbl>      <dbl>
1 Prosjek povjerenja      4.87        5.12
2 SD povjerenja           1.98        2.09
3 Udio portal             0.304       0.31
```

Uzorak od 100 osoba daje procjene koje su blizu populacijskim vrijednostima, ali nisu identične. Razlika između parametra i statistike naziva se **pogreška uzorkovanja** (sampling error). Ovo nije greška u smislu da smo nešto krivo napravili. To je neizbježna posljedica toga što radimo s dijelom umjesto s cjelinom.

---

4 Što se događa kad ponovimo uzorkovanje?

Ključan uvid je da bismo, da smo uzeli drugi uzorak od 100 osoba, dobili malo drugačije rezultate. Svaki uzorak je drugačiji. Statistika varira od uzorka do uzorka.

```
set.seed(1)
u1 <- pop |> slice_sample(n = 100) |> summarise(M = round(mean(media_trust), 2)) |> pull(M

set.seed(2)
u2 <- pop |> slice_sample(n = 100) |> summarise(M = round(mean(media_trust), 2)) |> pull(M

set.seed(3)
u3 <- pop |> slice_sample(n = 100) |> summarise(M = round(mean(media_trust), 2)) |> pull(M

cat("Populacijski prosjek:", round(mean(pop$media_trust), 2), "\n")
```

```
Populacijski prosjek: 4.87
```

```
cat("Uzorak 1 (n=100):", u1, "\n")
```

```
Uzorak 1 (n=100): 4.73
```

```
cat("Uzorak 2 (n=100):", u2, "\n")
```

```
Uzorak 2 (n=100): 4.87
```

```
cat("Uzorak 3 (n=100):", u3, "\n")
```

```
Uzorak 3 (n=100): 4.8
```

Svaki uzorak daje malo drugačiji prosjek. Ovo je normalno i neizbježno. Ključna pitanja su sljedeća — koliko ti prosjeci variraju? I kako ta varijacija ovisi o veličini uzorka?

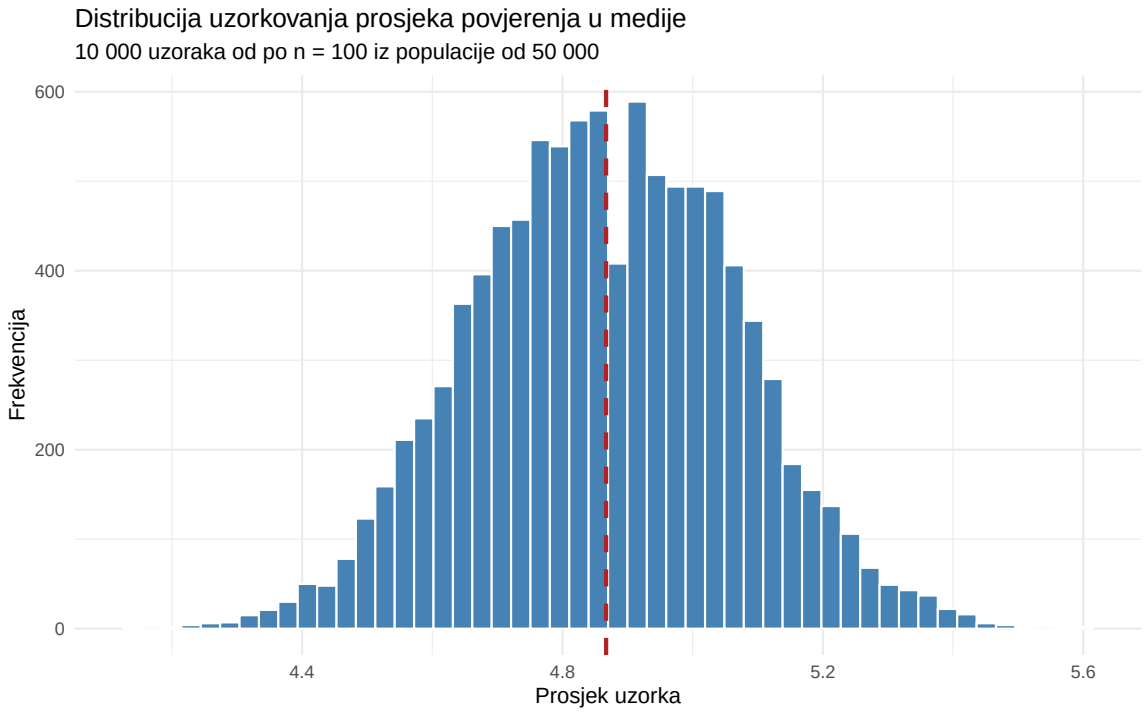
---

5 Distribucija uzorkovanja

Distribucija uzorkovanja je distribucija statistike (npr. prosjeka) kroz mnogo ponovljenih uzoraka. Zamislite da uzimate 10 000 uzoraka od po 100 osoba iz populacije. Svaki uzorak daje jedan prosjek. Distribucija tih 10 000 prosjeka je distribucija uzorkovanja.

```
set.seed(42)

# 10 000 uzoraka od po 100 osoba
sampling_dist <- tibble(
  uzorak = 1:10000,
  prosjek = map_dbl(1:10000, \(i) {
    pop |> slice_sample(n = 100) |> pull(media_trust) |> mean()
  })
)

# Populacijski prosjek
mu <- mean(pop$media_trust)

sampling_dist |>
  ggplot(aes(x = prosjek)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
```

```
  geom_vline(xintercept = mu, color = "firebrick", linewidth = 1, linetype = "dashed") +
  labs(
    title = "Distribucija uzorkovanja prosjeka povjerenja u medije",
    subtitle = "10 000 uzoraka od po n = 100 iz populacije od 50 000",
    x = "Prosjek uzorka",
    y = "Frekvencija"
  ) +
  theme_minimal()
```



Tri stvari su odmah uočljive. Prvo, distribucija je centrirana oko pravog populacijskog prosjeka (crvena crta). Ovo znači da je prosjek uzorka **nepristrana procjena** populacijskog prosjeka, što znači da ne precjenjuje ni podcjenjuje sustavno. Drugo, distribucija je približno normalna (zvonolika), čak i bez pretpostavke o obliku izvorne distribucije. Treće, distribucija je mnogo uža od distribucije pojedinačnih opažanja, gdje prosjeci manje variraju od pojedinačnih vrijednosti.

```
# Usporedba varijabilnosti
cat("SD pojedinačnih opažanja ( ):", round(sd(pop$media_trust), 2), "\n")
```

```
SD pojedinačnih opažanja ( ): 1.98
```

```
cat("SD distribucije uzorkovanja:", round(sd(sampling_dist$prosjek), 3), "\n")
```

```
SD distribucije uzorkovanja: 0.2
```

```r
cat("Omjer:", round(sd(pop$media_trust) / sd(sampling_dist$prosjek), 1), "\n")
```

```
Omjer: 9.9
```

SD distribucije uzorkovanja je otprilike 10 puta manji od SD pojedinačnih opažanja. Taj omjer nije slučajan, već je približno jednak $\sqrt{n} = \sqrt{100} = 10$. Ovo nas vodi do ključnog koncepta.

---

6 Standardna pogreška

Standardna pogreška (standard error, SE) je standardna devijacija distribucije uzorkovanja. Ona mjeri koliko tipično prosjeci uzoraka variraju oko populacijskog prosjeka.

Za prosjek, standardna pogreška se računa formulom:

$$SE = \frac{\sigma}{\sqrt{n}}$$

gdje je standardna devijacija populacije, a n veličina uzorka. U praksi ne znamo pa koristimo procjenu iz uzorka (s):

$$SE \approx \frac{s}{\sqrt{n}}$$

```r
# Teorijska SE za n = 100
sigma <- sd(pop$media_trust)
se_teorijska <- sigma / sqrt(100)

# Procijenjena SE iz jednog uzorka
set.seed(42)
uzorak <- pop |> slice_sample(n = 100)
se_procijenjena <- sd(uzorak$media_trust) / sqrt(100)

# Empirijska SE (iz 10 000 uzoraka)
se_empirijska <- sd(sampling_dist$prosjek)

cat("Teorijska SE:", round(se_teorijska, 3), "\n")
```

```
Teorijska SE: 0.198
```

```r
cat("Procijenjena SE (iz jednog uzorka):", round(se_procijenjena, 3), "\n")
```

```
Procijenjena SE (iz jednog uzorka): 0.209
```

```r
cat("Empirijska SE (iz simulacije):", round(se_empirijska, 3), "\n")
```

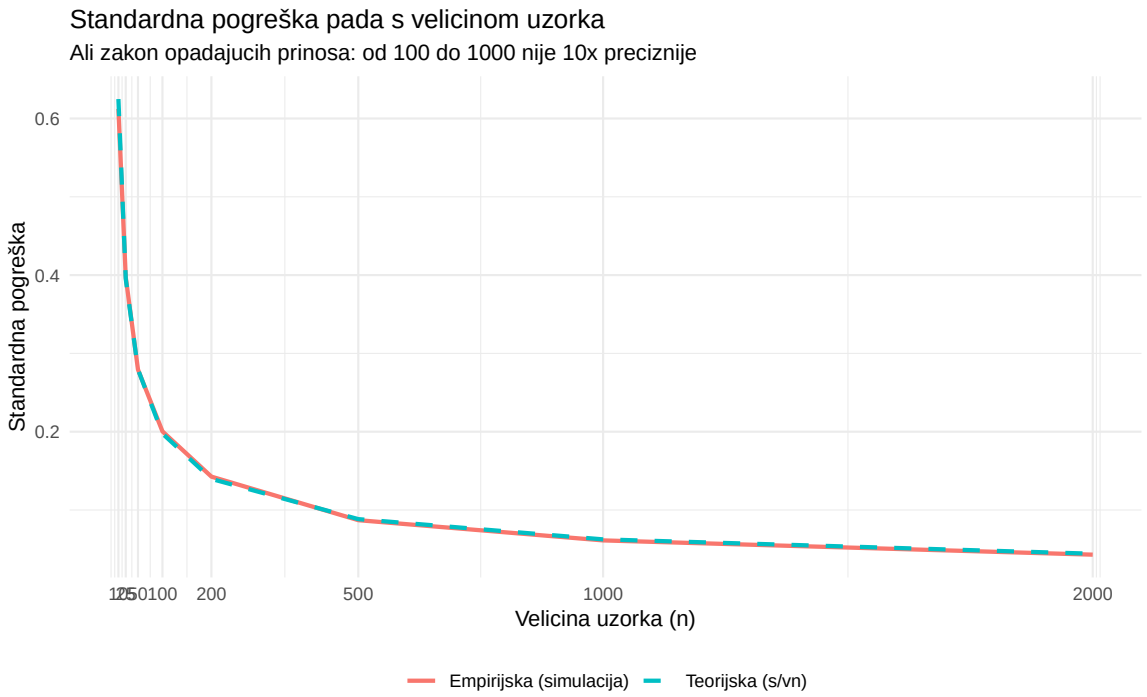
```
Empirijska SE (iz simulacije): 0.2
```

Sve tri vrijednosti su blizu jedna drugoj. Ovo potvrđuje da formula SE = s/√n dobro procjenjuje stvarnu varijabilnost prosjeka uzoraka.

6.1 Veličina uzorka i preciznost

Formula SE = /√n odmah otkriva nešto fundamentalno. Preciznost procjene raste s korijenom veličine uzorka, ne linearno.

```r
set.seed(42)

velicine <- c(10, 25, 50, 100, 200, 500, 1000, 2000)

sim_se <- map_df(velicine, \(n) {
  prosjeci <- map_dbl(1:2000, \(i) {
    pop |> slice_sample(n = n) |> pull(media_trust) |> mean()
  })
  tibble(n = n, se_empirijska = sd(prosjeci), se_formula = sigma / sqrt(n))
})

sim_se |>
  ggplot(aes(x = n)) +
  geom_line(aes(y = se_empirijska, color = "Empirijska (simulacija)"), linewidth = 1) +
  geom_line(aes(y = se_formula, color = "Teorijska ( /√n)"), linewidth = 1, linetype = "das
  scale_x_continuous(breaks = velicine) +
  labs(
    title = "Standardna pogreška pada s veličinom uzorka",
    subtitle = "Ali zakon opadajućih prinosa: od 100 do 1000 nije 10x preciznije",
    x = "Veličina uzorka (n)",
    y = "Standardna pogreška",
    color = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Pad je strm na početku, gdje od n=10 do n=100 je ogromno poboljšanje, ali se postepeno usporava. Da biste prepolovili SE, morate učetverostručiti uzorak. To objašnjava zašto su anketni uzorci obično između 500 i 2000, jer povećanje iznad toga donosi malo dodatne preciznosti u odnosu na trošak.

```
sim_se |>
  mutate(
    se_formula = round(se_formula, 3),
    se_empirijska = round(se_empirijska, 3),
    raspon_95 = paste0("±", round(1.96 * se_formula, 2))
  ) |>
  select(n, se_formula, raspon_95)
```

```
# A tibble: 8 x 3
      n se_formula raspon_95
  <dbl>      <dbl> <chr>
1    10      0.625 ±1.23
2    25      0.395 ±0.77
3    50      0.279 ±0.55
4   100      0.198 ±0.39
5   200      0.14  ±0.27
6   500      0.088 ±0.17
7  1000      0.062 ±0.12
8  2000      0.044 ±0.09
```

Stupac `raspon_95` pokazuje koliko širok je 95% interval oko prosjeka. S uzorkom od 100, prosjek povjerenja je precizan na ±0.39 bodova. S uzorkom od 1000, preciznost je ±0.12 bodova. U praksi, odlučujete kolika je vam preciznost dovoljna i na temelju toga birate veličinu uzorka.

> **Praktični savjet**
>
> Kad čitate medijske izvještaje o anketama, uvijek tražite veličinu uzorka i marginu pogreške. Anketa s n = 500 ima marginu pogreške oko ±4.4 postotna boda za proporcije (na 95% razini). Anketa s n = 1000 ima oko ±3.1. Kad novinar kaže "stranka A ima 32% a stranka B 29%", razlika od 3 postotna boda je unutar margine pogreške za većinu anketa. Zaključak "A vodi" iz takve ankete nije opravdan.

---

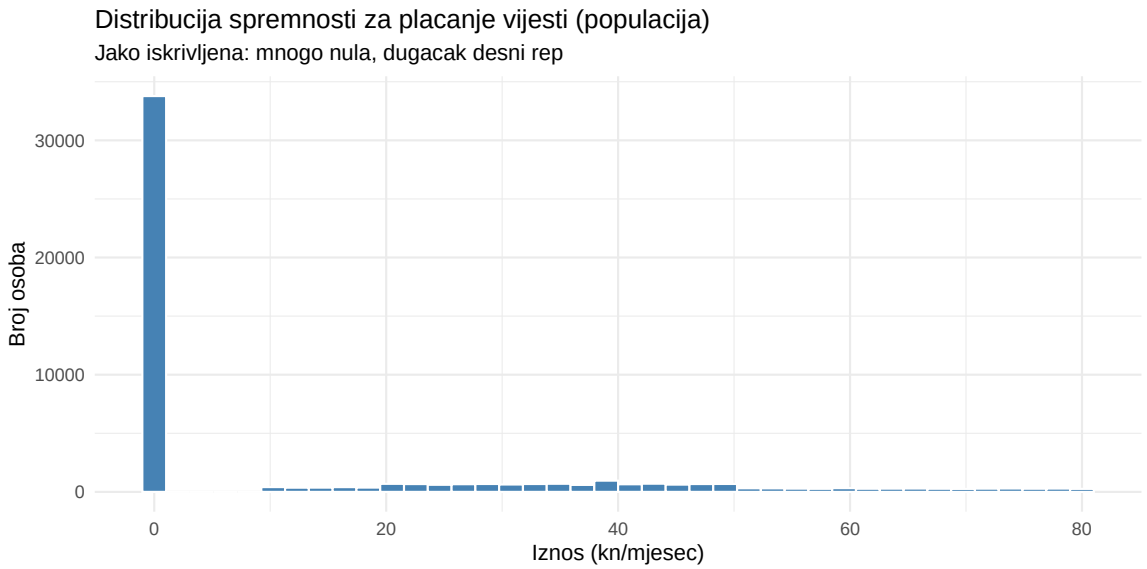
7 Centralni granični teorem

Centralni granični teorem (Central Limit Theorem, CLT) je najvažniji teorem u cijeloj statistici. On kaže da je distribucija uzorkovanja prosjeka približno normalna, neovisno o obliku izvorne distribucije, pod uvjetom da je uzorak dovoljno velik.

Ovo je izuzetno moćno jer znači da možemo koristiti normalnu distribuciju za donošenje zaključaka o prosjecima čak i kad izvorna varijabla nije normalna.

Demonstrirajmo to na varijabli `willing_to_pay` koja je jako iskrivljena (mnogo nula i dugačak desni rep).

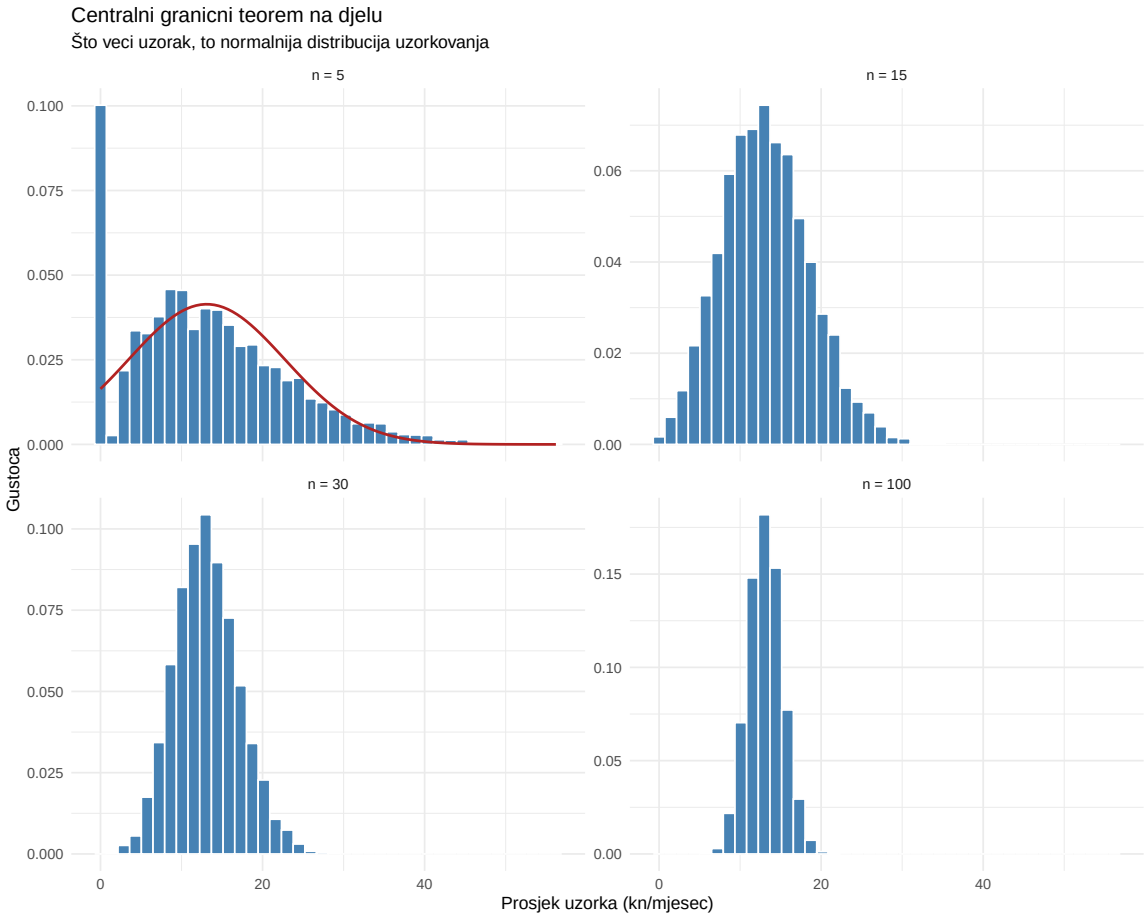
```
# Izvorna distribucija: daleko od normalne
pop |>
  ggplot(aes(x = willing_to_pay)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 40) +
  labs(
    title = "Distribucija spremnosti za plaćanje vijesti (populacija)",
    subtitle = "Jako iskrivljena: mnogo nula, dugačak desni rep",
    x = "Iznos (kn/mjesec)",
    y = "Broj osoba"
  ) +
  theme_minimal()
```



Ovo definitivno nije normalna distribucija. Većina ljudi nije spremna platiti ništa, a oni koji jesu spremni plaćaju različite iznose.

```
set.seed(42)

# Distribucije uzorkovanja za različite veličine uzorka
clt_sim <- map_df(c(5, 15, 30, 100), \(n) {
  prosjeci <- map_dbl(1:5000, \(i) {
    pop |> slice_sample(n = n) |> pull(willing_to_pay) |> mean()
  })
  tibble(n_label = paste("n =", n), n = n, prosjek = prosjeci)
}) |>
  mutate(n_label = fct_reorder(n_label, n))

clt_sim |>
  ggplot(aes(x = prosjek)) +
  geom_histogram(aes(y = after_stat(density)), fill = "steelblue", color = "white", bins =
  stat_function(fun = dnorm,
    args = list(mean = mean(pop$willing_to_pay),
                sd = sd(pop$willing_to_pay) / sqrt(5)),
    data = clt_sim |> filter(n == 5),
    color = "firebrick", linewidth = 0.8) +
  facet_wrap(~n_label, scales = "free_y") +
  labs(
    title = "Centralni granični teorem na djelu",
    subtitle = "Što veći uzorak, to normalnija distribucija uzorkovanja",
    x = "Prosjek uzorka (kn/mjesec)",
    y = "Gustoća"
  ) +
```

```
  theme_minimal()
```



Rezultat je zapanjujuć. S n = 5, distribucija prosjeka je još uvijek iskrivljena (jer je izvorni oblik još dominantan). S n = 15, već se počinje formirati zvonoliki oblik. S n = 30, distribucija je gotovo savršeno normalna. S n = 100, normalna aproksimacija je izvrsna.

Pravilo palca je da je n 30 obično dovoljno za CLT, ali za jako iskrivljene distribucije može trebati i više. Za približno normalne izvorne distribucije, i n = 10 može biti dovoljno.

7.1 Zašto je CLT toliko važan?

CLT je razlog zašto većina statističkih testova radi. T-test pretpostavlja da je distribucija prosjeka približno normalna. Ne pretpostavlja da su individualna opažanja normalna. Zahvaljujući CLT-u, distribucija prosjeka je (približno) normalna čak i kad individualna opažanja nisu, pod uvjetom da je uzorak dovoljno velik. Ovo daje statistici ogromnu moć jer možemo koristiti iste alate (normalnu distribuciju) na gotovo svaku vrstu podataka.

---

8 Pristranosti u uzorkovanju

CLT i formula za SE pretpostavljaju da je uzorak **slučajan**, što znači da svaka osoba u populaciji ima jednaku šansu biti odabrana. U praksi je ta pretpostavka često narušena, što uzrokuje veće probleme od male veličine uzorka.

8.1 Convenience sampling (prigodan uzorak)

Najčešća pristranost u komunikološkim istraživanjima. Anketirate studente na svom kolegiju jer su dostupni. Ali studenti nisu reprezentativni za opću populaciju ni po dobi, ni po obrazovanju, ni po medijskim navikama.

```
set.seed(42)

# "Populacija" = svi
pop_prosjek_trust <- round(mean(pop$media_trust), 2)

# "Prigodan uzorak" = samo mladi (18-24) s visokim obrazovanjem
pristrani_uzorak <- pop |>
  filter(age <= 24, education %in% c("viša/prvostupnik", "magistar/doktor")) |>
  slice_sample(n = 100)

# "Slučajni uzorak" iste veličine
slucajni_uzorak <- pop |> slice_sample(n = 100)

tibble(
  izvor = c("Populacija", "Slučajni uzorak (n=100)", "Pristrani uzorak (n=100)"),
  prosjek_trust = c(
    round(mean(pop$media_trust), 2),
    round(mean(slucajni_uzorak$media_trust), 2),
    round(mean(pristrani_uzorak$media_trust), 2)
  ),
  udio_portal = c(
    round(mean(pop$primary_news_source == "portal"), 3),
    round(mean(slucajni_uzorak$primary_news_source == "portal"), 3),
    round(mean(pristrani_uzorak$primary_news_source == "portal"), 3)
  )
)
```

```
# A tibble: 3 x 3
  izvor                    prosjek_trust udio_portal
  <chr>                            <dbl>       <dbl>
1 Populacija                        4.87       0.304
2 Slučajni uzorak (n=100)           4.79       0.38
3 Pristrani uzorak (n=100)          4.33       0.27
```

Pristrani uzorak daje drugačije procjene od populacijskih vrijednosti. Mladi visokoobrazovani ljudi imaju drugačije medijske navike od opće populacije. Nijedna količina povećanja uzorka ne može ispraviti ovu pristranost, jer 10 000 studenata i dalje nije reprezentativno za opću populaciju.

8.2 Online ankete i self-selection bias

Online ankete, koje su izuzetno popularne u komunikološkim istraživanjima, pate od posebnog oblika pristranosti. Odgovaraju samo ljudi koji su online, koji su na toj platformi, koji su vidjeli poziv na anketu i koji su motivirani odgovoriti. Svaki od ovih koraka filtrira populaciju.

```
# Simulacija: online anketa privlači neproporcijalno mlade korisnike mreža
online_uzorak <- pop |>
  mutate(
    vjerojatnost_odgovora = case_when(
      age < 30 & primary_news_source == "društvene mreže" ~ 0.15,
      age < 30 ~ 0.08,
      age < 50 & primary_news_source %in% c("portal", "društvene mreže") ~ 0.06,
      age < 50 ~ 0.03,
      age >= 50 & primary_news_source %in% c("portal", "društvene mreže") ~ 0.02,
      .default = 0.005
    )
  ) |>
  mutate(odgovorio = runif(n()) < vjerojatnost_odgovora) |>
  filter(odgovorio)

cat("Veličina online uzorka:", nrow(online_uzorak), "\n\n")
```

```
Veličina online uzorka: 2713
```

```
# Usporedba
tribble(
  ~karakteristika, ~populacija, ~online_uzorak,
  "Prosjek dobi", round(mean(pop$age), 1), round(mean(online_uzorak$age), 1),
  "Udio mlađih od 30", round(mean(pop$age < 30) * 100, 1), round(mean(online_uzorak$age <
  "Udio portal kao primarni", round(mean(pop$primary_news_source == "portal") * 100, 1), r
  "Udio društvene mreže", round(mean(pop$primary_news_source == "društvene mreže") * 100,
  "Prosjek povjerenja", round(mean(pop$media_trust), 2), round(mean(online_uzorak$media_tr
)
```

```
# A tibble: 5 x 3
  karakteristika           populacija online_uzorak
  <chr>                         <dbl>         <dbl>
```

```
1 Prosjek dobi                  42.7          31.5
2 Udio mlađih od 30             25.7          56.1
3 Udio portal kao primarni      30.4          32.1
4 Udio društvene mreže          27            48.7
5 Prosjek povjerenja             4.87          4.35
```

Online uzorak je mlađi, koristi više digitalne medije i ima drugačije povjerenje u medije. Čak i s velikim uzorkom, ove procjene su pristrane jer mehanizam uzorkovanja nije slučajan.

> **!** Važna napomena
>
> Veličina uzorka i kvaliteta uzorka su dva različita problema. Velik pristran uzorak je gori od malog slučajnog uzorka. Čuveni primjer je anketa Literary Digesta iz 1936. koja je imala 2.4 milijuna odgovora ali je pogrešno predvidjela američke predsjedničke izbore jer je uzorak bio pristran (bogatiji birači). Gallup je s uzorkom od samo 50 000 pogodio rezultat jer je koristio slučajno uzorkovanje. Veličina bez reprezentativnosti ne vrijedi ništa.

---

9 Procjena proporcija

Do sada smo se fokusirali na procjenu prosjeka. Ali u komunikologiji često procjenjujemo i proporcije (udjele). Koliki je udio ljudi koji portale koriste kao primarni izvor vijesti? Koliki je udio čitatelja koji kliknu na oglas?

Standardna pogreška za proporciju ima drugačiju formulu:

$$SE_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

gdje je $\hat{p}$ procijenjena proporcija iz uzorka.

```r
set.seed(42)

# Pravi udio korisnika portala
p_populacija <- mean(pop$primary_news_source == "portal")

# Procjena iz uzorka od 500
uzorak_500 <- pop |> slice_sample(n = 500)
p_hat <- mean(uzorak_500$primary_news_source == "portal")

se_p <- sqrt(p_hat * (1 - p_hat) / 500)
```

```r
cat("Populacijski udio portala:", round(p_populacija, 3), "\n")
```

```
Populacijski udio portala: 0.304
```

```r
cat("Procjena iz uzorka (n=500):", round(p_hat, 3), "\n")
```

```
Procjena iz uzorka (n=500): 0.318
```

```r
cat("SE proporcije:", round(se_p, 3), "\n")
```

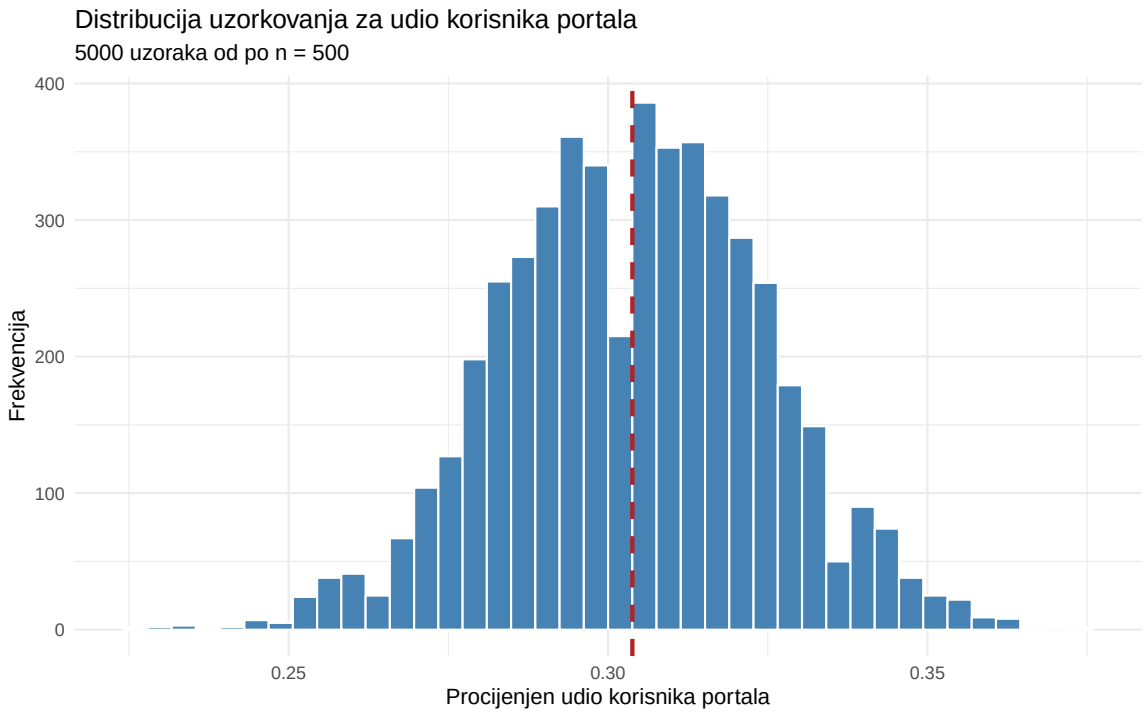
```
SE proporcije: 0.021
```

```r
cat("Margina pogreške (95%):", round(1.96 * se_p, 3), "\n")
```

```
Margina pogreške (95%): 0.041
```

Margina pogreške za proporcije ovisi o samoj proporciji. Najveća je kad je $\hat{p} = 0.5$ (maksimalna neizvjesnost) i smanjuje se kako se $\hat{p}$ približava 0 ili 1 (veća izvjesnost). Zato medijske ankete navode "marginu pogreške $\pm 3\%$" koja zapravo vrijedi samo za proporcije oko 50%.

```r
set.seed(42)

# 5000 uzoraka od po 500 osoba
prop_sim <- tibble(
  uzorak = 1:5000,
  p_hat = map_dbl(1:5000, \(i) {
    pop |> slice_sample(n = 500) |>
      pull(primary_news_source) |>
      (\(x) mean(x == "portal"))()
  })
)

prop_sim |>
  ggplot(aes(x = p_hat)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 40) +
  geom_vline(xintercept = p_populacija, color = "firebrick", linewidth = 1, linetype = "da
  labs(
    title = "Distribucija uzorkovanja za udio korisnika portala",
    subtitle = "5000 uzoraka od po n = 500",
    x = "Procijenjen udio korisnika portala",
    y = "Frekvencija"
  ) +
  theme_minimal()
```



Distribucija proporcija uzorka je također normalna (zahvaljujući CLT) i centrirana oko prave populacijske proporcije. Ovo nam omogućuje konstrukciju intervala pouzdanosti za proporcije, što je temelj za interpretaciju anketnih rezultata.

---

10 Interval pouzdanosti: osnovna ideja

Kad kažemo "prosječno povjerenje u medije je 4.87", to je točkasta procjena (point estimate). Problem s točkastom procjenom je da ne govori ništa o tome koliko je precizna. Je li pravi prosjek negdje između 4.5 i 5.2? Ili između 4.85 i 4.89?

Interval pouzdanosti (confidence interval, CI) daje raspon vrijednosti unutar kojeg se, s određenom vjerojatnošću, nalazi pravi populacijski parametar.

Za prosjek, 95% interval pouzdanosti je:

$$CI_{95\%} = \bar{x} \pm 1.96 \times SE$$

```
set.seed(42)
uzorak <- pop |> slice_sample(n = 200)

x_bar <- mean(uzorak$media_trust)
se <- sd(uzorak$media_trust) / sqrt(200)
```

```r
ci_lower <- x_bar - 1.96 * se
ci_upper <- x_bar + 1.96 * se

cat("Prosjek uzorka:", round(x_bar, 2), "\n")
```

```
Prosjek uzorka: 5.07
```

```r
cat("SE:", round(se, 3), "\n")
```

```
SE: 0.145
```

```r
cat("95% CI: [", round(ci_lower, 2), ",", round(ci_upper, 2), "]\n")
```

```
95% CI: [ 4.79 , 5.35 ]
```

```r
cat("Pravi populacijski prosjek:", round(mean(pop$media_trust), 2), "\n")
```

```
Pravi populacijski prosjek: 4.87
```

Interval pouzdanosti pokriva pravi populacijski prosjek u ovom slučaju. Ali ne mora uvijek, jer 5% intervala iz ponovljenih uzoraka neće pokriti pravi parametar. Zato se zove 95% interval, ne 100%.

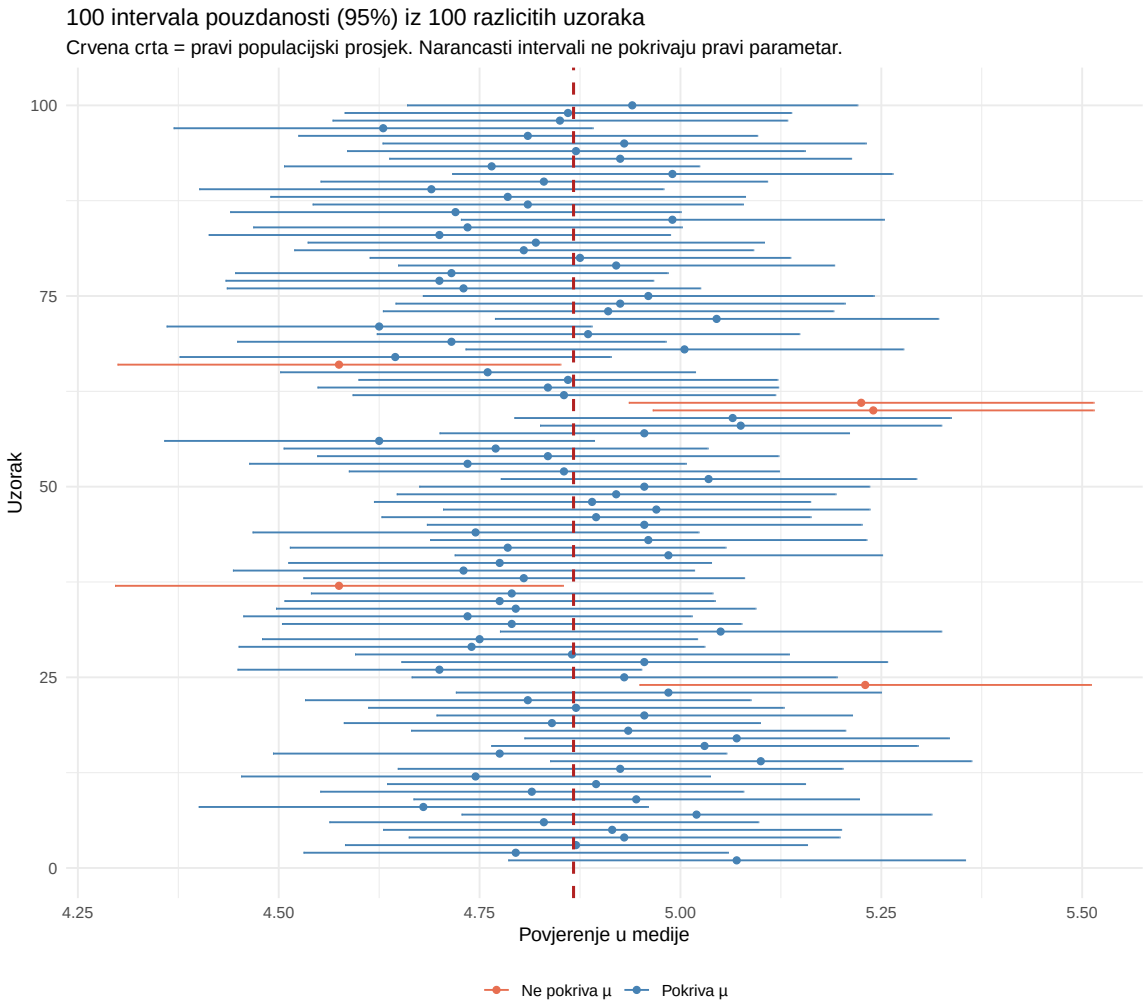
10.1 Vizualizacija: 100 intervala pouzdanosti

```r
set.seed(42)

mu_pop <- mean(pop$media_trust)

ci_sim <- map_df(1:100, \(i) {
  u <- pop |> slice_sample(n = 200)
  xbar <- mean(u$media_trust)
  se <- sd(u$media_trust) / sqrt(200)
  tibble(
    uzorak = i,
    xbar = xbar,
    ci_lo = xbar - 1.96 * se,
    ci_hi = xbar + 1.96 * se,
    pokriva_mu = ci_lo <= mu_pop & ci_hi >= mu_pop
  )
```

```
})

cat("Intervala koji pokrivaju pravi prosjek:", sum(ci_sim$pokriva_mu), "od 100\n")
```

```
Intervala koji pokrivaju pravi prosjek: 95 od 100
```

```
ci_sim |>
  ggplot(aes(x = xbar, y = uzorak, color = pokriva_mu)) +
  geom_point(size = 1.5) +
  geom_errorbarh(aes(xmin = ci_lo, xmax = ci_hi), height = 0) +
  geom_vline(xintercept = mu_pop, color = "firebrick", linewidth = 0.8, linetype = "dashed
  scale_color_manual(values = c("TRUE" = "steelblue", "FALSE" = "#e76f51"),
                     labels = c("TRUE" = "Pokriva  ", "FALSE" = "Ne pokriva  ")) +
  labs(
    title = "100 intervala pouzdanosti (95%) iz 100 različitih uzoraka",
    subtitle = "Crvena crta = pravi populacijski prosjek. Narančasti intervali ne pokrivaj
    x = "Povjerenje u medije",
    y = "Uzorak",
    color = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Ovaj graf je jedna od najvažnijih vizualizacija u cijelom kolegiju. Svaki horizontalni interval je jedan 95% CI iz zasebnog uzorka. Većina, oko 95, pokriva pravi prosjek (crvena crta). Nekolicina, oko 5, ne pokriva. Ovo je precizno značenje 95% intervala pouzdanosti. 95% takvih intervala, konstruiranih iz ponovljenih uzoraka, pokrit će pravi parametar.

! Važna napomena

Česta pogrešna interpretacija je "postoji 95% šansa da je pravi prosjek unutar ovog intervala." Ispravna interpretacija — "ako bismo ponovili uzorkovanje mnogo puta i svaki put konstruirali 95% CI, 95% tih intervala bi pokrilo pravi prosjek." Razlika zvuči suptilno, ali je konceptualno važna. Pravi prosjek je fiksan broj (nije slučajan). Interval je slučajan (jer ovisi o uzorku). Vjerojatnost se odnosi na postupak, ne na parametar.

> **Podsjetnik**
>
> U prvom dijelu predavanja naučili smo da distribucija uzorkovanja prosjeka ima oblik normalne distribucije (CLT), da se njezina širina mjeri standardnom pogreškom SE = s/√n i da 95% interval pouzdanosti pokriva prosjek ± 1.96 × SE. U ovom dijelu prelazimo na alate koji se koriste u praksi, poput t-distribucije, funkcije `t.test()` i planiranja veličine uzorka.

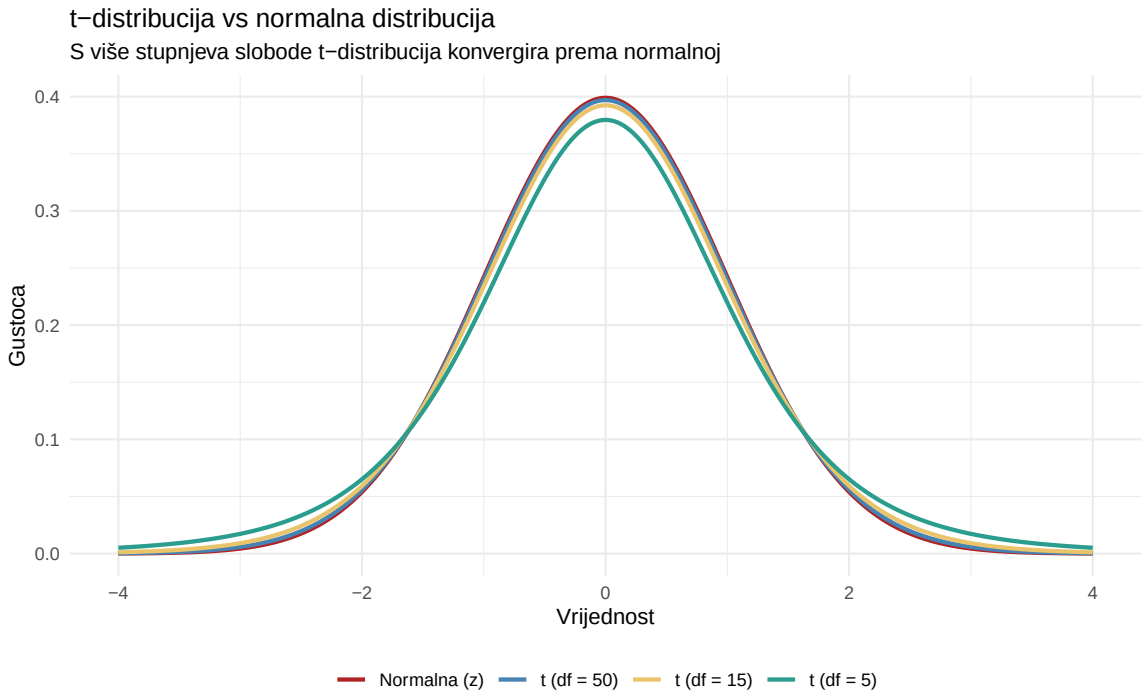
11 Od z do t: mali uzorci

Do sada smo koristili z = 1.96 za 95% CI. To je točno kad poznajemo populacijski ili kad je uzorak velik (n > 100). Ali u praksi obično ne poznajemo pa ga procjenjujemo iz uzorka pomoću s. Za male uzorke, ta dodatna nesigurnost znači da trebamo širi interval.

t-distribucija rješava ovaj problem. Izgleda poput normalne distribucije, ali ima deblje repove, gdje je veća vjerojatnost ekstremnijih vrijednosti. Oblik t-distribucije ovisi o **stupnjevima slobode** (degrees of freedom, df), koji su za jedan prosjek df = n − 1.

```
x <- seq(-4, 4, length.out = 300)

t_usporedba <- tibble(x = x) |>
  mutate(
    `Normalna (z)` = dnorm(x),
    `t (df = 5)` = dt(x, df = 5),
    `t (df = 15)` = dt(x, df = 15),
    `t (df = 50)` = dt(x, df = 50)
  ) |>
  pivot_longer(-x, names_to = "distribucija", values_to = "gustoca") |>
  mutate(distribucija = fct_relevel(distribucija,
    "Normalna (z)", "t (df = 50)", "t (df = 15)", "t (df = 5)"))

t_usporedba |>
  ggplot(aes(x = x, y = gustoca, color = distribucija)) +
  geom_line(linewidth = 1) +
  scale_color_manual(values = c(
    "Normalna (z)" = "firebrick",
    "t (df = 5)" = "#2a9d8f",
    "t (df = 15)" = "#e9c46a",
    "t (df = 50)" = "steelblue"
  )) +
  labs(
    title = "t-distribucija vs normalna distribucija",
    subtitle = "S više stupnjeva slobode t-distribucija konvergira prema normalnoj",
    x = "Vrijednost",
```

```
    y = "Gustoća",
    color = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



S $df = 5$ (uzorak od 6), t-distribucija ima znatno deblje repove od normalne. S $df = 50$, razlika je jedva vidljiva. Praktična posljedica je da za male uzorke koristimo veći multiplikator od 1.96.

```
tibble(
  df = c(5, 10, 15, 25, 50, 100, Inf),
  n = df + 1,
  t_95 = round(qt(0.975, df), 3),
  t_99 = round(qt(0.995, df), 3)
) |>
  mutate(n = if_else(is.infinite(df), "∞ (normalna)", as.character(n)))
```

```
# A tibble: 7 x 4
     df n              t_95  t_99
  <dbl> <chr>         <dbl> <dbl>
1     5 6              2.57  4.03
2    10 11             2.23  3.17
3    15 16             2.13  2.95
```

```
4    25 26            2.06  2.79
5    50 51            2.01  2.68
6   100 101           1.98  2.63
7   Inf ∞ (normalna)  1.96  2.58
```

Za n = 6 (df = 5), kritična vrijednost za 95% CI je 2.571 umjesto 1.960. Interval je značajno širi jer imamo manje podataka pa moramo biti oprezniji. Za n > 100, razlika između t i z je zanemariva i u praksi se često ignorira.

---

12 `t.test()`: sve u jednoj funkciji

R ima ugrađenu funkciju `t.test()` koja automatski računa t-interval pouzdanosti. Za sada je koristimo samo za CI (ne za testiranje hipoteza, to dolazi sljedeći tjedan).

```r
set.seed(42)
uzorak_200 <- pop |> slice_sample(n = 200)

# CI za prosjek povjerenja u medije
rezultat <- t.test(uzorak_200$media_trust, conf.level = 0.95)
rezultat
```

```
    One Sample t-test

data:  uzorak_200$media_trust
t = 34.961, df = 199, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 4.784028 5.355972
sample estimates:
mean of x
     5.07
```

Funkcija `t.test()` vraća mnogo informacija odjednom. Za interval pouzdanosti nas zanima `conf.int` i `estimate`.

```r
# Pristup pojedinim elementima
cat("Prosjek uzorka:", round(rezultat$estimate, 3), "\n")
```

```
Prosjek uzorka: 5.07
```

```r
cat("95% CI: [", round(rezultat$conf.int[1], 3), ",", round(rezultat$conf.int[2], 3), "]\n
```

```
95% CI: [ 4.784 , 5.356 ]
```

```r
cat("Stupnjevi slobode:", rezultat$parameter, "\n")
```

```
Stupnjevi slobode: 199
```

```r
# Usporedba s populacijom
cat("\nProvi populacijski prosjek:", round(mean(pop$media_trust), 3), "\n")
```

```
Provi populacijski prosjek: 4.867
```

```r
cat("Pokriva CI pravi prosjek?",
    mean(pop$media_trust) >= rezultat$conf.int[1] &
    mean(pop$media_trust) <= rezultat$conf.int[2], "\n")
```

```
Pokriva CI pravi prosjek? TRUE
```

12.1 Mijenjanje razine pouzdanosti

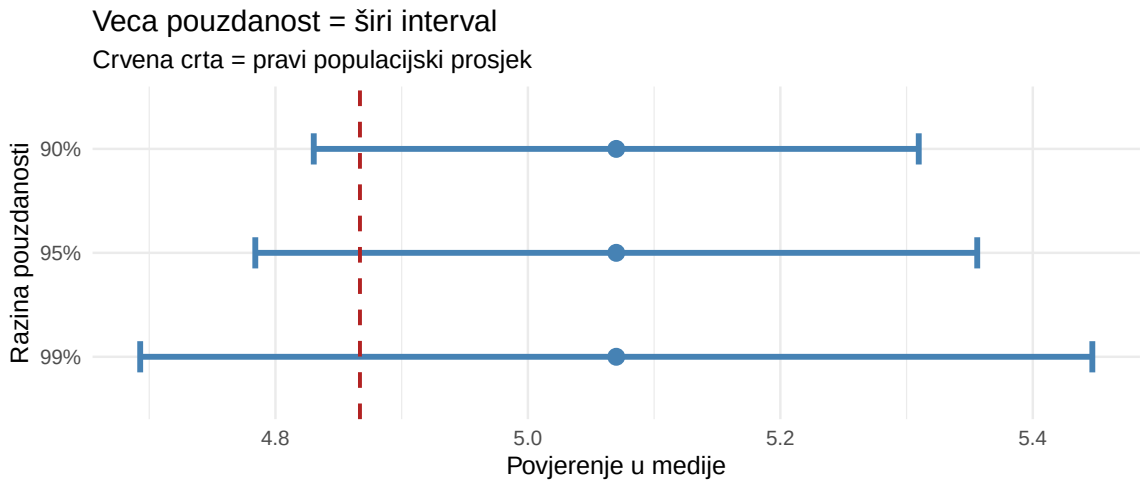
Možemo tražiti i 90% ili 99% interval.

```r
ci_90 <- t.test(uzorak_200$media_trust, conf.level = 0.90)$conf.int
ci_95 <- t.test(uzorak_200$media_trust, conf.level = 0.95)$conf.int
ci_99 <- t.test(uzorak_200$media_trust, conf.level = 0.99)$conf.int

xbar <- mean(uzorak_200$media_trust)

tibble(
  razina = c("90%", "95%", "99%"),
  donja = round(c(ci_90[1], ci_95[1], ci_99[1]), 3),
  gornja = round(c(ci_90[2], ci_95[2], ci_99[2]), 3),
  sirina = round(c(diff(ci_90), diff(ci_95), diff(ci_99)), 3)
)
```

```
# A tibble: 3 x 4
  razina donja gornja sirina
  <chr>  <dbl>  <dbl>  <dbl>
1 90%     4.83   5.31  0.479
2 95%     4.78   5.36  0.572
3 99%     4.69   5.45  0.754
```

Veća pouzdanost znači širi interval. 99% CI je širi od 95% jer morate pokriti više mogućih vrijednosti. Postoji kompromis između pouzdanosti i preciznosti. 100% CI bi bio od $-\infty$ do $+\infty$, što je potpuno beskorisno ali potpuno sigurno. U praksi se najčešće koristi 95% kao konvencija.

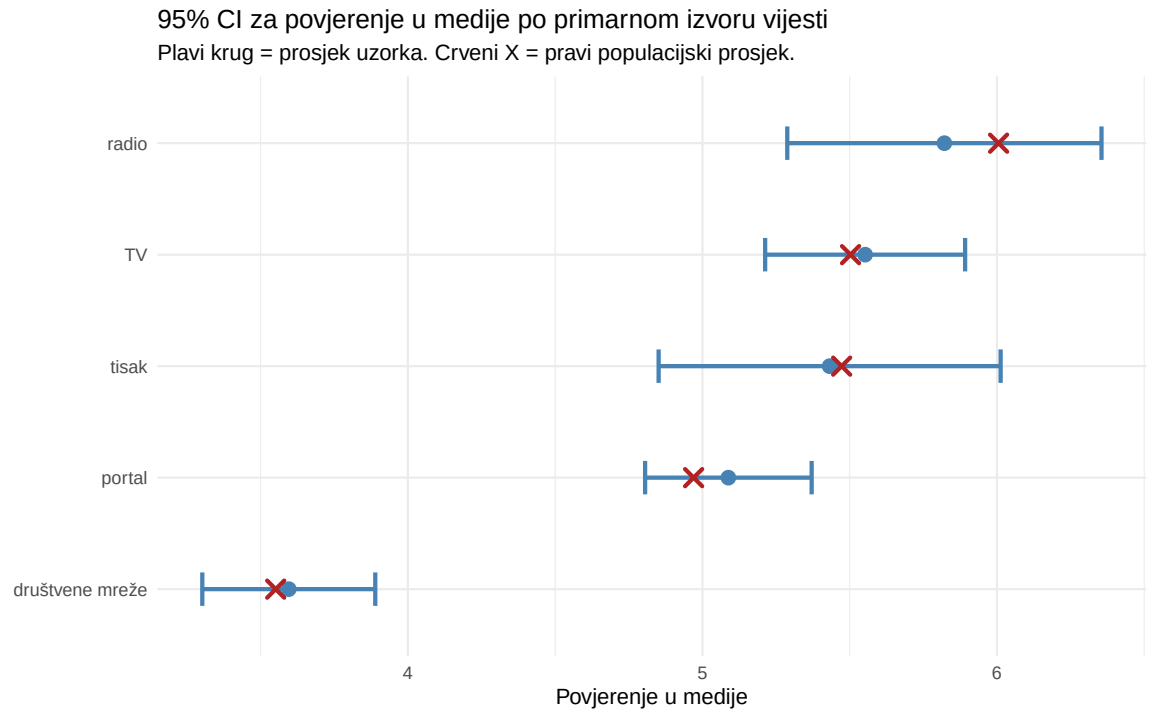
```
mu_pop <- mean(pop$media_trust)

tibble(
  razina = factor(c("90%", "95%", "99%"), levels = c("99%", "95%", "90%")),
  lo = c(ci_90[1], ci_95[1], ci_99[1]),
  hi = c(ci_90[2], ci_95[2], ci_99[2]),
  xbar = xbar
) |>
  ggplot(aes(y = razina)) +
  geom_errorbarh(aes(xmin = lo, xmax = hi), height = 0.3, linewidth = 1.2, color = "steelb
  geom_point(aes(x = xbar), size = 3, color = "steelblue") +
  geom_vline(xintercept = mu_pop, color = "firebrick", linewidth = 0.8, linetype = "dashed
  labs(
    title = "Veća pouzdanost = širi interval",
    subtitle = "Crvena crta = pravi populacijski prosjek",
    x = "Povjerenje u medije",
    y = "Razina pouzdanosti"
  ) +
  theme_minimal()
```



12.2 CI za podgrupe

U praksi nas često zanima CI za specifične podgrupe, ne samo za cijeli uzorak.

```r
set.seed(42)
uzorak_500 <- pop |> slice_sample(n = 500)

ci_po_izvoru <- uzorak_500 |>
  group_by(primary_news_source) |>
  filter(n() >= 20) |>
  summarise(
    n = n(),
    prosjek = mean(media_trust),
    se = sd(media_trust) / sqrt(n()),
    ci_lo = prosjek - qt(0.975, n() - 1) * se,
    ci_hi = prosjek + qt(0.975, n() - 1) * se,
    .groups = "drop"
  )

# Pravi populacijski prosjeci za usporedbu
pop_prosjeci <- pop |>
  group_by(primary_news_source) |>
  summarise(mu = mean(media_trust), .groups = "drop")

ci_po_izvoru |>
  left_join(pop_prosjeci, by = "primary_news_source") |>
  mutate(primary_news_source = fct_reorder(primary_news_source, prosjek)) |>
  ggplot(aes(y = primary_news_source)) +
  geom_errorbarh(aes(xmin = ci_lo, xmax = ci_hi), height = 0.3, linewidth = 1, color = "st
  geom_point(aes(x = prosjek), size = 3, color = "steelblue") +
  geom_point(aes(x = mu), size = 3, shape = 4, color = "firebrick", stroke = 1.5) +
  labs(
    title = "95% CI za povjerenje u medije po primarnom izvoru vijesti",
    subtitle = "Plavi krug = prosjek uzorka. Crveni X = pravi populacijski prosjek.",
    x = "Povjerenje u medije",
    y = NULL
  ) +
  theme_minimal()
```



Ovaj graf je izuzetno koristan za prezentaciju rezultata. Kad se intervali dviju grupa ne preklapaju, to sugerira statistički značajnu razliku, što ćemo formalno obraditi na predavanju o t-testu. Korisnici radija imaju najviše povjerenje, a korisnici društvenih mreža najmanje.

> **Praktični savjet**
>
> Kad prezentirate rezultate istraživanja, graf s intervalima pouzdanosti govori mnogo više od tablice prosjeka. Uključuje i veličinu uzorka (uži interval = više podataka) i nesigurnost procjene (širi interval = manje sigurni u točnu vrijednost). Naviknite se koristiti ovaj tip grafa.

---

13 Interval pouzdanosti za proporcije

Za proporcije (udjele), CI se računa malo drugačije jer je SE za proporciju $\sqrt{(\hat{p}(1-\hat{p})/n)}$.

```
set.seed(42)
uzorak_500 <- pop |> slice_sample(n = 500)

# Udio koji koristi portal kao primarni izvor
p_hat <- mean(uzorak_500$primary_news_source == "portal")
se_p <- sqrt(p_hat * (1 - p_hat) / 500)
```

```r
ci_lo <- p_hat - 1.96 * se_p
ci_hi <- p_hat + 1.96 * se_p

cat("Procjena p̂:", round(p_hat, 3), "\n")
```

```
Procjena p̂: 0.318
```

```r
cat("SE:", round(se_p, 3), "\n")
```

```
SE: 0.021
```

```r
cat("95% CI: [", round(ci_lo, 3), ",", round(ci_hi, 3), "]\n")
```

```
95% CI: [ 0.277 , 0.359 ]
```

```r
cat("Pravi populacijski udio:", round(mean(pop$primary_news_source == "portal"), 3), "\n")
```

```
Pravi populacijski udio: 0.304
```

13.1 `prop.test()` za proporcije

R ima funkciju `prop.test()` koja računa CI za proporcije. Koristi malo drugačiju metodu (Wilsonov interval) koja je preciznija za male uzorke i ekstremne proporcije.

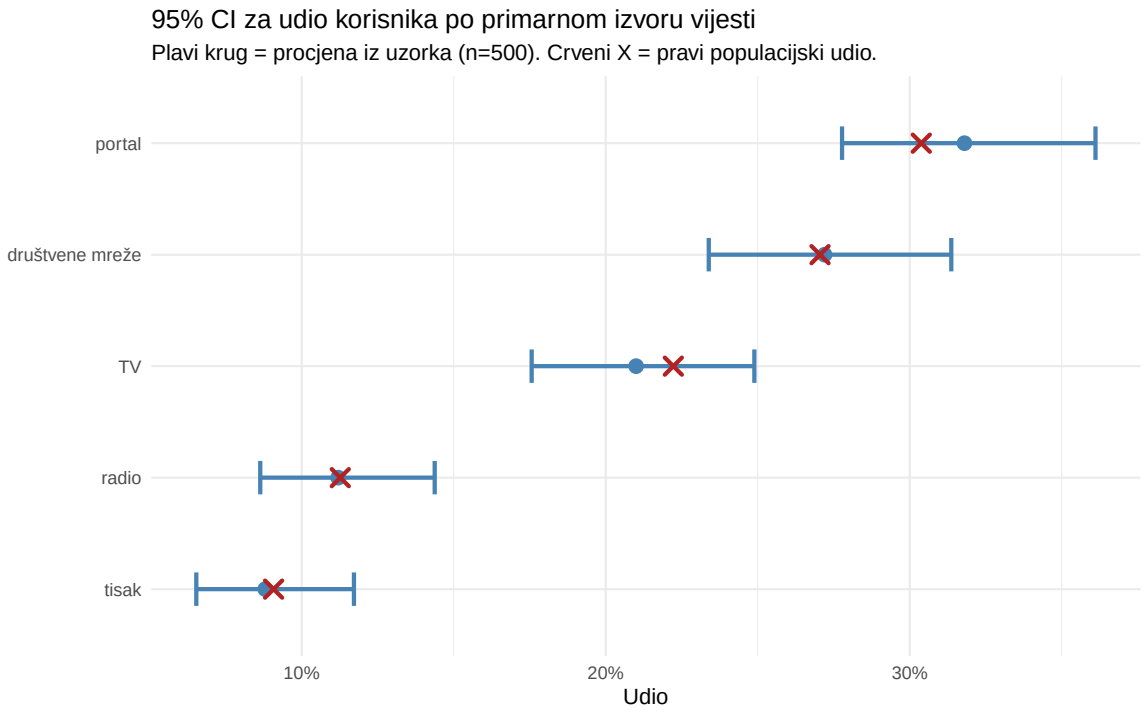
```r
# Koliko koristi portal od 500 ispitanika?
n_portal <- sum(uzorak_500$primary_news_source == "portal")

prop_rez <- prop.test(n_portal, n = 500, conf.level = 0.95)

cat("Procjena:", round(prop_rez$estimate, 3), "\n")
```

```
Procjena: 0.318
```

```r
cat("95% CI: [", round(prop_rez$conf.int[1], 3), ",", round(prop_rez$conf.int[2], 3), "]\n
```

```
95% CI: [ 0.278 , 0.361 ]
```

```r
# CI za sve izvore vijesti
izvore <- unique(pop$primary_news_source)

ci_izvore <- map_df(izvore, \(izvor) {
  n_da <- sum(uzorak_500$primary_news_source == izvor)
  test <- prop.test(n_da, n = 500, conf.level = 0.95)
  tibble(
    izvor = izvor,
    p_hat = test$estimate,
    ci_lo = test$conf.int[1],
    ci_hi = test$conf.int[2]
  )
})

# Pravi populacijski udjeli
pop_udjeli <- pop |>
  count(primary_news_source) |>
  mutate(udio = n / sum(n))

ci_izvore |>
  left_join(pop_udjeli, by = c("izvor" = "primary_news_source")) |>
  mutate(izvor = fct_reorder(izvor, p_hat)) |>
  ggplot(aes(y = izvor)) +
  geom_errorbarh(aes(xmin = ci_lo, xmax = ci_hi), height = 0.3, linewidth = 1, color = "st
  geom_point(aes(x = p_hat), size = 3, color = "steelblue") +
  geom_point(aes(x = udio), size = 3, shape = 4, color = "firebrick", stroke = 1.5) +
  scale_x_continuous(labels = scales::label_percent()) +
  labs(
    title = "95% CI za udio korisnika po primarnom izvoru vijesti",
    subtitle = "Plavi krug = procjena iz uzorka (n=500). Crveni X = pravi populacijski udi
    x = "Udio",
    y = NULL
  ) +
  theme_minimal()
```



Ovaj graf otkriva nešto što medijske ankete rijetko prikazuju - nesigurnost oko svakog broja. Portal i društvene mreže se ne mogu jasno razlučiti jer se intervali preklapaju. TV i radio se mogu jasno razlučiti jer se intervali ne preklapaju. Zato je prikaz intervala pouzdanosti uvijek pošteniji od samih postotaka.

---

14 Margina pogreške i planiranje uzorka

U medijskim izvještajima o anketama čujete izraz "margina pogreške ±3%". Što to znači i kako se računa?

Margina pogreške (margin of error, MoE) je pola širine intervala pouzdanosti. Za proporcije:

$$MoE = z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ako ne znamo p̂ unaprijed, koristimo najgori slučaj p̂ = 0.5 (koji daje najširu marginu):

$$MoE_{max} = \frac{z^*}{2\sqrt{n}}$$

Za 95% CI: MoE_max = 1.96 / (2√n) 1/√n.

```r
tibble(
  n = c(100, 200, 400, 500, 800, 1000, 1500, 2000),
  MoE_95 = round(1.96 * sqrt(0.25 / n) * 100, 1)
) |>
  mutate(opis = paste0("±", MoE_95, "%"))
```

```
# A tibble: 8 x 3
      n MoE_95 opis
  <dbl>  <dbl> <chr>
1   100    9.8 ±9.8%
2   200    6.9 ±6.9%
3   400    4.9 ±4.9%
4   500    4.4 ±4.4%
5   800    3.5 ±3.5%
6  1000    3.1 ±3.1%
7  1500    2.5 ±2.5%
8  2000    2.2 ±2.2%
```

Ovo objašnjava zašto su većina medijskih anketa u rasponu 500 do 1500 ispitanika. S n = 1000, margina je oko ±3.1%. S n = 2000, pada na ±2.2%. Poboljšanje je malo u odnosu na dodatni trošak i vrijeme.

14.1 Obrnuto: koliki uzorak trebam?

Češće pitanje u praksi je obrnuto — imam ciljanu marginu pogreške, koliki mi uzorak treba?

$$n = \frac{z^{*2} \times \hat{p}(1-\hat{p})}{MoE^2}$$

Za najgori slučaj ($\hat{p} = 0.5$):

$$n = \frac{z^{*2}}{4 \times MoE^2}$$

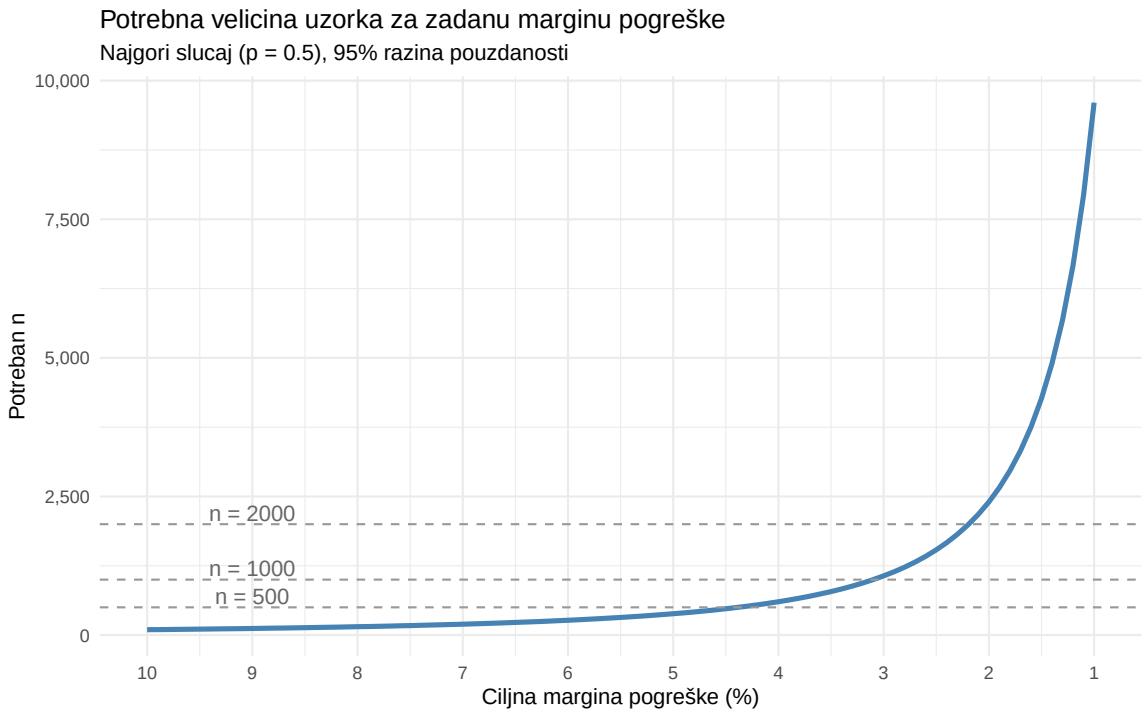
```r
# Funkcija za izračun potrebne veličine uzorka
velicina_uzorka <- function(moe, conf = 0.95, p = 0.5) {
  z <- qnorm(1 - (1 - conf) / 2)
  ceiling(z^2 * p * (1 - p) / moe^2)
}

tibble(
  ciljna_MoE = c("±5%", "±4%", "±3%", "±2%", "±1%"),
```

```
  moe = c(0.05, 0.04, 0.03, 0.02, 0.01),
  n_potreban = map_int(moe, velicina_uzorka)
)
```

```
# A tibble: 5 x 3
  ciljna_MoE   moe n_potreban
  <chr>      <dbl>      <int>
1 ±5%         0.05        385
2 ±4%         0.04        601
3 ±3%         0.03       1068
4 ±2%         0.02       2401
5 ±1%         0.01       9604
```

Za marginu od ±3% trebate 1068 ispitanika. Za ±2% trebate 2401. Za ±1% trebate čak 9604. Ovo ponovno potvrđuje zakon opadajućih prinosa, gdje je svako sljedeće poboljšanje sve skuplje.

```
tibble(
  moe = seq(0.01, 0.10, by = 0.001)
) |>
  mutate(n = map_dbl(moe, velicina_uzorka)) |>
  ggplot(aes(x = moe * 100, y = n)) +
  geom_line(color = "steelblue", linewidth = 1.2) +
  geom_hline(yintercept = c(500, 1000, 2000), linetype = "dashed", color = "grey60") +
  annotate("text", x = 9, y = c(500, 1000, 2000) + 200,
           label = c("n = 500", "n = 1000", "n = 2000"), color = "grey40") +
  scale_x_reverse(breaks = seq(1, 10, 1)) +
  scale_y_continuous(labels = scales::label_comma()) +
  labs(
    title = "Potrebna veličina uzorka za zadanu marginu pogreške",
    subtitle = "Najgori slučaj (p = 0.5), 95% razina pouzdanosti",
    x = "Ciljna margina pogreške (%)",
    y = "Potreban n"
  ) +
  theme_minimal()
```



Praktični savjet

Kad planirate istraživanje, odlučite o margini pogreške PRIJE nego počnete prikupljati podatke. Pitajte se koja razlika je praktično važna u vašem kontekstu. Ako vas zanima razlikuje li se popularnost dviju platformi za 5 postotnih bodova, trebate marginu manju od 5%, što znači uzorak od barem 400. Ako trebate razlučiti razlike od 2 postotna boda, trebate barem 2400 ispitanika.

---

15 Čitanje medijskih anketa kritički

Naučeno dosad daje nam alate za kritičku evaluaciju medijskih izvještaja o anketama. Pogledajmo tipičan primjer.

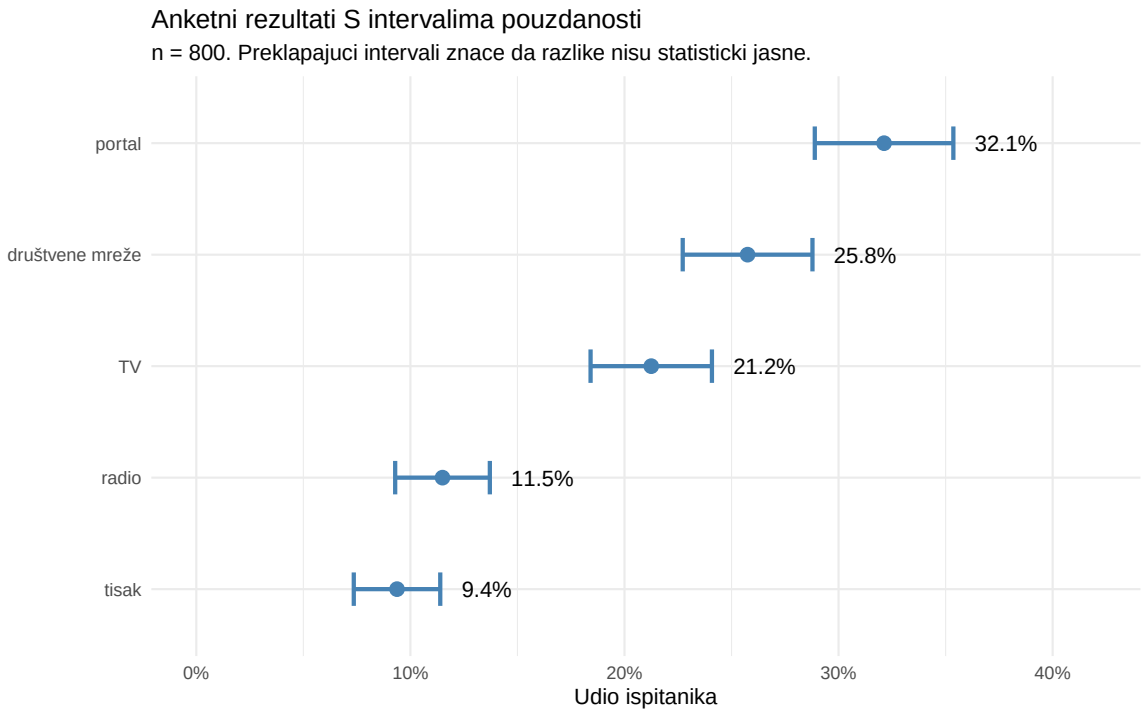
```
# Simulacija: medijska anketa o primarnom izvoru vijesti
set.seed(123)
anketa <- pop |> slice_sample(n = 800)

rezultati <- anketa |>
  count(primary_news_source) |>
  mutate(
```

```
    udio = n / sum(n),
    se = sqrt(udio * (1 - udio) / sum(n)),
    moe = 1.96 * se,
    ci_lo = udio - moe,
    ci_hi = udio + moe
  ) |>
  arrange(desc(udio))

rezultati |>
  mutate(across(c(udio, se, moe, ci_lo, ci_hi), \(x) round(x * 100, 1)))
```

```
# A tibble: 5 x 7
  primary_news_source     n  udio    se   moe ci_lo ci_hi
  <chr>               <int> <dbl> <dbl> <dbl> <dbl> <dbl>
1 portal                257  32.1   1.7   3.2  28.9  35.4
2 društvene mreže       206  25.8   1.5   3    22.7  28.8
3 TV                    170  21.2   1.4   2.8  18.4  24.1
4 radio                  92  11.5   1.1   2.2   9.3  13.7
5 tisak                  75   9.4   1     2     7.4  11.4
```

Novinar piše da je portal najpopularniji izvor vijesti (31%), a društvene mreže su na drugom mjestu (26%). Tehnički je to točno, ali zanemaruje intervale pouzdanosti. Kad uzmemo u obzir marginu pogreške, intervali za portal i društvene mreže se preklapaju. Ne možemo sa sigurnošću tvrditi da je portal popularniji od društvenih mreža.

```
rezultati |>
  mutate(primary_news_source = fct_reorder(primary_news_source, udio)) |>
  ggplot(aes(y = primary_news_source)) +
  geom_errorbarh(aes(xmin = ci_lo, xmax = ci_hi), height = 0.3, linewidth = 1, color = "st
  geom_point(aes(x = udio), size = 3, color = "steelblue") +
  geom_text(aes(x = ci_hi + 0.01, label = paste0(round(udio * 100, 1), "%")), hjust = 0) +
  scale_x_continuous(labels = scales::label_percent(), limits = c(0, 0.42)) +
  labs(
    title = "Anketni rezultati S intervalima pouzdanosti",
    subtitle = "n = 800. Preklapajući intervali znače da razlike nisu statistički jasne.",
    x = "Udio ispitanika",
    y = NULL
  ) +
  theme_minimal()
```



15.1 Kontrolna lista za čitanje anketa

Kad sretnete medijski izvještaj o anketi, postavite sedam pitanja. Koliki je uzorak? Ako ga ne navode, rezultati su sumnjivi. Kako su odabrali ispitanike? Slučajno telefonsko pozivanje ili online panel? Kolika je margina pogreške? Ako je navode samo na dnu stranice, obratite posebnu pažnju. Jesu li razlike veće od margine pogreške? Ako su razlike manje od dvostruke margine, zaključci su na klimavim nogama. Kad je anketa provedena? Stavovi se mogu promijeniti brzo. Tko je naručio anketu? Naručitelj može utjecati na formulaciju pitanja. Koliki je odaziv? Nizak odaziv (ispod 30%) sugerira self-selection bias.

---

16 Bootstrapping: alternativni pristup

Ponekad ne možemo pretpostaviti normalnost distribucije uzorkovanja, bilo zato što je uzorak premalen ili zato što nas zanima statistika za koju nemamo jednostavnu formulu za SE (medijan, omjer medijana i prosjeka, razlika između 90. i 10. percentila). **Bootstrap** je računalna metoda koja rješava ovaj problem.

Ideja je elegantna. Budući da ne možemo uzimati nove uzorke iz populacije (jer nemamo pristup cijeloj populaciji), uzimamo nove uzorke iz uzorka, s vraćanjem (with replacement).

```r
set.seed(42)
uzorak_50 <- pop |> slice_sample(n = 50)

# 5000 bootstrap uzoraka
boot_prosjeci <- map_dbl(1:5000, \(i) {
  uzorak_50 |>
    slice_sample(n = 50, replace = TRUE) |>
    pull(media_trust) |>
    mean()
})

# Bootstrap CI (percentilna metoda)
boot_ci <- quantile(boot_prosjeci, probs = c(0.025, 0.975))

# Usporedba s t-testom
t_ci <- t.test(uzorak_50$media_trust)$conf.int

cat("Bootstrap 95% CI: [", round(boot_ci[1], 3), ",", round(boot_ci[2], 3), "]\n")
```

```
Bootstrap 95% CI: [ 4.02 , 5.26 ]
```

```r
cat("t-test   95% CI: [", round(t_ci[1], 3), ",", round(t_ci[2], 3), "]\n")
```

```
t-test   95% CI: [ 4.02 , 5.26 ]
```

```r
cat("Pravi prosjek:   ", round(mean(pop$media_trust), 3), "\n")
```

```
Pravi prosjek:    4.867
```

Bootstrap i t-test daju vrlo slične rezultate kad su pretpostavke t-testa zadovoljene. Prednost bootstrapa je njegova fleksibilnost — možemo ga koristiti za bilo koju statistiku.

```r
# Bootstrap CI za MEDIJAN (za koji nema jednostavne formule za SE)
boot_medijani <- map_dbl(1:5000, \(i) {
  uzorak_50 |>
    slice_sample(n = 50, replace = TRUE) |>
    pull(daily_media_min) |>
    median()
})

boot_ci_medijan <- quantile(boot_medijani, probs = c(0.025, 0.975))

cat("Medijan uzorka:", median(uzorak_50$daily_media_min), "\n")
```

```
Medijan uzorka: 190
```

```r
cat("Bootstrap 95% CI za medijan: [", boot_ci_medijan[1], ",", boot_ci_medijan[2], "]\n")
```

```
Bootstrap 95% CI za medijan: [ 166.5 , 200 ]
```

```r
cat("Pravi populacijski medijan:", median(pop$daily_media_min), "\n")
```

```
Pravi populacijski medijan: 172
```

> **Kada koristiti bootstrap?**
>
> Koristite bootstrap kad nemate formulu za SE željene statistike, sumnjate u normalnost distribucije uzorkovanja, imate mali uzorak i tražite robusniju metodu, ili želite CI za medijan, percentile, omjere ili druge nestandardne mjere. Za prosjeke s $n > 30$, t-test je jednako dobar i jednostavniji.

---

17 Potpuna analiza: povjerenje u medije po demografskim skupinama

Spojimo sve naučeno u jednu koherentnu analizu. Situacija je sljedeća — provedena je anketa na 600 ispitanika o medijskim navikama. Trebamo procijeniti povjerenje u medije ukupno i po ključnim podgrupama te interpretirati rezultate.

```r
set.seed(2025)
anketa <- pop |> slice_sample(n = 600)

cat("Veličina uzorka:", nrow(anketa), "\n\n")
```

```
Veličina uzorka: 600
```

```r
# Korak 1: Ukupna procjena
ukupno <- t.test(anketa$media_trust)
cat("UKUPNO POVJERENJE U MEDIJE\n")
```

```
UKUPNO POVJERENJE U MEDIJE
```

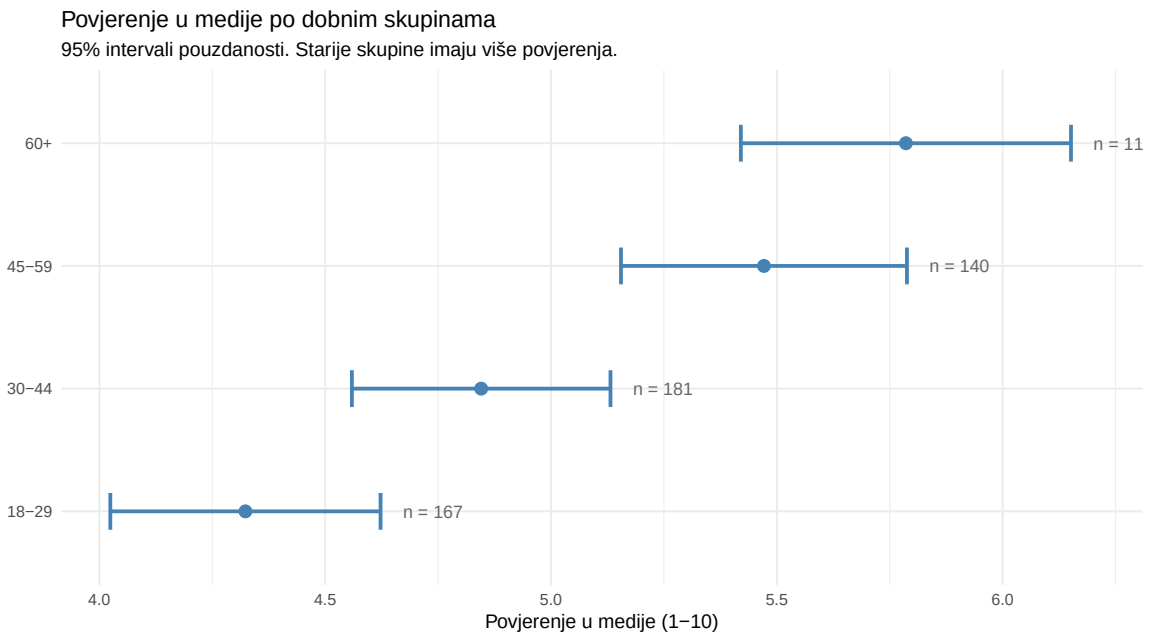
```r
cat("Prosjek:", round(ukupno$estimate, 2), "\n")
```

```
Prosjek: 5.02
```

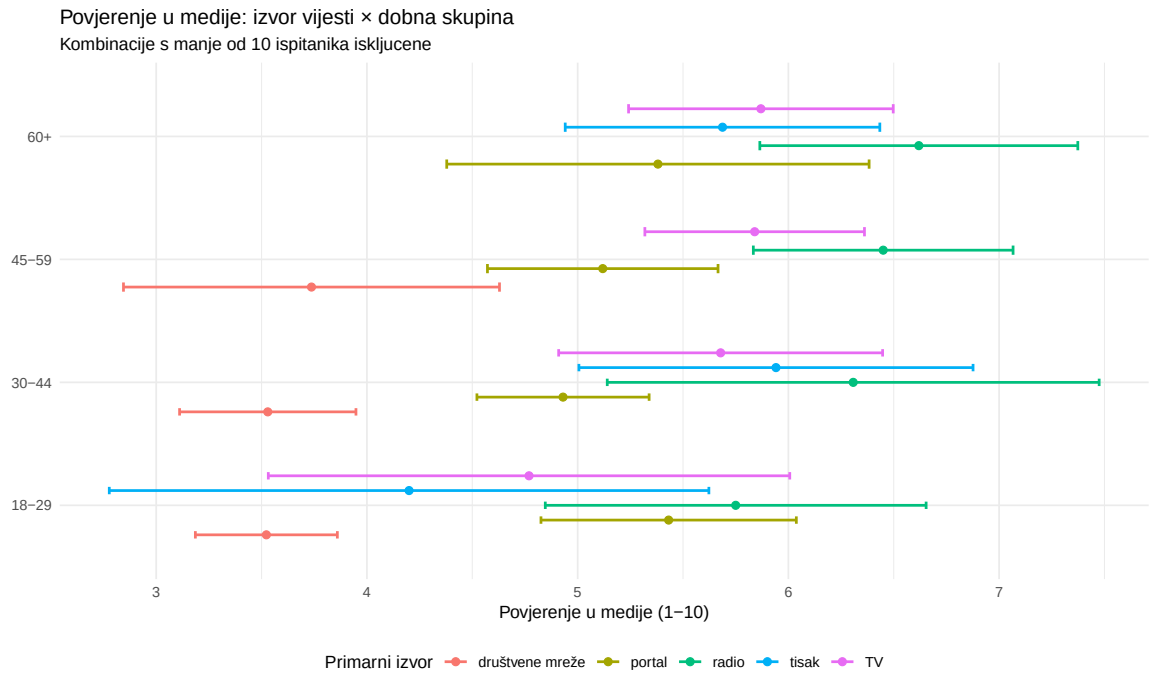
```r
cat("95% CI: [", round(ukupno$conf.int[1], 2), ",", round(ukupno$conf.int[2], 2), "]\n\n")
```

```
95% CI: [ 4.86 , 5.18 ]
```

```r
# Korak 2: CI po dobnim skupinama
anketa <- anketa |>
  mutate(dobna_skupina = case_when(
    age < 30 ~ "18-29",
    age < 45 ~ "30-44",
    age < 60 ~ "45-59",
    .default = "60+"
  ))

ci_dob <- anketa |>
  group_by(dobna_skupina) |>
  summarise(
    n = n(),
    prosjek = mean(media_trust),
    se = sd(media_trust) / sqrt(n()),
    ci_lo = prosjek - qt(0.975, n() - 1) * se,
    ci_hi = prosjek + qt(0.975, n() - 1) * se,
    .groups = "drop"
  )

ci_dob |>
  ggplot(aes(y = dobna_skupina)) +
  geom_errorbarh(aes(xmin = ci_lo, xmax = ci_hi), height = 0.3, linewidth = 1, color = "st
  geom_point(aes(x = prosjek), size = 3, color = "steelblue") +
  geom_text(aes(x = ci_hi + 0.05, label = paste0("n = ", n)), hjust = 0, size = 3.5, color
  labs(
    title = "Povjerenje u medije po dobnim skupinama",
    subtitle = "95% intervali pouzdanosti. Starije skupine imaju više povjerenja.",
    x = "Povjerenje u medije (1-10)",
    y = NULL
  ) +
  theme_minimal()
```

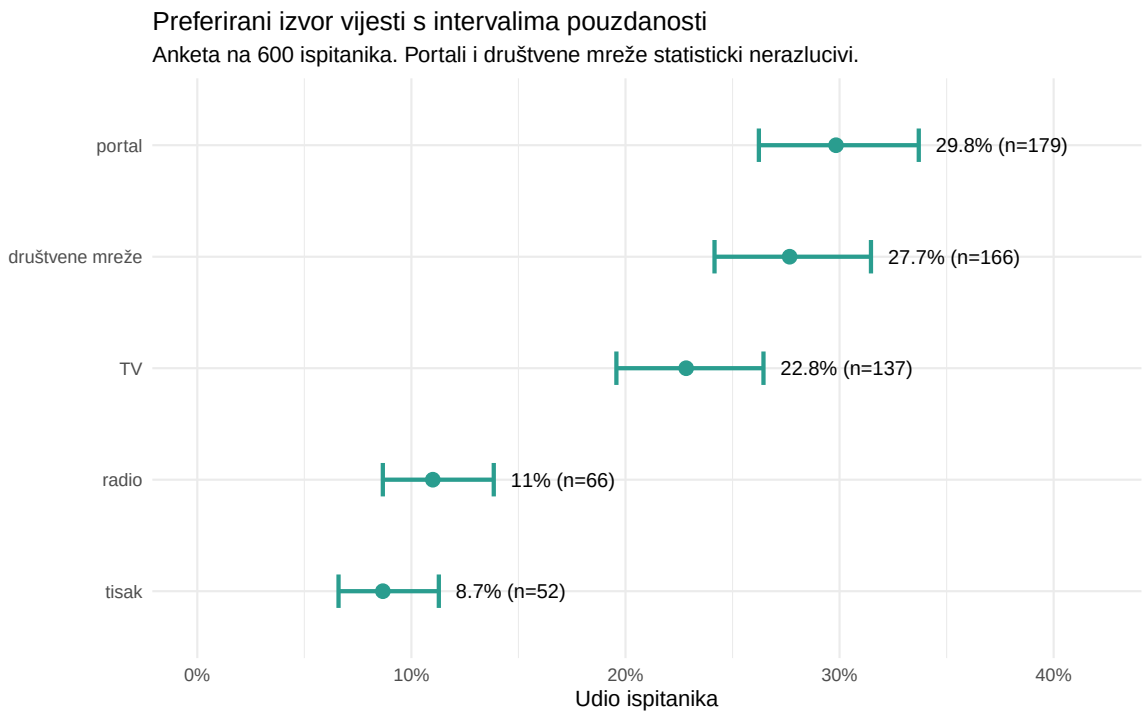



```
# Korak 3: Unakrsna analiza: izvor vijesti × dobna skupina
ci_krizno <- anketa |>
  group_by(primary_news_source, dobna_skupina) |>
  filter(n() >= 10) |>
  summarise(
    n = n(),
    prosjek = mean(media_trust),
    se = sd(media_trust) / sqrt(n()),
    ci_lo = prosjek - qt(0.975, n() - 1) * se,
    ci_hi = prosjek + qt(0.975, n() - 1) * se,
    .groups = "drop"
  )

ci_krizno |>
  ggplot(aes(y = dobna_skupina, color = primary_news_source)) +
  geom_errorbarh(aes(xmin = ci_lo, xmax = ci_hi), height = 0.3, linewidth = 0.8,
                 position = position_dodge(0.6)) +
  geom_point(aes(x = prosjek), size = 2, position = position_dodge(0.6)) +
  labs(
    title = "Povjerenje u medije: izvor vijesti × dobna skupina",
    subtitle = "Kombinacije s manje od 10 ispitanika isključene",
    x = "Povjerenje u medije (1-10)",
    y = NULL,
    color = "Primarni izvor"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```




```
# Korak 4: Proporcije po izvoru vijesti s CI
prop_rezultati <- anketa |>
  count(primary_news_source) |>
  mutate(
    p_hat = n / sum(n),
    test = map2(n, sum(n), \(x, nn) prop.test(x, nn, conf.level = 0.95)),
    ci_lo = map_dbl(test, \(t) t$conf.int[1]),
    ci_hi = map_dbl(test, \(t) t$conf.int[2])
  ) |>
  select(-test) |>
  arrange(desc(p_hat))

prop_rezultati |>
  mutate(primary_news_source = fct_reorder(primary_news_source, p_hat)) |>
  ggplot(aes(y = primary_news_source)) +
  geom_errorbarh(aes(xmin = ci_lo, xmax = ci_hi), height = 0.3, linewidth = 1, color = "#2
  geom_point(aes(x = p_hat), size = 3, color = "#2a9d8f") +
  geom_text(aes(x = ci_hi + 0.008, label = paste0(round(p_hat * 100, 1), "% (n=", n, ")"))
            hjust = 0, size = 3.5) +
  scale_x_continuous(labels = scales::label_percent(), limits = c(0, 0.42)) +
  labs(
    title = "Preferirani izvor vijesti s intervalima pouzdanosti",
    subtitle = "Anketa na 600 ispitanika. Portali i društvene mreže statistički nerazlučiv
    x = "Udio ispitanika",
    y = NULL
```

```
  ) +
  theme_minimal()
```




```
# Korak 5: Sažetak za klijenta
cat("=== SAŽETAK REZULTATA ANKETE ===\n\n")
```

```
=== SAŽETAK REZULTATA ANKETE ===
```

```
cat("Uzorak: n =", nrow(anketa), "ispitanika\n")
```

```
Uzorak: n = 600 ispitanika
```

```
cat("Margina pogreške za proporcije: ±", round(1.96 * sqrt(0.25 / nrow(anketa)) * 100, 1),
```

```
Margina pogreške za proporcije: ± 4 %
```

```
cat("1. Ukupno povjerenje u medije:", round(ukupno$estimate, 2),
    "(95% CI:", round(ukupno$conf.int[1], 2), "-", round(ukupno$conf.int[2], 2), ")\n")
```

```
1. Ukupno povjerenje u medije: 5.02 (95% CI: 4.86 - 5.18 )
```

```r
cat("   Na ljestvici od 1-10, to je ispod sredine.\n\n")
```

```
   Na ljestvici od 1-10, to je ispod sredine.
```

```r
cat("2. Najpopularniji izvori vijesti:\n")
```

```
2. Najpopularniji izvori vijesti:
```

```r
for (i in 1:nrow(prop_rezultati)) {
  cat("   ", prop_rezultati$primary_news_source[i], ":",
      round(prop_rezultati$p_hat[i] * 100, 1), "% (",
      round(prop_rezultati$ci_lo[i] * 100, 1), "-",
      round(prop_rezultati$ci_hi[i] * 100, 1), "%)\n")
}
```

```
    portal : 29.8 % ( 26.2 - 33.7 %)
    društvene mreže : 27.7 % ( 24.2 - 31.5 %)
    TV : 22.8 % ( 19.6 - 26.4 %)
    radio : 11 % ( 8.7 - 13.8 %)
    tisak : 8.7 % ( 6.6 - 11.3 %)
```

```r
cat("\n3. Napomena: razlika između portala i društvenih mreža je unutar margine\n")
```

```
3. Napomena: razlika između portala i društvenih mreža je unutar margine
```

```r
cat("   pogreške i ne može se smatrati statistički značajnom.\n")
```

```
   pogreške i ne može se smatrati statistički značajnom.
```

---

18 Uobičajene pogreške pri interpretaciji CI

Intervali pouzdanosti su intuitivno privlačni ali se često krivo interpretiraju. Evo najčešćih grešaka i ispravnih verzija.

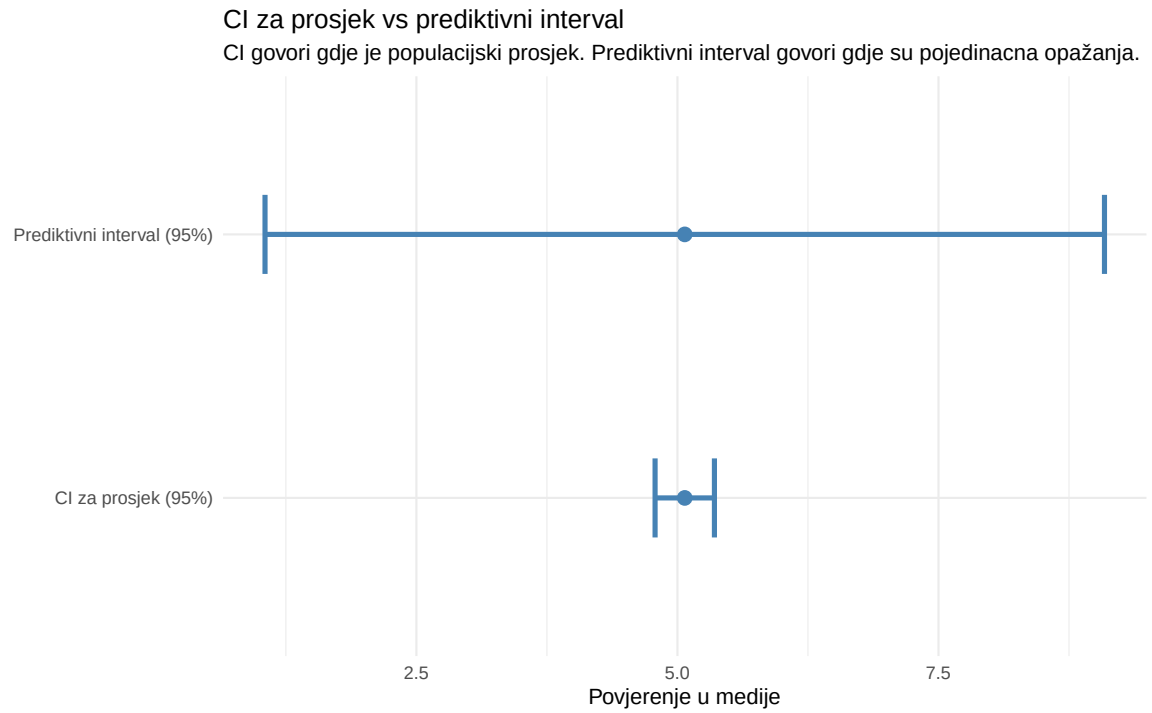
Pogrešno: “Postoji 95% šansa da je pravi prosjek unutar intervala [4.67, 5.01].” **Ispravno**: Pravi prosjek je fiksan broj. On ili jest ili nije unutar intervala. 95% se odnosi na postupak — 95% intervala konstruiranih ovom metodom pokrit će pravi parametar.

Pogrešno: “95% podataka pada unutar intervala [4.67, 5.01].” **Ispravno**: CI se odnosi na parametar (prosjek), ne na pojedinačna opažanja. Pojedinačne vrijednosti pokriva prediktivni interval, koji je mnogo širi.

Pogrešno: “Ako dva CI-a ne uključuju nulu, razlika je značajna.” **Ispravno**: Nula nije relevantna za pojedinačne CI-e. Preklapanje dvaju CI-a govori o mogućoj razlici, ali formalni test zahtijeva CI za razliku, što će biti obrađeno na predavanju o t-testu.

Pogrešno: “Širok CI znači da je mjerenje loše provedeno.” **Ispravno**: Širok CI obično znači mali uzorak ili veliku varijabilnost u podacima. To nije greška, nego realnost podataka.

```
set.seed(42)
uzorak_200 <- pop |> slice_sample(n = 200)

xbar <- mean(uzorak_200$media_trust)
se <- sd(uzorak_200$media_trust) / sqrt(200)
s <- sd(uzorak_200$media_trust)

tibble(
  tip = factor(c("CI za prosjek (95%)", "Prediktivni interval (95%)"),
               levels = c("CI za prosjek (95%)", "Prediktivni interval (95%)")),
  lo = c(xbar - 1.96 * se, xbar - 1.96 * s),
  hi = c(xbar + 1.96 * se, xbar + 1.96 * s),
  xbar = xbar
) |>
  ggplot(aes(y = tip)) +
  geom_errorbarh(aes(xmin = lo, xmax = hi), height = 0.3, linewidth = 1.2, color = "steelb
  geom_point(aes(x = xbar), size = 3, color = "steelblue") +
  labs(
    title = "CI za prosjek vs prediktivni interval",
    subtitle = "CI govori gdje je populacijski prosjek. Prediktivni interval govori gdje s
    x = "Povjerenje u medije",
    y = NULL
  ) +
  theme_minimal()
```



CI za prosjek je uzak (±0.28). Prediktivni interval je širok (±3.9). Ovo su dva potpuno različita pitanja gdje se razmatra gdje je pravi prosjek, odnosno gdje će pasti sljedeće opažanje. Ne miješajte ih.

! Ključni zaključci

1. Populacija je cjelina o kojoj zaključujemo. Uzorak je dio koji mjerimo. Parametri (,) opisuju populaciju. Statistike ($\bar{x}$, s) opisuju uzorak i procjenjuju parametre.
2. Svaki uzorak daje malo drugačiju procjenu. Distribucija tih procjena kroz ponovljene uzorke naziva se distribucija uzorkovanja. Njezina standardna devijacija je standardna pogreška (SE).
3. Standardna pogreška SE = s/√n mjeri koliko prosjeci uzoraka tipično variraju. Preciznost raste s korijenom veličine uzorka — da biste prepolovili SE, morate učetverostručiti n.
4. Centralni granični teorem kaže da je distribucija uzorkovanja prosjeka približno normalna za dovoljno velik n (pravilo palca n 30), neovisno o obliku izvorne distribucije.
5. Pristranost uzorka (convenience sampling, self-selection) je veći problem od male veličine uzorka. Velik pristran uzorak daje sustavno pogrešne rezultate koje ne može ispraviti nijedna statistička metoda.

6. t-distribucija se koristi umjesto normalne kad procjenjujemo iz uzorka. Za male uzorke daje šire intervale (deblje repove). Za n > 100, razlika je zanemariva.

7. `t.test()` računa interval pouzdanosti za prosjek. `prop.test()` računa CI za proporciju. Obje funkcije automatski koriste ispravne formule.

8. 95% CI znači da 95% ovako konstruiranih intervala pokriva pravi parametar. NE znači "95% šansa da je parametar unutar intervala." Parametar je fiksan. Interval je slučajan.

9. Margina pogreške za proporcije je približno $1/\sqrt{n}$, što daje za n = 1000 oko ±3.1%, a za n = 400 oko ±4.9%. To objašnjava uobičajene veličine uzoraka u anketama.

10. Kad planirate istraživanje, odredite ciljanu preciznost unaprijed i iz nje izvedite potrebnu veličinu uzorka koristeći formulu $n = z^{*2} \times p(1-p) / MoE^2$.

11. Bootstrap je računalna alternativa za CI kad nemamo formulu za SE željene statistike. Ideja je sljedeća — uzorkuj iz uzorka s vraćanjem, izračunaj statistiku, ponovi mnogo puta.

12. Kod čitanja medijskih anketa uvijek provjerite veličinu uzorka, metodu uzorkovanja, marginu pogreške i jesu li prijavljivane razlike veće od margine. Ako razlika između dvaju postotaka nije veća od dvostruke margine pogreške, zaključak da je neka opcija "u vodstvu" nije opravdan.

---

19 Zadaci za pripremu

1. Učitajte `media_population.csv` i izračunajte pravi populacijski prosjek za `daily_media_min`. Zatim uzmite 50 slučajnih uzoraka veličine n = 300 i za svaki izračunajte 95% CI pomoću `t.test()`. Koliki postotak intervala pokriva pravi prosjek?

2. Izračunajte potrebnu veličinu uzorka za anketu o preferencijama izvora vijesti s marginom pogreške ±2.5% na razini pouzdanosti 99%.

3. Napišite funkciju `bootstrap_ci(x, stat_fn, n_boot = 5000, conf = 0.95)` koja prima vektor `x`, funkciju `stat_fn` (npr. `mean` ili `median`) i vraća bootstrap CI. Testirajte je na varijabli `daily_media_min`.

---

20 Dodatno čitanje

Obavezno

Navarro, D. (2018). *Learning Statistics with R*, Chapter 10 (Estimating Unknown Quantities from a Sample). Besplatno dostupno na learningstatisticswithr.com. Pokriva uzorkovanje, CLT i intervale pouzdanosti s R kodom i odličnim objašnjenjima.

Diez, D., Çetinkaya-Rundel, M., & Barr, C. (2019). *OpenIntro Statistics* (4th edition), Chapter 5. Besplatno dostupno na openintro.org/book/os. Odličan vizualni pregled distribucije uzorkovanja i CI-a.

Preporučeno

Wheelan, C. (2013). *Naked Statistics*. W. W. Norton. Poglavlja 8 i 9 pokrivaju centralni granični teorem i uzorkovanje na izuzetno pristupačan način.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7-29. Članak argumentira zašto intervali pouzdanosti i veličine učinka trebaju zamijeniti p-vrijednosti kao primarni način izvještavanja rezultata.

---

21 Pojmovnik

Pojam	Objašnjenje
Populacija	Cjelokupni skup jedinica o kojima želimo donijeti zaključak.
Uzorak	Podskup populacije koji zaista mjerimo.
Parametar	Mjera populacije. Označava se grčkim slovima (, ,). U praksi nepoznat.
Statistika	Mjera uzorka. Označava se latinskim slovima ($\bar{x}$, s, $\hat{p}$). Procjena parametra.
Pogreška uzorkovanja	Razlika između statistike i parametra. Neizbježna posljedica rada s uzorkom.
Distribucija uzorkovanja	Distribucija statistike kroz mnogo ponovljenih uzoraka. Osnova za statističko zaključivanje.
Standardna pogreška (SE)	Standardna devijacija distribucije uzorkovanja. Za prosjek: SE = $s/\sqrt{n}$.
Centralni granični teorem (CLT)	Distribucija uzorkovanja prosjeka je približno normalna za dovoljno velik n, neovisno o obliku izvorne distribucije.

Pojam	Objašnjenje
t-distribucija	Distribucija slična normalnoj ali s debljim repovima. Koristi se kad procjenjujemo iz uzorka.
Stupnjevi slobode (df)	Parametar t-distribucije. Za jedan prosjek: df = n − 1. Više df = bliže normalnoj.
Interval pouzdanosti (CI)	Raspon vrijednosti koji s određenom vjerojatnošću pokriva pravi parametar.
Razina pouzdanosti	Postotak intervala koji bi pokrio parametar u ponovljenom uzorkovanju (obično 95% ili 99%).
Margina pogreške (MoE)	Pola širine intervala pouzdanosti. Za proporcije $1/\sqrt{n}$.
Nepristrana procjena	Statistika čija distribucija uzorkovanja je centrirana oko pravog parametra.
Convenience sampling	Uzorkovanje iz dostupne (ali nereprezentativne) skupine. Uvodi pristranost.
Self-selection bias	Pristranost kad ispitanici sami odlučuju hoće li sudjelovati. Tipično za online ankete.
Proporcija uzorka ($\hat{p}$)	Udio uzorka koji ima neku karakteristiku. Procjena populacijske proporcije .
Bootstrap	Računalna metoda za procjenu SE i CI ponovljenim uzorkovanjem iz uzorka s vraćanjem.
Prediktivni interval	Interval koji pokriva buduća pojedinačna opažanja. Mnogo širi od CI za prosjek.
Točkasta procjena	Jedna brojčana vrijednost kao procjena parametra (npr. $\bar{x} = 4.87$). Ne govori o preciznosti.
`t.test()`	R funkcija za t-test i t-interval pouzdanosti za prosjek.
`prop.test()`	R funkcija za test i CI za proporcije. Koristi Wilsonov interval.
`slice_sample()`	dplyr funkcija za slučajno uzorkovanje redova iz tibble. Argument `replace = TRUE` za bootstrap.
`qt()`	R funkcija za kritične vrijednosti t-distribucije. `qt(0.975, df)` za 95% CI.
`qnorm()`	R funkcija za kritične vrijednosti normalne distribucije. `qnorm(0.975)` = 1.96.