Tjedan 9: Testiranje hipoteza

Kako donijeti odluku na temelju podataka

2025-04-26

Table of contents

1 Jesu li carouseli zaista bolji? 2

2 Logika testiranja hipoteza 4

3 Od hipoteze do odluke 4

4 Jednouzorački t-test 5
4.1 Testna statistika 5
4.2 P-vrijednost 6
4.3 `t.test()` obavlja sve za vas 8

5 Dvosmjerni i jednosmjerni test 9

6 Dvouzorački t-test: natrag na Instagram 10
6.1 Welchov t-test: default koji ne trebate mijenjati 12

7 Simulacija: što p-vrijednost zapravo znači 13
7.1 Kad razlika zaista postoji 15

8 Dvije vrste pogrešaka 16

9 P-vrijednost: raščistimo zablude 18

10 Veličina učinka: Cohenov d 19
10.1 Što znači mali, srednji i veliki učinak 20

11 Statistička snaga: hoće li vaš test uopće nešto naći? 22
11.1 Koliki uzorak trebam? 23
11.2 Kolika je snaga našeg Instagram testa? 25

12 Upareni t-test: kad iste jedinice mjerite dva puta 26

13 Statistička značajnost nije isto što i praktična važnost **29**

14 Sve zajedno: izvještaj za urednicu **31**

15 ASA izjava i problem višestrukog testiranja **39**
15.1 Višestruko testiranje: kad testirate mnogo toga, nešto će "ispasti značajno" . 39

16 Pregled svih t-testova **41**

17 Zadaci za pripremu **42**

18 Dodatno čitanje **42**

19 Pojmovnik **43**

```
library(tidyverse)
```

Ishodi učenja

Nakon ovog predavanja moći ćete:

1. Formulirati nultu i alternativnu hipotezu za istraživačko pitanje.
2. Objasniti logiku testiranja hipoteza kroz analogiju sa suđenjem.
3. Izračunati i interpretirati testnu statistiku i p-vrijednost.
4. Provesti jednosmjerni i dvosmjerni t-test u R-u pomoću `t.test()`.
5. Objasniti razliku između greške tipa I i greške tipa II.
6. Izračunati i interpretirati Cohenov d kao mjeru veličine učinka.
7. Objasniti koncept statističke snage i faktore koji na nju utječu.
8. Kritički ocijeniti statističku značajnost u kontekstu praktične važnosti.

1 Jesu li carouseli zaista bolji?

Radite kao analitičarka u redakciji medijske kuće. Vaš Instagram profil objavljuje sadržaj u dva formata — carousel (objave s više slika koje korisnik lista) i obične slike (single image). Urednica vas jednog jutra zaustavi u hodniku i pita: "Imam osjećaj da carousel objave generiraju više angažmana. Imamo li za to dokaz?"

Vi znate odgovoriti na to pitanje. Otvarate podatke i gledate prosjeke — carousel objave imaju engagement rate od 10.1%, a obične slike 7.5%. Razlika postoji. Ali urednica nije pitala "je li prosjek različit u uzorku" — ona pita "možemo li se osloniti na tu razliku kad planiramo strategiju." A to je sasvim drugo pitanje. Možda je razlika realna i stabilna. Ali možda je samo artefakt — slučajni šum u podacima koji bi nestao kad bismo ponovili usporedbu na novom skupu objava.

Ovo je temeljno pitanje testiranja hipoteza — je li opažena razlika dovoljno velika da isključimo slučajnost kao objašnjenje? Drugim riječima, koliko bismo bili iznenađeni ovakvom razlikom da carousel zapravo *nije* bolji?

```
ig <- read_csv("../resources/datasets/instagram_ab_test.csv")
glimpse(ig)
```

```
Rows: 500
Columns: 11
$ post_id         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
$ format          <chr> "carousel", "carousel", "carousel", "carousel", "carou~
$ topic           <chr> "sport", "tech", "sport", "vijesti", "kultura", "kultu~
$ time_of_day     <chr> "poslijepodne", "jutro", "večer", "poslijepodne", "pod~
$ has_cta         <lgl> TRUE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, TRU~
$ reach           <dbl> 803, 1028, 4570, 1915, 4539, 3620, 3058, 2818, 4432, 2~
$ likes           <dbl> 48, 62, 200, 87, 428, 151, 61, 132, 212, 159, 210, 65,~
$ comments        <dbl> 14, 10, 94, 21, 47, 61, 27, 32, 108, 26, 42, 4, 30, 21~
$ shares          <dbl> 7, 8, 22, 25, 45, 50, 21, 27, 4, 14, 26, 5, 12, 14, 6,~
$ saves           <dbl> 16, 23, 49, 27, 67, 38, 30, 56, 84, 51, 28, 3, 12, 26,~
$ engagement_rate <dbl> 0.1059, 0.1002, 0.0799, 0.0836, 0.1293, 0.0829, 0.0455~
```

```
ig |>
  group_by(format) |>
  summarise(
    n = n(),
    M_engagement = round(mean(engagement_rate) * 100, 2),
    SD_engagement = round(sd(engagement_rate) * 100, 2),
    M_likes = round(mean(likes), 1),
    M_comments = round(mean(comments), 1),
    .groups = "drop"
  )
```

```
# A tibble: 2 x 6
  format           n M_engagement SD_engagement M_likes M_comments
  <chr>        <int>        <dbl>         <dbl>   <dbl>      <dbl>
1 carousel       236         10.1          2.11    180.       35.7
2 single_image   264          7.5          1.77    149.       24.4
```

Carousel objave imaju viši angažman u prosjeku. Ali svaka grupa ima i vlastitu varijabilnost — unutar carousela postoje sjajne i loše objave, isto kao i unutar običnih slika. Pitanje je — kolika je šansa da bismo vidjeli ovakvu ili veću razliku čak i da carousel zapravo nije bolji?

2 Logika testiranja hipoteza

Testiranje hipoteza slijedi logiku koja je iznenađujuće slična suđenju u kaznenom pravu. Na sudu, optuženik je nevin dok se ne dokaže krivnja. Ne morate dokazati nevinost — morate dokazati krivnju, i to izvan razumne sumnje. Ako dokazi nisu dovoljno jaki, presuda nije "nevin" nego "nije dokazano."

U statistici, uloge su analogne. Početna pretpostavka je da nema učinka — nema razlike, nema veze, nema efekta. Ovu pretpostavku zovemo nulta hipoteza i označavamo je s H . Istraživač pokušava prikupiti dovoljno dokaza da odbaci nultu hipotezu u korist alternativne hipoteze (H), koja tvrdi da učinak postoji.

Za naš Instagram primjer, hipoteze izgledaju kao što su sljedeće — H i H .

H — Nema razlike u angažmanu između carousel i single image formata. Svaka opažena razlika je posljedica slučajnosti.

H — Postoji razlika u angažmanu između dva formata. Opažena razlika odražava stvarnu razliku u populaciji.

U matematičkom jeziku to izražavamo na sljedeći način.

$$H_0 : \mu_{carousel} = \mu_{single}$$
$$H_1 : \mu_{carousel} \neq \mu_{single}$$

> **!** Nulta hipoteza uvijek sadrži jednakost
>
> Nulta hipoteza uvijek sadrži znak jednakosti (= ili ili). Alternativna hipoteza sadrži znak nejednakosti (ili > ili <). Nikad obrnuto. Mi testiramo nultu hipotezu i tražimo dokaze *protiv* nje — baš kao što tužitelj traži dokaze protiv pretpostavke nevinosti.

3 Od hipoteze do odluke

Cijeli postupak testiranja hipoteza možete sažeti u pet koraka. Prvi — postavite hipoteze, jasno formulirajte H i H prije nego pogledate podatke. Drugi — odaberite razinu značajnosti , prag ispod kojeg ćete odbaciti H (konvencija je = 0.05, odnosno 5%). Treći — izračunajte testnu statistiku iz podataka, broj koji kvantificira koliko se vaši podaci razlikuju od očekivanih pod H . Četvrti — izračunajte p-vrijednost, vjerojatnost da biste dobili ovako ekstremnu ili ekstremniju testnu statistiku kad bi H bila istinita. Peti — donesite odluku, ako je p < , odbacujete H ; ako je p , ne možete je odbaciti.

Krenimo korak po korak na jednostavnijem primjeru prije nego se vratimo na Instagram podatke.

4 Jednouzorački t-test

Najjednostavniji oblik t-testa uspoređuje prosjek jednog uzorka s nekom poznatom ili pretpostavljenom vrijednošću. Evo konkretne situacije — medijska kuća tvrdi da njihovi korisnici provode prosječno 3 minute čitajući članak. Vi ste skeptični — vaš osjećaj je da je stvarno vrijeme kraće. Provedete mjerenje na uzorku od 45 članaka.

```
set.seed(42)

# Simulirani podaci: stvarno prosječno vrijeme je 2.6 minuta
vrijeme_citanja <- tibble(
  clanak_id = 1:45,
  minuta = round(rnorm(45, mean = 2.6, sd = 0.9), 1)
)

# Opisna statistika
vrijeme_citanja |>
  summarise(
    n = n(),
    M = round(mean(minuta), 2),
    SD = round(sd(minuta), 2),
    SE = round(sd(minuta) / sqrt(n()), 3)
  )
```

```
# A tibble: 1 x 4
      n     M    SD    SE
  <int> <dbl> <dbl> <dbl>
1    45  2.54  1.06 0.159
```

Prosjek uzorka je ispod 3 minute. Ali je li dovoljno daleko od 3 da možemo odbaciti tvrdnju medijske kuće? Možda je razlika samo slučajni šum.

4.1 Testna statistika

Testna statistika za jednouzorački t-test mjeri koliko je prosjek uzorka udaljen od pretpostavljene vrijednosti, izraženo u jedinicama standardne pogreške — formalno, to je:

$$t = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Raspakujmo formulu. U brojniku je razlika između prosjeka uzorka ($\bar{x}$) i vrijednosti koju testiramo ($\mu_0 = 3$ minute). U nazivniku je standardna pogreška (SE), koja vam govori koliko prosjek uzorka tipično varira od uzorka do uzorka. Cijeli razlomak, dakle, kaže: "koliko

standardnih pogrešaka je moj prosjek udaljen od pretpostavljene vrijednosti?" Što je taj broj veći po apsolutnoj vrijednosti, to su podaci neobičniji pod nultom hipotezom.

```
x_bar <- mean(vrijeme_citanja$minuta)
mu_0 <- 3  # tvrdnja medijske kuće
s <- sd(vrijeme_citanja$minuta)
n <- nrow(vrijeme_citanja)
se <- s / sqrt(n)

t_stat <- (x_bar - mu_0) / se

cat("x̄ =", round(x_bar, 2), "\n")
```

```
x̄ = 2.54
```

```
cat("  =", mu_0, "\n")
```

```
  = 3
```

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

```
SE = 0.159
```

```
cat("t =", round(t_stat, 3), "\n")
```

```
t = -2.911
```

```
cat("df =", n - 1, "\n")
```

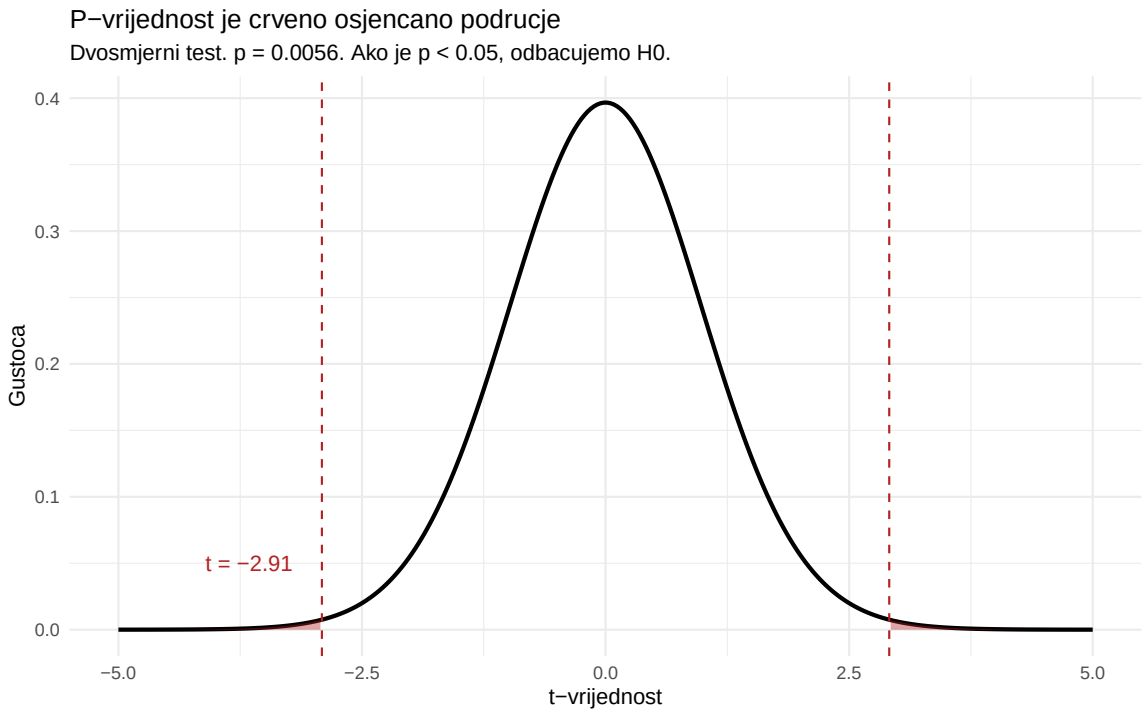
```
df = 44
```

Testna statistika t je negativna jer je prosjek uzorka manji od pretpostavljene vrijednosti. Apsolutna vrijednost |t| govori koliko standardnih pogrešaka je prosjek udaljen od . Što je ta udaljenost veća, to su jači dokazi protiv H .

4.2 P-vrijednost

Sada dolazi ključni korak. P-vrijednost je vjerojatnost da biste dobili testnu statistiku jednako ekstremnu ili ekstremniju od opažene, *pod pretpostavkom da je H istinita*. Ovo je suptilno ali ključno — p-vrijednost vam ne govori koliko je vjerojatno da je H istinita. Ona vam govori koliko bi vaši podaci bili neobični u svijetu gdje je H istinita.

```r
# Dvosmjerni test: gledamo obje strane
p_value <- 2 * pt(abs(t_stat), df = n - 1, lower.tail = FALSE)

x_vals <- seq(-5, 5, length.out = 300)

t_data <- tibble(x = x_vals, density = dt(x_vals, df = n - 1))

ggplot(t_data, aes(x = x, y = density)) +
  geom_line(linewidth = 1) +
  geom_area(data = t_data |> filter(x <= -abs(t_stat)),
            fill = "firebrick", alpha = 0.4) +
  geom_area(data = t_data |> filter(x >= abs(t_stat)),
            fill = "firebrick", alpha = 0.4) +
  geom_vline(xintercept = c(-abs(t_stat), abs(t_stat)),
             color = "firebrick", linetype = "dashed") +
  annotate("text", x = t_stat - 0.3, y = 0.05,
           label = paste("t =", round(t_stat, 2)), color = "firebrick", hjust = 1) +
  labs(
    title = "P-vrijednost je crveno osjenčano područje",
    subtitle = paste0("Dvosmjerni test. p = ", round(p_value, 4),
                      ". Ako je p < 0.05, odbacujemo H ."),
    x = "t-vrijednost",
    y = "Gustoća"
  ) +
  theme_minimal()
```




```
cat("t-statistika:", round(t_stat, 3), "\n")
```

```
t-statistika: -2.911
```

```
cat("P-vrijednost (dvosmjerni):", round(p_value, 4), "\n")
```

```
P-vrijednost (dvosmjerni): 0.0056
```

```
cat("  = 0.05\n")
```

```
  = 0.05
```

```
cat("Odluka:", if_else(p_value < 0.05, "Odbacujemo H ", "Ne možemo odbaciti H "), "\n")
```

```
Odluka: Odbacujemo H
```

4.3 `t.test()` obavlja sve za vas

U praksi ne trebate ručno računati t-statistiku i p-vrijednost. Funkcija `t.test()` sve radi u jednom pozivu.

```r
t.test(vrijeme_citanja$minuta, mu = 3)
```

```
    One Sample t-test

data:  vrijeme_citanja$minuta
t = -2.9114, df = 44, p-value = 0.005629
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.217817 2.857739
sample estimates:
mean of x
 2.537778
```

Funkcija vraća testnu statistiku, stupnjeve slobode, p-vrijednost, 95% interval pouzdanosti i prosjek uzorka. P-vrijednost je ispod 0.05, što znači da imamo dovoljno dokaza da odbacimo tvrdnju medijske kuće — prosječno vrijeme čitanja je statistički značajno različito od 3 minute.

> CI i testiranje hipoteza su dva lica istog novčića
>
> Pogledajte 95% interval pouzdanosti iz gornjeg rezultata. Ne sadrži vrijednost 3. To nije slučajnost — odbacivanje H na razini = 0.05 je matematički ekvivalentno tome da 95% CI ne sadrži testiranu vrijednost . Ovo su dva načina gledanja na isti problem — i obje perspektive su korisne.

5 Dvosmjerni i jednosmjerni test

U prethodnom primjeru koristili smo dvosmjerni test (two-tailed), što znači da smo testirali je li prosjek *različit* od 3, u bilo kojem smjeru. Hipoteze su bile H : = 3 nasuprot H : 3.

Ponekad unaprijed znate smjer. Ako Instagram tim očekuje da su carousel objave *bolje* (ne samo različite), može koristiti jednosmjerni test (one-tailed) s hipotezama H : _carousel _single nasuprot H : _carousel > _single.

```r
# Jednosmjerni: je li prosjek MANJI od 3?
t.test(vrijeme_citanja$minuta, mu = 3, alternative = "less")
```

```
    One Sample t-test

data:  vrijeme_citanja$minuta
t = -2.9114, df = 44, p-value = 0.002814
```

```
alternative hypothesis: true mean is less than 3
95 percent confidence interval:
     -Inf 2.804532
sample estimates:
mean of x
 2.537778
```

P-vrijednost jednosmjernog testa je točno pola dvosmjernog (kad je smjer u skladu s podacima). Jednosmjerni test je osjetljiviji u tom smjeru, ali potpuno slijep za razliku u suprotnom smjeru.

> ⚠ Smjer morate odrediti prije nego pogledate podatke
>
> Jednosmjerni test koristite samo ako ste smjer hipoteze odredili *prije* nego ste pogledali podatke. Ako pogledate podatke, vidite da je prosjek manji od 3, pa onda odlučite testirati "je li manji" — to je pristranost istraživača. Kad ste u sumnji, koristite dvosmjerni test. Velika većina objavljenih istraživanja koristi dvosmjerne testove upravo iz ovog razloga.

6 Dvouzorački t-test: natrag na Instagram

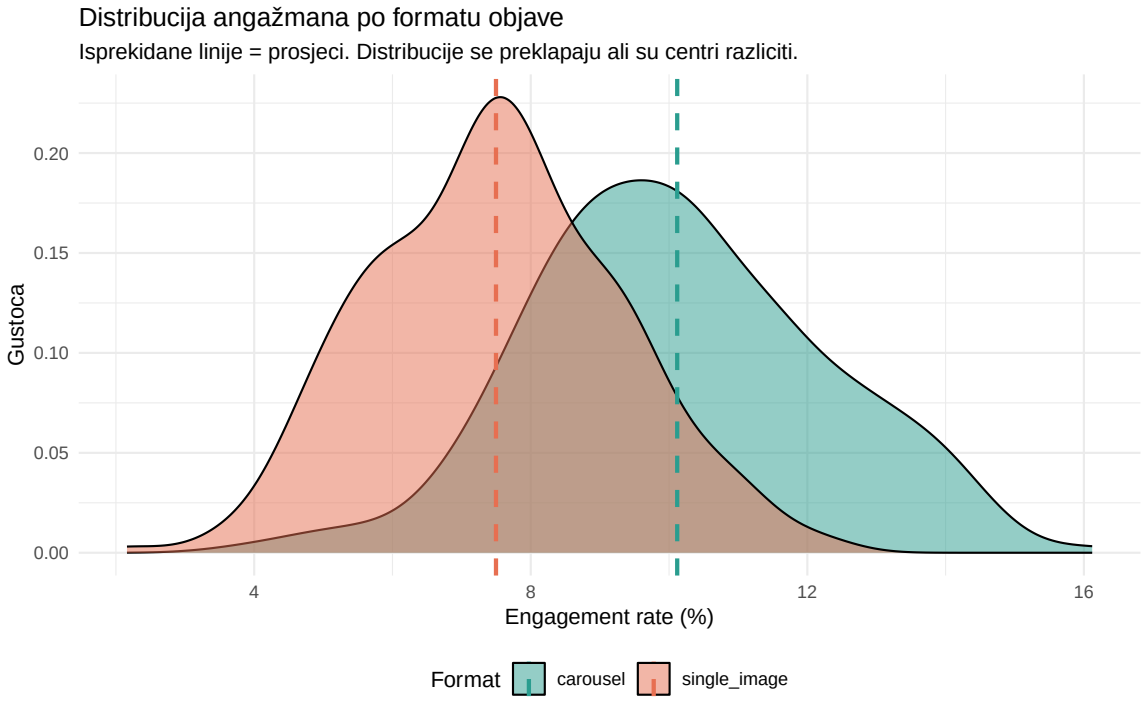
Sada se vraćamo na naš motivacijski primjer. Želimo testirati razlikuje li se angažman između carousel i single image objava. Budući da uspoređujemo prosjeke dviju nezavisnih skupina (carousel objave su jedne, single image su druge, i nema parenja), koristimo dvouzorački t-test.

$$H_0 : \mu_{carousel} = \mu_{single}$$
$$H_1 : \mu_{carousel} \neq \mu_{single}$$

Prije svega, pogledajmo distribucije.

```
ig |>
  ggplot(aes(x = engagement_rate * 100, fill = format)) +
  geom_density(alpha = 0.5) +
  geom_vline(data = ig |> group_by(format) |> summarise(M = mean(engagement_rate) * 100),
             aes(xintercept = M, color = format), linewidth = 1, linetype = "dashed") +
  scale_fill_manual(values = c("carousel" = "#2a9d8f", "single_image" = "#e76f51")) +
  scale_color_manual(values = c("carousel" = "#2a9d8f", "single_image" = "#e76f51")) +
  labs(
    title = "Distribucija angažmana po formatu objave",
    subtitle = "Isprekidane linije = prosjeci. Distribucije se preklapaju ali su centri ra
    x = "Engagement rate (%)",
    y = "Gustoća",
```

```
    fill = "Format", color = "Format"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Distribucije se preklapaju — postoje single image objave s visokim angažmanom i carousel objave s niskim — ali carousel distribucija je pomaknuta udesno. Testna statistika za dvouzorački t-test mjeri razliku prosjeka u jedinicama zajedničke standardne pogreške — ta statistika je:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE_{razlika}}$$

```
carousel <- ig |> filter(format == "carousel") |> pull(engagement_rate)
single <- ig |> filter(format == "single_image") |> pull(engagement_rate)

rezultat <- t.test(carousel, single)
rezultat
```

```
    Welch Two Sample t-test

data:  carousel and single
t = 14.942, df = 461.55, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 0.02276184 0.02965581
sample estimates:
 mean of x  mean of y
0.10117966 0.07497083
```

```
cat("Razlika prosjeka:", round((mean(carousel) - mean(single)) * 100, 2), "postotnih bodov
```

```
Razlika prosjeka: 2.62 postotnih bodova
```

```
cat("t-statistika:", round(rezultat$statistic, 2), "\n")
```

```
t-statistika: 14.94
```

```
cat("P-vrijednost:", format(rezultat$p.value, scientific = TRUE), "\n")
```

```
P-vrijednost: 1.856828e-41
```

```
cat("95% CI za razliku: [", round(rezultat$conf.int[1] * 100, 2), ",",
    round(rezultat$conf.int[2] * 100, 2), "] postotnih bodova\n")
```

```
95% CI za razliku: [ 2.28 , 2.97 ] postotnih bodova
```

P-vrijednost je iznimno mala — mnogo, mnogo manja od 0.05. Imamo snažne dokaze da se angažman zaista razlikuje između dva formata. Carousel objave imaju statistički značajno viši angažman.

6.1 Welchov t-test: default koji ne trebate mijenjati

R po defaultu koristi Welchov t-test, koji ne pretpostavlja jednake varijance u dvjema skupinama. Usporedimo ga s klasičnim Studentovim t-testom da vidite zašto je ovo mudar default.

```
# Welchov (default)
welch <- t.test(carousel, single, var.equal = FALSE)

# Studentov (pretpostavlja jednake varijance)
student <- t.test(carousel, single, var.equal = TRUE)

tibble(
  test = c("Welch (default)", "Student (var.equal=TRUE)"),
```

```
  t = round(c(welch$statistic, student$statistic), 3),
  df = round(c(welch$parameter, student$parameter), 1),
  p = format(c(welch$p.value, student$p.value), scientific = TRUE, digits = 3)
)
```

```
# A tibble: 2 x 4
  test                          t    df p
  <chr>                     <dbl> <dbl> <chr>
1 Welch (default)            14.9  462. 1.86e-41
2 Student (var.equal=TRUE)   15.1  498  1.27e-42
```

Rezultati su slični, ali Welchov test ima nerunde stupnjeve slobode jer ih prilagođava za razliku u varijancama. Kad su varijance jednake, oba testa daju gotovo identične rezultate. Kad varijance nisu jednake, Welchov je točniji. Zaključak je jednostavan — koristite Welchov test uvijek, jer ne zahtijeva dodatnu pretpostavku i nikad nije lošiji.

7 Simulacija: što p-vrijednost zapravo znači

P-vrijednost je jedan od najčešće pogrešno shvaćenih koncepata u cijeloj statistici. Simulacija pomaže izgraditi ispravnu intuiciju na način na koji teorijsko objašnjenje ne može.

Zamislimo svijet u kojem je H istinita — carousel i single image imaju identičan angažman, nema nikakve razlike. Ako u tom svijetu mnogo puta uzorkujemo i testiramo, koliko ćemo često *slučajno* dobiti $p < 0.05$?

```
set.seed(42)

# Simulacija: H je ISTINITA (isti prosjek za obje grupe)
sim_p <- map_dbl(1:10000, \(i) {
  grupa_a <- rnorm(100, mean = 0.08, sd = 0.02)
  grupa_b <- rnorm(100, mean = 0.08, sd = 0.02)  # ISTI prosjek!
  t.test(grupa_a, grupa_b)$p.value
})

cat("H je ISTINITA. Od 10 000 testova:\n")
```

```
H je ISTINITA. Od 10 000 testova:
```

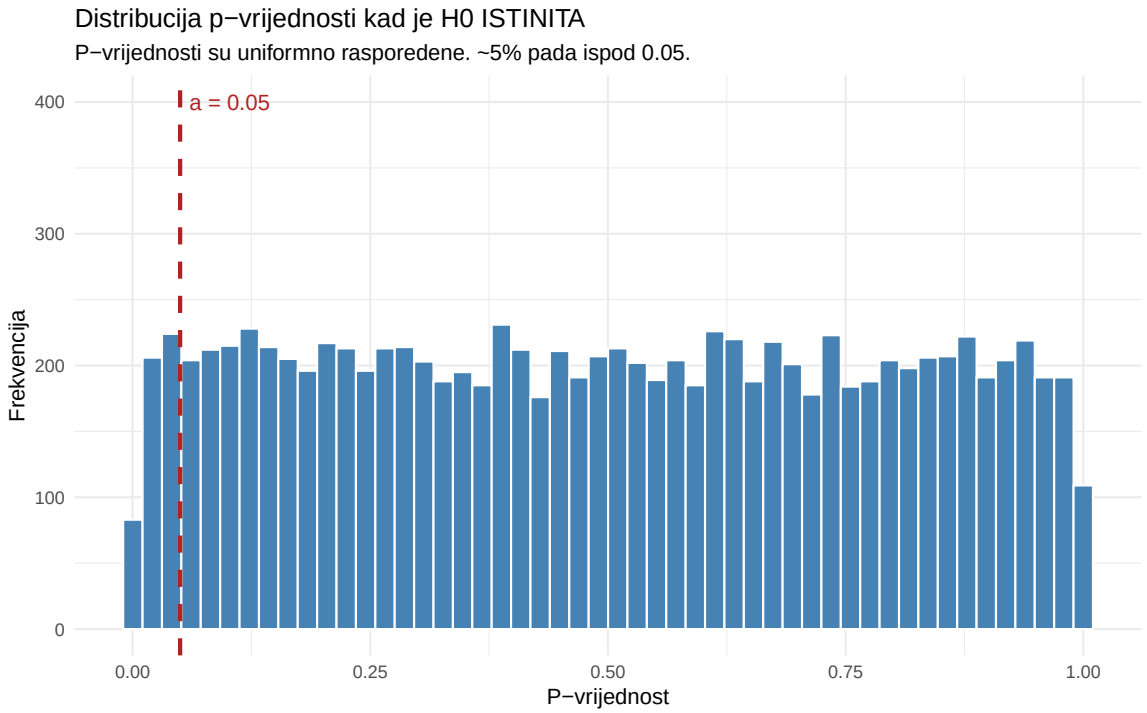
```
cat("p < 0.05:", sum(sim_p < 0.05), "(", round(mean(sim_p < 0.05) * 100, 1), "%)\n")
```

```
p < 0.05: 497 ( 5 %)
```

```r
cat("p < 0.01:", sum(sim_p < 0.01), "(", round(mean(sim_p < 0.01) * 100, 1), "%)\n")
```

```
p < 0.01: 79 ( 0.8 %)
```

```r
tibble(p = sim_p) |>
  ggplot(aes(x = p)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  geom_vline(xintercept = 0.05, color = "firebrick", linewidth = 1, linetype = "dashed") +
  annotate("text", x = 0.06, y = 400, label = "  = 0.05", color = "firebrick", hjust = 0) +
  labs(
    title = "Distribucija p-vrijednosti kad je H  ISTINITA",
    subtitle = "P-vrijednosti su uniformno raspoređene. ~5% pada ispod 0.05.",
    x = "P-vrijednost",
    y = "Frekvencija"
  ) +
  theme_minimal()
```



Ovo je ključan uvid. Kad je H istinita, p-vrijednosti su uniformno raspoređene između 0 i 1. Točno 5% pada ispod 0.05 — po definiciji . To znači da ćemo u 5% slučajeva pogrešno odbaciti H *čak i kad je istinita.* Ovo je greška tipa I, lažno pozitivni rezultat, i potpuno je neizbježna posljedica toga da smo postavili prag na 5%.

7.1 Kad razlika zaista postoji

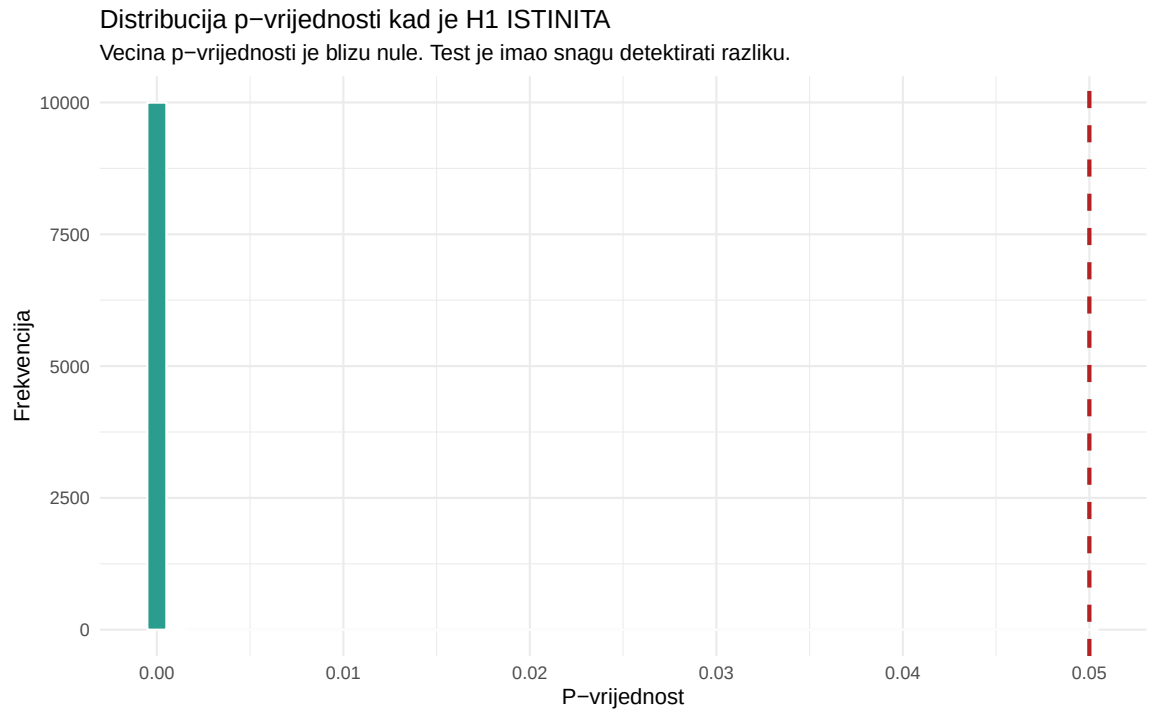
```
set.seed(42)

# Simulacija: H  je ISTINITA (razlika postoji)
sim_p_h1 <- map_dbl(1:10000, \(i) {
  grupa_a <- rnorm(100, mean = 0.10, sd = 0.02)
  grupa_b <- rnorm(100, mean = 0.08, sd = 0.02)  # RAZLIČIT prosjek
  t.test(grupa_a, grupa_b)$p.value
})

cat("H  je ISTINITA (razlika = 0.02). Od 10 000 testova:\n")
```

```
H  je ISTINITA (razlika = 0.02). Od 10 000 testova:
```

```
cat("p < 0.05:", sum(sim_p_h1 < 0.05), "(", round(mean(sim_p_h1 < 0.05) * 100, 1), "%)\n")
```

```
p < 0.05: 10000 ( 100 %)
```

```
tibble(p = sim_p_h1) |>
  ggplot(aes(x = p)) +
  geom_histogram(fill = "#2a9d8f", color = "white", bins = 50) +
  geom_vline(xintercept = 0.05, color = "firebrick", linewidth = 1, linetype = "dashed") +
  labs(
    title = "Distribucija p-vrijednosti kad je H  ISTINITA",
    subtitle = "Većina p-vrijednosti je blizu nule. Test je imao snagu detektirati razliku
    x = "P-vrijednost",
    y = "Frekvencija"
  ) +
  theme_minimal()
```



Slika je potpuno drugačija. Kad razlika zaista postoji, p-vrijednosti su koncentrirane blizu nule. Većina testova uspješno detektira razliku. Ali ne svi — neki testovi daju p 0.05 unatoč tome što razlika postoji. Postotak testova koji uspješno detektiraju pravu razliku zove se statistička snaga (power). Testovi koji je propuste čine grešku tipa II, lažno negativni rezultat.

8 Dvije vrste pogrešaka

Kad donosite odluku na temelju testa, možete pogriješiti na dva načina. Razumijevanje ovih dviju vrsta pogrešaka ključno je za mudru interpretaciju rezultata.

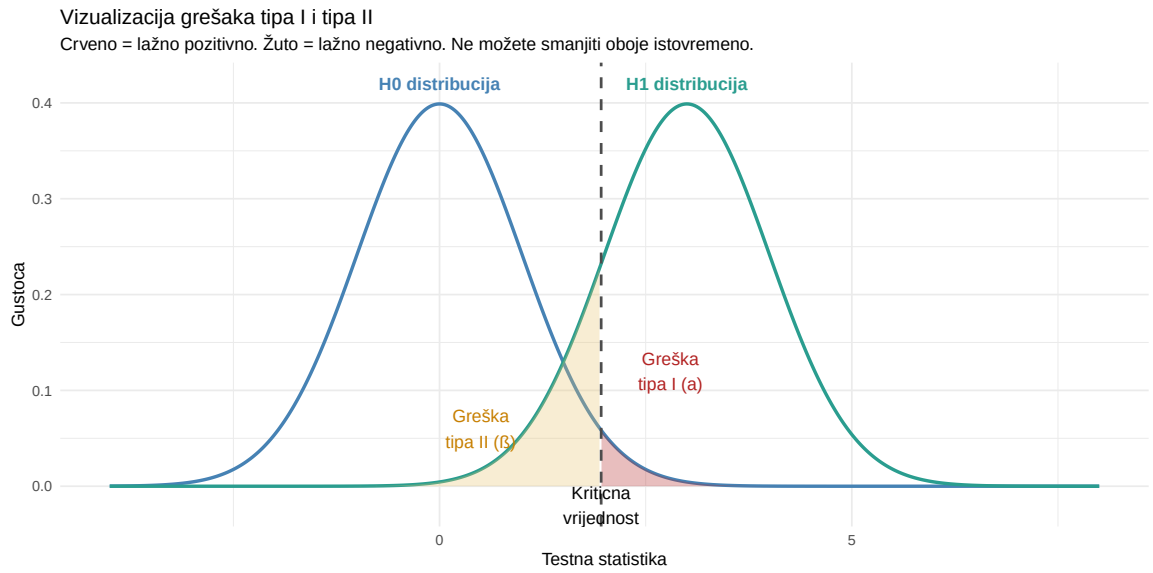
```
tribble(
  ~` `, ~`H  je istinita`, ~`H  je lažna`,
  "Ne odbacujemo H ", "  Ispravna odluka (1 -  )", "  Greška tipa II ( )",
  "Odbacujemo H ", "  Greška tipa I ( )", "  Ispravna odluka (snaga = 1 -  )"
)
```

```
# A tibble: 2 x 3
  ` `              `H  je istinita`           `H  je lažna`
  <chr>            <chr>                      <chr>
1 Ne odbacujemo H   Ispravna odluka (1 -  )   Greška tipa II ( )
2 Odbacujemo H      Greška tipa I ( )          Ispravna odluka (snaga = 1 -  )
```

Greška tipa I ( ) nastaje kad odbacite H iako je istinita — zaključite da razlika postoji kad je zapravo nema. Kontrolirate je postavljanjem (obično 0.05). U analogiji sa suđenjem, to je osuda nevine osobe.

Greška tipa II ( ) nastaje kad ne odbacite H iako je lažna — propustite pravu razliku. Ovisi o veličini uzorka, veličini učinka i razini . U analogiji sa suđenjem, to je oslobađanje krivca.

```
x <- seq(-4, 8, length.out = 500)
h0 <- dnorm(x, mean = 0, sd = 1)
h1 <- dnorm(x, mean = 3, sd = 1)

crit <- qnorm(0.975)

error_data <- tibble(x = x, H0 = h0, H1 = h1)

ggplot(error_data, aes(x = x)) +
  # H0 distribucija
  geom_line(aes(y = H0), color = "steelblue", linewidth = 1) +
  geom_area(data = error_data |> filter(x >= crit), aes(y = H0),
            fill = "firebrick", alpha = 0.3) +
  # H1 distribucija
  geom_line(aes(y = H1), color = "#2a9d8f", linewidth = 1) +
  geom_area(data = error_data |> filter(x < crit), aes(y = H1),
            fill = "#e9c46a", alpha = 0.3) +
  # Kritična vrijednost
  geom_vline(xintercept = crit, color = "grey30", linewidth = 0.8, linetype = "dashed") +
  annotate("text", x = 0, y = 0.42, label = "H  distribucija", color = "steelblue", fontfa
  annotate("text", x = 3, y = 0.42, label = "H  distribucija", color = "#2a9d8f", fontface
  annotate("text", x = 2.8, y = 0.12, label = "Greška\ntipa I ( )", color = "firebrick") +
  annotate("text", x = 0.5, y = 0.06, label = "Greška\ntipa II ( )", color = "#c77f00") +
  annotate("text", x = crit, y = -0.02, label = "Kritična\nvrijednost", hjust = 0.5) +
  labs(
    title = "Vizualizacija grešaka tipa I i tipa II",
    subtitle = "Crveno = lažno pozitivno. Žuto = lažno negativno. Ne možete smanjiti oboje
    x = "Testna statistika",
    y = "Gustoća"
  ) +
  theme_minimal()
```



Ovaj graf pokazuje ključan kompromis. Ako pomaknete kritičnu vrijednost udesno (stroži), smanjujete crveno područje (manje lažno pozitivnih) ali povećavate žuto (više lažno negativnih). Jedini način da smanjite oboje istovremeno je povećati uzorak (što razdvaja dvije distribucije) ili imati veći učinak.

> ❗ “Ne možemo odbaciti” nije isto što i “prihvaćamo”
>
> Odsutnost dokaza nije dokaz odsutnosti. Kad test daje p 0.05, ne kažemo “prihvaćamo H ” — kažemo “ne možemo odbaciti H na temelju dostupnih podataka.” Možda razlika postoji, ali naš uzorak je premalen da je detektira. Možda razlika postoji, ali je toliko mala da nije vidljiva s ovom količinom podataka. Zato nikad, nikad ne zaključujte “dokazali smo da nema razlike.”

9 P-vrijednost: raščistimo zablude

P-vrijednost je jedan od najčešće korištenih ali i najčešće pogrešno interpretiranih koncepata u cijeloj statistici. Potrebno je nekoliko minuta da precizno razjasnimo što ona jest, a što nije.

P-vrijednost *jest* vjerojatnost dobivanja testne statistike jednako ekstremne ili ekstremnije od opažene, pod pretpostavkom da je H istinita. Koliko biste bili iznenađeni ovakvim podacima da H zaista vrijedi?

P-vrijednost *nije* vjerojatnost da je H istinita. Ne možete reći “postoji samo 3% šanse da nema razlike.” P-vrijednost govori o podacima s obzirom na hipotezu, ne o hipotezi s obzirom na podatke. Ova razlika može djelovati kao cjepidlačenje, ali je zapravo fundamentalna.

P-vrijednost *nije* vjerojatnost da ste pogriješili. Mala p-vrijednost znači da su podaci neobični pod H . Ne znači da ste sigurno u pravu.

I ono najvažnije — p-vrijednost *nije* mjera veličine učinka. Vrijednost p = 0.001 ne znači da je razlika velika. Velik uzorak može proizvesti sićušnu p-vrijednost za trivijalno malu razliku. Pogledajmo to na primjeru.

```
set.seed(42)

# Mali uzorak, velik učinak
mali_uzorak <- t.test(rnorm(20, 10.5, 2), mu = 10)

# Velik uzorak, sićušan učinak
velik_uzorak <- t.test(rnorm(10000, 10.02, 2), mu = 10)

tibble(
  scenarij = c("Mali uzorak (n=20), velik učinak", "Velik uzorak (n=10000), sićušan učinak
  n = c(20, 10000),
  razlika = c("0.5 bodova", "0.02 boda"),
  p_vrijednost = c(round(mali_uzorak$p.value, 4), round(velik_uzorak$p.value, 4)),
  znacajno = c(mali_uzorak$p.value < 0.05, velik_uzorak$p.value < 0.05)
)
```

```
# A tibble: 2 x 5
  scenarij                                    n razlika    p_vrijednost znacajno
  <chr>                                   <dbl> <chr>             <dbl> <lgl>
1 Mali uzorak (n=20), velik učinak           20 0.5 bodova        0.149 FALSE
2 Velik uzorak (n=10000), sićušan učinak 10000 0.02 boda          0.843 FALSE
```

S 10 000 opažanja, razlika od 0.02 boda — praktički beznačajna — može biti statistički značajna. S 20 opažanja, razlika od 0.5 bodova — potencijalno važna — možda neće biti statistički značajna. Ovo jasno pokazuje zašto p-vrijednost sama nije dovoljna za donošenje odluka. Uvijek trebate i mjeru veličine učinka.

> **Gdje smo, kamo idemo**
>
> U prvom dijelu naučili smo logiku testiranja hipoteza, formuliranje H i H , jednouzorački i dvouzorački t-test, p-vrijednost i greške tipa I i II. U nastavku prelazimo na pitanje koje je jednako važno kao statistička značajnost — koliko je učinak zapravo velik?

10 Veličina učinka: Cohenov d

P-vrijednost odgovara na pitanje "postoji li učinak?" ali šuti o tome koliko je taj učinak velik. Za to vam treba mjera veličine učinka. Najčešća za razliku dvaju prosjeka je Cohenov d, koji izražava razliku u jedinicama zajedničke standardne devijacije — formalno:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$

Zašto dijeliti sa standardnom devijacijom? Zato što vam razlika od 2.6 postotnih bodova ne znači ništa dok ne znate koliko pojedinačne objave variraju. Ako sve objave imaju engagement rate između 7% i 11%, razlika od 2.6 bodova je ogromna. Ako variraju od 0% do 50%, ista razlika je zanemariva. Cohenov d stavlja razliku u kontekst varijabilnosti.

```
ig <- read_csv("../resources/datasets/instagram_ab_test.csv")

carousel <- ig |> filter(format == "carousel") |> pull(engagement_rate)
single <- ig |> filter(format == "single_image") |> pull(engagement_rate)

# Ručni izračun
n1 <- length(carousel)
n2 <- length(single)
s_pooled <- sqrt(((n1 - 1) * sd(carousel)^2 + (n2 - 1) * sd(single)^2) / (n1 + n2 - 2))
d <- (mean(carousel) - mean(single)) / s_pooled

cat("Razlika prosjeka:", round((mean(carousel) - mean(single)) * 100, 2), "postotnih bodov
```

```
Razlika prosjeka: 2.62 postotnih bodova
```

```
cat("Pooled SD:", round(s_pooled * 100, 2), "postotnih bodova\n")
```

```
Pooled SD: 1.94 postotnih bodova
```

```
cat("Cohenov d:", round(d, 3), "\n")
```

```
Cohenov d: 1.351
```

10.1 Što znači mali, srednji i veliki učinak

Cohen (1988) je predložio smjernice za interpretaciju koje su postale konvencija u društvenim znanostima. One izgledaju ovako:

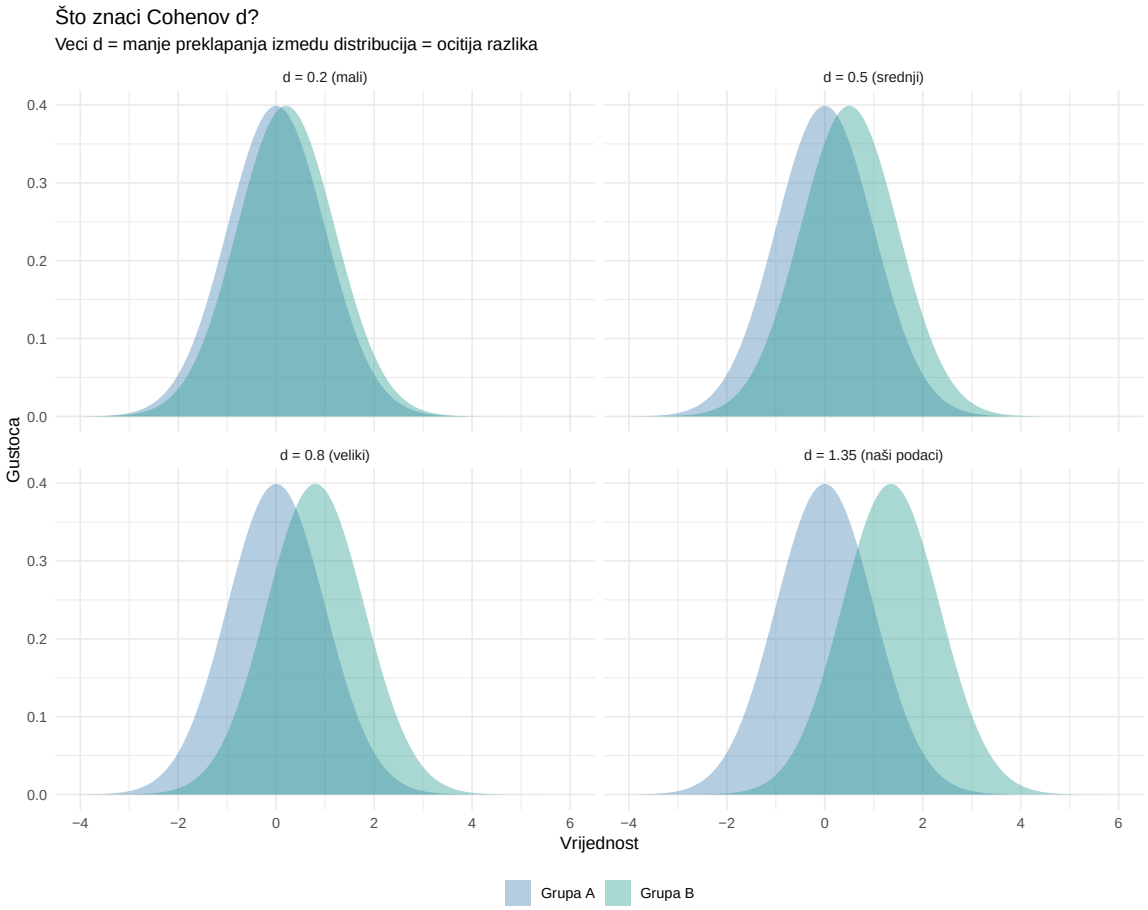
```
tribble(
  ~d, ~interpretacija, ~primjer,
  "0.2", "Mali učinak", "Jedva primjetna razlika u praksi",
  "0.5", "Srednji učinak", "Razlika vidljiva prostim okom",
  "0.8", "Veliki učinak", "Razlika očita i praktično važna"
)
```

```
# A tibble: 3 x 3
  d     interpretacija primjer
  <chr> <chr>          <chr>
1 0.2   Mali učinak    Jedva primjetna razlika u praksi
2 0.5   Srednji učinak Razlika vidljiva prostim okom
3 0.8   Veliki učinak  Razlika očita i praktično važna
```

Naš d je veliki učinak. Carousel objave generiraju značajno viši angažman, i to u praktično važnoj mjeri — ovo je informacija koju p-vrijednost sama ne može dati.

Vizualizirajmo što različite veličine učinka *izgledaju* kao preklapanje dviju distribucija.

```
# Vizualizacija: što znači d = 0.2, 0.5, 0.8, 1.3
d_values <- c(0.2, 0.5, 0.8, round(d, 2))
d_labels <- c("d = 0.2 (mali)", "d = 0.5 (srednji)", "d = 0.8 (veliki)",
              paste0("d = ", round(d, 2), " (naši podaci)"))

x <- seq(-4, 6, length.out = 300)

d_viz <- map_df(seq_along(d_values), \(i) {
  tibble(
    panel = d_labels[i],
    x = x,
    Grupa_A = dnorm(x, 0, 1),
    Grupa_B = dnorm(x, d_values[i], 1)
  ) |>
    pivot_longer(c(Grupa_A, Grupa_B), names_to = "grupa", values_to = "gustoca")
}) |>
  mutate(panel = factor(panel, levels = d_labels))

d_viz |>
  ggplot(aes(x = x, y = gustoca, fill = grupa)) +
  geom_area(alpha = 0.4, position = "identity") +
  facet_wrap(~panel, ncol = 2) +
  scale_fill_manual(values = c("Grupa_A" = "steelblue", "Grupa_B" = "#2a9d8f"),
                    labels = c("Grupa A", "Grupa B")) +
  labs(
    title = "Što znači Cohenov d?",
    subtitle = "Veći d = manje preklapanja između distribucija = očitija razlika",
    x = "Vrijednost", y = "Gustoća", fill = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



S d = 0.2, distribucije se gotovo potpuno preklapaju — razliku biste teško primijetili u praksi. S d = 0.8, razdvajanje je očito. Naš d oko 1.3 pokazuje vrlo jasno razdvajanje — carousel i single image su očigledno različite kategorije po angažmanu.

> **Uvijek izvještavajte veličinu učinka**
>
> Umjesto "razlika je statistički značajna ($p < 0.001$)", napišite: "carousel objave imaju značajno viši angažman od single image objava (razlika = 2.6 postotnih bodova, d = 1.34, $p < 0.001$)." Ovo daje čitatelju informaciju i o postojanju i o veličini razlike — sve u jednoj rečenici.

11 Statistička snaga: hoće li vaš test uopće nešto naći?

Statistička snaga (power) je vjerojatnost da test odbaci H kad je H istinita — jednostavnije rečeno, vjerojatnost da ćete detektirati pravu razliku ako ona postoji.

Snaga ovisi o četiri faktora. Veličina učinka — veću razliku je lakše detektirati. Veličina uzorka — više podataka daje veću snagu. Razina značajnosti — veći daje veću snagu, ali

i više lažno pozitivnih. Varijabilnost podataka — manja varijabilnost znači čišći signal. To su faktori koji snagu određuju.

Konvencija kaže da snaga treba biti barem 0.80 (80%). To znači da ako razlika postoji, želite je detektirati barem u 8 od 10 pokušaja.

11.1 Koliki uzorak trebam?

Najčešća primjena analize snage je planiranje istraživanja *prije* nego prikupite podatke. Ključno pitanje je — koliki uzorak trebate da biste detektirali očekivanu veličinu učinka s 80% snagom?

```
# power.t.test() za dvouzorački test
# Koliki uzorak trebam za srednji učinak (d = 0.5)?
power.t.test(
  delta = 0.5,       # očekivana razlika u SD jedinicama (Cohenov d)
  sd = 1,            # standardizirano na 1
  sig.level = 0.05,  #
  power = 0.80,      # željena snaga
  type = "two.sample",
  alternative = "two.sided"
)
```

```
     Two-sample t test power calculation

              n = 63.76576
          delta = 0.5
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Za detektiranje srednjeg učinka (d = 0.5) s 80% snagom potrebno je otprilike 64 ispitanika po grupi, ukupno 128. Pogledajmo kako se potreban uzorak mijenja s veličinom učinka.

```
d_values <- c(0.2, 0.3, 0.5, 0.8, 1.0, 1.3)

power_tablica <- map_df(d_values, \(d_val) {
  rez <- power.t.test(delta = d_val, sd = 1, sig.level = 0.05, power = 0.80,
                      type = "two.sample", alternative = "two.sided")
  tibble(
    cohenov_d = d_val,
```

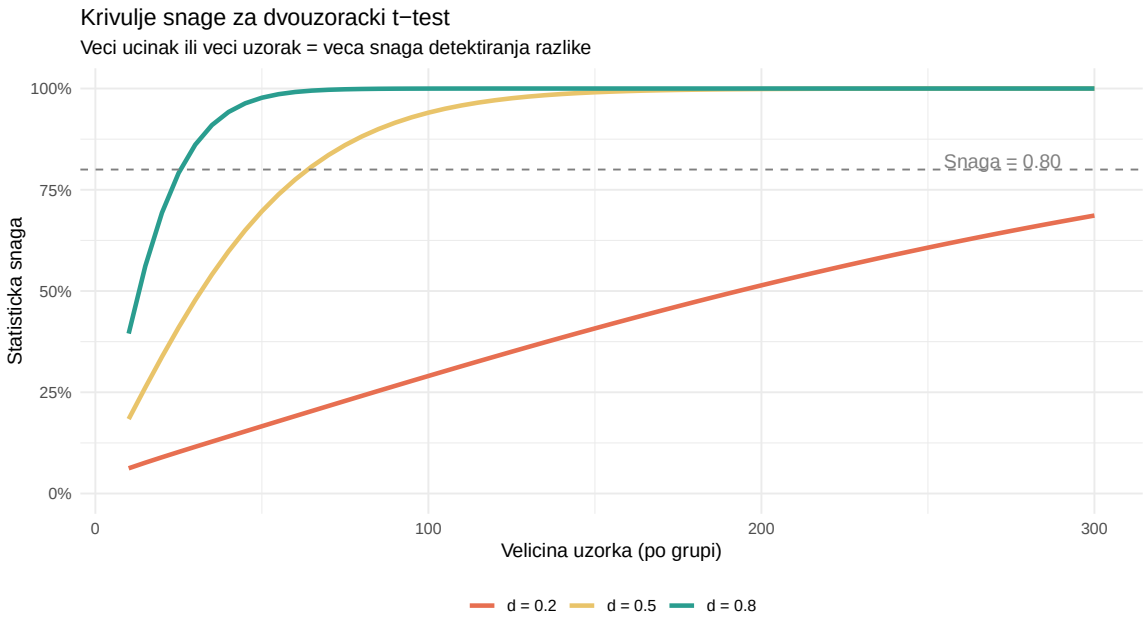
```
    n_po_grupi = ceiling(rez$n),
    ukupno_n = ceiling(rez$n) * 2
  )
})

power_tablica
```

```
# A tibble: 6 x 3
  cohenov_d n_po_grupi ukupno_n
      <dbl>      <dbl>    <dbl>
1       0.2        394      788
2       0.3        176      352
3       0.5         64      128
4       0.8         26       52
5       1           17       34
6       1.3         11       22
```

Brojke su poučne. Za mali učinak (d = 0.2) trebate skoro 400 ispitanika po grupi — ukupno 800. Za veliki učinak (d = 0.8) trebate samo 26 po grupi. Ovo je razlog zašto je planiranje unaprijed ključno — morate imati realistična očekivanja o veličini učinka da biste znali koliko podataka trebate prikupiti.

```
# Krivulja snage: kako snaga raste s veličinom uzorka
n_range <- seq(10, 300, by = 5)

power_curves <- map_df(c(0.2, 0.5, 0.8), \(d_val) {
  map_df(n_range, \(n_val) {
    p <- power.t.test(n = n_val, delta = d_val, sd = 1, sig.level = 0.05,
                      type = "two.sample", alternative = "two.sided")$power
    tibble(n = n_val, power = p, d = paste("d =", d_val))
  })
})

power_curves |>
  ggplot(aes(x = n, y = power, color = d)) +
  geom_line(linewidth = 1.2) +
  geom_hline(yintercept = 0.80, linetype = "dashed", color = "grey50") +
  annotate("text", x = 290, y = 0.82, label = "Snaga = 0.80", color = "grey50", hjust = 1)
  scale_y_continuous(labels = scales::label_percent(), limits = c(0, 1)) +
  scale_color_manual(values = c("d = 0.2" = "#e76f51", "d = 0.5" = "#e9c46a", "d = 0.8" =
  labs(
    title = "Krivulje snage za dvouzorački t-test",
    subtitle = "Veći učinak ili veći uzorak = veća snaga detektiranja razlike",
    x = "Veličina uzorka (po grupi)",
```

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



Za mali učinak (d = 0.2, crvena), snaga sporo raste i ni s 300 ispitanika po grupi ne dostiže 100%. Za veliki učinak (d = 0.8, zelena), snaga brzo raste i s 30 po grupi je već blizu 80%.

11.2 Kolika je snaga našeg Instagram testa?

```
# Kolika je snaga našeg testa s d   1.34 i n   250 po grupi?
power_ig <- power.t.test(
  n = min(n1, n2),
  delta = d,
  sd = 1,
  sig.level = 0.05,
  type = "two.sample",
  alternative = "two.sided"
)

cat("Snaga našeg testa:", round(power_ig$power, 4), "\n")
```

```
Snaga našeg testa: 1
```

Snaga je gotovo 100%. S ovakvom veličinom učinka i ovakvim uzorkom, gotovo je nemoguće da bismo propustili ovu razliku. Test je bio više nego adekvatno snažan — u praksi, mogli smo detektirati ovu razliku s mnogo manje podataka.

```
# Koliki minimalni uzorak bi bio dovoljan?
min_n <- power.t.test(
  delta = d,
  sd = 1,
  sig.level = 0.05,
  power = 0.80,
  type = "two.sample"
)

cat("Minimalni n po grupi za 80% snagu:", ceiling(min_n$n), "\n")
```

```
Minimalni n po grupi za 80% snagu: 10
```

```
cat("Mi smo imali:", min(n1, n2), "po grupi\n")
```

```
Mi smo imali: 236 po grupi
```

12 Upareni t-test: kad iste jedinice mjerite dva puta

Dosad smo uspoređivali dvije nezavisne skupina — carousel objave su jedne, single image su druge, i nema nikakve veze između pojedinačnih objava u dvjema grupama. Ali ponekad mjerite istu jedinicu u dva uvjeta. Na primjer — angažman istih pratitelja prije i poslije redizajna profila, ili ocjene istih članaka od strane dva različita urednika.

Kad su opažanja u parovima, koristite upareni t-test. Umjesto da uspoređujete dva prosjeka, on računa razliku za svaki par i testira je li prosjek tih razlika različit od nule. Ovo je daleko osjetljiviji pristup jer uklanja varijabilnost *između* parova i fokusira se samo na varijabilnost *unutar* parova.

```
set.seed(42)

# Simulacija: 30 članaka, svaki ocjenjen od 2 urednika
urednicki_rating <- tibble(
  clanak_id = 1:30,
  urednik_A = round(rnorm(30, mean = 6.5, sd = 1.2), 1),
  urednik_B = round(urednik_A + rnorm(30, mean = 0.5, sd = 0.8), 1)
)

# Urednik B ocjenjuje sustavno više
```

```r
urednicki_rating <- urednicki_rating |>
  mutate(razlika = urednik_B - urednik_A)

urednicki_rating |>
  summarise(
    M_A = round(mean(urednik_A), 2),
    M_B = round(mean(urednik_B), 2),
    M_razlika = round(mean(razlika), 2),
    SD_razlika = round(sd(razlika), 2)
  )
```

```
# A tibble: 1 x 4
    M_A   M_B M_razlika SD_razlika
  <dbl> <dbl>     <dbl>      <dbl>
1  6.59  6.99      0.41       0.84
```

```r
# Upareni t-test
t.test(urednicki_rating$urednik_B, urednicki_rating$urednik_A, paired = TRUE)
```

```
    Paired t-test

data:  urednicki_rating$urednik_B and urednicki_rating$urednik_A
t = 2.648, df = 29, p-value = 0.01296
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.09257456 0.72075877
sample estimates:
mean difference
      0.4066667
```

Usporedimo što se dogodi kad na istim podacima pokrenemo upareni i neupareni test.

```r
# Usporedba: upareni vs nepareni test na istim podacima
paired_p <- t.test(urednicki_rating$urednik_B, urednicki_rating$urednik_A, paired = TRUE)$
unpaired_p <- t.test(urednicki_rating$urednik_B, urednicki_rating$urednik_A, paired = FALS

cat("Upareni test p-vrijednost:  ", round(paired_p, 5), "\n")
```

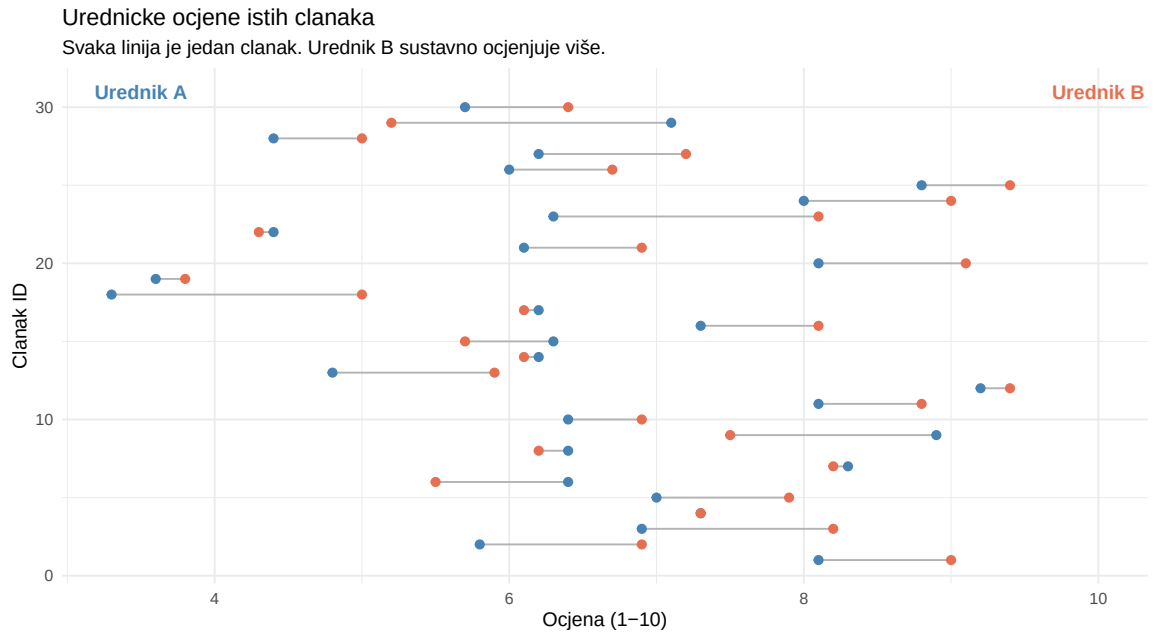
```
Upareni test p-vrijednost:   0.01296
```

```r
cat("Nepareni test p-vrijednost: ", round(unpaired_p, 5), "\n")
```

```
Nepareni test p-vrijednost:  0.30785
```

Upareni test daje manju p-vrijednost jer je osjetljiviji. Zašto? Zato što uklanja varijabilnost između parova. Neki članci su općenito bolji, neki lošiji — to je varijabilnost koja nema veze s razlikom između urednika. Kad tu varijabilnost kontrolirate (parenjem), preostaje samo varijabilnost u razlici između dva urednika, koja je mnogo manja.

```
urednicki_rating |>
  ggplot() +
  geom_segment(aes(x = urednik_A, xend = urednik_B, y = clanak_id, yend = clanak_id),
               color = "grey70", linewidth = 0.5) +
  geom_point(aes(x = urednik_A, y = clanak_id), color = "steelblue", size = 2) +
  geom_point(aes(x = urednik_B, y = clanak_id), color = "#e76f51", size = 2) +
  annotate("text", x = 3.5, y = 31, label = "Urednik A", color = "steelblue", fontface = "
  annotate("text", x = 10, y = 31, label = "Urednik B", color = "#e76f51", fontface = "bol
  labs(
    title = "Uredničke ocjene istih članaka",
    subtitle = "Svaka linija je jedan članak. Urednik B sustavno ocjenjuje više.",
    x = "Ocjena (1-10)",
    y = "Članak ID"
  ) +
  theme_minimal()
```



Svaka vodoravna linija predstavlja jedan članak. Plava točka je ocjena urednika A, crvena urednika B. Većina linija ide udesno, što znači da urednik B dosljedno ocjenjuje više.

❗ Koji test za koju situaciju?

Nezavisni (neupareni) t-test — dvije različite skupina bez veze. Primjeri: muškarci vs žene, kontrolna vs eksperimentalna grupa, carousel vs single image.
Upareni t-test — ista jedinica mjerena dva puta. Primjeri: prije i poslije intervencije, isti sadržaj na dva kanala, isti ispitanik u dva uvjeta. Ključno pitanje: možete li smisleno spariti opažanja? Ako da, koristite upareni test — bit će osjetljiviji.

13 Statistička značajnost nije isto što i praktična važnost

Ovo je možda najvažnija lekcija cijelog predavanja. Statistička značajnost ($p < 0.05$) govori da razlika vjerojatno nije slučajnost. Ali ne govori vam je li ta razlika dovoljno velika da na nju trebate reagirati. Ovo razlikovanje je ključno za svakoga tko donosi poslovne ili istraživačke odluke na temelju podataka.

```
set.seed(42)

# Scenarij 1: Statistički značajno ali praktički beznačajno
# Novi dizajn naslovnice povećava CTR s 2.00% na 2.05%
n_velik <- 50000
ctr_stari <- rbinom(n_velik, 1, 0.0200)
ctr_novi <- rbinom(n_velik, 1, 0.0205)

test1 <- t.test(ctr_novi, ctr_stari)

cat("=== Scenarij 1: Velik uzorak, sićušna razlika ===\n")
```

```
=== Scenarij 1: Velik uzorak, sićušna razlika ===
```

```
cat("Razlika CTR:", round((mean(ctr_novi) - mean(ctr_stari)) * 100, 3), "postotnih bodova\
```

```
Razlika CTR: -0.014 postotnih bodova
```

```
cat("P-vrijednost:", round(test1$p.value, 4), "\n")
```

```
P-vrijednost: 0.8762
```

```
cat("Statistički značajno:", test1$p.value < 0.05, "\n")
```

```
Statistički značajno: FALSE
```

```r
cat("Isplati li se redizajn? Vjerojatno ne.\n\n")
```

```
Isplati li se redizajn? Vjerojatno ne.
```

```r
# Scenarij 2: Statistički neznačajno ali potencijalno praktički važno
# Novi format povećava CTR s 2.0% na 3.5% ali mali uzorak
n_mali <- 80
ctr_stari2 <- rbinom(n_mali, 1, 0.020)
ctr_novi2 <- rbinom(n_mali, 1, 0.035)

test2 <- t.test(ctr_novi2, ctr_stari2)

cat("=== Scenarij 2: Mali uzorak, veća razlika ===\n")
```

```
=== Scenarij 2: Mali uzorak, veća razlika ===
```

```r
cat("Razlika CTR:", round((mean(ctr_novi2) - mean(ctr_stari2)) * 100, 2), "postotnih bodov
```

```
Razlika CTR: 0 postotnih bodova
```

```r
cat("P-vrijednost:", round(test2$p.value, 4), "\n")
```

```
P-vrijednost: 1
```

```r
cat("Statistički značajno:", test2$p.value < 0.05, "\n")
```

```
Statistički značajno: FALSE
```

```r
cat("Zaslužuje li daljnje istraživanje? Vjerojatno da.\n")
```

```
Zaslužuje li daljnje istraživanje? Vjerojatno da.
```

Ova dva scenarija savršeno ilustriraju zašto p-vrijednost sama nije dovoljna. U prvom, razlika od 0.05 postotnih bodova je statistički značajna (jer imate 50 000 opažanja), ali praktički beznačajna — redizajn koji donosi toliko poboljšanje se ne isplati. U drugom, razlika od 1.5 postotnih bodova nije statistički značajna (jer imate samo 80 opažanja), ali je potencijalno vrlo važna — i zaslužuje daljnje istraživanje s većim uzorkom.

Donošenje odluka zahtijeva da razmotrite veličinu učinka, praktične posljedice, interval pouzdanosti i kontekst. Sljedeća tablica sažima četiri moguća scenarija.

```
tribble(
  ~` `, ~`Praktički važno`, ~`Praktički nevažno`,
  "Statistički značajno (p < 0.05)", "  Djeluj! Razlika postoji i važna je.", "  Razlika po
  "Statistički neznačajno (p   0.05)", "  Možda nemaš dovoljno podataka. Povećaj uzorak.",
)
```

```
# A tibble: 2 x 3
  ` `                              `Praktički važno`        `Praktički nevažno`
  <chr>                            <chr>                    <chr>
1 Statistički značajno (p < 0.05)   "  Djeluj! Razlika post~  Razlika postoji ~
2 Statistički neznačajno (p   0.05) "\U0001f50d Možda nemaš~  Nema učinka i to~
```

14 Sve zajedno: izvještaj za urednicu

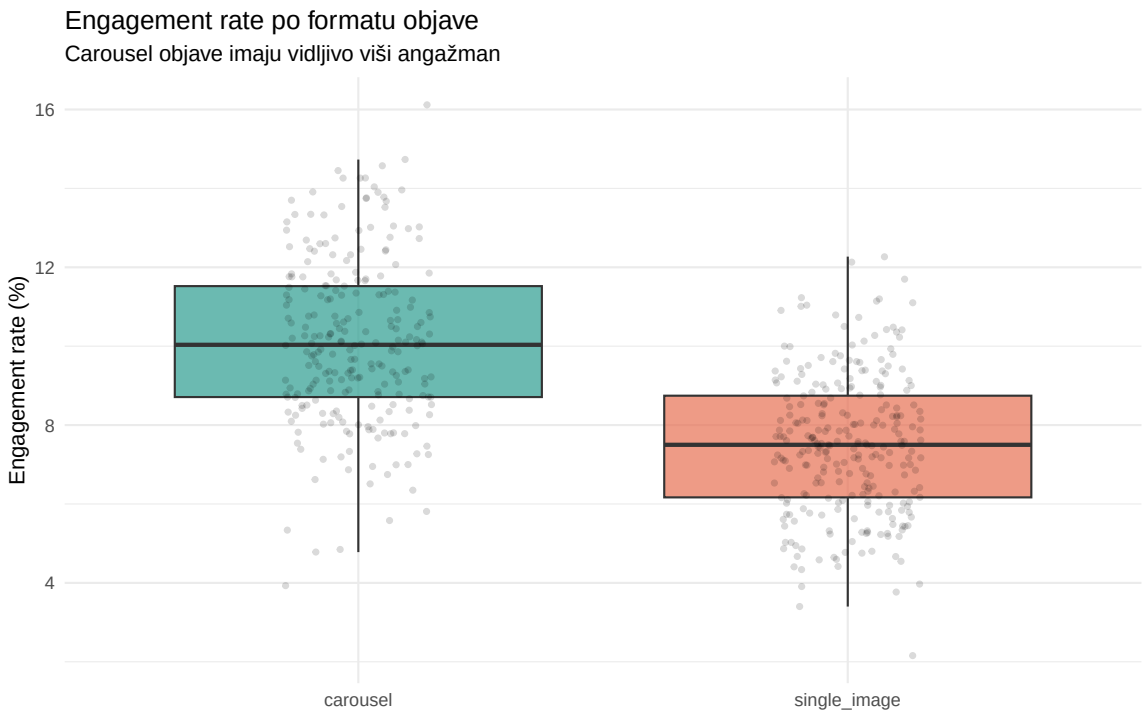
Spojimo sve u koherentan izvještaj. Slijedimo strukturu koju ćete koristiti u svakoj budućoj analizi, gdje su ključni koraci sljedeći — opisna statistika, vizualizacija, statistički test, veličina učinka, podanalize po podgrupama, zaključak s preporukom.

```
# Korak 1: Opisna statistika po formatu
ig |>
  group_by(format) |>
  summarise(
    n = n(),
    M_engagement = round(mean(engagement_rate) * 100, 2),
    SD_engagement = round(sd(engagement_rate) * 100, 2),
    M_likes = round(mean(likes), 0),
    M_comments = round(mean(comments), 0),
    M_shares = round(mean(shares), 0),
    M_saves = round(mean(saves), 0),
    .groups = "drop"
  )
```

```
# A tibble: 2 x 8
  format        n M_engagement SD_engagement M_likes M_comments M_shares M_saves
  <chr>     <int>        <dbl>         <dbl>   <dbl>      <dbl>    <dbl>   <dbl>
1 carousel    236         10.1          2.11     180         36       25      42
2 single_i~   264          7.5          1.77     149         24       20      30
```

```
ig |>
  ggplot(aes(x = format, y = engagement_rate * 100, fill = format)) +
  geom_boxplot(alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.15, alpha = 0.15, size = 1) +
```

```
  scale_fill_manual(values = c("carousel" = "#2a9d8f", "single_image" = "#e76f51")) +
  labs(
    title = "Engagement rate po formatu objave",
    subtitle = "Carousel objave imaju vidljivo viši angažman",
    x = NULL,
    y = "Engagement rate (%)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```




```
# Korak 2: Statistički test
test_ig <- t.test(carousel, single)

# Korak 3: Veličina učinka
s_pooled <- sqrt(((n1 - 1) * sd(carousel)^2 + (n2 - 1) * sd(single)^2) / (n1 + n2 - 2))
d_ig <- (mean(carousel) - mean(single)) / s_pooled

cat("=== REZULTAT DVOUZORAČKOG T-TESTA ===\n")
```

```
=== REZULTAT DVOUZORAČKOG T-TESTA ===
```

```r
cat("t(", round(test_ig$parameter, 1), ") = ", round(test_ig$statistic, 2), "\n", sep = ""
```

```
t(461.6) = 14.94
```

```r
cat("p < 0.001\n")
```

```
p < 0.001
```

```r
cat("Razlika prosjeka: ", round((mean(carousel) - mean(single)) * 100, 2), " postotnih bod
```

```
Razlika prosjeka: 2.62 postotnih bodova
```

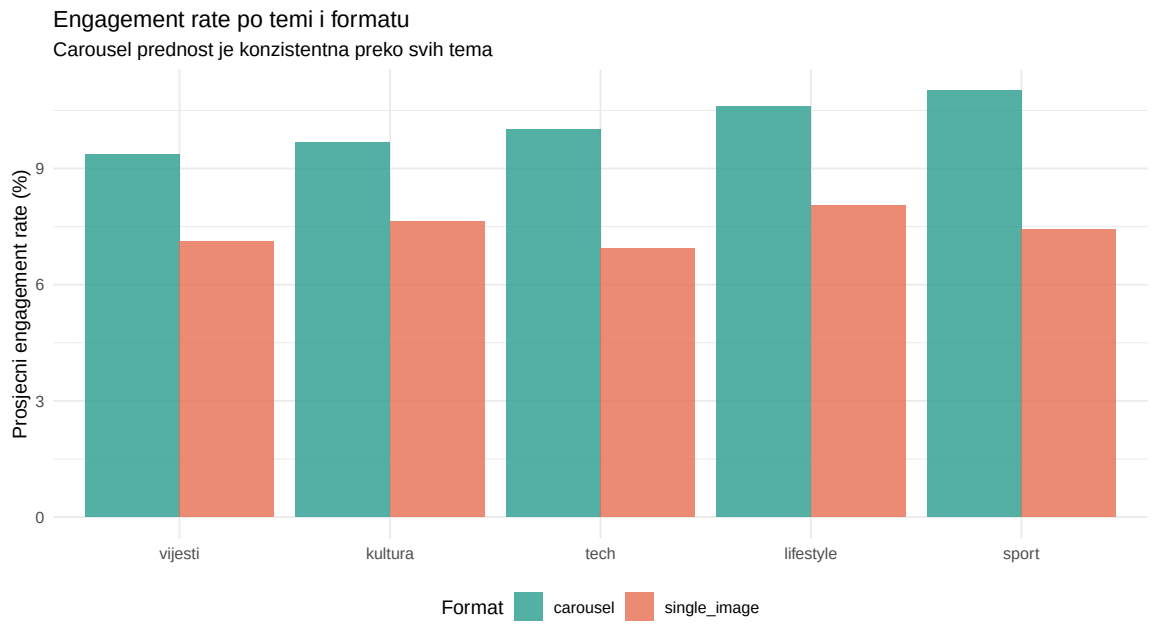
```r
cat("95% CI za razliku: [", round(test_ig$conf.int[1] * 100, 2), ", ",
    round(test_ig$conf.int[2] * 100, 2), "] postotnih bodova\n", sep = "")
```

```
95% CI za razliku: [2.28, 2.97] postotnih bodova
```

```r
cat("Cohenov d:", round(d_ig, 2), "(veliki učinak)\n")
```

```
Cohenov d: 1.35 (veliki učinak)
```

```r
# Korak 4: Je li prednost carousela konzistentna po temama?
ig |>
  group_by(topic, format) |>
  summarise(M = mean(engagement_rate) * 100, .groups = "drop") |>
  ggplot(aes(x = fct_reorder(topic, M, .fun = max), y = M, fill = format)) +
  geom_col(position = "dodge", alpha = 0.8) +
  scale_fill_manual(values = c("carousel" = "#2a9d8f", "single_image" = "#e76f51")) +
  labs(
    title = "Engagement rate po temi i formatu",
    subtitle = "Carousel prednost je konzistentna preko svih tema",
    x = NULL,
    y = "Prosječni engagement rate (%)",
    fill = "Format"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

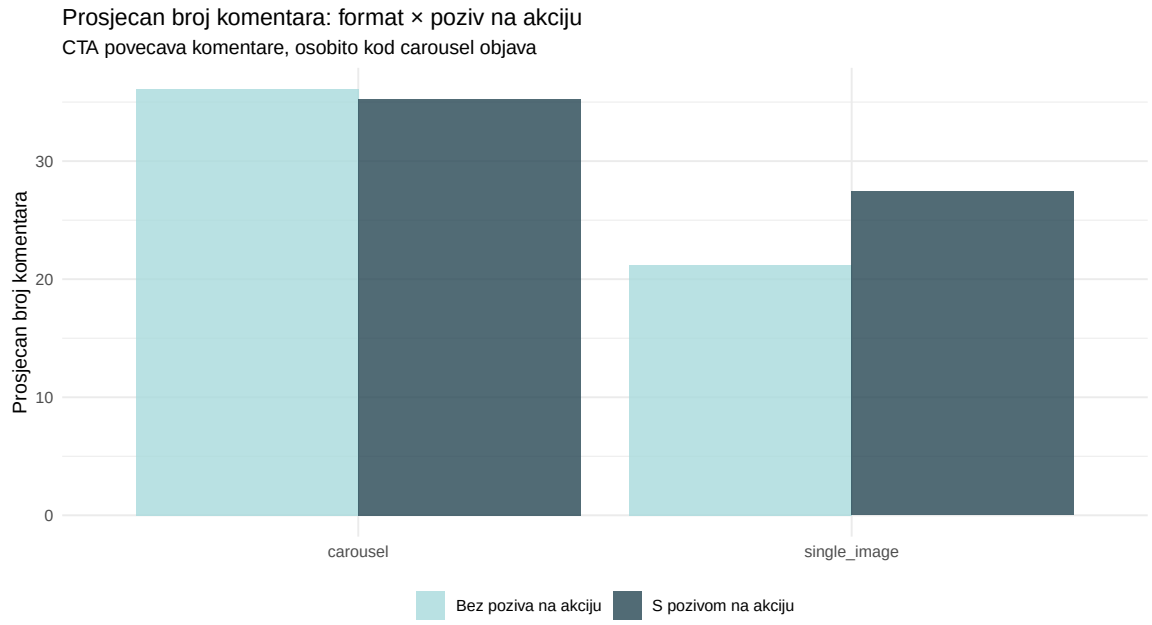



```
# Statistički testovi po temi
ig |>
  group_by(topic) |>
  summarise(
    n_carousel = sum(format == "carousel"),
    n_single = sum(format == "single_image"),
    M_carousel = round(mean(engagement_rate[format == "carousel"]) * 100, 2),
    M_single = round(mean(engagement_rate[format == "single_image"]) * 100, 2),
    razlika = round(M_carousel - M_single, 2),
    p = round(t.test(
      engagement_rate[format == "carousel"],
      engagement_rate[format == "single_image"]
    )$p.value, 4),
    znacajno = p < 0.05,
    .groups = "drop"
  )
```

```
# A tibble: 5 x 8
  topic     n_carousel n_single M_carousel M_single razlika     p znacajno
  <chr>          <int>    <int>      <dbl>    <dbl>   <dbl> <dbl> <lgl>
1 kultura            35       49       9.66     7.64    2.02     0 TRUE
2 lifestyle          61       68      10.6      8.04    2.56     0 TRUE
3 sport              46       53      11        7.42    3.58     0 TRUE
4 tech               25       22      10.0      6.93    3.09     0 TRUE
5 vijesti            69       72       9.37     7.12    2.25     0 TRUE
```

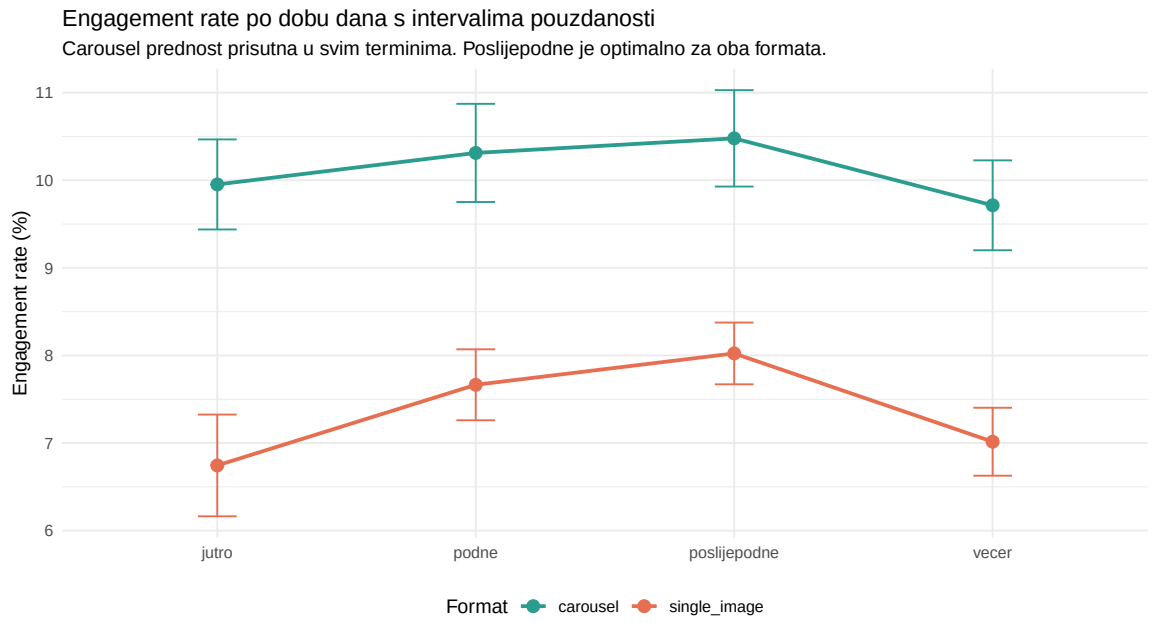
Prednost carousela je statistički značajna za sve teme. Konzistentnost učinka kroz podgrupe pojačava povjerenje u zaključak — ovo nije artefakt jedne specifične teme.

```
# Korak 5: Utjecaj CTA na komentare
ig |>
  group_by(format, has_cta) |>
  summarise(M_comments = mean(comments), .groups = "drop") |>
  mutate(has_cta = if_else(has_cta, "S pozivom na akciju", "Bez poziva na akciju")) |>
  ggplot(aes(x = format, y = M_comments, fill = has_cta)) +
  geom_col(position = "dodge", alpha = 0.8) +
  scale_fill_manual(values = c("S pozivom na akciju" = "#264653", "Bez poziva na akciju" =
  labs(
    title = "Prosječan broj komentara: format × poziv na akciju",
    subtitle = "CTA povećava komentare, osobito kod carousel objava",
    x = NULL,
    y = "Prosječan broj komentara",
    fill = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```




```
# Korak 6: Engagement po dobu dana
ig |>
  mutate(time_of_day = factor(time_of_day, levels = c("jutro", "podne", "poslijepodne", "v
  group_by(time_of_day, format) |>
  summarise(
```

```
    M = mean(engagement_rate) * 100,
    SE = sd(engagement_rate) / sqrt(n()) * 100,
    .groups = "drop"
  ) |>
  ggplot(aes(x = time_of_day, y = M, color = format, group = format)) +
  geom_line(linewidth = 1) +
  geom_point(size = 3) +
  geom_errorbar(aes(ymin = M - 1.96 * SE, ymax = M + 1.96 * SE), width = 0.15) +
  scale_color_manual(values = c("carousel" = "#2a9d8f", "single_image" = "#e76f51")) +
  labs(
    title = "Engagement rate po dobu dana s intervalima pouzdanosti",
    subtitle = "Carousel prednost prisutna u svim terminima. Poslijepodne je optimalno za
    x = NULL,
    y = "Engagement rate (%)",
    color = "Format"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```




```
# Korak 7: Izvještaj za urednicu
cat("                                \n")
```

```
cat("   IZVJEŠTAJ: A/B TEST FORMATA INSTAGRAM OBJAVA\n")
```

```
   IZVJEŠTAJ: A/B TEST FORMATA INSTAGRAM OBJAVA
```

```
cat("                                \n\n")
```

```
cat("UZORAK: ", nrow(ig), " objava (", n1, " carousel, ", n2, " single image)\n\n", sep =
```

```
UZORAK: 500 objava (236 carousel, 264 single image)
```

```
cat("GLAVNI NALAZ:\n")
```

```
GLAVNI NALAZ:
```

```
cat("Carousel objave generiraju statistički značajno viši angažman\n")
```

```
Carousel objave generiraju statistički značajno viši angažman
```

```
cat("od single image objava.\n\n")
```

```
od single image objava.
```

```
cat("BROJKE:\n")
```

```
BROJKE:
```

```
cat("  Carousel engagement:    ", round(mean(carousel) * 100, 2), "% (SD = ",
    round(sd(carousel) * 100, 2), "%)\n", sep = "")
```

```
  Carousel engagement:    10.12% (SD = 2.11%)
```

```
cat("  Single image engagement:", round(mean(single) * 100, 2), "% (SD = ",
    round(sd(single) * 100, 2), "%)\n", sep = "")
```

```
  Single image engagement:7.5% (SD = 1.77%)
```

```
cat("  Razlika: ", round((mean(carousel) - mean(single)) * 100, 2),
    " postotnih bodova\n", sep = "")
```

```
  Razlika: 2.62 postotnih bodova
```

```
cat("  95% CI:  [", round(test_ig$conf.int[1] * 100, 2), ", ",
    round(test_ig$conf.int[2] * 100, 2), "] postotnih bodova\n\n", sep = "")
```

```
  95% CI:  [2.28, 2.97] postotnih bodova
```

```
cat("STATISTIKA:\n")
```

```
STATISTIKA:
```

```
cat("  t(", round(test_ig$parameter, 1), ") = ", round(test_ig$statistic, 2),
    ", p < 0.001\n", sep = "")
```

```
  t(461.6) = 14.94, p < 0.001
```

```
cat("  Cohenov d = ", round(d_ig, 2), " (veliki učinak)\n\n", sep = "")
```

```
  Cohenov d = 1.35 (veliki učinak)
```

```
cat("DODACI:\n")
```

```
DODACI:
```

```
cat("  * Prednost je konzistentna preko svih tema i svih doba dana.\n")
```

```
  * Prednost je konzistentna preko svih tema i svih doba dana.
```

```
cat("  * Poziv na akciju dodatno pojačava komentare (+15%).\n")
```

```
  * Poziv na akciju dodatno pojačava komentare (+15%).
```

```
cat("  * Poslijepodne je optimalno vrijeme za objavu oba formata.\n\n")
```

```
  * Poslijepodne je optimalno vrijeme za objavu oba formata.
```

```
cat("PREPORUKA:\n")
```

```
PREPORUKA:
```

```
cat("  Prebacite što veći udio objava na carousel format,\n")
```

```
  Prebacite što veći udio objava na carousel format,
```

```
cat("  osobito za lifestyle i sportski sadržaj koji i inače\n")
```

```
  osobito za lifestyle i sportski sadržaj koji i inače
```

```
cat("  generiraju najviši angažman. Kombinirajte s pozivom\n")
```

```
  generiraju najviši angažman. Kombinirajte s pozivom
```

```
cat("  na akciju za maksimalne komentare.\n")
```

```
  na akciju za maksimalne komentare.
```

15 ASA izjava i problem višestrukog testiranja

Američka statistička asocijacija (ASA) je 2016. izdala službenu izjavu o p-vrijednostima — prvi put u svojoj 177-godišnjoj povijesti da se oglasila o konkretnom statističkom konceptu. Šest principa iz te izjave vrijedi zapamtiti.

P-vrijednosti mogu pokazati koliko su podaci nekompatibilni sa specificiranim statističkim modelom. Ne mjere vjerojatnost da je hipoteza istinita, niti vjerojatnost da su podaci nastali samo slučajnošću. Znanstveni zaključci ne bi se trebali temeljiti samo na tome prelazi li p-vrijednost specifičan prag. Ispravno zaključivanje zahtijeva puno izvještavanje i transparentnost. P-vrijednost ne mjeri veličinu učinka niti važnost rezultata. I sama p-vrijednost ne pruža dobru mjeru dokaza za ili protiv hipoteze.

15.1 Višestruko testiranje: kad testirate mnogo toga, nešto će “ispasti značajno”

Kad provodite mnogo testova istovremeno, povećava se šansa da barem jedan bude lažno pozitivan — čak i kad nijedan pravi učinak ne postoji.

```r
# Simulacija: 20 testova, SVI pod H  (nema pravih razlika)
set.seed(42)

sim_20_testova <- map_df(1:20, \(i) {
  a <- rnorm(50, mean = 5, sd = 2)
  b <- rnorm(50, mean = 5, sd = 2)  # isti prosjek!
  test <- t.test(a, b)
  tibble(test_broj = i, p = round(test$p.value, 4), znacajno = test$p.value < 0.05)
})

cat("Od 20 testova (svi H  istiniti):\n")
```

```
Od 20 testova (svi H  istiniti):
```

```r
cat("Statistički značajnih:", sum(sim_20_testova$znacajno), "\n\n")
```

```
Statistički značajnih: 1
```

```r
sim_20_testova |> filter(znacajno)
```

```
# A tibble: 1 x 3
  test_broj      p znacajno
      <int>  <dbl> <lgl>
1         9 0.0377 TRUE
```

Čak i kad nijedan učinak ne postoji, jedan ili više testova ispada "statistički značajan." Kad biste izvijestili samo te značajne rezultate i prešutjeli ostalih 18 ili 19 testova, to bi bila obmana. Postoje korekcije za ovaj problem — najjednostavnija je Bonferronijeva, koja dijeli s brojem testova.

```r
# Bonferronijeva korekcija
sim_20_testova |>
  mutate(
    p_korigirana = p.adjust(p, method = "bonferroni"),
    znacajno_korigirano = p_korigirana < 0.05
  ) |>
  filter(znacajno | znacajno_korigirano) |>
  select(test_broj, p, znacajno, p_korigirana, znacajno_korigirano)
```

```
# A tibble: 1 x 5
  test_broj      p znacajno p_korigirana znacajno_korigirano
      <int>  <dbl> <lgl>           <dbl> <lgl>
1         9 0.0377 TRUE            0.754 FALSE
```

Nakon Bonferronijeve korekcije, nijedan test nije značajan. Korekcija je konzervativna (može propustiti prave učinke), ali štiti od lažno pozitivnih kad provodite mnogo testova. Benjamini-Hochberg (BH) korekcija je manje konzervativna alternativa koju ćete često sresti u literaturi.

⚠ P-hacking: ono što ne smijete raditi

Ako u istraživanju testirate mnogo varijabli i izvijestite samo one koje su značajne — to se zove p-hacking ili cherry-picking. Rezultati dobiveni na taj način nisu pouzdani jer ne uzimaju u obzir višestruko testiranje. Uvijek izvijestite koliko ste testova proveli, ne samo one koji su dali $p < 0.05$. Transparentnost nije opcija, nego temelj pouzdane znanosti.

16 Pregled svih t-testova

```
tribble(
  ~test, ~situacija, ~R_kod, ~primjer,
  "Jednouzorački", "Jedan uzorak vs poznata vrijednost", "t.test(x, mu = 5)", "Je li prosj
  "Dvouzorački (nezavisni)", "Dvije nezavisne skupine", "t.test(x, y)", "Carousel vs singl
  "Upareni", "Iste jedinice, dva mjerenja", "t.test(x, y, paired = TRUE)", "Ocjene istih č
)
```

```
# A tibble: 3 x 4
  test                   situacija                          R_kod       primjer
  <chr>                  <chr>                              <chr>       <chr>
1 Jednouzorački          Jedan uzorak vs poznata vrijednost t.test(x, ~ Je li ~
2 Dvouzorački (nezavisni) Dvije nezavisne skupine           t.test(x, ~ Carous~
3 Upareni                Iste jedinice, dva mjerenja        t.test(x, ~ Ocjene~
```

Sva tri testa dijele istu logiku — postavljate H , računate t-statistiku, gledate p-vrijednost i donosite odluku. Razlikuju se u formulaciji H i načinu izračuna standardne pogreške. Funkcija `t.test()` pokriva sva tri slučaja.

❗ Ključni zaključci

Testiranje hipoteza počinje od pretpostavke da nema učinka. Postavljate nultu hipotezu (H : nema razlike) i tražite dokaze protiv nje. Ako su podaci dovoljno neobični pod H (p <), odbacujete H .
Testna statistika mjeri neobičnost podataka pod H . Za t-test: t = razlika / SE. Veći |t| znači jači dokaz protiv H .
P-vrijednost je vjerojatnost podataka pod H , ne vjerojatnost hipoteze. Nije vjerojatnost da je H istinita. Nije mjera veličine učinka. Mala p-vrijednost znači da su

podaci neobični u svijetu gdje H vrijedi.
`t.test()` pokriva sve tri varijante. Jednouzorački (`mu = vrijednost`), dvouzorački (dva vektora) i upareni (`paired = TRUE`). Welchov test (default) ne pretpostavlja jednake varijance i uvijek je dobar izbor.
Greška tipa I je lažno pozitivni rezultat ( ). Greška tipa II je propuštena prava razlika ( ). Snaga = 1 - trebala bi biti barem 0.80.
Cohenov d stavlja razliku u kontekst varijabilnosti. d = 0.2 mali, 0.5 srednji, 0.8 veliki učinak. Uvijek ga izvijestite uz p-vrijednost.
Planirajte uzorak unaprijed. `power.t.test()` računa koliko podataka trebate za zadanu snagu i veličinu učinka. Ovo radite *prije* prikupljanja podataka.
Upareni t-test je osjetljiviji od nezavisnog jer kontrolira varijabilnost između parova. Koristite ga kad iste jedinice mjerite dva puta.
Statistička značajnost nije praktična važnost. Velik uzorak može detektirati trivijalne razlike. Mali uzorak može propustiti važne. Uvijek razmotrite i veličinu učinka i kontekst.
Višestruko testiranje zahtijeva korekciju. Ako provodite mnogo testova, koristite Bonferroni ili BH korekciju — ili barem transparentno izvijestite koliko ste testova proveli.
"Ne možemo odbaciti H " nije isto što i "H je istinita." Odsutnost dokaza nije dokaz odsutnosti. Možda samo nemate dovoljno podataka.

17 Zadaci za pripremu

1. Učitajte `instagram_ab_test.csv`. Testirajte razlikuje li se prosječan broj `saves` između carousel i single image objava. Izračunajte Cohenov d i interpretirajte ga.
2. Odredite minimalnu veličinu uzorka po grupi potrebnu za detektiranje srednjeg učinka (d = 0.5) s 90% snagom na razini = 0.01.
3. Simulirajte 1000 t-testova gdje su oba uzorka iz iste distribucije (H istinita). Koliki postotak p-vrijednosti je ispod 0.05? Nacrtajte histogram p-vrijednosti i usporedite s uniformnom distribucijom.

18 Dodatno čitanje

Obavezno

Navarro, D. (2018). *Learning Statistics with R*, Chapter 11 (Hypothesis Testing). Besplatno dostupno na learningstatisticswithr.com. Pokriva logiku testiranja hipoteza, t-test i p-vrijednost s R kodom.

Preporučeno

Diez, D., Çetinkaya-Rundel, M., & Barr, C. (2019). *OpenIntro Statistics* (4th edition), Chapter 7. Besplatno dostupno na openintro.org/book/os. Jasne vizualizacije logike testiranja.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2), 129-133. Službena izjava o pravilnoj upotrebi i interpretaciji p-vrijednosti.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, 112(1), 155-159. Klasičan kratki pregled veličina učinka i analize snage.

19 Pojmovnik

Pojam	Objašnjenje
Nulta hipoteza (H)	Početna pretpostavka da nema učinka ili razlike. Sadrži znak jednakosti.
Alternativna hipoteza (H)	Tvrdnja da učinak ili razlika postoji. Sadrži znak nejednakosti.
Testna statistika	Broj koji mjeri koliko su podaci neobični pod H . Za t-test: t = razlika / SE.
P-vrijednost	Vjerojatnost podataka (ili ekstremnijih) pod pretpostavkom da je H istinita.
Razina značajnosti ( )	Prag za odbacivanje H . Obično 0.05. Kontrolira grešku tipa I.
Greška tipa I	Odbacimo H kad je istinita. Lažno pozitivni rezultat. Vjerojatnost = .
Greška tipa II	Ne odbacimo H kad je lažna. Propušteni pravi učinak. Vjerojatnost = .
Statistička snaga	Vjerojatnost detektiranja pravog učinka. Snaga = 1 − . Cilj 0.80.
Cohenov d	Standardizirana mjera veličine učinka: razlika prosjeka / pooled SD. d = 0.2 mali, 0.5 srednji, 0.8 veliki.
Pooled SD	Zajednička standardna devijacija dviju grupa, ponderirana njihovim veličinama.
Jednouzorački t-test	Usporedba jednog prosjeka s poznatom vrijednošću. `t.test(x, mu = ...)`.
Dvouzorački t-test	Usporedba prosjeka dviju nezavisnih skupina. `t.test(x, y)`.
Welchov t-test	Default u R-u. Ne pretpostavlja jednake varijance. Robusniji od Studentovog.
Upareni t-test	Usporedba parova (iste jedinice, dva mjerenja). `t.test(x, y, paired = TRUE)`.
Dvosmjerni test	H : . Testira razliku u oba smjera. Default u R-u.

Pojam	Objašnjenje
Jednosmjerni test	H : $>$ ili $<$. Osjetljiviji u jednom smjeru ali slijep za drugi.
Višestruko testiranje	Provođenje mnogo testova istovremeno. Inflacionira grešku tipa I.
Bonferronijeva korekcija	Dijeljenje s brojem testova. Konzervativna ali jednostavna korekcija.
P-hacking	Selektivno izvještavanje značajnih rezultata iz mnogo provedenih testova. Neprihvatljiva praksa.
`power.t.test()`	R funkcija za analizu snage: izračun potrebnog n, snage ili detektabilnog učinka.
`p.adjust()`	R funkcija za korekciju p-vrijednosti za višestruko testiranje (Bonferroni, BH, itd.).