Tjedan 7: Uvod u vjerojatnost

Slučajnost, distribucije i zašto ništa nije sigurno

2025-04-05

Table of contents

```
library(tidyverse)
```

Ishodi učenja

Nakon ovog predavanja moći ćete

1. Objasniti što je vjerojatnost i zašto je temelj sve statističke analize.
2. Opisati razliku između frekvencijskog i bayesijanskog pristupa vjerojatnosti.
3. Primijeniti osnovna pravila vjerojatnosti, uključujući komplementarno pravilo, pravilo zbrajanja i pravilo množenja.
4. Izračunati vjerojatnost nezavisnih i zavisnih događaja.
5. Objasniti što je binomna distribucija i prepoznati situacije u kojima se primjenjuje.
6. Koristiti `dbinom()`, `pbinom()` i `rbinom()` za izračun i simulaciju binomnih vjerojatnosti.
7. Vizualizirati distribucije vjerojatnosti u ggplot2.
8. Povezati koncept vjerojatnosti s praktičnim pitanjima iz komunikologije (viralnost sadržaja, stopa otvaranja emaila, konverzija).

1 Zašto vjerojatnost?

Do sada smo se bavili opisivanjem podataka koji postoje. Izračunali smo prosjeke, napravili grafove, očistili neuredne datasete. Ali statistika ne služi samo za opisivanje onoga što znamo.

Služi i za donošenje zaključaka o onome što ne znamo. A za to nam treba vjerojatnost.

Zamislite da radite A/B test naslova na portalu. Varijanta A ima click-through rate (CTR) od 4.2%, varijanta B ima 4.8%. Je li B zaista bolja ili je razlika samo slučajnost? Odgovor na to pitanje zahtijeva razumijevanje vjerojatnosti. Kolika je vjerojatnost da bismo vidjeli ovakvu ili veću razliku čistom slučajnošću, čak i da su naslovi jednako dobri? Ako je ta vjerojatnost mala, zaključujemo da B vjerojatno zaista jest bolji. Ako je velika, zaključujemo da nemamo dovoljno dokaza.

Ovo je logika koja stoji iza svakog statističkog testa koji ćemo učiti u nastavku kolegija. Govorimo o t-testovima, hi-kvadrat testovima, ANOVA-i i regresiji. Svi oni koriste vjerojatnost kao temelj za donošenje zaključaka. Bez razumijevanja vjerojatnosti, ti testovi su crne kutije u koje ubacujete brojeve i dobivate misterioznu p-vrijednost. S razumijevanjem vjerojatnosti, ti testovi postaju logični alati s jasnom interpretacijom.

Ovo je možda najkonceptualnije predavanje na kolegiju. Nema mnogo koda, nema čišćenja podataka, nema dugačkih pipeline. Umjesto toga, gradimo intuiciju o slučajnosti i distribucijama koja će nam služiti kroz ostatak kolegija.

---

2 Naši podaci: objave na društvenim mrežama

Za ilustraciju vjerojatnosnih koncepata koristimo dataset od 2000 objava na društvenim mrežama. Za svaku objavu imamo platformu, tip sadržaja, broj pratitelja, lajkova, dijeljenja, komentara i oznaku je li objava postala viralna.

```
posts <- read_csv("../resources/datasets/social_posts.csv")
glimpse(posts)
```

```
Rows: 2,000
Columns: 11
$ post_id      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
$ platform     <chr> "Instagram", "Instagram", "Instagram", "TikTok", "Twitter~
$ content_type <chr> "slika", "carousel", "story", "reel", "video", "tekst", "~
$ followers    <dbl> 2398, 19468, 1386, 1221, 26918, 9229, 1603, 17998, 4187, ~
$ likes        <dbl> 202, 255, 75, 55, 0, 44, 132, 721, 439, 37, 3018, 37, 217~
$ shares       <dbl> 7, 20, 9, 13, 0, 7, 20, 28, 29, 4, 431, 7, 10, 89, 45, 6,~
$ comments     <dbl> 11, 8, 3, 3, 0, 4, 0, 9, 46, 1, 22, 0, 15, 118, 8, 5, 0, ~
$ hashtags     <dbl> 14, 10, 6, 4, 0, 15, 3, 14, 16, 10, 9, 2, 15, 4, 8, 1, 3,~
$ post_hour    <dbl> 11, 8, 18, 12, 11, 23, 11, 7, 17, 20, 19, 13, 9, 14, 16, ~
$ day_of_week  <chr> "utorak", "subota", "srijeda", "ponedjeljak", "utorak", "~
$ is_viral     <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
```

```r
posts |>
  count(platform, sort = TRUE)
```

```
# A tibble: 6 x 2
  platform      n
  <chr>     <int>
1 Instagram   530
2 TikTok      505
3 Facebook    361
4 Twitter/X   247
5 YouTube     230
6 LinkedIn    127
```

```r
# Koliko je objava viralno?
posts |>
  count(is_viral) |>
  mutate(udio = round(n / sum(n), 3))
```

```
# A tibble: 2 x 3
  is_viral     n  udio
  <lgl>    <int> <dbl>
1 FALSE     1972 0.986
2 TRUE        28 0.014
```

Od 2000 objava, samo mali postotak je viralan. Ovo nam daje savršen kontekst za razmišljanje o vjerojatnosti. Kolika je šansa da objava postane viralna? Ovisi li to o platformi? O tipu sadržaja? O broju pratitelja? Ova pitanja ćemo istraživati kroz predavanje.

---

3 Što je vjerojatnost?

Intuitivan odgovor je da je vjerojatnost broj koji izražava koliko je nešto izvjesno. Ako kažemo da je vjerojatnost kiše sutra 70%, to znači da smo prilično sigurni da će padati, ali ne potpuno. Ako kažemo da je vjerojatnost da novčić padne na glavu 50%, to znači da su oba ishoda jednako vjerovatna.

Formalno, vjerojatnost je broj između 0 i 1. Vrijednost 0 znači da se događaj sigurno neće dogoditi. Vrijednost 1 znači da će se sigurno dogoditi. Sve između izražava stupanj neizvjesnosti.

$$P(\text{događaj}) \in [0, 1]$$

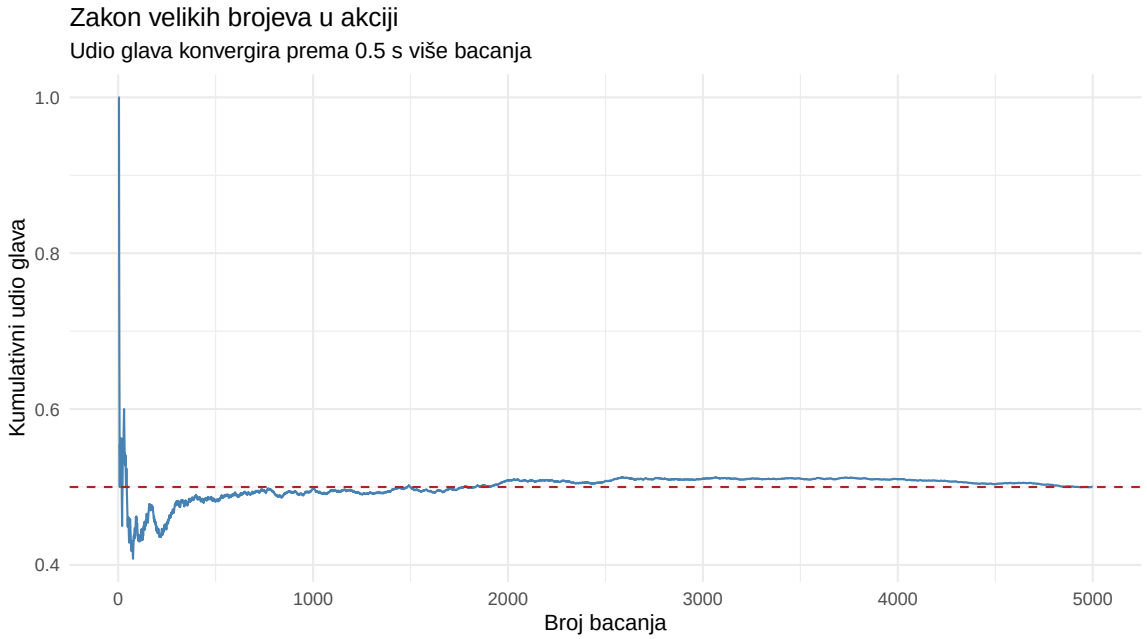
Ponekad se vjerojatnost izražava kao postotak (0% do 100%), ali u statistici i R-u koristimo razlomke (0 do 1).

3.1 Frekvencijski pristup

Frekvencijski (ili klasični) pristup definira vjerojatnost kao dugoročnu relativnu frekvenciju. Ako bacite novčić 10 000 puta, otprilike 5 000 puta će pasti na glavu. Omjer 5000/10000 = 0.5 je vjerojatnost.

Pokažimo to simulacijom.

```
set.seed(42)

# Simulacija bacanja novčića
n_bacanja <- 5000
bacanja <- sample(c("glava", "pismo"), size = n_bacanja, replace = TRUE)

# Kumulativni udio glava nakon svakog bacanja
kum_udio <- cumsum(bacanja == "glava") / seq_along(bacanja)

tibble(bacanje = 1:n_bacanja, udio_glava = kum_udio) |>
  ggplot(aes(x = bacanje, y = udio_glava)) +
  geom_line(color = "steelblue") +
  geom_hline(yintercept = 0.5, linetype = "dashed", color = "firebrick") +
  labs(
    title = "Zakon velikih brojeva u akciji",
    subtitle = "Udio glava konvergira prema 0.5 s više bacanja",
    x = "Broj bacanja",
    y = "Kumulativni udio glava"
  ) +
  theme_minimal()
```



Na početku, udio skače gore-dolje jer je uzorak mali. Ali s više bacanja, udio se stabilizira oko 0.5. Ovo je **zakon velikih brojeva**. S dovoljno ponavljanja, relativna frekvencija konvergira prema pravoj vjerojatnosti.

Funkcija `set.seed(42)` fiksira generator slučajnih brojeva da bismo svaki put dobili iste rezultate. Bez nje, svako pokretanje koda bi dalo malo drugačiji graf. U ponovljivoj analizi, uvijek postavljamo seed.

3.2 Bayesijanski pristup (kratki osvrt)

Bayesijanski pristup definira vjerojatnost kao stupanj uvjerenja. Umjesto da govori o dugoročnoj frekvenciji, bayesijanska statistika kaže da na temelju onoga što znamo, naša uvjerenost da će se X dogoditi je Y.

Ova razlika je filozofska i nema praktične posljedice za većinu analiza koje ćemo raditi. Frekvencijski pristup je dominantan u komunikologiji i društvenim znanostima, pa ćemo ga koristiti. Ali vrijedi znati da postoji alternativni pristup, posebno jer bayesijanska statistika postaje sve popularnija u istraživanjima.

Na ovom kolegiju koristimo frekvencijski pristup. U tjednu 15 ćemo se kratko osvrnuti na bayesijanski kao pogled naprijed.

---

4 Osnovna pravila vjerojatnosti

Postoji nekoliko temeljnih pravila koja vrijede za sve vjerojatnosti. Naučimo ih na primjerima iz našeg dataseta.

4.1 Komplementarno pravilo

Ako je P(A) vjerojatnost da se događaj A dogodi, tada je vjerojatnost da se A ne dogodi jednaka 1 - P(A).

$$P(\text{nije A}) = 1 - P(\text{A})$$

```
# Vjerojatnost da je objava viralna
p_viral <- mean(posts$is_viral)
cat("P(viralno) =", round(p_viral, 3), "\n")
```

```
P(viralno) = 0.014
```

```
# Vjerojatnost da NIJE viralna (komplement)
p_ne_viral <- 1 - p_viral
cat("P(nije viralno) =", round(p_ne_viral, 3), "\n")
```

```
P(nije viralno) = 0.986
```

Ovo zvuči trivijalno, ali komplementarno pravilo je izuzetno korisno u praksi. Ponekad je lakše izračunati vjerojatnost da se nešto ne dogodi pa oduzeti od 1. Na primjer, ako želimo znati vjerojatnost da barem jedna od 10 objava postane viralna, lakše je izračunati vjerojatnost da nijedna ne postane viralna i oduzeti od 1.

4.2 Pravilo zbrajanja (ILI)

Vjerojatnost da se dogodi A ILI B ovisi o tome jesu li događaji međusobno isključivi.

Ako su **međusobno isključivi** (ne mogu se dogoditi istovremeno), jednostavno zbrajamo.

$$P(\text{A ili B}) = P(\text{A}) + P(\text{B})$$

```
# Vjerojatnost da je objava na Instagramu ILI TikToku
# (svaka objava je samo na jednoj platformi, pa su isključivi)
p_ig <- mean(posts$platform == "Instagram")
p_tt <- mean(posts$platform == "TikTok")

cat("P(Instagram) =", round(p_ig, 3), "\n")
```

```
P(Instagram) = 0.265
```

```r
cat("P(TikTok) =", round(p_tt, 3), "\n")
```

```
P(TikTok) = 0.252
```

```r
cat("P(Instagram ILI TikTok) =", round(p_ig + p_tt, 3), "\n")
```

```
P(Instagram ILI TikTok) = 0.518
```

```r
# Provjera
mean(posts$platform %in% c("Instagram", "TikTok"))
```

```
[1] 0.5175
```

Ako **nisu međusobno isključivi** (mogu se dogoditi istovremeno), moramo oduzeti presjek jer ga inače računamo dvaput.

$$P(\text{A ili B}) = P(\text{A}) + P(\text{B}) - P(\text{A i B})$$

```r
# Vjerojatnost da je objava video ILI viralna
# (objava može biti oboje istovremeno)
p_video <- mean(posts$content_type == "video")
p_viral <- mean(posts$is_viral)
p_video_i_viral <- mean(posts$content_type == "video" & posts$is_viral)

cat("P(video) =", round(p_video, 3), "\n")
```

```
P(video) = 0.224
```

```r
cat("P(viralno) =", round(p_viral, 3), "\n")
```

```
P(viralno) = 0.014
```

```r
cat("P(video I viralno) =", round(p_video_i_viral, 4), "\n")
```

```
P(video I viralno) = 0.009
```

```r
cat("P(video ILI viralno) =", round(p_video + p_viral - p_video_i_viral, 3), "\n")
```

```
P(video ILI viralno) = 0.23
```

```r
# Provjera
mean(posts$content_type == "video" | posts$is_viral)
```

```
[1] 0.2295
```

4.3 Pravilo množenja (I)

Vjerojatnost da se dogode i A i B ovisi o tome jesu li događaji nezavisni.

Ako su **nezavisni** (jedan ne utječe na drugi), koristimo formulu

$$P(\text{A i B}) = P(\text{A}) \times P(\text{B})$$

```r
# Jesu li platforma i viralnost nezavisni?
# Ako jesu, P(Instagram I viralno) = P(Instagram) * P(viralno)

p_ig <- mean(posts$platform == "Instagram")
p_viral <- mean(posts$is_viral)

cat("P(Instagram) * P(viralno) =", round(p_ig * p_viral, 4), "\n")
```

```
P(Instagram) * P(viralno) = 0.0037
```

```r
# Stvarna zajednička vjerojatnost
p_ig_viral <- mean(posts$platform == "Instagram" & posts$is_viral)
cat("P(Instagram I viralno) stvarno =", round(p_ig_viral, 4), "\n")
```

```
P(Instagram I viralno) stvarno = 0.002
```

Ako se ove dvije vrijednosti razlikuju, događaji nisu potpuno nezavisni. To znači da platforma utječe na vjerojatnost viralnosti (ili obrnuto). Ovo je važan konceptualni most. Statistički testovi koje ćemo učiti u nastavku kolegija upravo testiraju je li neka razlika rezultat zavisnosti ili čiste slučajnosti.

4.4 Uvjetna vjerojatnost

Uvjetna vjerojatnost je vjerojatnost jednog događaja DADO da se drugi već dogodio. Piše se P(A|B) i čita "vjerojatnost A dado B".

$$P(\text{A}|\text{B}) = \frac{P(\text{A i B})}{P(\text{B})}$$

```
# Vjerojatnost viralnosti DADO da je objava na TikToku
p_viral_dado_tt <- posts |>
  filter(platform == "TikTok") |>
  summarise(p = mean(is_viral)) |>
  pull(p)

cat("P(viralno | TikTok) =", round(p_viral_dado_tt, 4), "\n")
```

```
P(viralno | TikTok) = 0.0356
```

```
# Usporedba s ukupnom vjerojatnošću viralnosti
cat("P(viralno) ukupno =", round(mean(posts$is_viral), 4), "\n")
```

```
P(viralno) ukupno = 0.014
```

```
# Uvjetne vjerojatnosti po platformi
posts |>
  group_by(platform) |>
  summarise(
    n = n(),
    n_viral = sum(is_viral),
    p_viral = round(mean(is_viral), 4),
    .groups = "drop"
  ) |>
  arrange(desc(p_viral))
```

```
# A tibble: 6 x 4
  platform      n n_viral p_viral
  <chr>     <int>   <int>   <dbl>
1 TikTok      505      18  0.0356
2 YouTube     230       4  0.0174
3 Instagram   530       4  0.0075
4 Facebook    361       2  0.0055
5 LinkedIn    127       0  0
6 Twitter/X   247       0  0
```

Ako je P(viralno | TikTok) različit od P(viralno), to znači da platforma i viralnost nisu nezavisni. Ovo je temelj za testove koje ćemo raditi u tjednu 11 (hi-kvadrat test) gdje ćemo formalno testirati jesu li kategoričke varijable nezavisne.

Praktični savjet

U komunikologiji, uvjetna vjerojatnost je sveprisutna. Kolika je vjerojatnost konverzije DADO da je korisnik kliknuo na oglas? Kolika je vjerojatnost otvaranja emaila DADO da je poslan utorkom ujutro? Kolika je vjerojatnost dijeljenja DADO da je sadržaj video? Kad god analizirate performanse po segmentima, zapravo računate uvjetne vjerojatnosti.

---

5 Distribucije vjerojatnosti: od podataka do modela

Do sada smo računali vjerojatnosti iz stvarnih podataka (empirijske vjerojatnosti). Ali u statistici koristimo i **teorijske distribucije** koje opisuju kakvi bi podaci trebali izgledati pod određenim pretpostavkama. Dvije najvažnije su binomna i normalna distribucija.

Zašto nam trebaju teorijske distribucije? Zato što nam omogućuju izračun vjerojatnosti za događaje koje nismo opazili. Iz naših podataka možemo izračunati da je 1.4% objava viralno. Ali što ako želimo znati kolika je vjerojatnost da od sljedećih 100 objava točno 5 bude viralno? Ili kolika je vjerojatnost da nijedna ne bude viralna? Za te izračune koristimo distribuciju vjerojatnosti.

---

6 Binomna distribucija

Binomna distribucija opisuje broj uspjeha u fiksnom broju nezavisnih pokušaja, gdje svaki pokušaj ima istu vjerojatnost uspjeha. Ovo je jedna od najvažnijih distribucija u statistici jer modelira mnogo realnih situacija.

Zamislite da imate 20 objava na Instagramu i svaka ima istu vjerojatnost od 2% da postane viralna (nezavisno jedna od druge). Koliko ćete viralnih objava imati? Možda 0. Možda 1. Možda 2. Teorijski čak i svih 20, ali to je izuzetno malo vjerovatno. Binomna distribucija nam daje točnu vjerojatnost za svaki od tih ishoda.

6.1 Parametri binomne distribucije

Binomna distribucija ima dva parametra.

n je broj pokušaja (u našem primjeru, 20 objava). **p** je vjerojatnost uspjeha u jednom pokušaju (u našem primjeru, 0.02 ili 2%).

Piše se X ~ Binomial(n, p) i čita "X slijedi binomnu distribuciju s n pokušaja i vjerojatnošću p".

6.2 Izračun u R-u: dbinom()

Funkcija `dbinom(x, size, prob)` daje točnu vjerojatnost da dobijemo točno x uspjeha od size pokušaja s vjerojatnošću prob.

```
# Vjerojatnost da NIJEDNA od 20 objava ne postane viralna
dbinom(x = 0, size = 20, prob = 0.02)
```

```
[1] 0.667608
```

```
# Vjerojatnost da TOČNO 1 od 20 postane viralna
dbinom(x = 1, size = 20, prob = 0.02)
```

```
[1] 0.272493
```

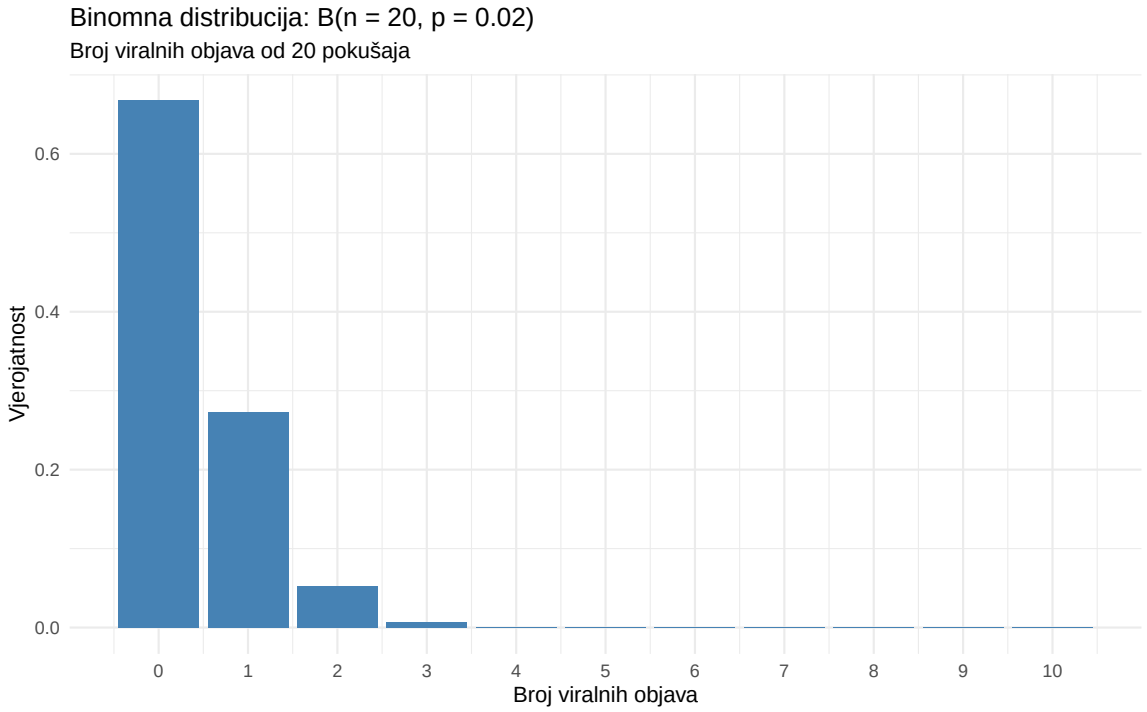
```
# Vjerojatnost da TOČNO 2 od 20 postanu viralne
dbinom(x = 2, size = 20, prob = 0.02)
```

```
[1] 0.05283029
```

Šansa da nijedna objava ne postane viralna je oko 67%. Šansa za točno jednu viralnu je oko 27%. Šansa za točno dvije je oko 5%. Brzo pada. Ovo ima smisla jer kad je vjerojatnost uspjeha samo 2%, većinu vremena nećete imati nijedan uspjeh u 20 pokušaja.

6.3 Vizualizacija binomne distribucije

```r
tibble(
  x = 0:10,
  vjerojatnost = dbinom(x, size = 20, prob = 0.02)
) |>
  ggplot(aes(x = x, y = vjerojatnost)) +
  geom_col(fill = "steelblue") +
  scale_x_continuous(breaks = 0:10) +
  labs(
    title = "Binomna distribucija: B(n = 20, p = 0.02)",
    subtitle = "Broj viralnih objava od 20 pokušaja",
    x = "Broj viralnih objava",
    y = "Vjerojatnost"
  ) +
  theme_minimal()
```



Graf jasno pokazuje da je najvjerojatniji ishod 0 viralnih objava, zatim 1, zatim 2. Ishodi s 3 ili više su tako malo vjerovatni da ih jedva vidimo.

6.4 Kumulativna vjerojatnost: pbinom()

Funkcija `pbinom(q, size, prob)` daje kumulativnu vjerojatnost, odnosno P(X q). To je vjerojatnost da dobijemo q ili manje uspjeha.

```r
# P(X <= 1): vjerojatnost 0 ili 1 viralne objave od 20
pbinom(q = 1, size = 20, prob = 0.02)
```

```
[1] 0.940101
```

```r
# P(X <= 3): vjerojatnost 3 ili manje
pbinom(q = 3, size = 20, prob = 0.02)
```

```
[1] 0.9994003
```

```r
# P(X >= 2): vjerojatnost 2 ili više (komplement od P(X <= 1))
1 - pbinom(q = 1, size = 20, prob = 0.02)
```

```
[1] 0.05989898
```

Vjerojatnost da ćemo imati jednu ili nijednu viralnu objavu je oko 94%. Vjerojatnost da ćemo imati barem 2 je samo oko 6%. Ovi izračuni su ključni za postavljanje realnih očekivanja u digitalnom marketingu.

6.5 Simulacija: rbinom()

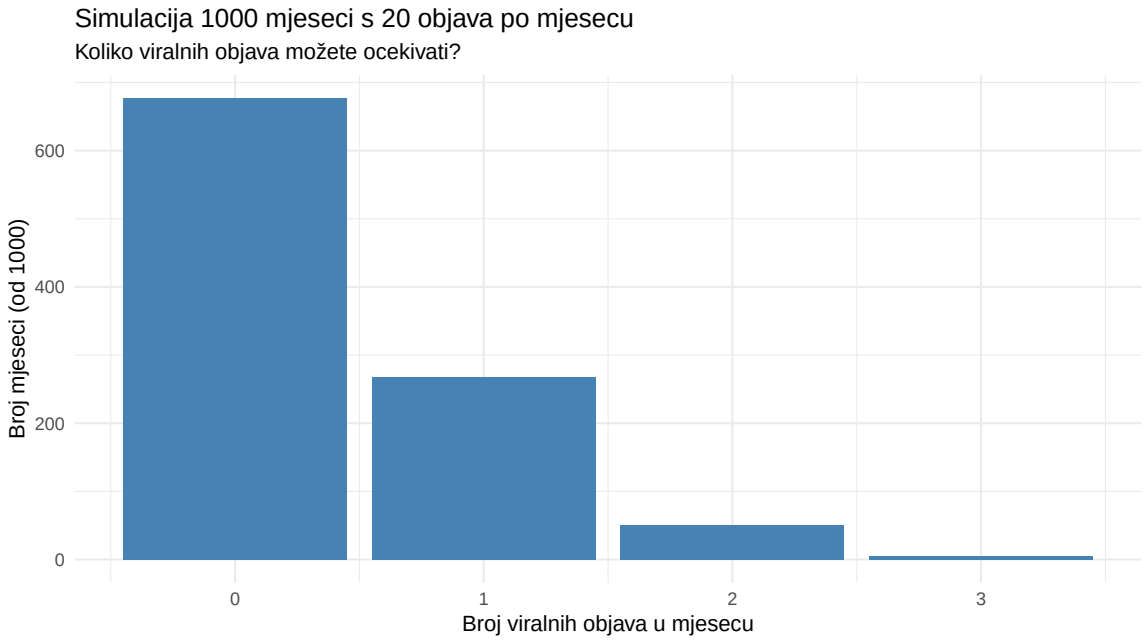
Funkcija `rbinom(n, size, prob)` generira slučajne uzorke iz binomne distribucije. Ovo je korisno za simulaciju scenarija.

```r
set.seed(42)

# Simuliraj 1000 "mjeseci" u kojima imaš po 20 objava
simulacija <- tibble(
  mjesec = 1:1000,
  n_viralnih = rbinom(n = 1000, size = 20, prob = 0.02)
)

# Distribucija rezultata
simulacija |>
  count(n_viralnih) |>
  mutate(udio = round(n / sum(n), 3))
```

```
# A tibble: 4 x 3
  n_viralnih     n  udio
       <int> <int> <dbl>
1          0   677 0.677
```

```
2           1   268 0.268
3           2    50 0.05
4           3     5 0.005
```

```
simulacija |>
  ggplot(aes(x = n_viralnih)) +
  geom_bar(fill = "steelblue") +
  scale_x_continuous(breaks = 0:8) +
  labs(
    title = "Simulacija 1000 mjeseci s 20 objava po mjesecu",
    subtitle = "Koliko viralnih objava možete očekivati?",
    x = "Broj viralnih objava u mjesecu",
    y = "Broj mjeseci (od 1000)"
  ) +
  theme_minimal()
```

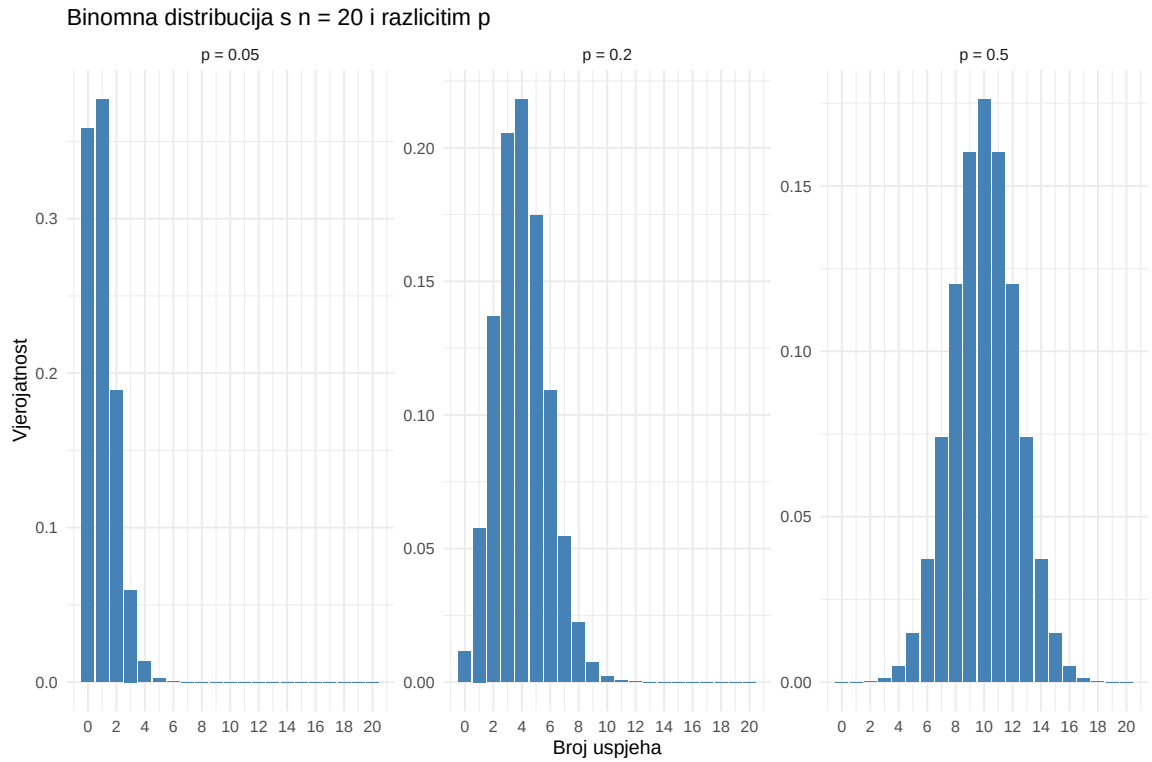


Simulacija potvrđuje teoriju. U većini mjeseci nećete imati nijednu viralnu objavu. Ponekad jednu. Rijetko dvije. I vrlo rijetko tri ili više. Ovo je korisna informacija za postavljanje KPI-jeva (key performance indicators) u komunikacijskim kampanjama.

6.6 Kako p mijenja distribuciju

Pogledajmo kako se distribucija mijenja s različitim vjerojatnostima uspjeha.

```r
# Tri različite vjerojatnosti uspjeha
expand_grid(
  p = c(0.05, 0.20, 0.50),
  x = 0:20
) |>
  mutate(
    vjerojatnost = dbinom(x, size = 20, prob = p),
    p_label = paste0("p = ", p)
  ) |>
  ggplot(aes(x = x, y = vjerojatnost)) +
  geom_col(fill = "steelblue") +
  facet_wrap(~p_label, scales = "free_y") +
  scale_x_continuous(breaks = seq(0, 20, by = 2)) +
  labs(
    title = "Binomna distribucija s n = 20 i različitim p",
    x = "Broj uspjeha",
    y = "Vjerojatnost"
  ) +
  theme_minimal()
```



Kad je p malo (0.05), distribucija je jako iskrivljena udesno i koncentrirana oko 0 i 1. Kad je p umjereno (0.20), distribucija se širi i centar se pomiče udesno. Kad je $p = 0.50$, distribucija

je simetrična i izgleda gotovo poput zvona. Ova simetrija kod p = 0.5 nas vodi prema najvažnijoj distribuciji u cijeloj statistici, a to je normalna.

6.7 Primjena: A/B test emaila

Zamislite da testirate dva naslova za newsletter. Naslov A ima open rate 22%, naslov B ima 28%. Poslali ste svaki naslov na uzorak od 50 pretplatnika. Naslov B je imao 14 otvaranja od 50, dok je A imao 11. Je li ovo uvjerljiva razlika?

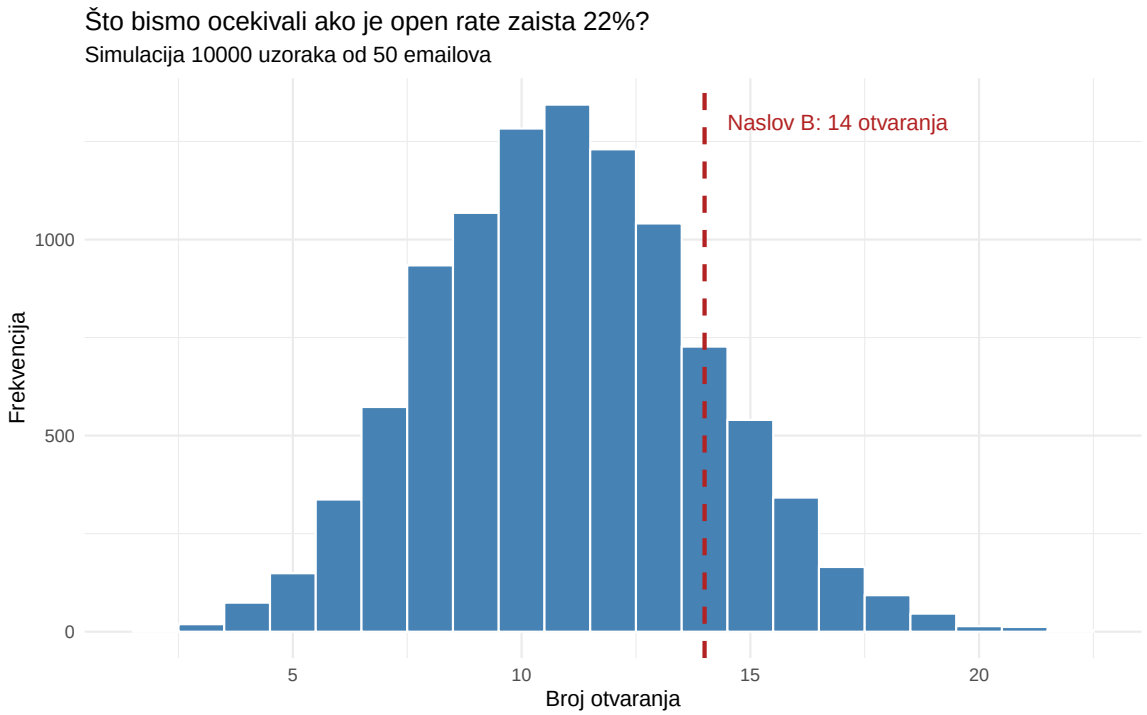
```
# Ako je B zaista isti kao A (p = 0.22), kolika je šansa da vidimo 14 ili više otvaranja?
p_14_ili_vise <- 1 - pbinom(q = 13, size = 50, prob = 0.22)
cat("P(X >= 14 | p = 0.22) =", round(p_14_ili_vise, 3), "\n")
```

```
P(X >= 14 | p = 0.22) = 0.194
```

Šansa da vidimo 14 ili više otvaranja ako je pravi open rate samo 22% iznosi oko 19%. To nije zanemarivo. Ne možemo s velikom sigurnošću tvrditi da je B bolji samo na temelju ovog jednog uzorka.

Ovo je srž statističkog razmišljanja i prethodnica formalnih testova hipoteza koje ćemo učiti u tjednu 10. Pitanje je uvijek isto — koliko je vjerovatno vidjeti ovakav ili ekstremniji rezultat čistom slučajnošću?

```
set.seed(42)

# Simulacija: ako je pravi open rate 22%, koliko otvaranja bismo dobili u 50 emailova?
sim_a <- tibble(
  otvaranja = rbinom(10000, size = 50, prob = 0.22)
)

sim_a |>
  ggplot(aes(x = otvaranja)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "white") +
  geom_vline(xintercept = 14, color = "firebrick", linetype = "dashed", linewidth = 1) +
  annotate("text", x = 14.5, y = 1300, label = "Naslov B: 14 otvaranja",
           hjust = 0, color = "firebrick") +
  labs(
    title = "Što bismo očekivali ako je open rate zaista 22%?",
    subtitle = "Simulacija 10000 uzoraka od 50 emailova",
    x = "Broj otvaranja",
    y = "Frekvencija"
  ) +
  theme_minimal()
```



Crvena crta pokazuje rezultat naslova B (14 otvaranja). Vidimo da je to na desnom repu distribucije ali nije ekstreman rezultat. Dobar dio simuliranih uzoraka ima 14 ili više otvaranja čak i kad je pravi open rate samo 22%. Ovo sugerira da razlika možda nije statistički značajna. Trebamo ili veći uzorak ili veću razliku da bismo donijeli sigurnije zaključke.

! Važna napomena

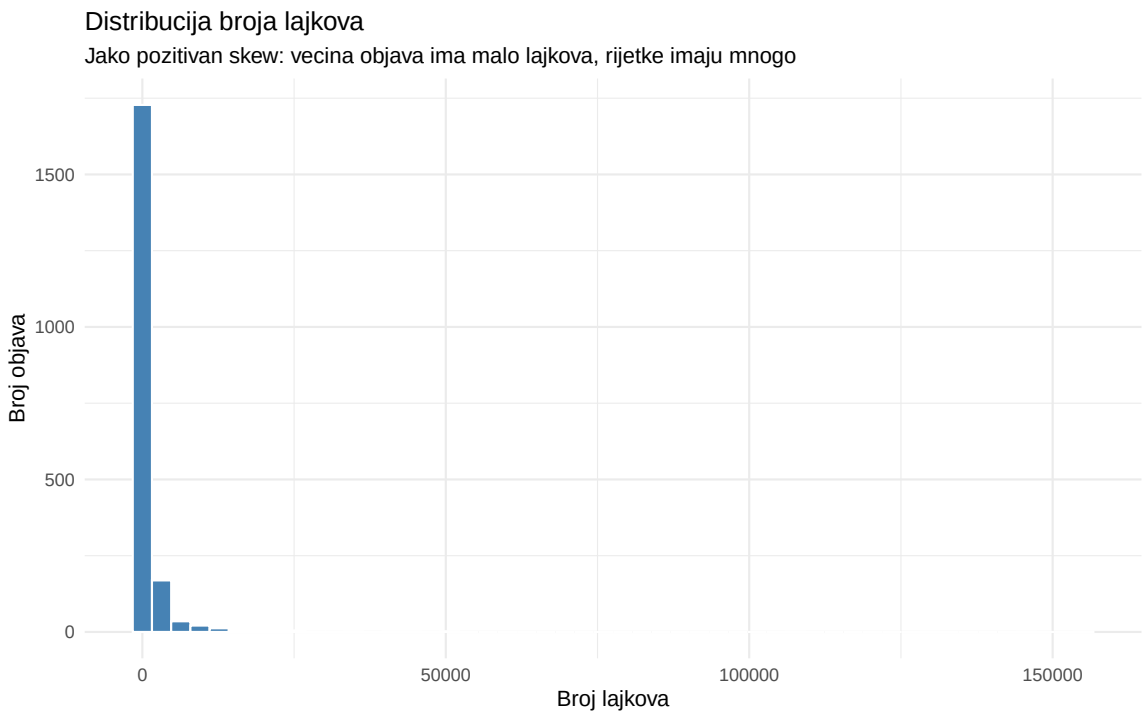
Binomna distribucija se primjenjuje kad imate fiksni broj nezavisnih pokušaja, svaki s istom vjerojatnošću uspjeha. Primjeri iz komunikologije uključuju broj lajkova (svaki pratitelj nezavisno odlučuje), broj otvaranja emaila (svaki pretplatnik nezavisno odlučuje), broj konverzija na landing stranici (svaki posjetitelj nezavisno odlučuje). Pretpostavka nezavisnosti je važna jer ako jedan lajk poveća vidljivost pa uzrokuje sljedeći lajk (viralnost), stroga binomna pretpostavka je narušena.

---

7 Distribucija u stvarnim podacima

Pogledajmo distribucije u našem datasetu i povežimo ih s teorijskim konceptima.

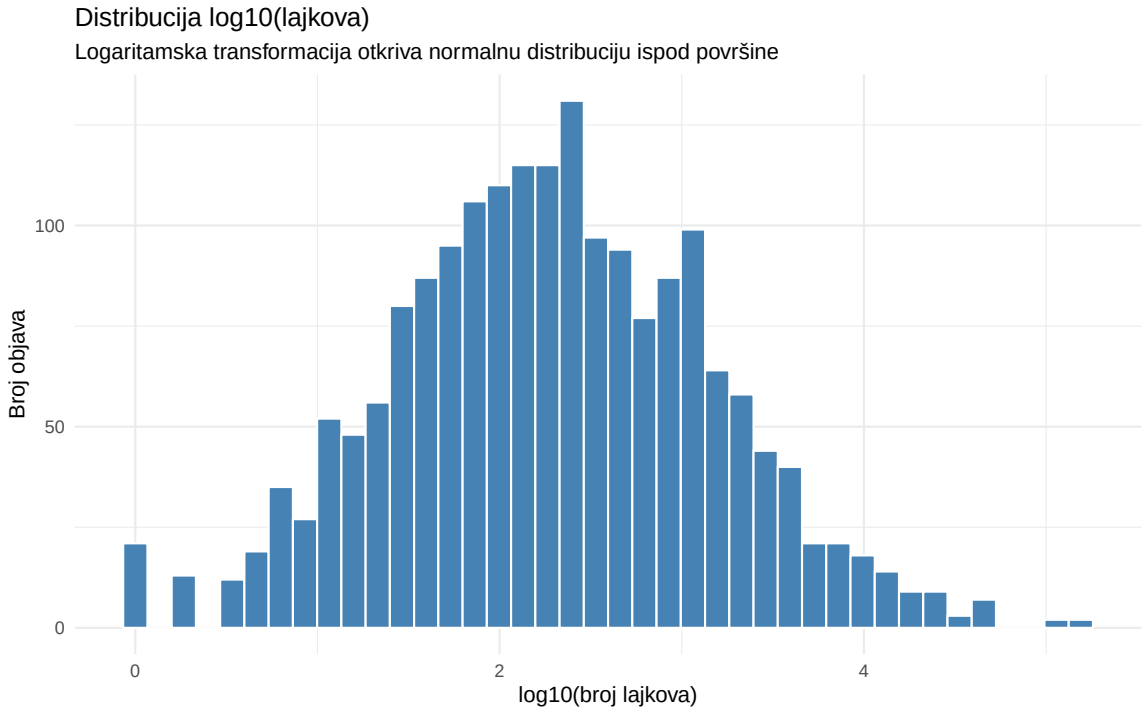
```r
posts |>
  ggplot(aes(x = likes)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  labs(
    title = "Distribucija broja lajkova",
    subtitle = "Jako pozitivan skew: većina objava ima malo lajkova, rijetke imaju mnogo",
    x = "Broj lajkova",
    y = "Broj objava"
  ) +
  theme_minimal()
```



Ova distribucija je jako iskrivljena udesno. Većina objava ima relativno malo lajkova, ali postoji dugačak rep objava s tisućama ili stotinama tisuća lajkova. Ovo je tipično za metrike angažmana na društvenim mrežama i zove se **power law** ili **log-normalna** distribucija.

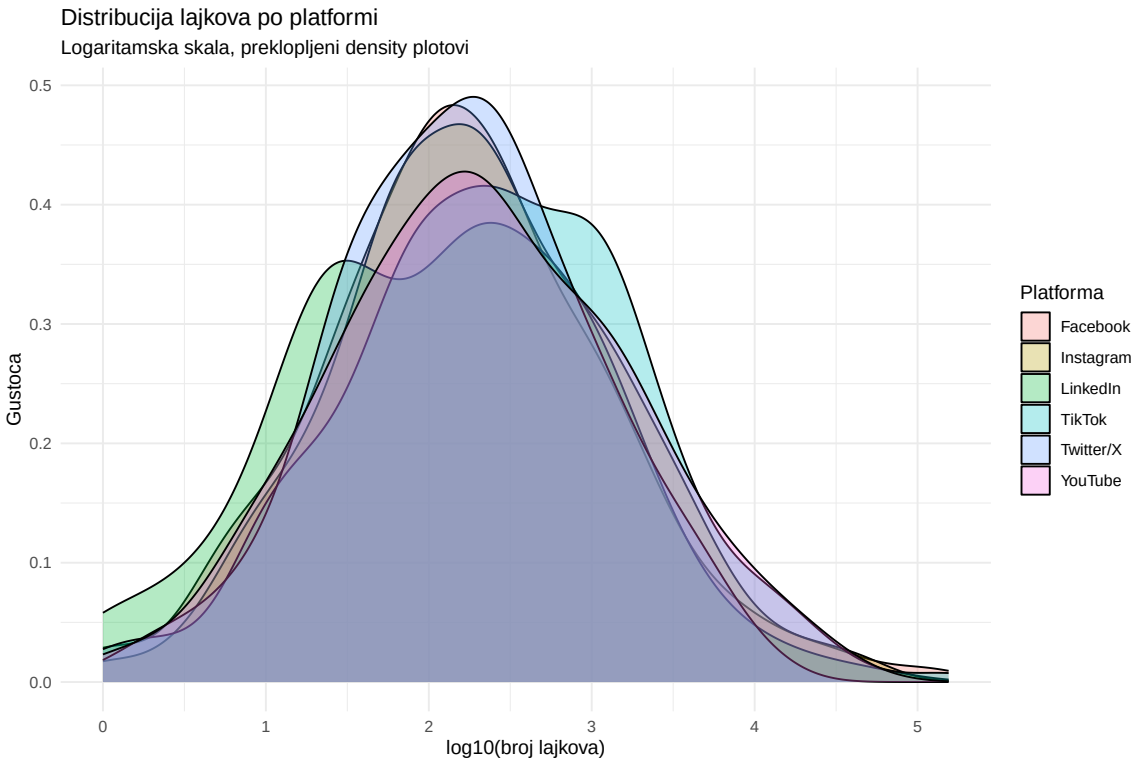
```r
# Logaritmirana distribucija izgleda puno "normalnije"
posts |>
  filter(likes > 0) |>
  ggplot(aes(x = log10(likes))) +
  geom_histogram(fill = "steelblue", color = "white", bins = 40) +
  labs(
    title = "Distribucija log10(lajkova)",
    subtitle = "Logaritamska transformacija otkriva normalnu distribuciju ispod površine",
    x = "log10(broj lajkova)",
```

```
    y = "Broj objava"
  ) +
  theme_minimal()
```



Kad primijenimo logaritamsku transformaciju, distribucija počinje nalikovati na zvonoliku krivulju. Ovo je važno zapažanje jer mnoge varijable u komunikologiji koje izgledaju neprirodno iskrivljene zapravo su log-normalno distribuirane. Na logaritamskoj skali, postaju normalne. Normalnu distribuciju ćemo detaljno obraditi u drugom dijelu predavanja.

```
posts |>
  filter(likes > 0) |>
  ggplot(aes(x = log10(likes), fill = platform)) +
  geom_density(alpha = 0.3) +
  labs(
    title = "Distribucija lajkova po platformi",
    subtitle = "Logaritamska skala, preklopljeni density plotovi",
    x = "log10(broj lajkova)",
    y = "Gustoća",
    fill = "Platforma"
  ) +
  theme_minimal()
```



Platforme se razlikuju po distribuciji angažmana. YouTube i TikTok imaju širu distribuciju (veća varijabilnost, češće ekstremne vrijednosti), dok LinkedIn ima užu i pomaknutu ulijevo (manji ali konzistentniji angažman). Ovo odražava fundamentalne razlike u mehanici platformi.

Podsjetnik

U prvom dijelu naučili smo osnovna pravila vjerojatnosti (komplement, zbrajanje, množenje, uvjetna vjerojatnost) i binomnu distribuciju za modeliranje diskretnih ishoda (uspjeh/neuspjeh). U ovom dijelu prelazimo na kontinuirane varijable i upoznajemo najvažniju distribuciju u cijeloj statistici, a to je normalna.

8 Normalna distribucija

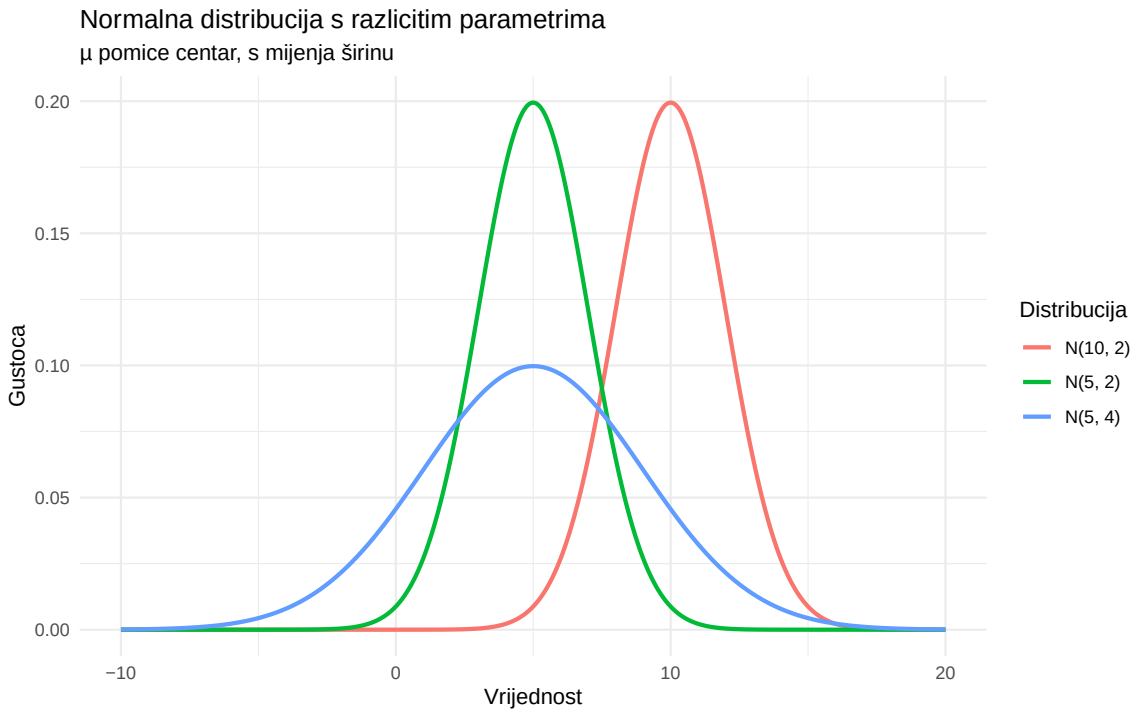
Normalna distribucija (ili Gaussova krivulja, ili zvonolika krivulja) je najvažnija distribucija u statistici. Razlog nije samo u tome što mnoge varijable u prirodi imaju približno normalan oblik. Važniji razlog je **centralni granični teorem** koji kaže da prosjek dovoljno velikog uzorka ima približno normalnu distribuciju, neovisno o obliku izvorne distribucije. Ovo čini normalnu distribuciju temeljom gotovo svih statističkih testova.

8.1 Parametri normalne distribucije

Normalna distribucija ima dva parametra. **(mi)** je srednja vrijednost (prosjek), koja određuje centar distribucije. **(sigma)** je standardna devijacija, koja određuje širinu distribucije.

Piše se X ~ N(,) i čita "X slijedi normalnu distribuciju s prosjekom i standardnom devijacijom ."

```
# Vizualizacija normalne distribucije s različitim parametrima
tibble(x = seq(-10, 20, length.out = 500)) |>
  mutate(
    `N(5, 2)` = dnorm(x, mean = 5, sd = 2),
    `N(5, 4)` = dnorm(x, mean = 5, sd = 4),
    `N(10, 2)` = dnorm(x, mean = 10, sd = 2)
  ) |>
  pivot_longer(-x, names_to = "distribucija", values_to = "gustoca") |>
  ggplot(aes(x = x, y = gustoca, color = distribucija)) +
  geom_line(linewidth = 1) +
  labs(
    title = "Normalna distribucija s različitim parametrima",
    subtitle = "  pomiče centar,  mijenja širinu",
    x = "Vrijednost",
    y = "Gustoća",
    color = "Distribucija"
  ) +
  theme_minimal()
```



N(5, 2) i N(10, 2) imaju istu širinu ($= 2$) ali različite centre ($= 5$ vs $= 10$). N(5, 2) i N(5, 4) imaju isti centar ali različite širine. Veća standardna devijacija znači širu, plošniju krivulju. Manja znači užu i višu.

8.2 Pravilo 68-95-99.7

Jedno od najkorisnijih svojstava normalne distribucije je da uvijek isti postotak podataka pada unutar istog broja standardnih devijacija od prosjeka.

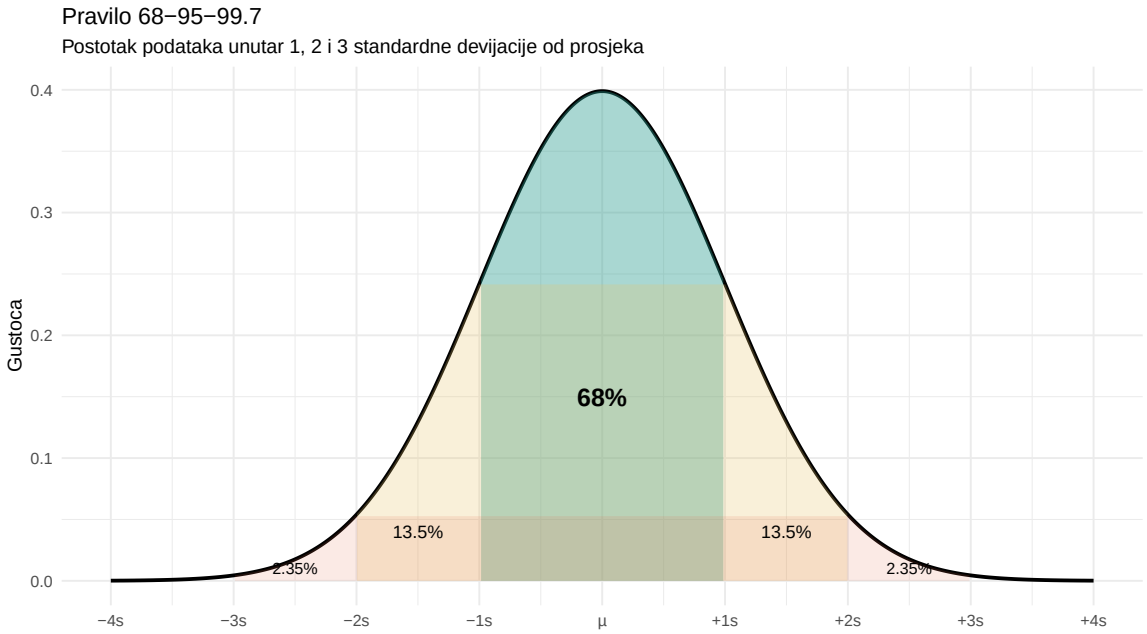
Otprilike **68%** podataka je unutar 1 standardne devijacije od prosjeka (± 1).

Otprilike **95%** podataka je unutar 2 standardne devijacije (± 2).

Otprilike **99.7%** podataka je unutar 3 standardne devijacije (± 3).

```
# Vizualizacija pravila 68-95-99.7
mu <- 0
sigma <- 1

norm_data <- tibble(x = seq(-4, 4, length.out = 500), y = dnorm(x, mu, sigma))

ggplot(norm_data, aes(x = x, y = y)) +
  geom_line(linewidth = 1) +
  geom_area(data = norm_data |> filter(x >= -1, x <= 1), alpha = 0.4, fill = "#2a9d8f") +
  geom_area(data = norm_data |> filter(x >= -2, x <= -1 | x >= 1, x <= 2), alpha = 0.25, f
  geom_area(data = norm_data |> filter(x >= -3, x <= -2 | x >= 2, x <= 3), alpha = 0.15, f
```

```
  annotate("text", x = 0, y = 0.15, label = "68%", size = 5, fontface = "bold") +
  annotate("text", x = 1.5, y = 0.04, label = "13.5%", size = 3.5) +
  annotate("text", x = -1.5, y = 0.04, label = "13.5%", size = 3.5) +
  annotate("text", x = 2.5, y = 0.01, label = "2.35%", size = 3) +
  annotate("text", x = -2.5, y = 0.01, label = "2.35%", size = 3) +
  scale_x_continuous(breaks = -4:4, labels = c("-4 ", "-3 ", "-2 ", "-1 ", " ", "+1 ", "+2 ",
  labs(
    title = "Pravilo 68-95-99.7",
    subtitle = "Postotak podataka unutar 1, 2 i 3 standardne devijacije od prosjeka",
    x = NULL,
    y = "Gustoća"
  ) +
  theme_minimal()
```



Ovo pravilo ima ogromnu praktičnu korist. Ako znate prosjek i standardnu devijaciju, odmah znate raspone u kojima se nalazi većina podataka. Na primjer, ako je prosječno vrijeme čitanja članka 80 sekundi sa SD od 25, tada je 95% čitatelja unutar raspona 80 $\pm$ 50 sekundi, dakle između 30 i 130 sekundi. Netko tko čita 200 sekundi je daleko izvan normalnog raspona.

8.3 Provjera pravila na stvarnim podacima

```
# Koristimo log-transformirane lajkove koji su približno normalni
log_likes <- posts |>
  filter(likes > 0) |>
```

```
  pull(likes) |>
  log10()

mu <- mean(log_likes)
sigma <- sd(log_likes)

cat("Prosjek log10(likes):", round(mu, 2), "\n")
```

```
Prosjek log10(likes): 2.29
```

```
cat("SD log10(likes):", round(sigma, 2), "\n\n")
```

```
SD log10(likes): 0.88
```

```
# Koliko podataka pada unutar 1, 2, 3 SD?
unutar_1sd <- mean(log_likes >= mu - sigma & log_likes <= mu + sigma)
unutar_2sd <- mean(log_likes >= mu - 2*sigma & log_likes <= mu + 2*sigma)
unutar_3sd <- mean(log_likes >= mu - 3*sigma & log_likes <= mu + 3*sigma)

cat("Unutar 1 SD:", round(unutar_1sd * 100, 1), "% (teorijski: 68%)\n")
```

```
Unutar 1 SD: 69.1 % (teorijski: 68%)
```

```
cat("Unutar 2 SD:", round(unutar_2sd * 100, 1), "% (teorijski: 95%)\n")
```

```
Unutar 2 SD: 95.1 % (teorijski: 95%)
```

```
cat("Unutar 3 SD:", round(unutar_3sd * 100, 1), "% (teorijski: 99.7%)\n")
```

```
Unutar 3 SD: 99.8 % (teorijski: 99.7%)
```

Rezultati su blizu teorijskih vrijednosti, što potvrđuje da su logaritmirani lajkovi približno normalno distribuirani. Poklapanje nije savršeno jer nijedna stvarna varijabla nije savršeno normalna, ali je dovoljno dobro za praktičnu primjenu.

---

9 Z-score: standardizacija

Z-score (standardizirani rezultat) izražava koliko je neka vrijednost udaljena od prosjeka, mjereno u standardnim devijacijama.

$$z = \frac{x - \mu}{\sigma}$$

Ako je z = 0, vrijednost je na prosjeku. Ako je z = 1, vrijednost je jednu standardnu devijaciju iznad prosjeka. Ako je z = -2, vrijednost je dvije standardne devijacije ispod prosjeka.

```
# Z-score za lajkove (na log skali)
posts_z <- posts |>
  filter(likes > 0) |>
  mutate(
    log_likes = log10(likes),
    z_likes = (log_likes - mean(log_likes)) / sd(log_likes)
  )

# Objave s najvišim z-scoreom (najneobičajniji angažman)
posts_z |>
  select(post_id, platform, content_type, likes, z_likes) |>
  arrange(desc(z_likes)) |>
  head(10)
```

```
# A tibble: 10 x 5
   post_id platform  content_type  likes z_likes
     <dbl> <chr>     <chr>         <dbl>   <dbl>
 1     595 TikTok    video        155132    3.31
 2    1756 TikTok    reel         144871    3.27
 3    1508 Facebook  tekst        129142    3.22
 4    1520 Facebook  slika        108320    3.13
 5     525 Facebook  reel          50834    2.76
 6    1970 Instagram reel          50748    2.75
 7     378 Instagram reel          46499    2.71
 8     254 TikTok    reel          44184    2.69
 9    1569 Instagram story         43805    2.68
10    1984 LinkedIn  tekst         41559    2.66
```

Objave s z-scoreom većim od 2 ili 3 su statistički neobične. U normalnoj distribuciji, samo oko 5% podataka ima z > 2 (ili z < -2), a samo 0.3% ima z > 3 (ili z < -3). Ovo čini z-score korisnim alatom za identifikaciju outliera.

9.1 Z-score za usporedbu nepovezanih varijabli

Velika prednost z-scorea je što omogućuje usporedbu varijabli na potpuno različitim skalama. Recimo da želite usporediti koliko je neka objava neobična po broju lajkova i po broju komentara.

```
posts_z2 <- posts |>
  filter(likes > 0, comments > 0) |>
  mutate(
    z_likes = scale(log10(likes))[,1],
    z_comments = scale(log10(comments))[,1]
  )

# Koje objave imaju neproporcionalno više komentara nego lajkova?
posts_z2 |>
  mutate(razlika = z_comments - z_likes) |>
  select(post_id, platform, content_type, likes, comments, z_likes, z_comments, razlika) |
  arrange(desc(razlika)) |>
  head(8)
```

```
# A tibble: 8 x 8
  post_id platform  content_type likes comments z_likes z_comments razlika
    <dbl> <chr>     <chr>        <dbl>    <dbl>   <dbl>      <dbl>   <dbl>
1     659 LinkedIn  tekst           35        5  -1.38      -0.579   0.796
2    1174 LinkedIn  tekst           15        2  -1.89      -1.10    0.787
3    1097 TikTok    reel           144       22  -0.523      0.263   0.785
4    1273 Instagram slika           29        4  -1.49      -0.706   0.783
5     260 Facebook  story          133       20  -0.571      0.209   0.779
6     914 TikTok    reel             8        1  -2.27      -1.49    0.772
7    1212 Instagram slika           57        8  -1.08      -0.312   0.769
8    1706 Facebook  slika          224       34  -0.256      0.510   0.766
```

Funkcija `scale()` u R-u automatski izračunava z-score (oduzima prosjek i dijeli sa SD). `[,1]` na kraju je tehnički detalj koji pretvara matricu u vektor.

Objave s velikom pozitivnom razlikom (z_comments » z_likes) su one koje su generirale neproporcionalno mnogo diskusije s obzirom na ukupni angažman. To su često kontroverzni sadržaji ili sadržaji koji potiču na odgovor. Ovo je primjer kako statistička standardizacija otkriva obrasce koji nisu očiti iz sirovih brojeva.

---

10 R funkcije za normalnu distribuciju

R ima četiri funkcije za normalnu distribuciju, organizirane prema istom obrascu kao binomna (d/p/q/r).

10.1 dnorm(): gustoća

```
# Gustoća u točki x za standardnu normalnu N(0,1)
dnorm(0)       # Gustoća na prosjeku (vrh krivulje)
```

```
[1] 0.3989423
```

```
dnorm(1)       # Gustoća na 1 SD iznad prosjeka
```

```
[1] 0.2419707
```

```
dnorm(2)       # Gustoća na 2 SD iznad prosjeka
```

```
[1] 0.05399097
```

```
# Gustoća za nestandardnu normalnu
dnorm(100, mean = 80, sd = 25)  # Koliko je "normalan" rezultat od 100 sekundi?
```

```
[1] 0.01158766
```

Za razliku od binomne, `dnorm()` ne daje vjerojatnost nego gustoću. U kontinuiranoj distribuciji, vjerojatnost jedne specifične vrijednosti je 0 (jer postoji beskonačno mnogo mogućih vrijednosti). Gustoća nam govori koliko je ta vrijednost relativno česta u odnosu na druge.

10.2 pnorm(): kumulativna vjerojatnost

`pnorm()` je najkorisnija od četiri funkcije. Daje vjerojatnost P(X x), odnosno postotak distribucije koji je ispod zadane vrijednosti.

```
# Za standardnu normalnu N(0,1):
pnorm(0)    # 50% distribucije je ispod prosjeka
```

```
[1] 0.5
```

```r
pnorm(1)    # ~84% je ispod 1 SD iznad prosjeka
```

```
[1] 0.8413447
```

```r
pnorm(-1)   # ~16% je ispod 1 SD ispod prosjeka
```

```
[1] 0.1586553
```

```r
# Koliki postotak čitatelja provede manje od 60 sekundi?
# (ako je vrijeme ~ N(80, 25))
pnorm(60, mean = 80, sd = 25)
```

```
[1] 0.2118554
```

Oko 21% čitatelja bi provelo manje od 60 sekundi, pod pretpostavkom normalne distribucije s prosjekom 80 i SD 25.

```r
# Postotak čitatelja između 60 i 120 sekundi
pnorm(120, mean = 80, sd = 25) - pnorm(60, mean = 80, sd = 25)
```

```
[1] 0.7333453
```

```r
# Postotak čitatelja iznad 150 sekundi (gornji rep)
1 - pnorm(150, mean = 80, sd = 25)
```

```
[1] 0.00255513
```

```r
# Postotak s z-scoreom između -1 i 1 (provjera pravila 68%)
pnorm(1) - pnorm(-1)
```

```
[1] 0.6826895
```

Razlika `pnorm(120, ...) - pnorm(60, ...)` daje vjerojatnost između dva praga. Komplement `1 - pnorm(...)` daje gornji rep. Ovi izračuni su temelj za p-vrijednosti koje ćemo učiti u tjednu 10.

10.3 qnorm(): kvantili (obrnuta funkcija)

`qnorm()` je obrnuta funkcija od `pnorm()`. Daje vrijednost ispod koje se nalazi zadani postotak distribucije.

```r
# Ispod koje vrijednosti je 95% distribucije?
qnorm(0.95, mean = 80, sd = 25)
```

```
[1] 121.1213
```

```r
# Ispod koje je 5%? (donji kvintil)
qnorm(0.05, mean = 80, sd = 25)
```

```
[1] 38.87866
```

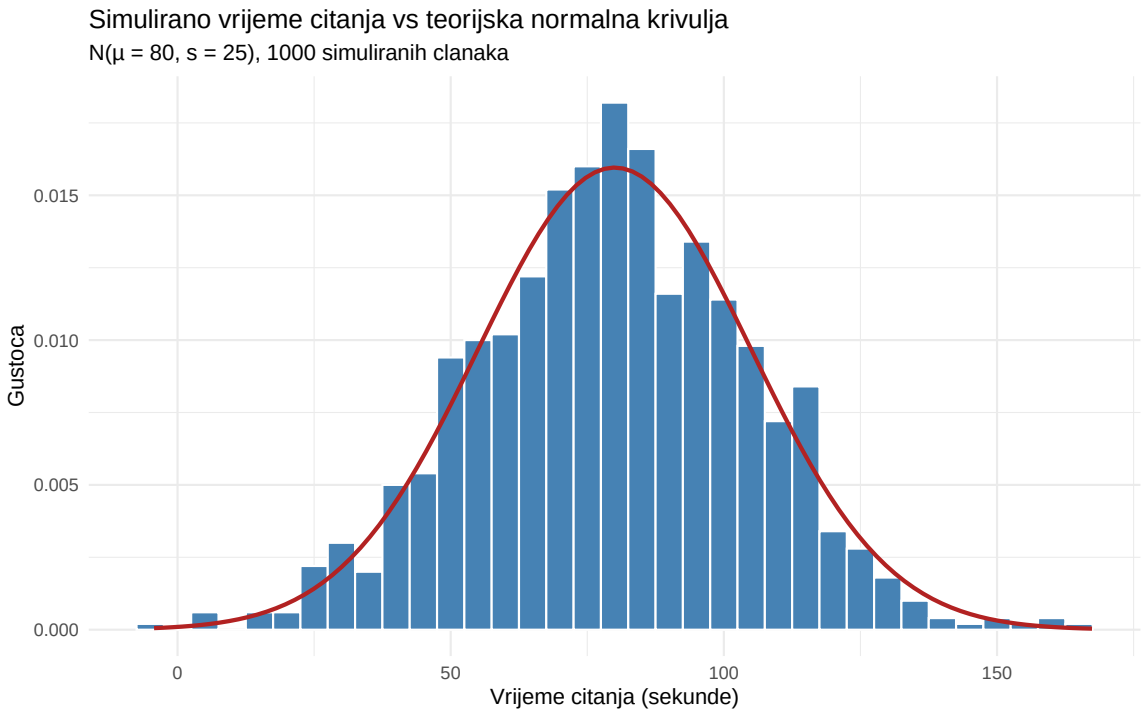
```r
# Top 1% čitatelja (oni koji čitaju najduže)
qnorm(0.99, mean = 80, sd = 25)
```

```
[1] 138.1587
```

Top 1% najdediciranijih čitatelja provodi više od 138 sekundi na članku. Ovo je korisno za postavljanje pragova. Na primjer, možete definirati "super čitatelja" kao onoga koji je u top 5% po vremenu čitanja.

10.4 rnorm(): simulacija

```r
set.seed(42)

# Simulacija 1000 članaka s prosječnim vremenom čitanja ~ N(80, 25)
sim_vrijeme <- tibble(
  clanak = 1:1000,
  vrijeme = rnorm(1000, mean = 80, sd = 25)
)

sim_vrijeme |>
  ggplot(aes(x = vrijeme)) +
  geom_histogram(aes(y = after_stat(density)),
                 fill = "steelblue", color = "white", binwidth = 5) +
  stat_function(fun = dnorm, args = list(mean = 80, sd = 25),
                color = "firebrick", linewidth = 1) +
  labs(
    title = "Simulirano vrijeme čitanja vs teorijska normalna krivulja",
    subtitle = "N(  = 80,   = 25), 1000 simuliranih članaka",
    x = "Vrijeme čitanja (sekunde)",
    y = "Gustoća"
  ) +
  theme_minimal()
```



Crvena krivulja je teorijska normalna distribucija. Histogram prikazuje simulirane podatke. Poklapanje je dobro, posebno u sredini distribucije. Na repovima uvijek postoji malo odstupanja jer je uzorak konačan.

Funkcija `stat_function()` je elegantan način za dodavanje teorijske krivulje na ggplot graf. Prima ime distribucijske funkcije i njezine argumente.

> **Praktični savjet**
>
> Obrazac d/p/q/r vrijedi za sve distribucije u R-u. `d` daje gustoću (ili vjerojatnost za diskretne), `p` daje kumulativnu vjerojatnost, `q` daje kvantile (obrnuto od p), `r` generira slučajne uzorke. Za binomnu: `dbinom`, `pbinom`, `qbinom`, `rbinom`. Za normalnu: `dnorm`, `pnorm`, `qnorm`, `rnorm`. Za t-distribuciju (tjedan 12): `dt`, `pt`, `qt`, `rt`. Naučite obrazac jednom, primijenite svugdje.

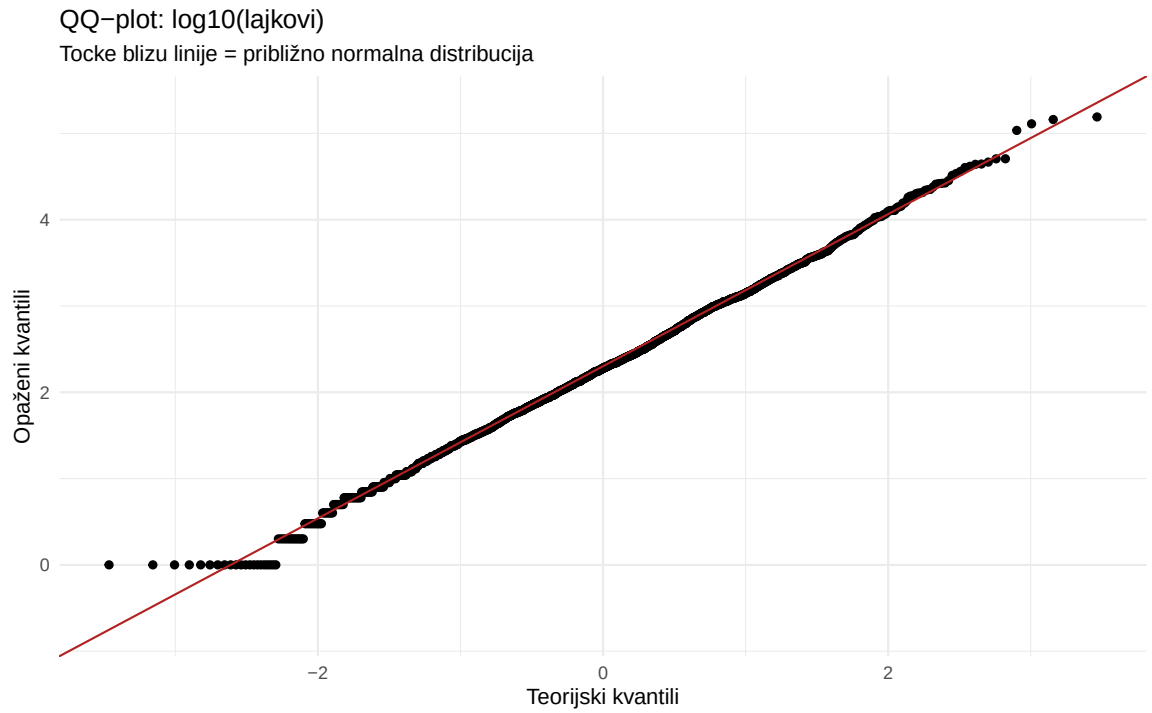
---

11 QQ-plot: je li moja varijabla normalno distribuirana?

Vizualna provjera normalnosti je važan korak u mnogim analizama jer mnogi statistički testovi pretpostavljaju (približno) normalnu distribuciju. QQ-plot (quantile-quantile plot) je standardni alat za ovu provjeru.

QQ-plot uspoređuje kvantile vaših podataka s kvantilima teorijske normalne distribucije. Ako su podaci normalno distribuirani, točke padaju na ravnu liniju. Odstupanja od linije ukazuju na nenormalnost.

```
# QQ-plot za log-transformirane lajkove (trebali bi biti približno normalni)
posts |>
  filter(likes > 0) |>
  ggplot(aes(sample = log10(likes))) +
  stat_qq() +
  stat_qq_line(color = "firebrick") +
  labs(
    title = "QQ-plot: log10(lajkovi)",
    subtitle = "Točke blizu linije = približno normalna distribucija",
    x = "Teorijski kvantili",
    y = "Opaženi kvantili"
  ) +
  theme_minimal()
```

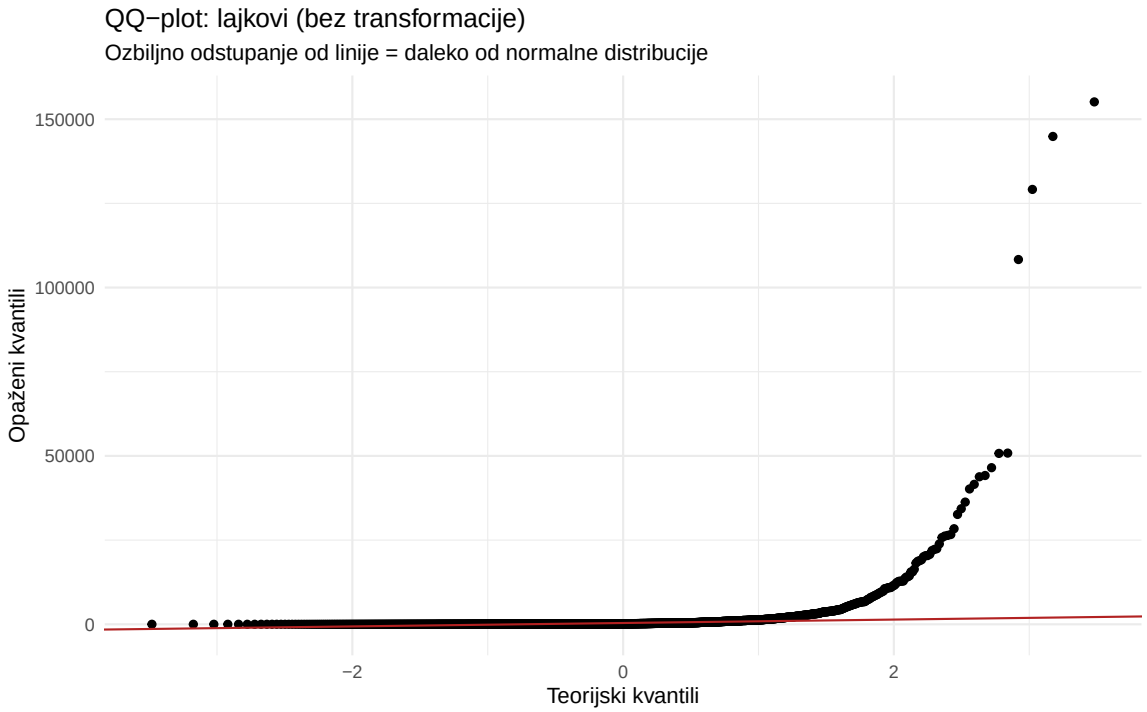


Točke uglavnom prate crvenu liniju, s malim odstupanjima na repovima. Ovo je tipičan rezultat za stvarne podatke i smatra se prihvatljivo normalnim za većinu statističkih testova.

Za usporedbu, pogledajmo QQ-plot za netransformirane lajkove.

```
# QQ-plot za netransformirane lajkove (jako iskrivljeni)
posts |>
```

```r
ggplot(aes(sample = likes)) +
stat_qq() +
stat_qq_line(color = "firebrick") +
labs(
  title = "QQ-plot: lajkovi (bez transformacije)",
  subtitle = "Ozbiljno odstupanje od linije = daleko od normalne distribucije",
  x = "Teorijski kvantili",
  y = "Opaženi kvantili"
) +
theme_minimal()
```

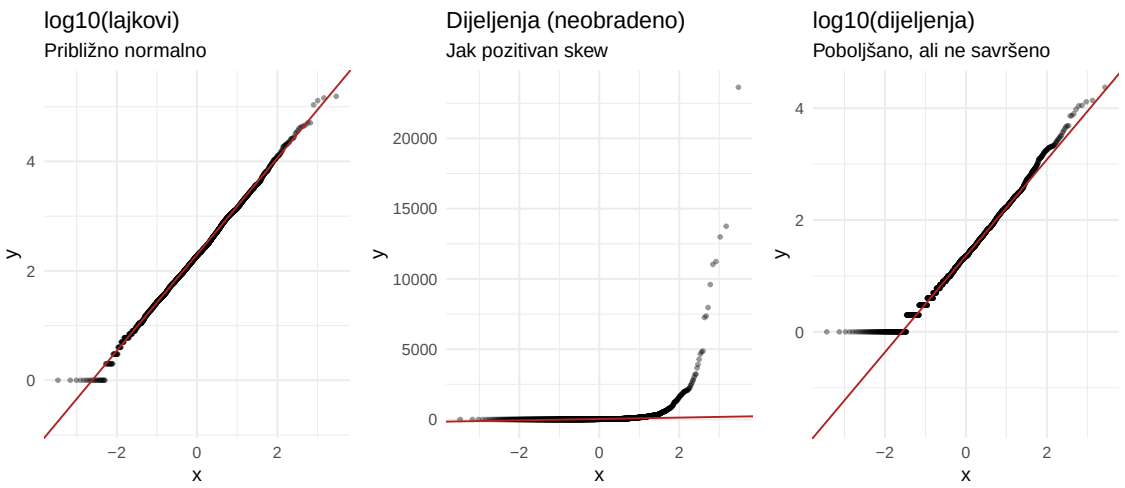


Razlika je dramatična. Netransformirani lajkovi jako odstupaju od normalne distribucije. Desni rep se savija strmo prema gore, što znači da postoje mnogo veće vrijednosti nego što bi normalna distribucija predvidjela. Ovo je vizualni potpis za pozitivno iskrivljenu (right-skewed) distribuciju.

11.1 Čitanje QQ-plota

Različiti obrasci na QQ-plotu govore različite priče. Točke na ravnoj liniji znače normalnu distribuciju. Točke koje se savijaju prema gore na desnom kraju znače pozitivan skew (dugačak desni rep). Točke koje se savijaju prema dolje na lijevom kraju znače negativan skew (dugačak lijevi rep). Točke koje se savijaju prema gore na oba kraja (oblik slova S) znače "teže repove" od normalne distribucije (više ekstrema nego što normalna predviđa).

```r
library(patchwork)

p_qq1 <- posts |>
  filter(likes > 0) |>
  ggplot(aes(sample = log10(likes))) +
  stat_qq(size = 0.8, alpha = 0.4) +
  stat_qq_line(color = "firebrick") +
  labs(title = "log10(lajkovi)", subtitle = "Približno normalno") +
  theme_minimal()

p_qq2 <- posts |>
  ggplot(aes(sample = shares)) +
  stat_qq(size = 0.8, alpha = 0.4) +
  stat_qq_line(color = "firebrick") +
  labs(title = "Dijeljenja (neobrađeno)", subtitle = "Jak pozitivan skew") +
  theme_minimal()

p_qq3 <- posts |>
  filter(shares > 0) |>
  ggplot(aes(sample = log10(shares))) +
  stat_qq(size = 0.8, alpha = 0.4) +
  stat_qq_line(color = "firebrick") +
  labs(title = "log10(dijeljenja)", subtitle = "Poboljšano, ali ne savršeno") +
  theme_minimal()

p_qq1 + p_qq2 + p_qq3
```



Usporedba tri QQ-plota pokazuje transformacijsku strategiju. Sirovi podaci o dijeljenjima su daleko od normalnih. Log-transformacija ih značajno približava normalnosti, ali ne savršeno.

U praksi, “dovoljno normalno” je uglavnom prihvatljivo za statističke testove, posebno s velikim uzorcima.

---

12 Praktična primjena: postavljanje pragova i identifikacija outliera

Normalna distribucija i z-score daju nam objektivne alate za donošenje odluka koje bi inače bile proizvoljne.

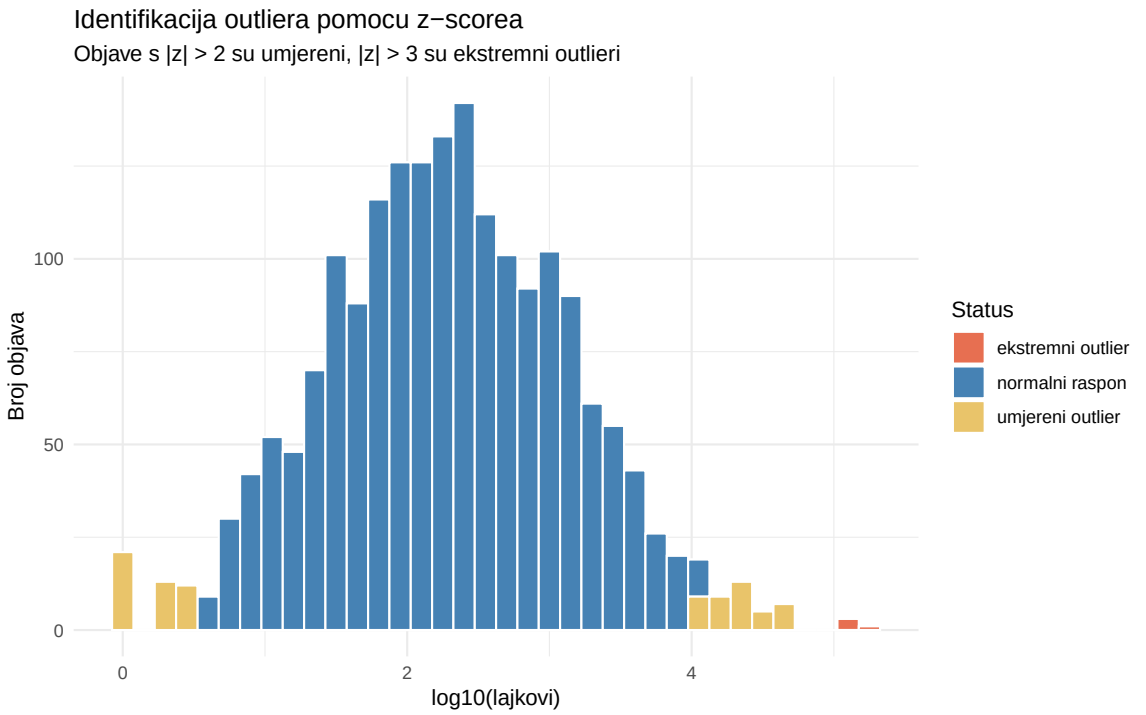
12.1 Definiranje “neobičnog” rezultata

```
# Identifikacija outliera u engagement metrikama
posts_analiza <- posts |>
  filter(likes > 0) |>
  mutate(
    log_likes = log10(likes),
    z_likes = scale(log_likes)[,1],
    outlier_status = case_when(
      abs(z_likes) > 3 ~ "ekstremni outlier",
      abs(z_likes) > 2 ~ "umjereni outlier",
      .default = "normalni raspon"
    )
  )

posts_analiza |>
  count(outlier_status) |>
  mutate(udio = round(n / sum(n), 3))
```

```
# A tibble: 3 x 3
  outlier_status        n  udio
  <chr>             <int> <dbl>
1 ekstremni outlier     4 0.002
2 normalni raspon    1795 0.951
3 umjereni outlier     89 0.047
```

```
posts_analiza |>
  ggplot(aes(x = log_likes, fill = outlier_status)) +
  geom_histogram(binwidth = 0.15, color = "white") +
  scale_fill_manual(values = c(
    "normalni raspon" = "steelblue",
```

```
    "umjereni outlier" = "#e9c46a",
    "ekstremni outlier" = "#e76f51"
  )) +
  labs(
    title = "Identifikacija outliera pomoću z-scorea",
    subtitle = "Objave s |z| > 2 su umjereni, |z| > 3 su ekstremni outlieri",
    x = "log10(lajkovi)",
    y = "Broj objava",
    fill = "Status"
  ) +
  theme_minimal()
```



12.2 Planiranje: kolika je šansa za uspjeh kampanje?

Normalna distribucija omogućuje izračun vjerojatnosti za buduće kampanje na temelju povijesnih podataka.

```
# Pretpostavimo da je open rate newsletter kampanja ~ N(0.25, 0.07)
# (prosjek 25%, SD 7%)

# Kolika je vjerojatnost da kampanja ima open rate iznad 30%?
p_iznad_30 <- 1 - pnorm(0.30, mean = 0.25, sd = 0.07)
cat("P(open rate > 30%) =", round(p_iznad_30, 3), "\n")
```

```
P(open rate > 30%) = 0.238
```

```
# Kolika je vjerojatnost da padne ispod 15%? (loš rezultat)
p_ispod_15 <- pnorm(0.15, mean = 0.25, sd = 0.07)
cat("P(open rate < 15%) =", round(p_ispod_15, 3), "\n")
```

```
P(open rate < 15%) = 0.077
```

```
# Koji open rate je na granici top 10% kampanja?
top_10 <- qnorm(0.90, mean = 0.25, sd = 0.07)
cat("Prag za top 10%:", round(top_10 * 100, 1), "%\n")
```

```
Prag za top 10%: 34 %
```

Ovi izračuni omogućuju objektivno postavljanje ciljeva. Umjesto proizvoljnog "cilj nam je 30% open rate", možemo reći "30% open rate je u top 24% naših kampanja, što je ambiciozan ali realan cilj."

12.3 Usporedba platformi na zajedničkoj skali

Z-score omogućuje usporedbu angažmana između platformi koje imaju potpuno različite skale.

```
# Z-score lajkova UNUTAR svake platforme
posts_platform_z <- posts |>
  filter(likes > 0) |>
  mutate(log_likes = log10(likes)) |>
  group_by(platform) |>
  mutate(
    z_likes = (log_likes - mean(log_likes)) / sd(log_likes)
  ) |>
  ungroup()

# Top 5 objava po z-scoreu unutar svake platforme
posts_platform_z |>
  group_by(platform) |>
  slice_max(z_likes, n = 1) |>
  select(platform, content_type, likes, followers, z_likes) |>
  arrange(desc(z_likes))
```

```
# A tibble: 6 x 5
# Groups:   platform [6]
  platform  content_type  likes followers z_likes
  <chr>     <chr>         <dbl>     <dbl>   <dbl>
1 Facebook  tekst        129142   1616076    3.26
2 TikTok    video        155132    903229    3.09
3 Instagram reel          50748    633458    2.81
4 LinkedIn  tekst         41559    553174    2.73
5 YouTube   reel          36288    225790    2.49
6 Twitter/X slika         12775    261180    2.43
```

Objava s 500 lajkova na LinkedInu može biti neobičnija (viši z-score) nego objava s 50 000 lajkova na TikToku, jer su skale potpuno različite. Z-score normalizira tu razliku i omogućuje poštenu usporedbu.

---

13 Od vjerojatnosti do statističkog zaključivanja

Sve što smo naučili danas je temelj za ono što dolazi. Pogledajmo kako se koncepti povezuju.

Binomna distribucija ćemo koristiti u tjednu 11 za hi-kvadrat testove (je li distribucija kategorija onakva kakvu očekujemo?) i u tjednu 10 za razumijevanje logike testiranja hipoteza.

Normalna distribucija je temelj za t-testove (tjedan 12), ANOVA-u (tjedan 13) i regresiju (tjedan 14) jer pretpostavljaju normalnu distribuciju reziduala.

Z-score je temelj za standardizirane veličine učinka i za razumijevanje p-vrijednosti.

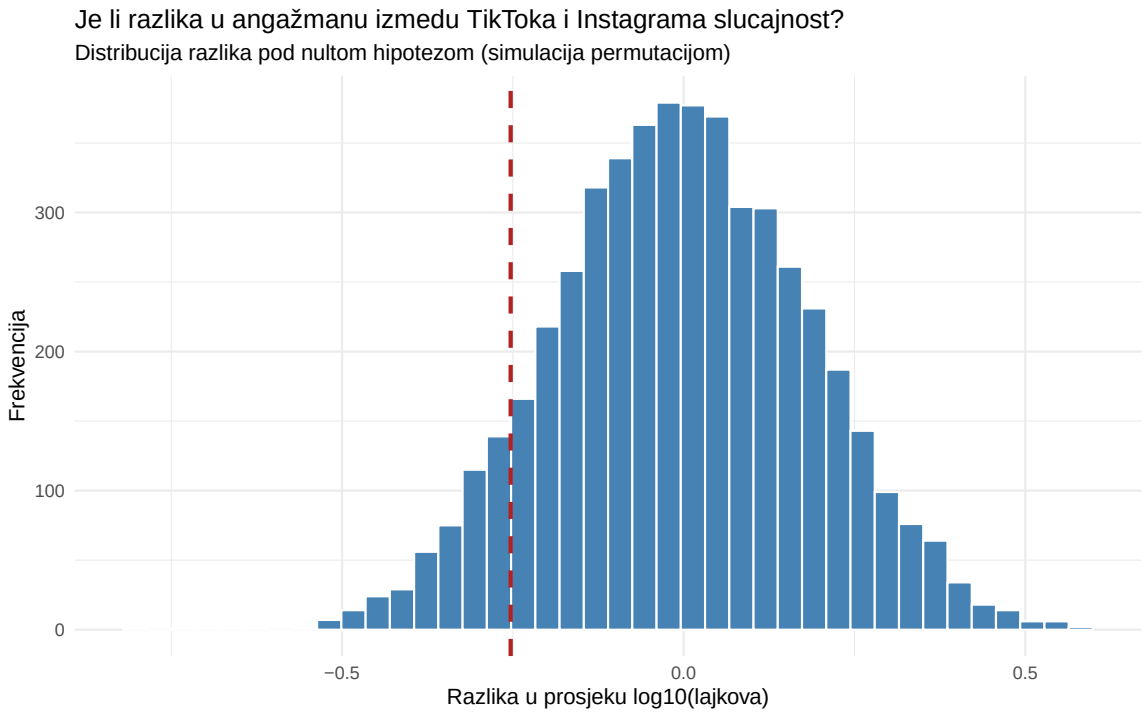
Uvjetna vjerojatnost je logika iza svakog statističkog testa. Kolika je vjerojatnost vidjeti ovakav rezultat DADO da je nulta hipoteza istinita?

```r
# Primjer: je li prosječni angažman na TikToku zaista veći nego na Instagramu?
set.seed(42)

# Uzmimo uzorke od 50 objava s svake platforme
uzorak_tt <- posts |>
  filter(platform == "TikTok", likes > 0) |>
  slice_sample(n = 50) |>
  pull(likes) |>
  log10()

uzorak_ig <- posts |>
  filter(platform == "Instagram", likes > 0) |>
```

```
  slice_sample(n = 50) |>
  pull(likes) |>
  log10()

razlika <- mean(uzorak_tt) - mean(uzorak_ig)
cat("Razlika u prosjeku log10(lajkova):", round(razlika, 3), "\n")
```

```
Razlika u prosjeku log10(lajkova): -0.253
```

```
# Koliko je ova razlika neobična? Simulirajmo!
sim_razlike <- replicate(5000, {
  sve <- c(uzorak_tt, uzorak_ig)
  pomijesano <- sample(sve)
  mean(pomijesano[1:50]) - mean(pomijesano[51:100])
})

tibble(razlika_sim = sim_razlike) |>
  ggplot(aes(x = razlika_sim)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 40) +
  geom_vline(xintercept = razlika, color = "firebrick", linetype = "dashed", linewidth = 1
  labs(
    title = "Je li razlika u angažmanu između TikToka i Instagrama slučajnost?",
    subtitle = "Distribucija razlika pod nultom hipotezom (simulacija permutacijom)",
    x = "Razlika u prosjeku log10(lajkova)",
    y = "Frekvencija"
  ) +
  theme_minimal()
```



Crvena crta označava opaženu razliku. Histogram prikazuje distribuciju razlika koje bismo očekivali čistom slučajnošću (ako nema stvarne razlike između platformi). Ako je crvena crta daleko od centra histograma, razlika je neobična i vjerojatno nije slučajna. Ovo je, u suštini, logika t-testa koji ćemo formalno naučiti u tjednu 12, ali ovdje smo ju demonstrirali simulacijom.

> Statistika je u konačnici nauka o donošenju zaključaka u uvjetima neizvjesnosti. Vjerojatnost je jezik kojim tu neizvjesnost izražavamo. Svaki statistički test, svaki interval pouzdanosti, svaka p-vrijednost govori o vjerojatnosti. Razumjeti vjerojatnost znači razumjeti statistiku.

! Ključni zaključci

1. Vjerojatnost je broj između 0 i 1 koji izražava izvjesnost. Frekvencijski pristup definira ju kao dugoročnu relativnu frekvenciju.

2. Zakon velikih brojeva — s više ponavljanja, relativna frekvencija konvergira prema pravoj vjerojatnosti.

3. Osnovna pravila — komplement P(ne A) = 1 − P(A), zbrajanje za isključive P(A ili B) = P(A) + P(B), množenje za nezavisne P(A i B) = P(A) × P(B). Uvjetna vjerojatnost P(A|B) je temelj za segmentnu analizu.

4. Binomna distribucija modelira broj uspjeha u n nezavisnih pokušaja s vjerojat-

nošću p. R funkcije su `dbinom()`, `pbinom()`, `rbinom()`.

5. Normalna distribucija je definirana prosjekom i standardnom devijacijom . Pravilo 68-95-99.7 daje postotak podataka unutar 1, 2 i 3 SD od prosjeka.
6. Z-score z = (x −) / izražava koliko je vrijednost udaljena od prosjeka u jedinicama SD. Omogućuje usporedbu varijabli na različitim skalama.
7. R funkcije za normalnu uključuju `dnorm()` za gustoću, `pnorm()` za kumulativnu vjerojatnost (P(X x)), `qnorm()` za kvantile (obrnuto od pnorm) i `rnorm()` za simulaciju.
8. Obrazac d/p/q/r vrijedi za sve distribucije u R-u.
9. QQ-plot uspoređuje kvantile podataka s teorijskim. Točke na ravnoj liniji znače normalnost. Odstupanja ukazuju na skew ili teške repove.
10. Metrike angažmana na društvenim mrežama su tipično log-normalno distribuirane. Log-transformacija ih često pretvara u (približno) normalne.
11. Z-score služi za identifikaciju outliera ($|z| > 2$ ili 3) i za usporedbu opažanja između grupa s različitim skalama.
12. Svaki statistički test koji ćemo učiti temelji se na pitanju — kolika je vjerojatnost vidjeti ovakav ili ekstremniji rezultat čistom slučajnošću? Ovo predavanje daje konceptualni temelj za to pitanje.

⚠ Priprema za kolokvij (tjedan 8)

Sljedeći tjedan je **kolokvij** koji pokriva gradivo iz tjedana 1 do 7. Kolokvij će uključivati

1. Konceptualna pitanja o istraživačkom dizajnu, mjerenju i vrstama varijabli (tjedan 1).
2. Čitanje i pisanje R koda — tibbleovi, pipe, dplyr glagoli, ggplot2 (tjedni 2 do 5).
3. Interpretaciju deskriptivnih statistika i grafova (tjedan 4 i 5).
4. Razumijevanje funkcija i DRY principa (tjedan 6).
5. Izračun i interpretaciju vjerojatnosti, uključujući binomnu i normalnu distribuciju (tjedan 7).

Za pripremu trebate

1. Ponovite pojmovnike iz svakog tjedna. Svaki pojmovnik sadrži ključne koncepte.
2. Pokrenite sve primjere iz predavanja 2 do 7 i pokušajte ih modificirati.
3. Vježbajte čitanje R koda. Za zadani pipeline, opišite riječima što svaki korak radi.

4. Vježbajte interpretaciju. Za zadani graf ili tablicu, napišite 2 do 3 rečenice o tome što rezultat znači.
5. Vježbajte izračun. Za zadanu situaciju, izračunajte odgovarajuću vjerojatnost koristeći `pbinom()` ili `pnorm()`.

---

14 Dodatno čitanje

Obavezno

Navarro, D. (2018). *Learning Statistics with R*, Chapter 9 (Introduction to Probability). Besplatno dostupno na learningstatisticswithr.com. Pokriva sva pravila vjerojatnosti i distribucije detaljnije nego ovo predavanje.

Wickham, H. & Grolemund, G. (2023). *R for Data Science* (2nd edition), Section 26.4. Besplatno dostupno na r4ds.hadley.nz. Pregled generiranja slučajnih brojeva i simulacije.

Preporučeno

Diez, D., Çetinkaya-Rundel, M., & Barr, C. (2019). *OpenIntro Statistics* (4th edition), Chapters 3 i 4. Besplatno dostupno na openintro.org/book/os. Distribucije s mnogo grafičkih prikaza i primjera.

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Pelican Books. Popularnoznanstvena knjiga koja izvrsno objašnjava ulogu vjerojatnosti u donošenju odluka.

Ellenberg, J. (2014). *How Not to Be Wrong: The Power of Mathematical Thinking*. Penguin Press. Poglavlja o vjerojatnosti su izuzetno pristupačna i puna primjera iz svakodnevnog života.

---

15 Pojmovnik

Pojam	Objašnjenje
Vjerojatnost	Broj između 0 i 1 koji izražava izvjesnost događaja.
Frekvencijski pristup	Definira vjerojatnost kao dugoročnu relativnu frekvenciju.
Bayesijanski pristup	Definira vjerojatnost kao stupanj uvjerenja.
Zakon velikih brojeva	S više ponavljanja, relativna frekvencija konvergira prema pravoj vjerojatnosti.

Pojam	Objašnjenje
Komplementarno pravilo	P(ne A) = 1 − P(A).
Pravilo zbrajanja	Za isključive: P(A ili B) = P(A) + P(B). Za neisključive: oduzeti P(A i B).
Pravilo množenja	Za nezavisne: P(A i B) = P(A) × P(B).
Međusobno isključivi	Događaji koji se ne mogu dogoditi istovremeno.
Nezavisni događaji	Jedan ne utječe na vjerojatnost drugoga.
Uvjetna vjerojatnost	P(A|B) = P(A i B) / P(B). Vjerojatnost A dado da se B dogodio.
Distribucija vjerojatnosti	Funkcija koja dodjeljuje vjerojatnost svakom mogućem ishodu.
Binomna distribucija	Distribucija broja uspjeha u n nezavisnih pokušaja s vjerojatnošću p.
Normalna distribucija	Zvonolika, simetrična distribucija definirana prosjekom i standardnom devijacijom . Najvažnija distribucija u statistici.
Standardna normalna	Normalna distribucija s = 0 i = 1. Piše se N(0, 1).
Pravilo 68-95-99.7	U normalnoj distribuciji, 68% podataka je unutar ±1 , 95% unutar ±2 , 99.7% unutar ±3 od prosjeka.
Z-score	Standardizirani rezultat: z = (x −) / . Izražava udaljenost od prosjeka u jedinicama SD.
`scale()`	R funkcija koja izračunava z-score (oduzima prosjek, dijeli s SD).
`dbinom()`	Točna vjerojatnost binomne distribucije.
`pbinom()`	Kumulativna vjerojatnost binomne distribucije P(X q).
`rbinom()`	Generiranje slučajnih uzoraka iz binomne distribucije.
`dnorm()`	Gustoća normalne distribucije u zadanoj točki.
`pnorm()`	Kumulativna vjerojatnost normalne distribucije P(X x).
`qnorm()`	Kvantili normalne distribucije (obrnuta funkcija od pnorm).
`rnorm()`	Generiranje slučajnih uzoraka iz normalne distribucije.
`set.seed()`	Fiksira generator slučajnih brojeva za ponovljivost simulacija.
QQ-plot	Graf koji uspoređuje kvantile podataka s teorijskim kvantilima. Služi za provjeru normalnosti.

Pojam	Objašnjenje
`stat_qq()`	ggplot2 funkcija za QQ-plot. Kombinira se s `stat_qq_line()`.
Power law distribucija	Distribucija s dugačkim desnim repom. Tipična za metrike angažmana.
Log-normalna distribucija	Distribucija čiji logaritam je normalno distribuiran.
Outlier	Opažanje neobično udaljeno od ostatka podataka. Često definirano kao
Centralni granični teorem	Prosjek uzorka ima približno normalnu distribuciju, neovisno o obliku izvorne distribucije, kad je uzorak dovoljno velik.
Permutacijski test	Simulacijski pristup za testiranje razlike između grupa: miješa podatke i uspoređuje opaženu razliku s distribucijom pod nultom hipotezom.