Tjedan 5: Deskriptivna statistika

Kako brojkama opisati ono što podaci govore

2025-03-15

Table of contents

```
library(tidyverse)
library(scales)
```

Ishodi učenja

Nakon ovog predavanja moći ćete

1. Objasniti razliku između mjera centralne tendencije (aritmetička sredina, skraćena sredina, medijan, mod) i odabrati odgovarajuću mjeru za različite tipove podataka.
2. Izračunati i interpretirati mjere varijabilnosti (raspon, prosječno apsolutno odstupanje, varijanca, standardna devijacija, interkvartilni raspon) te objasniti zašto varijanca dijeli s $N - 1$, a ne s N.
3. Prepoznati asimetriju i zaobljenost distribucije te objasniti zašto su te karakteristike važne za izbor statističkih metoda.
4. Generirati cjeloviti sažetak varijable koristeći `summarise()` i `across()`.
5. Koristiti `group_by()` i `summarise()` za izračunavanje deskriptivnih statistika po grupama.
6. Izračunati i interpretirati standardne rezultate (z-scores) te objasniti zašto su korisni za usporedbu varijabli na različitim skalama.
7. Izračunati i interpretirati Pearsonov i Spearmanov koeficijent korelacije te razumjeti njihova ograničenja.
8. Prepoznati različite tipove nedostajućih vrijednosti i primijeniti odgovarajuće strategije za rad s njima.

1 Zašto su brojke same po sebi beskorisne

Zamislite da ste upravo završili veliko istraživanje o korištenju TikToka u Hrvatskoj. Proveli ste anketu na 300 ispitanika, prikupili podatke o tome koliko minuta dnevno svaka osoba

provodi na platformi i sada sjedite pred ogromnom tablicom punom brojki. Vaš urednik ili klijent vas pita jednostavno pitanje. Koliko ljudi zapravo koriste TikTok i koliko vremena tamo provode?

Mogli biste im poslati cijelu tablicu. Svih 300 redova. Ali to nitko neće čitati i, što je još važnije, nitko iz toga neće izvući nikakav zaključak. Ljudski mozak jednostavno nije dizajniran da iz stotina pojedinačnih brojki spontano prepozna obrasce. Upravo zato postoji deskriptivna statistika. Njezin posao je uzeti gomilu podataka i pretvoriti je u nekoliko smislenih brojki koje opisuju što se u podacima zapravo događa.

To zvuči jednostavno, i donekle jest, ali postoji jedna zamka o kojoj treba voditi računa od samog početka. **Svaki put kad sažmete podatke u jednu ili dvije brojke, nešto izgubite.** Aritmetička sredina od 65 minuta dnevno na TikToku zvuči informativno, ali skriva činjenicu da neki ljudi provode 140 minuta, a neki samo 7. Ta informacija o raspršenosti podataka jednako je važna kao i ta jedna prosječna vrijednost, a ponekad je i važnija. Upravo zato u deskriptivnoj statistici nikad ne gledamo samo jednu mjeru. Trebamo barem dvije stvari. Trebamo nešto što nam govori gdje se podaci nalaze (mjere centralne tendencije) i nešto što nam govori koliko su raspršeni (mjere varijabilnosti).

Navarro u svojoj knjizi koristi zgodni paralelizam. Zamislite da vam netko opisuje grupu ljudi riječima. Reći će vam nešto o prosječnoj osobi u grupi (to je centralna tendencija), ali će vam reći i koliko su ljudi u grupi slični jedni drugima ili se pak drastično razlikuju (to je varijabilnost). Trebate obje informacije da biste stvorili mentalnu sliku o čemu se radi. Upravo to ćemo naučiti danas.

Krenimo od podataka.

---

2 Naši podaci: anketa o korištenju TikToka

Tijekom ovog predavanja koristit ćemo simulirani dataset koji sadrži podatke iz ankete o korištenju TikToka. Anketa je provedena na 300 ispitanika različitih dobnih skupina, a prikupljene su informacije o dnevnom vremenu korištenja, broju pogledanih videozapisa tjedno, aktivnosti na platformi i povjerenju u sadržaj koji tamo pronalaze.

Učitajmo podatke i pogledajmo s čime radimo.

```
tiktok <- read_csv("../resources/datasets/tiktok_usage.csv")
glimpse(tiktok)
```

```
Rows: 300
Columns: 11
$ respondent_id         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1~
$ age                   <dbl> 19, 22, 20, 35, 28, 41, 19, 24, 31, 45, 21, 23, ~
```

```
$ age_group             <chr> "18-24", "18-24", "18-24", "25-34", "25-
34", "35~
$ gender                <chr> "female", "male", "female", "male", "female", "m~
$ daily_minutes         <dbl> 95, 78, 112, 45, 62, 22, 130, 88, 55, 18, 105, 7~
$ weekly_videos_watched <dbl> 320, 250, 410, 140, 200, 70, 480, 290, 175, 55, ~
$ likes_given           <dbl> 45, 30, 60, 15, 25, 8, 70, 35, 20, 5, 50, 28, 10~
$ comments_posted       <dbl> 3, 1, 5, 2, 3, 0, 8, 2, 1, 0, 4, 1, 1, 2, 0, 10,~
$ follows_creators      <dbl> 12, 8, 15, 5, 9, 3, 20, 10, 7, 2, 14, 7, 4, 8, 1~
$ trust_score           <dbl> 6, 5, 7, 4, 5, 3, 8, 6, 4, 3, 7, 5, 3, 5, 2, 8, ~
$ education             <chr> "student", "student", "student", "employed", "em~
```

Imamo 300 redova i 11 stupaca. Svaki red predstavlja jednog ispitanika. Varijabla `daily_minutes` bilježi koliko minuta dnevno osoba koristi TikTok, `age_group` svrstava ispitanike u dobne skupine, `trust_score` mjeri povjerenje u TikTok sadržaj na skali od 1 do 10, a `weekly_videos_watched` bilježi otprilike koliko videozapisa tjedno pogledaju.

Pogledajmo prvih desetak redova da stvorimo osjećaj za podatke.

```
tiktok |>
  select(respondent_id, age, age_group, daily_minutes, trust_score) |>
  head(10)
```

```
# A tibble: 10 x 5
   respondent_id   age age_group daily_minutes trust_score
           <dbl> <dbl> <chr>             <dbl>       <dbl>
 1             1    19 18-24                95           6
 2             2    22 18-24                78           5
 3             3    20 18-24               112           7
 4             4    35 25-34                45           4
 5             5    28 25-34                62           5
 6             6    41 35-44                22           3
 7             7    19 18-24               130           8
 8             8    24 18-24                88           6
 9             9    31 25-34                55           4
10            10    45 35-44                18           3
```

Na prvi pogled vidimo da mlađi ispitanici provode više vremena na TikToku od starijih. Ali koliko više? I koliko se ispitanici unutar iste dobne skupine razlikuju međusobno? Na ta pitanja odgovaraju deskriptivne statistike.

---

3 Mjere centralne tendencije

Kad netko kaže da želi znati koliko ljudi koriste TikTok, zapravo pita za neku vrstu tipične ili prosječne vrijednosti. U statistici to zovemo **mjerom centralne tendencije** jer tražimo središte oko kojeg se podaci grupiraju. Ideja je intuitvna, ali čim pokušate biti precizni, stvari postaju kompliciranije nego što biste očekivali. Postoji više načina da definirate središte skupa podataka, i ne daju svi iste rezultate. U ovom poglavlju obradit ćemo četiri mjere. To su aritmetička sredina, skraćena sredina, medijan i mod.

3.1 Aritmetička sredina

Aritmetička sredina je ono što većina ljudi misli kad kaže prosjek. Zbrojite sve vrijednosti i podijelite s brojem opažanja. Formula je jednostavna.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

gdje je N broj opažanja, a X_i svaka pojedinačna vrijednost. Matematička notacija može izgledati zastrašujuće ako je vidite prvi put, ali ideja je stvarno banalna. Zbrojite sve, podijelite s ukupnim brojem. To je to.

U R u to izračunavamo funkcijom `mean()`. Izračunajmo prosječno dnevno korištenje TikToka u cijelom uzorku.

```
tiktok |>
  summarise(
    prosjek_minuta = mean(daily_minutes),
    n = n()
  )
```

```
# A tibble: 1 x 2
  prosjek_minuta     n
           <dbl> <int>
1           56.9   300
```

Prosječni ispitanik u našem uzorku provodi otprilike 55 minuta dnevno na TikToku. Ali koliko je ta informacija korisna sama za sebe? Zamislite da pišete članak i u njemu navedete samo taj broj. Čitatelj bi mogao pomisliti da većina ljudi provodi oko sat vremena dnevno na TikToku, što je potpuno pogrešan zaključak jer ta sredina skriva enormnu razliku između dobnih skupina.

Aritmetička sredina ima jednu veliku prednost i jednu veliku manu. Prednost je da koristi svaku vrijednost u podacima, pa je u tom smislu najinformativnija mjera. U statistici se kaže da je sredina **dovoljni statistik** (sufficient statistic) za normalnu distribuciju, što znači

da sadrži svu informaciju o centralnoj tendenciji koju podaci nude. To zvuči apstraktno, ali praktična implikacija je važna. Ako su vaši podaci normalno distribuirani, aritmetička sredina je definitivno pravi izbor.

Mana je da je **osjetljiva na ekstremne vrijednosti** (outliers). Ako u vašem uzorku postoji jedna osoba koja koristi TikTok 8 sati dnevno (480 minuta), ta jedna osoba će pomaknuti prosjek za cijeli uzorak prema gore. To je razlog zašto prosječna plaća u nekoj zemlji može biti znatno viša od plaće koju prima većina zaposlenih. Nekolicina ljudi s ekstremno visokim primanjima vuče prosjek prema gore.

Evo konkretnog primjera koji pokazuje koliko jedna ekstremna vrijednost može utjecati na sredinu. Zamislite da imate pet korisnika koji TikTok koriste 20, 25, 30, 35 i 40 minuta dnevno. Prosjek je 30 minuta. Sada zamislite da šesti korisnik provodi 480 minuta (8 sati!) dnevno. Prosjek skače na 105 minuta, što nikako ne opisuje tipičnog korisnika u toj grupi.

```
bez_outliera <- c(20, 25, 30, 35, 40)
s_outlierom <- c(20, 25, 30, 35, 40, 480)

tibble(
  skup = c("Bez ekstremne vrijednosti", "S ekstremnom vrijednošću"),
  prosjek = c(mean(bez_outliera), mean(s_outlierom)),
  medijan = c(median(bez_outliera), median(s_outlierom))
)
```

```
# A tibble: 2 x 3
  skup                        prosjek medijan
  <chr>                         <dbl>   <dbl>
1 Bez ekstremne vrijednosti        30    30
2 S ekstremnom vrijednošću        105    32.5
```

Primijetite kako prosjek skoči sa 30 na 105, dok medijan ostaje stabilan. Na ovo ćemo se vratiti za trenutak.

> Praktični savjet
>
> Kad u medijskim izvještajima vidite izraz prosječna vrijednost bez dodatnog konteksta, uvijek se zapitajte postoje li u tim podacima ekstremne vrijednosti. Ako postoje, aritmetička sredina može biti zavaravajuća. Upravo zato odgovorni novinari uz prosjek uvijek navode i medijan, ili barem napomenu o rasponu podataka. Kad čitate da je prosječna plaća u Hrvatskoj, recimo, 1400 eura, zapitajte se koliki je medijan. Razlika vam govori koliko su plaće neravnomjerno raspodijeljene.

3.2 Skraćena sredina (trimmed mean)

Postoji kompromis između aritmetičke sredine (koja koristi sve podatke, ali je osjetljiva na outliere) i medijana (koji je robustan, ali ignorira većinu podataka). Taj kompromis

zove se **skraćena sredina** (trimmed mean). Ideja je jednostavna. Prije nego izračunamo prosjek, izbacimo određeni postotak najmanjih i najvećih vrijednosti. Najčešće se koristi 5% skraćivanje, što znači da izbacimo 5% najmanjih i 5% najvećih vrijednosti, pa izračunamo prosjek preostalih 90%.

U R u je to trivijalno jer funkcija `mean()` već ima argument `trim` koji prima proporciju (ne postotak!) koja se skraćuje sa svake strane.

```
tiktok |>
  summarise(
    prosjek = mean(daily_minutes),
    skracena_5 = mean(daily_minutes, trim = 0.05),
    skracena_10 = mean(daily_minutes, trim = 0.10),
    medijan = median(daily_minutes)
  )
```

```
# A tibble: 1 x 4
  prosjek skracena_5 skracena_10 medijan
    <dbl>      <dbl>       <dbl>   <dbl>
1    56.9       55.4        54.2      50
```

Vidimo da se skraćena sredina nalazi negdje između pune aritmetičke sredine i medijana. Što više skraćujemo, to se rezultat više približava medijanu. Zapravo, ako postavimo `trim = 0.5`, dobili bismo upravo medijan, jer bismo izbacili sve osim srednje vrijednosti.

Skraćena sredina je osobito korisna kad znate da vaši podaci imaju neke ekstremne vrijednosti, ali ne želite ih potpuno ignorirati. U praksi, 5% ili 10% skraćivanje obično dobro funkcionira. Navarro u knjizi napominje da se skraćena sredina pojavljuje iznenađujuće rijetko u objavljenim istraživanjima, što je šteta, jer je u mnogim situacijama bolji izbor od obične sredine.

3.3 Medijan

Medijan je vrijednost koja dijeli podatke na pola. Kad sve vrijednosti poredamo od najmanje do najveće, medijan je ona koja se nalazi točno na sredini. Ako imamo neparan broj opažanja, medijan je srednje opažanje. Ako imamo paran broj, uzimamo prosjek dvaju srednjih opažanja.

Ova definicija zvuči jednostavno, ali skriva nešto dublje. Medijan je zapravo odgovor na pitanje koje se razlikuje od pitanja na koje odgovara sredina. Aritmetička sredina minimizira zbroj kvadriranih odstupanja od sebe same. To zvuči apstraktno, ali praktično znači da sredina daje veliku težinu velikim odstupanjima. Medijan, s druge strane, minimizira zbroj **apsolutnih** odstupanja. To znači da medijan tretira sva odstupanja jednako, neovisno o tome koliko su velika. Zato je robustan.

Ključna razlika u odnosu na aritmetičku sredinu jest da medijan **ne ovisi o ekstremnim vrijednostima**. Ako jedna osoba koristi TikTok 480 umjesto 140 minuta dnevno, medijan se neće promijeniti ni za minutu jer ta osoba i dalje ostaje na istom kraju poretka. Vidjeli smo to u primjeru iznad. Zato kažemo da je medijan **robusna** mjera centralne tendencije.

Izračunajmo medijan za naše podatke, zajedno sa sredinom, kako bismo ih mogli usporediti.

```
tiktok |>
  summarise(
    prosjek = mean(daily_minutes),
    medijan = median(daily_minutes),
    razlika = mean(daily_minutes) - median(daily_minutes)
  )
```

```
# A tibble: 1 x 3
  prosjek medijan razlika
    <dbl>   <dbl>   <dbl>
1    56.9      50    6.85
```

Vidimo da su prosjek i medijan različiti. Kad je prosjek veći od medijana, to je signal da distribucija ima rep prema desno, odnosno da postoje neke veće vrijednosti koje vuku prosjek prema gore. U našem slučaju ta razlika postoji jer imamo velik broj mladih ispitanika koji koriste TikTok znatno više od ostalih, pa distribucija ukupnog uzorka ima pozitivnu asimetriju.

Odnos između sredine i medijana zapravo je brz dijagnostički alat za oblik distribucije. Ako je sredina otprilike jednaka medijanu, distribucija je vjerojatno prilično simetrična. Ako je sredina znatno veća od medijana, distribucija je pozitivno asimetrična (rep prema desno). Ako je sredina znatno manja od medijana, distribucija je negativno asimetrična (rep prema lijevo). Ovo nije egzaktan test, ali u praksi je korisna brza provjera.

3.4 Mod

Mod je najjednostavnija mjera centralne tendencije. To je jednostavno vrijednost koja se najčešće pojavljuje u podacima. Za kontinuirane podatke (poput minuta korištenja) mod nije osobito koristan jer svaka vrijednost može biti jedinstvena. Ali za kategoričke podatke, mod je savršena mjera. Zapravo, mod je **jedina** smislena mjera centralne tendencije za kategoričke podatke jer ne možete izračunati prosjek ili medijan kategorija poput spola ili vrste medija.

Na primjer, koji je najčešći tip korisnika u našem uzorku po dobi?

```
tiktok |>
  count(age_group, sort = TRUE)
```

```
# A tibble: 4 x 2
  age_group     n
  <chr>     <int>
1 18-24       102
2 25-34        86
3 35-44        62
4 45+          50
```

Modalna kategorija je 18 do 24 jer ta dobna skupina ima najviše ispitanika. To je logično jer su mladi ljudi dominantna publika TikToka, pa ih je u anketi bilo najlakše regrutirati.

Pogledajmo i mod za razinu obrazovanja i spol.

```
tiktok |>
  count(education, sort = TRUE)
```

```
# A tibble: 2 x 2
  education     n
  <chr>     <int>
1 employed    198
2 student     102
```

```
tiktok |>
  count(gender, sort = TRUE)
```

```
# A tibble: 2 x 2
  gender     n
  <chr>  <int>
1 female   157
2 male     143
```

Jedna stvar koju vrijedi napomenuti o modu jest da distribucija može imati više modova. Kad distribucija ima dva vrha, kažemo da je **bimodalna**. To se u praksi događa kad su u uzorku pomiješane dvije različite populacije. Na primjer, ako bismo gledali distribuciju dnevnog korištenja TikToka za cijeli uzorak (bez razdvajanja po dobi), mogli bismo vidjeti dva vrha. Jedan je za mlade korisnike (oko 100 minuta) i drugi za starije (oko 15 minuta). To je signal da ukupna distribucija zapravo skriva dvije različite grupe, što je izuzetno korisna informacija.

3.5 Kada koristiti koju mjeru?

Ovo je pitanje na koje studenti često žele jednostavan odgovor, ali odgovor zapravo ovisi o kontekstu. Ipak, postoje neka korisna pravila koja se izvode iz matematičkih svojstava svake mjere.

Za **numeričke podatke koji su približno simetrično distribuirani** (nemaju dugačke repove na jednoj strani), aritmetička sredina je sasvim dobra mjera. Ona koristi sve podatke i statistička teorija se u velikoj mjeri oslanja na nju. Ako nemate razloga za sumnju u ekstremne vrijednosti, koristite sredinu.

Za **numeričke podatke s izrazitim ekstremnim vrijednostima** (poput prihoda, cijena nekretnina, broja pratitelja na društvenim mrežama ili broja dijeljenja objave), medijan je pouzdaniji izbor. Alternativno, možete koristiti skraćenu sredinu koja je kompromis između robusnosti i informativnosti.

Za **kategoričke podatke**, mod je jedina smislena opcija jer ne možete izračunati prosjek spola ili vrste medija. To se čini očitim, ali iznenađujuće je koliko se često u izvještajima pokušavaju interpretirati prosjeci Likertove skale (na primjer, prosječan odgovor 3.7 na skali od 1 do 5) kao da su smisleni. Tehnički, Likertove skale su ordinalne varijable, i prosjek ordinalnih podataka je diskutabilan. U praksi se to ipak često radi, ali vrijedi biti svjestan ograničenja.

> **!** Važna napomena
>
> Nikada nemojte izvijestiti samo jednu mjeru centralne tendencije. Kad pišete izvještaj ili analizu, dobra praksa je navesti i prosjek i medijan za numeričke varijable. Ako su slični, distribucija je vjerojatno prilično simetrična. Ako se razlikuju, to je signal da se u podacima nešto zanimljivo događa i vrijedi istražiti dalje. U akademskim radovima iz komunikologije standardno se navode sredina i standardna devijacija za sve ključne varijable, obično u tablici deskriptivnih statistika.

---

4 Mjere varijabilnosti

Znati gdje se podaci nalaze je tek pola priče. Jednako je važno znati koliko su podaci raspršeni. Navarro u knjizi koristi lijep primjer. Zamislite da vam kažem da je prosječna temperatura u dva grada jednaka, recimo 15°C. Na temelju te informacije mogli biste pomisliti da su ta dva grada klimatski slična. Ali zamislite da u jednom gradu temperatura nikad ne padne ispod 10°C niti naraste iznad 20°C, dok u drugom temperatura varira od minus 20°C zimi do plus 45°C ljeti. Očito, to su potpuno različiti gradovi unatoč istoj prosječnoj temperaturi. Razlika je u varijabilnosti.

Isto vrijedi za medijske podatke. Zamislite dva medijska portala čiji članci u prosjeku dobivaju po 50 komentara. Na prvom portalu svaki članak dobiva između 45 i 55 komentara.

Na drugom portalu neki članci dobiju 0, a poneki 200. Prosjek je isti, ali situacija je potpuno drugačija. Upravo to razlikuju mjere varijabilnosti.

4.1 Raspon

Najjednostavnija mjera varijabilnosti je raspon. To je razlika između najveće i najmanje vrijednosti u podacima. Izračunajmo raspon dnevnog korištenja TikToka.

```
tiktok |>
  summarise(
    minimum = min(daily_minutes),
    maksimum = max(daily_minutes),
    raspon = max(daily_minutes) - min(daily_minutes)
  )
```

```
# A tibble: 1 x 3
  minimum maksimum raspon
    <dbl>    <dbl>  <dbl>
1       7      140    133
```

Raspon nam govori da se dnevno korištenje kreće od jedva desetak minuta do gotovo dva i pol sata, što je ogromna razlika. Međutim, raspon ima isti problem kao aritmetička sredina, samo s druge strane. Ovisi samo o dvije najekstremnije vrijednosti i potpuno ignorira sve ostale. Ako se u uzorku pojavi jedna osoba koja koristi TikTok 12 sati dnevno, raspon će eksplodirati, a svi ostali podaci ostaju isti. Raspon je koristan kao brzi orijentir, ali za ozbiljnu analizu trebamo nešto bolje.

4.2 Prosječno apsolutno odstupanje

Prije nego prijeđemo na varijancu, vrijedi se nakratko zadržati na jednoj mjeri koja je konceptualno jednostavnija, a to je **prosječno apsolutno odstupanje** (average absolute deviation, AAD). Ideja je jednostavna. Za svako opažanje izračunamo koliko se razlikuje od aritmetičke sredine (to je odstupanje ili devijacija), uzmemo apsolutnu vrijednost tog odstupanja (jer nas zanima veličina odstupanja, ne smjer), i izračunamo prosjek svih apsolutnih odstupanja.

$$\text{AAD} = \frac{1}{N}\sum_{i=1}^{N}|X_i - \bar{X}|$$

Zašto ovo uopće spominjemo? Zato što je AAD puno intuitivniji od varijance. Kad kažete da je prosječno apsolutno odstupanje 30 minuta, to doslovno znači da se prosječni ispitanik razlikuje od sredine za oko 30 minuta. To je vrlo jednostavno za interpretirati.

R nema ugrađenu funkciju za AAD, ali ga možemo lako izračunati.

```r
tiktok |>
  summarise(
    prosjek = mean(daily_minutes),
    aad = mean(abs(daily_minutes - mean(daily_minutes)))
  )
```

```
# A tibble: 1 x 2
  prosjek   aad
    <dbl> <dbl>
1    56.9  33.0
```

Problem s AAD-om je da apsolutna vrijednost matematički nije ugodna za rad. Nije diferencijabilna u nuli, što otežava izvođenje formula u statistici. Zato su statističari davno odlučili koristiti kvadrate umjesto apsolutnih vrijednosti, i tako smo dobili varijancu. Ta odluka ima duboke posljedice za cijelu statistiku, ali za naše potrebe dovoljno je znati da varijanca postoji zato što su kvadrati matematički elegantniji od apsolutnih vrijednosti, čak i ako su manje intuitivni.

4.3 Varijanca

Varijanca je sofisticiranija mjera raspršenosti. Ideja je sljedeća. Uzmemo svaku vrijednost u podacima i izračunamo koliko se razlikuje od aritmetičke sredine. To se zove **odstupanje** (deviation). Zatim ta odstupanja kvadriramo (jer bi se pozitivna i negativna inače poništila) i izračunamo njihov prosjek. Ili, točnije, gotovo prosjek.

Pogledajmo najprije formulu, pa ćemo razjasniti taj gotovo prosjek.

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

4.3.1 Zašto dijelimo s $N-1$? Besselova korekcija

Primijetite da dijelimo s $N-1$, a ne s N. Ovo je jedan od onih detalja koji studente redovito zbunjuju, a vrijedi razumjeti zašto je tako. Dijeljenje s $N-1$ zove se **Besselova korekcija** i postoji iz razloga koji su vezani uz procjenu populacijskih parametara iz uzorka.

Evo intuicije. Kad računamo varijancu uzorka, koristimo sredinu uzorka $\bar{X}$ kao procjenu populacijske sredine μ. Ali sredina uzorka je izračunata iz istih podataka iz kojih računamo odstupanja. To stvara suptilni problem. Odstupanja od sredine uzorka su sustavno manja nego što bi bila odstupanja od prave populacijske sredine, jer je sredina uzorka po definiciji najbliža moguća vrijednost tim konkretnim podacima. Dijeljenje s $N-1$ umjesto N ispravlja tu pristranost i daje **nepristranu procjenu** populacijske varijance.

Ako vam ovo zvuči apstraktno, ne brinite previše. Za velike uzorke (N > 30) razlika između dijeljenja s N i $N-1$ je minimalna. Ali za male uzorke može biti značajna, pa se konvencija $N-1$ koristi uvijek. R-ova funkcija `var()` automatski koristi $N-1$.

Navarro u knjizi posvećuje dosta prostora objašnjavanju ovog koncepta i iskreno kaže da je to jedan od najtežih dijelova uvodnog kolegija statistike. Mi ćemo se na ovu temu detaljno vratiti kad budemo govorili o uzorcima i populacijama u kasnijim tjednima. Za sada je dovoljno zapamtiti da `var()` u R u radi ono što treba.

```
tiktok |>
  summarise(
    varijanca = var(daily_minutes)
  )
```

```
# A tibble: 1 x 1
  varijanca
      <dbl>
1     1487.
```

Problem s varijancom je što je teško interpretirati. Mjerna jedinica varijance je kvadrat izvorne mjerne jedinice, dakle u našem slučaju minute na kvadrat. Što to uopće znači, minute na kvadrat? Ništa intuitivno. Zato postoji standardna devijacija.

4.3.2 Ručno izračunavanje varijance korak po korak

Korisno je barem jednom vidjeti kako se varijanca računa ručno, korak po korak, čak i ako to u praksi nikad nećemo raditi. Ovo pomaže izgraditi intuiciju.

```
# Uzmimo mali podskup podataka za demonstraciju
demo <- tiktok |>
  slice(1:6) |>
  select(respondent_id, daily_minutes)

demo |>
  mutate(
    sredina = mean(daily_minutes),
    odstupanje = daily_minutes - mean(daily_minutes),
    kvadrirano_odstupanje = odstupanje^2
  )
```

```
# A tibble: 6 x 5
  respondent_id daily_minutes sredina odstupanje kvadrirano_odstupanje
          <dbl>         <dbl>   <dbl>      <dbl>                 <dbl>
1             1            95      69         26                   676
```

2	2	78	69	9	81
3	3	112	69	43	1849
4	4	45	69	-24	576
5	5	62	69	-7	49
6	6	22	69	-47	2209

Varijanca je zbroj svih kvadriranih odstupanja podijeljen s $N-1$. Vidimo da veća odstupanja (pozitivna ili negativna) imaju neproporcionalno velik utjecaj na varijancu jer se kvadriraju. To je upravo razlog zašto je varijanca (i standardna devijacija) osjetljiva na ekstremne vrijednosti.

4.4 Standardna devijacija

Standardna devijacija je jednostavno korijen iz varijance.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Njezina prednost je što je u istim mjernim jedinicama kao i izvorni podaci. Ako je standardna devijacija dnevnog korištenja TikToka 40 minuta, to znači da se ispitanici u prosjeku razlikuju od aritmetičke sredine za otprilike 40 minuta. To nije savršena interpretacija (prisjetite se, standardna devijacija koristi kvadrate, ne apsolutne vrijednosti, pa nije identična prosječnom apsolutnom odstupanju), ali je dovoljno dobra za intuitivno razumijevanje.

```
tiktok |>
  summarise(
    prosjek = mean(daily_minutes),
    sd = sd(daily_minutes),
    medijan = median(daily_minutes),
    aad = mean(abs(daily_minutes - mean(daily_minutes)))
  )
```

```
# A tibble: 1 x 4
  prosjek    sd medijan   aad
    <dbl> <dbl>   <dbl> <dbl>
1    56.9  38.6      50  33.0
```

Primijetite da je standardna devijacija nešto veća od prosječnog apsolutnog odstupanja. To je uvijek tako, jer kvadriranje daje veću težinu velikim odstupanjima.

Standardna devijacija je daleko najčešće korištena mjera varijabilnosti u znanosti, uključujući komunikologiju. Kad u akademskom radu vidite tablicu deskriptivnih statistika, gotovo uvijek će sadržavati sredinu (M) i standardnu devijaciju (SD) za svaku varijablu.

4.4.1 Interpretacija standardne devijacije

Za podatke koji su približno normalno distribuirani, vrijedi korisno pravilo palca. Otprilike 68% podataka nalazi se unutar jedne standardne devijacije od sredine. Otprilike 95% podataka nalazi se unutar dvije standardne devijacije. Otprilike 99.7% unutar tri standardne devijacije.

Ovo se ponekad naziva pravilo 68-95-99.7 ili empirijsko pravilo. Ako je sredina 55 i SD 40, onda se otprilike 68% ispitanika nalazi između 15 i 95 minuta. Naravno, ovo pravilo vrijedi samo za normalno distribuirane podatke, a naši podaci o TikToku sigurno nisu savršeno normalno distribuirani. Ali i tada, pravilo daje koristan okvirni uvid.

4.5 Interkvartilni raspon

Baš kao što medijan ima prednost nad aritmetičkom sredinom kod ekstremnih vrijednosti, tako i **interkvartilni raspon** (IQR) ima prednost nad standardnom devijacijom. Da bismo razumjeli IQR, moramo najprije razumjeti percentile i kvartile.

4.5.1 Percentili i kvartili

Percentil je vrijednost ispod koje se nalazi određeni postotak podataka. 25. percentil (ili prvi kvartil, Q1) je vrijednost ispod koje se nalazi 25% podataka. 50. percentil je medijan. 75. percentil (treći kvartil, Q3) je vrijednost ispod koje se nalazi 75% podataka.

Kvartili dijele podatke na četiri jednaka dijela, baš kao što medijan dijeli na dva.

```
tiktok |>
  summarise(
    Q1 = quantile(daily_minutes, 0.25),
    medijan = quantile(daily_minutes, 0.50),
    Q3 = quantile(daily_minutes, 0.75),
    IQR = IQR(daily_minutes)
  )
```

```
# A tibble: 1 x 4
     Q1 medijan    Q3   IQR
  <dbl>   <dbl> <dbl> <dbl>
1    21      50    90    69
```

IQR je razlika između Q3 i Q1, dakle raspon unutar kojeg se nalaze srednjih 50% podataka. To je robusna mjera koja neće skočiti zbog jedne ekstremne vrijednosti.

Možemo izračunati i detaljnije percentile.

```r
percentili <- c(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)

tibble(
  percentil = percentili * 100,
  vrijednost = quantile(tiktok$daily_minutes, percentili)
)
```

```
# A tibble: 7 x 2
  percentil vrijednost
      <dbl>      <dbl>
1         5       9.95
2        10      11
3        25      21
4        50      50
5        75      90
6        90     114.
7        95     124.
```

Ova tablica nam daje bogatu sliku o distribuciji. Vidimo da 5% ispitanika koristi TikTok manje od pedesetak minuta (najniži percentil), a 5% koristi više od gornjeg percentila. Srednja polovica uzorka (od 25. do 75. percentila) pokriva raspon koji nam daje IQR.

> Praktični savjet
>
> Kad opisujete podatke u izvještaju ili radu, kombinirajte mjere centralne tendencije i varijabilnosti. Dobar opis bi glasio otprilike ovako. Ispitanici u prosjeku koriste TikTok M minuta dnevno (SD = X), a medijan je Y minuta. Srednja polovica ispitanika provodi na platformi između Q1 i Q3 minuta dnevno (IQR = Z). Ova kombinacija sredine, SD, medijana i IQR daje čitatelju bogatu sliku podataka u samo dvije rečenice.

4.6 Koja mjera varijabilnosti?

Izbor mjere varijabilnosti prati istu logiku kao izbor mjere centralne tendencije. Ako ste izabrali sredinu, standardna devijacija je prirodan par jer obje koriste iste matematičke principe (kvadrate odstupanja). Ako ste izabrali medijan, IQR je prirodan par jer su obje robusne mjere.

U praksi, u akademskim radovima gotovo uvijek vidite sredinu i SD. To je konvencija. Ali to ne znači da je to uvijek najbolji izbor. Za podatke koji su jako asimetrični (što je čest slučaj s medijskim metrikama poput broja dijeljenja, broja pratitelja, ili vremena na stranici), medijan i IQR su informativniji.

---

5 Ukupni sažetak varijable

U praksi, kad dobijete novi dataset, prva stvar koju želite napraviti je brzi pregled svih varijabli. Umjesto da za svaku varijablu posebno računate sredinu, SD, medijan i tako dalje, R nudi načine da to napravite odjednom.

Osnovna funkcija `summary()` daje brzi pregled.

```
tiktok |>
  select(daily_minutes, weekly_videos_watched, trust_score) |>
  summary()
```

```
 daily_minutes    weekly_videos_watched  trust_score
 Min.   :  7.00   Min.   : 20.0         Min.   :2.000
 1st Qu.: 21.00   1st Qu.: 65.0         1st Qu.:3.000
 Median : 50.00   Median :159.0         Median :4.000
 Mean   : 56.85   Mean   :189.6         Mean   :4.497
 3rd Qu.: 90.00   3rd Qu.:300.0         3rd Qu.:6.000
 Max.   :140.00   Max.   :500.0         Max.   :8.000
```

Funkcija `summary()` za numeričke varijable automatski ispisuje minimum, prvi kvartil, medijan, sredinu, treći kvartil i maksimum. To je tzv. **five-number summary** (plus sredina), i daje solidan pregled distribucije.

Još moćniji pristup je koristiti `summarise()` s `across()` da izračunamo točno one statistike koje želimo, za sve numeričke varijable odjednom.

```
tiktok |>
  summarise(
    across(
      where(is.numeric),
      list(
        prosjek = ~mean(.x, na.rm = TRUE),
        sd = ~sd(.x, na.rm = TRUE),
        medijan = ~median(.x, na.rm = TRUE)
      ),
      .names = "{.col}_{.fn}"
    )
  ) |>
  pivot_longer(
    everything(),
    names_to = c("varijabla", "statistika"),
    names_sep = "_(?=[^_]+$)",
    values_to = "vrijednost"
  ) |>
```

```
  pivot_wider(
    names_from = statistika,
    values_from = vrijednost
  )
```

```
# A tibble: 8 x 4
  varijabla             prosjek     sd medijan
  <chr>                   <dbl>  <dbl>   <dbl>
1 respondent_id          150.    86.7     150.
2 age                     32.1   10.6      30
3 daily_minutes           56.9   38.6      50
4 weekly_videos_watched  190.   139.      159
5 likes_given             23.5   20.2      17
6 comments_posted          1.91   2.23      1
7 follows_creators         7.08   5.39      6
8 trust_score              4.50   1.85      4
```

Ovaj kod izgleda komplicirano, ali radi nešto vrlo korisno. Za svaku numeričku varijablu izračunava sredinu, SD i medijan, te rezultate prikazuje u čitljivoj tablici. Funkcija `across()` primjenjuje iste izračune na sve odabrane stupce, a `pivot_longer()` i `pivot_wider()` preoblikuju rezultat u preglednu formu. Ovo je obrazac koji ćete koristiti iznova i iznova, pa ga vrijedi zapamtiti (ili još bolje, spremiti u skriptu za ponovnu upotrebu).

---

6 Deskriptivne statistike po grupama

Ukupne statistike su korisne, ali prava snaga deskriptivne analize dolazi do izražaja kad podatke razbijemo po grupama. U komunikologiji nas gotovo uvijek zanima usporedba. Koriste li žene i muškarci TikTok jednako? Razlikuju li se dobne skupine? Ovisi li povjerenje u sadržaj o intenzitetu korištenja?

Tidyverse čini ovu vrstu analize izuzetno elegantnom. Kombinacija `group_by()` i `summarise()` je jedan od najmoćnijih alata koje ćete naučiti u ovom kolegiju. Logika je jednostavna. `group_by()` podijeli podatke u grupe, a `summarise()` izračuna statistike za svaku grupu zasebno. U pozadini, R ponavlja identičan izračun za svaki podskup podataka i rezultate slaže u jednu tablicu.

Pogledajmo kako se korištenje TikToka razlikuje po dobnim skupinama.

```
tiktok |>
  group_by(age_group) |>
  summarise(
    n = n(),
    prosjek = mean(daily_minutes),
    sd = sd(daily_minutes),
    medijan = median(daily_minutes),
    min = min(daily_minutes),
    max = max(daily_minutes),
    .groups = "drop"
  ) |>
  arrange(desc(prosjek))
```

```
# A tibble: 4 x 7
  age_group     n prosjek    sd medijan   min   max
  <chr>     <int>   <dbl> <dbl>   <dbl> <dbl> <dbl>
1 18-24       102   104.  17.1    104      72   140
2 25-34        86    52.6  7.78    52.5    40    68
3 35-44        62    22.3  2.70    22      18    30
4 45+          50    10.7  2.17    11       7    15
```

Ovdje se jasno vidi ono što smo slutili iz sirovih podataka. Najmlađa skupina (18 do 24 godine) koristi TikTok u prosjeku preko 100 minuta dnevno, dok najstarija skupina (45+) provodi na platformi tek desetak minuta. Razlika je dramatična.

Primijetite i nešto važno. Standardna devijacija unutar svake grupe je znatno manja nego u ukupnom uzorku. To je zato što smo grupiranjem uklonili najveći izvor varijabilnosti, a to je upravo dob. Ova tehnika grupiranja ključna je za razumijevanje podataka. Kad god vidite veliku standardnu devijaciju u ukupnom uzorku, prvo što biste trebali učiniti jest pogledati postoji li neka grupna varijabla koja objašnjava tu varijabilnost. U našem slučaju, dob objašnjava najveći dio razlika u korištenju TikToka.

Možemo ići i korak dalje te kombinirati dva kriterija grupiranja. Pogledajmo korištenje po dobnoj skupini i spolu.

```
tiktok |>
  group_by(age_group, gender) |>
  summarise(
    n = n(),
    prosjek = round(mean(daily_minutes), 1),
    sd = round(sd(daily_minutes), 1),
    .groups = "drop"
  ) |>
  arrange(age_group, gender)
```

```
# A tibble: 8 x 5
  age_group gender     n prosjek    sd
  <chr>     <chr>  <int>   <dbl> <dbl>
1 18-24     female    58   117.   11.1
2 18-24     male      44    87.6   6
3 25-34     female    45    54.8   8.8
4 25-34     male      41    50.1   5.7
5 35-44     female    30    21.8   2.4
6 35-44     male      32    22.8   2.9
7 45+       female    24    10.7   2.2
8 45+       male      26    10.8   2.2
```

Vidimo da unutar svake dobne skupine žene u prosjeku koriste TikTok nešto više nego muškarci. Ta razlika je konzistentna kroz sve dobne skupine, ali je relativno mala u usporedbi s razlikom između samih dobnih skupina. To je upravo ona vrsta uvida koju dobivate kad pravilno razbijete podatke po relevantnim kategorijama.

> **!** Važna napomena
>
> Argument `.groups = "drop"` na kraju `summarise()` poziva služi tome da R ukloni grupiranje nakon izračuna. Bez njega, rezultirajući tibble bi ostao grupiran po `age_group` (jer je `summarise()` automatski uklonio samo zadnju razinu grupiranja, a `gender` je zadnja). To može uzrokovati neočekivano ponašanje u kasnijim operacijama. Dobra praksa je uvijek eksplicitno navesti ovaj argument kad koristite više od jedne grupirajuće varijable.

Pogledajmo i deskriptivne statistike za varijablu povjerenja u sadržaj (trust_score) po dobnim skupinama.

```
tiktok |>
  group_by(age_group) |>
  summarise(
    n = n(),
    prosjek_trust = round(mean(trust_score), 1),
    sd_trust = round(sd(trust_score), 1),
    medijan_trust = median(trust_score),
    .groups = "drop"
  )
```

```
# A tibble: 4 x 5
  age_group     n prosjek_trust sd_trust medijan_trust
  <chr>     <int>         <dbl>    <dbl>         <dbl>
1 18-24       102           6.7      0.8             7
2 25-34        86           4.5      0.5             4
3 35-44        62           3        0               3
4 45+          50           2        0               2
```

Zanimljivo je da povjerenje u sadržaj prati sličan obrazac kao i korištenje. Mladi korisnici imaju veće povjerenje u TikTok sadržaj. Čini se da što više vremena netko provodi na platformi, to više vjeruje sadržaju koji tamo pronalazi. Ovo bi moglo biti zanimljivo za daljnje istraživanje, ali budite oprezni s uzročno-posljedičnim zaključcima. Korelacija nije uzročnost, a do tog pojma stižemo uskoro.

---

7 Oblik distribucije: asimetrija i zaobljenost

Osim centralne tendencije i varijabilnosti, treća važna karakteristika podataka je **oblik distribucije**. Dva najvažnija aspekta oblika su asimetrija (skewness) i zaobljenost (kurtosis).

7.1 Asimetrija (skewness)

Distribucija je **simetrična** kad lijeva i desna strana izgledaju kao zrcalna slika. Normalna distribucija (ona poznata zvonolika krivulja) je savršeno simetrična. U praksi, savršena simetrija je rijetka.

Kad distribucija ima dugačak rep prema desno (više ekstremno visokih vrijednosti), kažemo da je **pozitivno asimetrična** (right-skewed ili positively skewed). Kad ima dugačak rep prema lijevo, kažemo da je **negativno asimetrična** (left-skewed ili negatively skewed).

Matematička definicija asimetrije koristi treći standardizirani moment

$$\text{skewness} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{s}\right)^3$$

Primijetite da je eksponent 3, a ne 2 kao kod varijance. Budući da kubiranje čuva predznak (negativan broj na treću ostaje negativan), ovaj izraz je pozitivan kad distribucija ima rep prema desno (jer veliki pozitivni residuali dominiraju) i negativan kad ima rep prema lijevo.

U komunikologiji se pozitivna asimetrija pojavljuje vrlo često. Broj pratitelja na društvenim mrežama, broj dijeljenja objave, prihod od oglašavanja, vrijeme provedeno na web stranici, broj komentara na članku, sve su to varijable koje tipično imaju pozitivnu asimetriju. Većina vrijednosti je relativno mala, ali postoji dugačak rep velikih vrijednosti. To je gotovo univerzalna karakteristika metrika angažmana u digitalnim medijima i vrijedi je zapamtiti kao opće pravilo.

Jednostavan način da provjerite asimetriju jest usporedba aritmetičke sredine i medijana. Ako je sredina veća od medijana, distribucija je vjerojatno pozitivno asimetrična. Ako je medijan veći, negativno asimetrična.

```r
tiktok |>
  summarise(
    across(
      c(daily_minutes, weekly_videos_watched, likes_given, comments_posted),
      list(
        prosjek = ~mean(.x),
        medijan = ~median(.x),
        razlika = ~mean(.x) - median(.x)
      ),
      .names = "{.col}_{.fn}"
    )
  ) |>
  pivot_longer(
    everything(),
    names_to = c("varijabla", "statistika"),
    names_sep = "_(?=[^_]+$)",
    values_to = "vrijednost"
  ) |>
  pivot_wider(names_from = statistika, values_from = vrijednost)
```

```
# A tibble: 4 x 4
  varijabla             prosjek medijan razlika
  <chr>                   <dbl>   <dbl>   <dbl>
1 daily_minutes           56.9       50   6.85
2 weekly_videos_watched  190.       159  30.6
3 likes_given             23.5       17   6.54
4 comments_posted          1.91       1   0.913
```

Pozitivna razlika između prosjeka i medijana za sve varijable potvrđuje da su distribucije pozitivno asimetrične. To je očekivano jer u uzorku postoji skupina mladih korisnika koji imaju izrazito visoke vrijednosti na svim metrikama korištenja.

7.2 Zaobljenost (kurtosis)

Zaobljenost opisuje koliko su repovi distribucije teški u usporedbi s normalnom distribucijom. Koristi četvrti standardizirani moment

$$\text{kurtosis} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{s}\right)^4$$

Za normalnu distribuciju, kurtosis iznosi 3. Zato se često koristi **višak zaobljenosti** (excess kurtosis) koji oduzima 3, pa normalna distribucija ima excess kurtosis jednak 0.

Distribucija s velikom zaobljenošću (leptokurtic, excess kurtosis > 0) ima teže repove i oštriji vrh od normalne. To znači da ima više ekstremnih vrijednosti nego što bismo očekivali. Distribucija s malom zaobljenošću (platykurtic, excess kurtosis < 0) ima lakše repove i zaobljeniji vrh.

Za svakodnevni rad u komunikologiji, zaobljenost je manje važna od asimetrije. Ipak, vrijedi ju poznavati jer se pojavljuje u izvještajima statističkog softvera i u akademskim radovima. Najvažnija praktična implikacija je da distribucija s velikom zaobljenošću ima više ekstremnih vrijednosti nego što bismo očekivali od normalne distribucije, što može utjecati na rezultate statističkih testova koji pretpostavljaju normalnost.

> **Praktični savjet**
>
> Asimetrija i zaobljenost postaju osobito važne kad počnemo raditi inferencijsku statistiku (t-testove, ANOVA-u, regresiju) jer mnogi od tih testova pretpostavljaju normalnost distribucije. U tom kontekstu, asimetrija i zaobljenost služe kao dijagnostički alati za provjeru pretpostavki. O tome ćemo detaljno govoriti u kasnijim tjednima.

---

8 Standardni rezultati (z-scores)

Ponekad trebamo usporediti vrijednosti na potpuno različitim skalama. Na primjer, kako usporediti nekoga tko provodi 120 minuta dnevno na TikToku s nekim tko ima trust_score 8? To su različite varijable s različitim mjernim jedinicama i različitim rasponima. Standardni rezultati (z-scores) rješavaju taj problem.

Standardni rezultat nam govori koliko se standardnih devijacija neka vrijednost nalazi iznad ili ispod aritmetičke sredine. Formula je

$$z_i = \frac{X_i - \bar{X}}{s}$$

Ako je z-score jednak 0, ta vrijednost je jednaka prosjeku. Ako je 1, ta vrijednost je jednu standardnu devijaciju iznad prosjeka. Ako je minus 2, ta je vrijednost dvije standardne devijacije ispod prosjeka.

Z-scores su korisni iz još jednog razloga. Kad pretvorite varijablu u z-scores, rezultirajuća varijabla uvijek ima sredinu 0 i standardnu devijaciju 1. To se zove **standardizacija** i često se koristi u naprednim statističkim metodama.

Izračunajmo z-score za dnevno korištenje TikToka. Možemo to napraviti ručno (koristeći formulu) ili pomoću ugrađene funkcije `scale()`.

```
tiktok |>
  mutate(
    z_rucno = (daily_minutes - mean(daily_minutes)) / sd(daily_minutes),
    z_scale = as.numeric(scale(daily_minutes))
  ) |>
  select(respondent_id, age_group, daily_minutes, z_rucno, z_scale) |>
  head(10)
```

```
# A tibble: 10 x 5
   respondent_id age_group daily_minutes z_rucno z_scale
           <dbl> <chr>              <dbl>   <dbl>   <dbl>
 1             1 18-24                 95  0.989   0.989
 2             2 18-24                 78  0.548   0.548
 3             3 18-24                112  1.43    1.43
 4             4 25-34                 45 -0.307  -0.307
 5             5 25-34                 62  0.133   0.133
 6             6 35-44                 22 -0.904  -0.904
 7             7 18-24                130  1.90    1.90
 8             8 18-24                 88  0.808   0.808
 9             9 25-34                 55 -0.0481 -0.0481
10            10 35-44                 18 -1.01   -1.01
```

Obje metode daju identične rezultate. Funkcija `scale()` vraća matricu, pa koristimo `as.numeric()` da pretvorimo rezultat u obični numerički vektor.

Sada vidimo da osoba s 95 minuta dnevno ima pozitivan z-score (iznad prosjeka), dok osoba s 22 minute ima negativan z-score (ispod prosjeka). Osoba čiji je z-score oko 2 nalazi se dvije standardne devijacije iznad prosjeka, što je prilično ekstremna vrijednost.

> Z-scores su poput zajedničkog jezika za različite varijable. Svaki put kad pretvorite podatke u z-scores, omogućujete usporedbu jabuka i naranči.

Pogledajmo kako z-scores izgledaju kad ih izračunamo unutar svake dobne skupine (što je ponekad smislenije nego ukupni z-score).

```
tiktok |>
  group_by(age_group) |>
  mutate(
    z_unutar_grupe = as.numeric(scale(daily_minutes))
  ) |>
  ungroup() |>
  select(respondent_id, age_group, daily_minutes, z_unutar_grupe) |>
  head(12)
```

```
# A tibble: 12 x 4
   respondent_id age_group daily_minutes z_unutar_grupe
           <dbl> <chr>             <dbl>          <dbl>
 1             1 18-24                95        -0.531
 2             2 18-24                78        -1.52
 3             3 18-24               112         0.463
 4             4 25-34                45        -0.972
 5             5 25-34                62         1.21
 6             6 35-44                22        -0.114
 7             7 18-24               130         1.52
 8             8 18-24                88        -0.940
 9             9 25-34                55         0.314
10            10 35-44                18        -1.60
11            11 18-24               105         0.0539
12            12 18-24                72        -1.88
```

Sada z-score govori koliko se osoba razlikuje od prosjeka **svoje vlastite dobne skupine**, što je ponekad informativnije od ukupnog z-scorea. Na primjer, osoba od 19 godina koja koristi TikTok 95 minuta dnevno možda ima negativan z-score unutar skupine 18 do 24 (jer je ispod prosjeka te skupine), ali bi imala pozitivan z-score u ukupnom uzorku. Kontekst je važan.

---

9 Korelacije

Do sada smo opisivali jednu varijablu po jednu. Ali u istraživanjima nas često zanima **veza između dviju varijabli**. Postoji li povezanost između dobi ispitanika i vremena koje provode na TikToku? Jesu li ljudi koji više koriste platformu ujedno i oni koji joj više vjeruju?

9.1 Kovarijanca: temelj korelacije

Prije nego uđemo u korelaciju, vrijedi razumjeti koncept koji stoji iza nje. To je **kovarijanca**. Kovarijanca mjeri u kojoj mjeri dvije varijable variraju zajedno. Ako su obje varijable iznadprosječne za istog ispitanika i ispodprosječne za istog ispitanika, one pozitivno kovariraju. Ako jedna tendira biti iznadprosječna kad je druga ispodprosječna, negativno kovariraju.

Formula za kovarijancu je

$$\text{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

Primijetite sličnost s varijancom. Varijanca je zapravo kovarijanca varijable same sa sobom. Jedina razlika je da umjesto kvadriranja odstupanja jedne varijable, množimo odstupanja dviju različitih varijabli.

Problem s kovarijancom je isti kao s varijancom. Rezultat ovisi o mjernim jedinicama varijabli. Kovarijanca između dobi (u godinama) i korištenja TikToka (u minutama) bit će potpuno drugačija od kovarijance između dobi (u mjesecima) i korištenja TikToka (u satima), čak i ako je veza identična. Zato trebamo standardiziranu mjeru, a to je Pearsonov koeficijent korelacije.

9.2 Pearsonov koeficijent korelacije

Najčešća mjera linearne povezanosti dviju numeričkih varijabli je **Pearsonov koeficijent korelacije**, označen s r. To je zapravo standardizirana kovarijanca, što znači da kovarijancu podijelimo s produktom standardnih devijacija obiju varijabli.

$$r_{XY} = \frac{\text{Cov}(X, Y)}{s_X \cdot s_Y} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

Ako prepoznajete z-scores u ovoj formuli, u pravu ste. Korelacija je zapravo prosjek umnožaka z-scores dviju varijabli. To je elegantna definicija jer pokazuje da korelacija zapravo mjeri u kojoj mjeri dvije varijable variraju zajedno, nakon što smo obje stavili na istu skalu.

Vrijednost korelacije kreće se od minus 1 do plus 1. Korelacija blizu plus 1 znači snažnu pozitivnu linearnu vezu (kad jedna varijabla raste, raste i druga). Korelacija blizu minus 1 znači snažnu negativnu linearnu vezu (kad jedna raste, druga pada). Korelacija blizu 0 znači da linearna veza ne postoji ili je vrlo slaba.

Izračunajmo korelaciju između dobi i dnevnog korištenja.

```
tiktok |>
  summarise(
    kovarijanca = cov(age, daily_minutes),
    korelacija = cor(age, daily_minutes)
  )
```

```
# A tibble: 1 x 2
  kovarijanca korelacija
        <dbl>      <dbl>
1       -380.     -0.925
```

Korelacija je snažno negativna, što znači da stariji ispitanici koriste TikTok manje. Kovarijanca je negativna i velika, ali njezina apsolutna vrijednost nam ne govori ništa korisno bez konteksta jer ovisi o mjernim jedinicama. Korelacija od minus 0.9 (ili koliko god iznosi) odmah nam govori da je veza snažna.

9.3 Interpretacija korelacija

Jedna od najčešćih pitanja je koliko velika mora biti korelacija da bismo je smatrali značajnom ili važnom. Cohen (1988) je predložio sljedeće smjernice koje se još uvijek široko koriste.

Za korelacije oko $|r| = 0.10$ kažemo da je veza **slaba** (small). Za korelacije oko $|r| = 0.30$ kažemo da je **umjerena** (medium). Za korelacije oko $|r| = 0.50$ ili više kažemo da je **snažna** (large).

No, Navarro s pravom upozorava da su ove smjernice samo grubi orijentiri i da ovise o kontekstu. U nekim područjima (na primjer, u predviđanju ponašanja) korelacija od 0.30 je zapravo prilično impresivna. U drugim (na primjer, u procjeni pouzdanosti testa) korelacija od 0.50 može biti nedovoljna. Kontekst je uvijek ključan.

Pogledajmo više korelacija odjednom.

```
tiktok |>
  summarise(
    r_dob_minute = cor(age, daily_minutes),
    r_minute_trust = cor(daily_minutes, trust_score),
    r_minute_videos = cor(daily_minutes, weekly_videos_watched),
    r_dob_trust = cor(age, trust_score),
    r_likes_comments = cor(likes_given, comments_posted)
  )
```

```
# A tibble: 1 x 5
  r_dob_minute r_minute_trust r_minute_videos r_dob_trust r_likes_comments
         <dbl>          <dbl>           <dbl>       <dbl>            <dbl>
1       -0.925          0.988           0.998      -0.935            0.965
```

Korelacija između dnevnog korištenja i povjerenja u sadržaj je pozitivna i prilično snažna. Ljudi koji više koriste TikTok ujedno iskazuju veće povjerenje u sadržaj koji tamo pronalaze. Korelacija između minuta i broja pogledanih videozapisa je gotovo savršena, što je logično jer su to dvije strane iste medalje. Korelacija između lajkova i komentara je umjereno pozitivna, što sugerira da aktivniji korisnici tendiraju biti aktivni na više načina.

9.4 Matrica korelacija

Kad imamo više numeričkih varijabli, korisno je izračunati korelacije između svih parova odjednom. To daje **matricu korelacija**.

```
tiktok |>
  select(age, daily_minutes, weekly_videos_watched, likes_given,
         comments_posted, follows_creators, trust_score) |>
  cor() |>
  round(2)
```

```
                          age daily_minutes weekly_videos_watched likes_given
age                      1.00         -0.92                 -0.90       -0.86
daily_minutes           -0.92          1.00                  1.00        0.98
weekly_videos_watched   -0.90          1.00                  1.00        0.99
likes_given             -0.86          0.98                  0.99        1.00
comments_posted         -0.77          0.93                  0.94        0.97
follows_creators        -0.87          0.98                  0.99        0.99
trust_score             -0.93          0.99                  0.98        0.96
                        comments_posted follows_creators trust_score
age                               -0.77            -0.87       -0.93
daily_minutes                      0.93             0.98        0.99
weekly_videos_watched              0.94             0.99        0.98
likes_given                        0.97             0.99        0.96
comments_posted                    1.00             0.98        0.91
follows_creators                   0.98             1.00        0.97
trust_score                        0.91             0.97        1.00
```

Matrica korelacija je simetrična (korelacija između X i Y je jednaka korelaciji između Y i X) i na dijagonali su uvijek jedinice (svaka varijabla je savršeno korelirana sama sa sobom). Ovo su svojstva koja slijede direktno iz matematičke definicije korelacije.

9.5 Spearmanov koeficijent korelacije

Pearsonov koeficijent mjeri linearnu vezu. Ali što ako veza između varijabli postoji, ali nije linearna? Na primjer, možda korištenje TikToka i povjerenje u sadržaj rastu zajedno, ali ne linearno nego u obliku krivulje. U tom slučaju Pearsonov r može podcijeniti snagu veze.

Za takve situacije postoji **Spearmanov koeficijent korelacije** (ρ ili r_s). Spearmanova korelacija funkcionira tako da najprije pretvori podatke u rangove, a zatim izračuna Pearsonov koeficijent na rangovima. Budući da rangovi čuvaju redoslijed ali ne i udaljenosti, Spearmanova korelacija mjeri **monotonost** veze, to jest koliko dosljedno jedna varijabla raste kad druga raste (ili pada), neovisno o tome je li veza linearna.

U R u, Spearmanova korelacija se računa jednostavno dodavanjem argumenta `method = "spearman"` u funkciju `cor()`.

```
tiktok |>
  summarise(
    pearson = cor(age, daily_minutes, method = "pearson"),
    spearman = cor(age, daily_minutes, method = "spearman")
  )
```

```
# A tibble: 1 x 2
  pearson spearman
    <dbl>    <dbl>
1  -0.925   -0.972
```

Kad su Pearsonov i Spearmanov koeficijent slični, to sugerira da je veza približno linearna. Kad se razlikuju (osobito kad je Spearmanov veći od Pearsonovog), to sugerira da postoji monotona, ali nelinearna veza.

Spearmanova korelacija ima i dodatnu prednost. Manje je osjetljiva na ekstremne vrijednosti nego Pearsonova, jer radi s rangovima, a ne izvornim vrijednostima. Zato se ponekad koristi kao robusna alternativa Pearsonovom koeficijentu.

> Praktični savjet
>
> Kad radite s podacima za koje sumnjate da imaju nelinearnu vezu ili ekstremne vrijednosti, izračunajte i Pearsonov i Spearmanov koeficijent. Ako su slični, veza je vjerojatno linearna i bez problematičnih outliera. Ako se razlikuju, istražite dalje (najčešće pomoću scatterplota, o čemu ćemo govoriti sljedećeg tjedna).

9.6 Ograničenja korelacije

Korelacija je izuzetno korisna mjera, ali ima nekoliko važnih ograničenja koja morate poznavati. Navarro im u knjizi posvećuje značajan prostor, i to s dobrim razlogom.

Korelacija mjeri samo linearnu vezu. Ako je veza između dviju varijabli zakrivljena (na primjer, performanse rastu s vježbom ali se onda stabiliziraju), Pearsonov r može biti nizak čak i kad je veza vrlo snažna. Čak i Spearmanov koeficijent zahtijeva monotonost. Ako je veza U-oblik (na primjer, zadovoljstvo je nisko i pri vrlo niskom i pri vrlo visokom radnom opterećenju), ni jedna korelacija neće to uhvatiti.

Korelacija nije uzročnost. Ovo je toliko važno da zaslužuje poseban odlomak. Činjenica da je korištenje TikToka korelirano s povjerenjem u sadržaj ne znači da korištenje TikToka uzrokuje veće povjerenje (niti obratno). Moguće je da treća varijabla, poput dobi, objašnjava obje pojave. Mladi ljudi i više koriste TikTok i općenito imaju drugačiji odnos prema digitalnim medijima. Ovo je toliko čest problem da ima i ime, **confounding** (zbunjivanje) varijabli.

Navarro koristi sjajan primjer. Broj utopljavanja koreliran je s prodajom sladoleda. To ne znači da sladoled uzrokuje utapljanje. Treća varijabla (vrućina) objašnjava oboje, jer kad je vruće, ljudi i kupuju sladoled i idu plivati, a plivanje povećava rizik od utapljanja.

Ekstremne vrijednosti mogu drastično utjecati na korelaciju. Jedna ili dvije ekstremne točke mogu stvoriti iluziju korelacije tamo gdje je zapravo nema, ili maskirati korelaciju koja zapravo postoji.

Ograničeni raspon smanjuje korelaciju. Ako vaš uzorak pokriva samo uzak raspon jedne varijable, korelacija će biti niža nego u populaciji. Na primjer, ako istražujete vezu između IQ-a i akademskog uspjeha, ali vaš uzorak uključuje samo studente na elitnom sveučilištu (gdje svi imaju visok IQ), korelacija će biti niska jer nema dovoljno varijabilnosti u IQ-u.

> ❗ Važna napomena
>
> Kad god izračunate korelaciju, obavezno napravite i scatterplot. Postoje poznati primjeri (poput Anscombeovog kvarteta, a u novije vrijeme i Datasaurus Dozen) u kojima potpuno različiti skupovi podataka imaju identičnu korelaciju, a vizualno su potpuno različiti. Grafiku ćemo detaljno raditi sljedeći tjedan, ali zapamtite ovo pravilo od sada. Brojke bez grafike mogu lako zavarati.

---

10 Rad s nedostajućim vrijednostima

U stvarnom svijetu podaci gotovo nikad nisu potpuni. Ispitanici preskoče pitanje u anketi, senzor prestane raditi, sustav ne zabilježi klik. R koristi oznaku `NA` (not available) za nedostajuće vrijednosti, i ove vrijednosti zahtijevaju posebnu pažnju.

10.1 Kako R tretira nedostajuće vrijednosti

Problem je u tome što većina R funkcija za deskriptivnu statistiku vraća `NA` ako u podacima postoji ijedna nedostajuća vrijednost.

```
primjer <- c(10, 20, NA, 40, 50)

# Ovo vraća NA
mean(primjer)
```

```
[1] NA
```

```
# Ovo ignorira NA i računa prosjek od preostalih vrijednosti
mean(primjer, na.rm = TRUE)
```

```
[1] 30
```

To je zapravo dobro ponašanje jer vas prisiljava da svjesno odlučite što ćete učiniti s nedostajućim podacima. Ali u praksi, najčešće rješenje je dodati argument `na.rm = TRUE` koji govori R u da ignorira nedostajuće vrijednosti.

Unutar tidyverse pipeline, `na.rm = TRUE` se stavlja unutar svake funkcije u `summarise()`.

```r
tiktok |>
  summarise(
    prosjek = mean(daily_minutes, na.rm = TRUE),
    sd = sd(daily_minutes, na.rm = TRUE),
    medijan = median(daily_minutes, na.rm = TRUE)
  )
```

```
# A tibble: 1 x 3
  prosjek    sd medijan
    <dbl> <dbl>   <dbl>
1    56.9  38.6      50
```

10.2 Tipovi nedostajućih vrijednosti

Navarro u knjizi ne ulazi duboko u ovu temu, ali za komunikologe je korisno poznavati osnove. Postoje tri tipa nedostajućih vrijednosti.

MCAR (Missing Completely at Random) znači da je nedostajanje potpuno nasumično, nepovezano ni s jednom varijablom u podacima. Na primjer, ispitanik je slučajno preskočio pitanje jer mu je kliznuo prst na touchscreenu. U tom slučaju, ignoriranje nedostajućih vrijednosti (`na.rm = TRUE`) ne uvodi nikakvu pristranost.

MAR (Missing at Random) znači da je nedostajanje povezano s drugim varijablama u podacima, ali ne s nedostajućom vrijednošću samom. Na primjer, stariji ispitanici češće preskaču pitanja o TikToku jer smatraju da se to na njih ne odnosi. Nedostajanje je povezano s dobi, ali ne izravno s korištenjem TikToka. U tom slučaju, ignoriranje nedostajućih vrijednosti može uvesti pristranost, ali postoje statističke metode za korekciju.

MNAR (Missing Not at Random) znači da je nedostajanje izravno povezano s vrijednošću koja nedostaje. Na primjer, ljudi koji provode izrazito mnogo vremena na TikToku možda preskaču to pitanje jer ih je sram priznati koliko vremena tamo provode. Ovo je najproblematičniji slučaj jer ga je teško detektirati i ispraviti.

Za uvodni kolegij, dovoljno je biti svjestan da nedostajuće vrijednosti nisu uvijek nasumične. Kad god imate nedostajuće podatke, razmislite zašto nedostaju i je li sigurno jednostavno ih ignorirati.

10.3 Provjera nedostajućih vrijednosti

Prije bilo kakve analize uvijek provjerite koliko nedostajućih vrijednosti imate.

```r
tiktok |>
  summarise(across(everything(), ~sum(is.na(.x)))) |>
  pivot_longer(everything(), names_to = "varijabla", values_to = "broj_NA")
```

```
# A tibble: 11 x 2
   varijabla             broj_NA
   <chr>                   <int>
 1 respondent_id               0
 2 age                         0
 3 age_group                   0
 4 gender                      0
 5 daily_minutes               0
 6 weekly_videos_watched       0
 7 likes_given                 0
 8 comments_posted             0
 9 follows_creators            0
10 trust_score                 0
11 education                   0
```

U našem datasetu nema nedostajućih vrijednosti jer su podaci simulirani. Ali u stvarnom radu ih gotovo sigurno hoćete imati. Navikavanje na `na.rm = TRUE` od početka će vam uštedjeti mnogo frustracije.

> Praktični savjet
>
> Ako neka varijabla ima velik postotak nedostajućih vrijednosti (recimo više od 20%), razmislite treba li uopće koristiti tu varijablu u analizi. Također, umjesto `na.rm = TRUE`, ponekad je bolje koristiti `tidyr::drop_na()` na početku analize kako biste radili s kompletnim opažanjima. Razlika je u tome što `na.rm = TRUE` radi po varijabli (svaka statistika koristi sva dostupna opažanja), dok `drop_na()` eliminira cijele retke koji imaju bilo koju nedostajuću vrijednost. Koja opcija je bolja ovisi o konkretnoj situaciji.

---

11 Sve zajedno: kompletna deskriptivna analiza

Da bismo zaokružili ovo predavanje, napravimo kompletnu deskriptivnu analizu našeg dataseta o korištenju TikToka. Ovo je obrazac koji ćete koristiti na početku gotovo svake analize. Učitajte podatke, pogledajte strukturu, izračunajte deskriptivne statistike ukupno i po grupama, provjerite korelacije.

```
tiktok |>
  group_by(age_group) |>
  summarise(
    n = n(),
    prosjek_min = round(mean(daily_minutes), 1),
    sd_min = round(sd(daily_minutes), 1),
```

```
    medijan_min = median(daily_minutes),
    prosjek_trust = round(mean(trust_score), 1),
    sd_trust = round(sd(trust_score), 1),
    .groups = "drop"
  )
```

```
# A tibble: 4 x 7
  age_group     n prosjek_min sd_min medijan_min prosjek_trust sd_trust
  <chr>     <int>       <dbl>  <dbl>       <dbl>         <dbl>    <dbl>
1 18-24       102       104.    17.1       104             6.7      0.8
2 25-34        86        52.6    7.8        52.5           4.5      0.5
3 35-44        62        22.3    2.7        22             3        0
4 45+          50        10.7    2.2        11             2        0
```

Ova tablica u šest redova sažima informacije koje bi vam inače trebale stranice i stranice sirovih podataka. Vidimo jasne obrasce. Mladi (18 do 24) koriste TikTok oko 100 minuta dnevno s umjerenom varijabilnošću. Srednja skupina (25 do 34) koristi ga upola manje. Starije skupine jedva ga koriste, ali su unutar tih skupina ispitanici prilično ujednačeni (mala standardna devijacija). Povjerenje u sadržaj prati isti obrazac jer je snažno korelirano s intenzitetom korištenja.

Dopunimo ovu tablicu korelacijama unutar svake skupine.

```
tiktok |>
  group_by(age_group) |>
  summarise(
    n = n(),
    r_minute_trust = round(cor(daily_minutes, trust_score), 2),
    r_minute_videos = round(cor(daily_minutes, weekly_videos_watched), 2),
    .groups = "drop"
  )
```

```
# A tibble: 4 x 4
  age_group     n r_minute_trust r_minute_videos
  <chr>     <int>          <dbl>           <dbl>
1 18-24       102           0.94            1
2 25-34        86           0.86            1
3 35-44        62          NA               0.99
4 45+          50          NA               0.99
```

Ovaj korak je važan jer korelacije mogu biti različite u podskupovima nego u ukupnom uzorku. Na primjer, ukupna korelacija između korištenja i povjerenja može biti visoka dijelom zato što obje varijable koreliraju s dobi (mlađi koriste više i imaju veće povjerenje). Korelacija unutar svake dobne skupine govori nam postoji li veza i nakon što smo kontrolirali

za dob. Ovo je uvod u koncept kontrole varijabli koji ćemo detaljno obraditi kad budemo radili regresiju.

! Ključni zaključci

1. Deskriptivna statistika sažima podatke u manji broj smislenih brojki, ali svako sažimanje znači i gubitak informacija.
2. Mjere centralne tendencije (sredina, skraćena sredina, medijan, mod) odgovaraju na pitanje gdje se podaci nalaze. Svaka ima svoja svojstva, uključujući sredinu koja koristi sve podatke ali je osjetljiva na outliere, medijan koji je robustan ali ignorira većinu podataka, i skraćenu sredinu kao kompromis.
3. Mjere varijabilnosti (raspon, AAD, varijanca, SD, IQR) odgovaraju na pitanje koliko su podaci raspršeni. Varijanca dijeli s $N-1$ umjesto N (Besselova korekcija) kako bi bila nepristrana procjena populacijske varijance.
4. Asimetrija i zaobljenost opisuju oblik distribucije. Medijske metrike gotovo uvijek imaju pozitivnu asimetriju (dugačak rep prema desno).
5. Z-scores standardiziraju varijable na zajedničku skalu (sredina 0, SD 1) i omogućuju usporedbu varijabli s različitim mjernim jedinicama.
6. Kombinacija `group_by()` i `summarise()` je temeljni alat za izračunavanje deskriptivnih statistika po grupama u R u.
7. Pearsonov koeficijent korelacije mjeri linearnu povezanost, a Spearmanov mjeri monotonu povezanost. Korelacija nije uzročnost i mjeri samo specifičan tip veze.
8. Nedostajuće vrijednosti (`NA`) zahtijevaju svjesnu odluku o tome kako ih tretirati. Različiti tipovi nedostajanja (MCAR, MAR, MNAR) impliciraju različite posljedice za analizu.
9. Uvijek kombinirajte numeričke statistike s vizualizacijom. Brojke bez grafike mogu zavarati (Anscombeov kvartet).

⚠ Priprema za sljedeći tjedan

Sljedeći tjedan bavimo se **vizualizacijom podataka s ggplot2**. To je prirodni nastavak ovog predavanja jer ćemo naučiti kako sve statistike koje smo danas izračunali prikazati grafički. Histogrami, boxplotovi, scatterplotovi i bar chartovi su alati koji daju život brojkama.

Za pripremu napravite sljedeće:

1. Ponovite današnje R primjere i eksperimentirajte s njima. Promijenite varijable u `summarise()` pozivu i pogledajte što se događa.
2. Razmislite o tome koje bi grafičke prikaze htjeli vidjeti za naše TikTok podatke. Histogram dnevnog korištenja? Boxplot po dobnim skupinama? Scatterplot dobi i korištenja?
3. Pročitajte poglavlje 3 iz knjige Kieran Healy, *Data Visualization* (besplatno dostupno online).
4. Instalirajte paket `patchwork` ako ga nemate sa `install.packages("patchwork")`. Koristit ćemo ga za kombiniranje grafika.

---

12 Dodatno čitanje

Obavezno

Navarro, D. (2018). *Learning Statistics with R*, Chapter 5: Descriptive Statistics. Besplatno dostupno na learningstatisticswithr.com. Poglavlje pokriva iste teme kao ovo predavanje, uključujući trimmed mean, kovarijancu i Spearmanovu korelaciju, ali s primjerima iz psihologije i u base R sintaksi.

Preporučeno

Wickham, H. & Grolemund, G. (2023). *R for Data Science* (2nd edition), Chapters 3 i 4. Besplatno dostupno na r4ds.hadley.nz. Odlično pokrivanje tidyverse pristupa manipulaciji i sažimanju podataka.

Healy, K. (2018). *Data Visualization: A Practical Introduction*. Besplatno dostupno na socviz.co. Poglavlje 1 daje izvrsnu motivaciju zašto je vizualizacija neodvojiva od deskriptivne statistike.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edition). Klasična referenca za interpretaciju veličine efekata, uključujući smjernice za korelacije.

---

13 Pojmovnik

Pojam	Objašnjenje
Aritmetička sredina (mean)	Zbroj svih vrijednosti podijeljen s brojem opažanja. Koristi sve podatke ali je osjetljiva na ekstremne vrijednosti.
Skraćena sredina (trimmed mean)	Aritmetička sredina izračunata nakon uklanjanja određenog postotka najmanjih i najvećih vrijednosti. Kompromis između sredine i medijana.
Medijan (median)	Srednja vrijednost kad se podaci poredaju po veličini. Robusna mjera centralne tendencije jer ne ovisi o ekstremnim vrijednostima.
Mod (mode)	Vrijednost koja se najčešće pojavljuje. Jedina smislena mjera centralne tendencije za kategoričke podatke.
Raspon (range)	Razlika između najveće i najmanje vrijednosti u podacima. Jednostavna ali nerobusna mjera varijabilnosti.
Prosječno apsolutno odstupanje (AAD)	Prosjek apsolutnih razlika svake vrijednosti od sredine. Intuitivnija ali matematički manje pogodna od varijance.
Varijanca (variance)	Prosječno kvadrirano odstupanje od aritmetičke sredine (s $N-1$ u nazivniku). Mjeri raspršenost podataka.
Standardna devijacija (SD)	Korijen iz varijance. Izražena u istim mjernim jedinicama kao izvorni podaci.
Besselova korekcija	Dijeljenje s $N-1$ umjesto N pri izračunu varijance uzorka, kako bi procjena populacijske varijance bila nepristrana.
Interkvartilni raspon (IQR)	Razlika između 75. i 25. percentila. Raspon unutar kojeg se nalaze srednjih 50% podataka. Robusna mjera varijabilnosti.
Percentil	Vrijednost ispod koje se nalazi određeni postotak podataka. 50. percentil je medijan.
Asimetrija (skewness)	Mjera simetrije distribucije. Pozitivna asimetrija znači dugačak rep prema desno, negativna prema lijevo.
Zaobljenost (kurtosis)	Mjera težine repova distribucije u usporedbi s normalnom distribucijom.
Standardni rezultat (z-score)	Broj standardnih devijacija za koji se neka vrijednost razlikuje od aritmetičke sredine. Omogućuje usporedbu varijabli na različitim skalama.

Pojam	Objašnjenje
Kovarijanca (covariance)	Mjera zajedničkog variranja dviju varijabli. Ovisi o mjernim jedinicama, pa se koristi korelacija kao standardizirana verzija.
Pearsonov koeficijent korelacije (r)	Standardizirana kovarijanca. Mjera linearne povezanosti dviju numeričkih varijabli. Kreće se od minus 1 do plus 1.
Spearmanov koeficijent korelacije (r_s)	Pearsonov koeficijent izračunat na rangovima. Mjeri monotonost veze i robusniji je od Pearsonovog koeficijenta.
MCAR, MAR, MNAR	Tri tipa nedostajućih vrijednosti: potpuno nasumično, nasumično uvjetovano drugim varijablama, te sustavno povezano s nedostajućom vrijednošću.