Tjedan 10: Kategorički podaci i hi-kvadrat testovi

Kada su varijable kategorije, ne brojevi

2025-05-03

Table of contents

```
library(tidyverse)
```

Ishodi učenja

Nakon ovog predavanja moći ćete:

1. Prepoznati situacije u kojima su hi-kvadrat testovi prikladni (kategoričke varijable).
2. Provesti i interpretirati hi-kvadrat test za dobrotu prilagodbe (goodness-of-fit).
3. Provesti i interpretirati hi-kvadrat test nezavisnosti za kontingencijsku tablicu.
4. Izračunati očekivane frekvencije i objasniti njihovo značenje.
5. Interpretirati standardizirane reziduale za otkrivanje specifičnih odstupanja.
6. Primijeniti Fisherov egzaktni test kad su očekivane frekvencije male.
7. Izračunati Cramérovo V kao mjeru veličine učinka za kategoričke podatke.
8. Kritički ocijeniti rezultate istraživanja koja koriste kategoričke varijable.

1 Generacijski jaz u medijskim navikama

Zamislite da vodite istraživački tim pri velikoj medijskoj kući. Uprava je upravo objavila strategiju za sljedećih pet godina i ključno pitanje glasi — na koje platforme trebamo usmjeriti resurse? Intuicija kaže da mladi gledaju streaming, a stariji i dalje sjede pred televizorom. Ali intuicija je jedna stvar, a podatci druga. Vaš zadatak je provesti anketu i dati odgovor utemeljen na dokazima.

Anketa je provedena na 800 ispitanika iz pet hrvatskih regija, iz četiriju dobnih skupina. Svaki ispitanik je naveo koji tip medija najčešće koristi, koju vrstu sadržaja preferira, koliko je zadovoljan ponudom i niz demografskih podataka. Varijable koje vas najviše zanimaju su kategoričke — na primjer dobna skupina, tip medija, regija i obrazovanje. To nisu brojevi

koje možete zbrajati ili iz kojih možete računati prosjeke. To su kutije u koje se ljudi razvrstavaju. A za kutije trebate drugačiji statistički alat.

Prošla dva tjedna bavili smo se pitanjima poput "je li prosječni angažman veći za jedan tip sadržaja nego za drugi." To su pitanja o prosjjecima numeričkih varijabli, i za njih smo koristili t-test. Ali kad vas zanima *postoji li veza između dobne skupine i preferiranog medija* — kad su obje varijable kategoričke — t-test nema što reći. Za takva pitanja koristimo hi-kvadrat testove.

2 Podaci

```
survey <- read_csv("../resources/datasets/media_survey_chi2.csv")
glimpse(survey)
```

```
Rows: 800
Columns: 10
$ respondent_id      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ~
$ age_group          <chr> "45-59", "18-29", "45-59", "45-59", "30-44", "30-
44~
$ gender             <chr> "ženski", "muški", "ženski", "muški", "ženski", "mu~
$ education          <chr> "osnovna", "srednja", "visoka", "osnovna", "visoka"~
$ region             <chr> "Zagreb", "Zagreb", "Primorje", "Zagreb", "Slavonij~
$ media_type         <chr> "TV", "podcast", "streaming", "TV", "web_portal", "~
$ content_preference <chr> "zabava", "edukacija", "vijesti", "vijesti", "kultu~
$ hours_per_week     <dbl> 10.4, 3.4, 10.6, 7.7, 3.1, 8.3, 5.7, 7.3, 6.4, 7.5,~
$ satisfaction       <dbl> 4, 4, 3, 4, 3, 2, 4, 4, 4, 4, 3, 3, 5, 4, 2, 4, 3, ~
$ recommends         <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE~
```

```
# Distribucija ključnih varijabli
survey |> count(age_group, sort = TRUE)
```

```
# A tibble: 4 x 2
  age_group     n
  <chr>     <int>
1 18-29       210
2 45-59       206
3 30-44       205
4 60+         179
```

```
survey |> count(media_type, sort = TRUE)
```

```
# A tibble: 6 x 2
  media_type     n
  <chr>      <int>
1 TV           199
2 streaming    177
3 web_portal   173
4 podcast      109
5 radio         97
6 tisak         45
```

Ključne kategoričke varijable su `age_group` (četiri kategorije poput 18-29, 30-44, 45-59, 60+) i `media_type` (šest kategorija — streaming, TV, web portal, radio, tisak, podcast). Temeljno pitanje za upravu glasi — postoji li veza između dobi i preferiranog medija? I ako postoji, koliko je jaka?

3 Kontingencijska tablica: prvi pogled na vezu

Kad imate dvije kategoričke varijable i želite vidjeti kako se njihove kategorije preklapaju, prvi korak je kontingencijska tablica (ponekad zvana i cross-tabulation). Ona prikazuje koliko ispitanika pripada svakoj kombinaciji kategorija — koliko mladih preferira streaming, koliko starijih bira televiziju, i tako dalje.

```
# Kontingencijska tablica: dob × tip medija
kont_tablica <- table(survey$age_group, survey$media_type)
kont_tablica
```

```
       podcast radio streaming tisak TV web_portal
  18-29      28     8        90     6 22         56
  30-44      48    21        51     6 24         55
  45-59      25    27        29    13 69         43
  60+         8    41         7    20 84         19
```

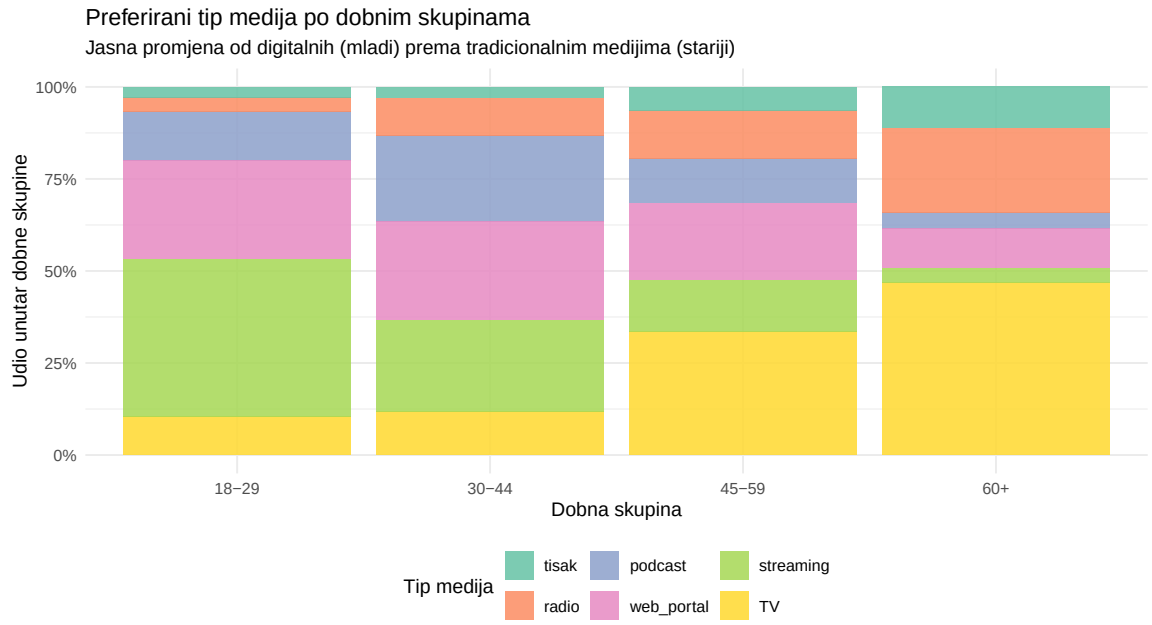
Apsolutne frekvencije su korisne, ali teško ih je uspoređivati kad dobne skupine nemaju jednaki broj ispitanika. Proporcije po retku rješavaju taj problem — svaka dobna skupina se normalizira na 100%, pa možete izravno uspoređivati strukture preferencija.

```
# Proporcije po retku (svaka dobna skupina = 100%)
round(prop.table(kont_tablica, margin = 1) * 100, 1)
```

```
       podcast radio streaming tisak   TV web_portal
  18-29    13.3   3.8      42.9   2.9 10.5       26.7
  30-44    23.4  10.2      24.9   2.9 11.7       26.8
  45-59    12.1  13.1      14.1   6.3 33.5       20.9
  60+       4.5  22.9       3.9  11.2 46.9       10.6
```

Proporcije govore jasnu priču. Među osobama 18-29, 43% preferira streaming i 27% web portale. Među osobama 60+, 47% preferira TV i 23% radio. Obrazac je očit — mlađi preferiraju digitalne medije, stariji tradicionalne. Ali je li ovaj obrazac statistički značajan, ili bi mogao nastati čistim slučajem u uzorku od 800 ljudi?

```
survey |>
  count(age_group, media_type) |>
  group_by(age_group) |>
  mutate(udio = n / sum(n)) |>
  ungroup() |>
  mutate(media_type = fct_reorder(media_type, udio, .fun = sum)) |>
  ggplot(aes(x = age_group, y = udio, fill = media_type)) +
  geom_col(position = "fill", alpha = 0.85) +
  scale_y_continuous(labels = scales::label_percent()) +
  scale_fill_brewer(palette = "Set2") +
  labs(
    title = "Preferirani tip medija po dobnim skupinama",
    subtitle = "Jasna promjena od digitalnih (mladi) prema tradicionalnim medijima (starij
    x = "Dobna skupina",
    y = "Udio unutar dobne skupine",
    fill = "Tip medija"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Vizualno, razlika je dramatična. Ali da bismo prešli s "izgleda različito" na "statistički je različito", trebamo formalni test.

4 Hi-kvadrat test za dobrotu prilagodbe

Prije nego se uhvatimo u koštac s vezom dviju varijabli, počnimo s jednostavnijim pitanjem — odgovara li distribucija jedne kategoričke varijable nekoj očekivanoj distribuciji? Ovo se zove test za dobrotu prilagodbe (goodness-of-fit test).

Evo konkretne situacije. Medijska kuća tvrdi da ima podjednaku publiku iz svih pet regija Hrvatske. Naša anketa pokazuje nešto drugačiju sliku. Želimo testirati odstupa li opažena distribucija regija značajno od uniformne distribucije (20% iz svake regije).

$$H_0 : \text{Distribucija regija u uzorku odgovara uniformnoj distribuciji (20\% svaka)}$$

$$H_1 : \text{Distribucija regija nije uniformna}$$

```
# Opažene frekvencije
opazene <- survey |> count(region) |> arrange(desc(n))
opazene
```

```
# A tibble: 5 x 2
  region           n
  <chr>        <int>
1 Zagreb         258
2 Slavonija      165
3 Dalmacija      139
4 Sjeverozapad   137
5 Primorje       101
```

```
# Hi-kvadrat test za dobrotu prilagodbe
# H : sve regije imaju jednaki udio (20% svaka)
gof_test <- chisq.test(opazene$n)
gof_test
```

```
    Chi-squared test for given probabilities

data:  opazene$n
X-squared = 88, df = 4, p-value < 2.2e-16
```

4.1 Što zapravo radi hi-kvadrat statistika?

Hi-kvadrat statistika mjeri koliko se vaši opaženi podaci razlikuju od onoga što biste očekivali da je nulta hipoteza istinita. Formula je:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Slovo O označava opaženu frekvenciju (observed), a E očekivanu frekvenciju (expected) za svaku kategoriju. Za svaku kategoriju izračunate razliku između opaženog i očekivanog, kvadrirate je (da se pozitivna i negativna odstupanja ne poništavaju) i podijelite s očekivanom frekvencijom (da normalizirate za veličinu kategorije). Zbrojite sve te doprinose i dobijete ukupni ^{2}. Što je veći, to su opaženi podaci udaljeniji od očekivanih pod H .

Raspakujmo to na našim podacima.

```
n_total <- nrow(survey)
n_kategorija <- 5

# Pod H  (uniformna distribucija), svaka regija ima n/5 ispitanika
ocekivane <- rep(n_total / n_kategorija, n_kategorija)

tibble(
  regija = opazene$region,
  O = opazene$n,
  E = ocekivane,
  razlika = O - E,
  doprinos_chi2 = round((O - E)^2 / E, 2)
) |>
  bind_rows(tibble(
    regija = "UKUPNO",
    O = sum(opazene$n),
    E = sum(ocekivane),
    razlika = 0,
    doprinos_chi2 = round(sum((opazene$n - ocekivane)^2 / ocekivane), 2)
  ))
```

```
# A tibble: 6 x 5
  regija           O     E razlika doprinos_chi2
  <chr>        <int> <dbl>   <dbl>         <dbl>
1 Zagreb         258   160      98         60.0
2 Slavonija      165   160       5          0.16
3 Dalmacija      139   160     -21          2.76
4 Sjeverozapad   137   160     -23          3.31
5 Primorje       101   160     -59         21.8
6 UKUPNO         800   800       0         88
```

Svaka kategorija doprinosi ukupnom ^{2} ovisno o tome koliko njezina opažena frekvencija odstupa od očekivane. Zadnji redak (UKUPNO) je testna statistika — to je ona ista ^{2} vrijednost koju je `chisq.test()` izračunao automatski.

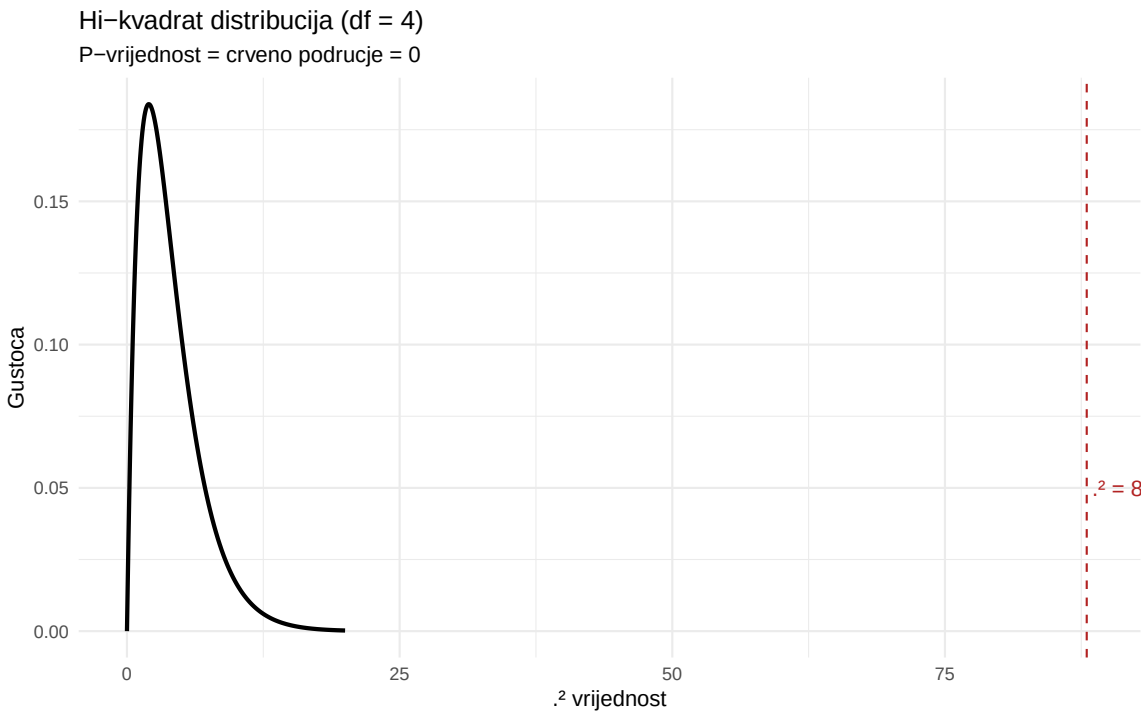
Da bismo dobili p-vrijednost, uspoređujemo našu ^{2} statistiku s hi-kvadrat distribucijom. Ta distribucija ima jedan parametar — stupnjeve slobode — koji se za goodness-of-fit test računaju kao broj kategorija minus 1.

```r
# Vizualizacija: hi-kvadrat distribucija
df_gof <- n_kategorija - 1  # stupnjevi slobode = broj kategorija - 1

x_vals <- seq(0, 20, length.out = 300)

chi_data <- tibble(x = x_vals, density = dchisq(x_vals, df = df_gof))

ggplot(chi_data, aes(x = x, y = density)) +
  geom_line(linewidth = 1) +
  geom_area(data = chi_data |> filter(x >= gof_test$statistic), fill = "firebrick", alpha
  geom_vline(xintercept = gof_test$statistic, color = "firebrick", linetype = "dashed") +
  annotate("text", x = gof_test$statistic + 0.5, y = 0.05,
           label = paste0("χ² = ", round(gof_test$statistic, 2)),
           color = "firebrick", hjust = 0) +
  labs(
    title = "Hi-kvadrat distribucija (df = 4)",
    subtitle = paste0("P-vrijednost = crveno područje = ", round(gof_test$p.value, 4)),
    x = "χ² vrijednost",
    y = "Gustoća"
  ) +
  theme_minimal()
```



Crveno područje desno od naše testne statistike je p-vrijednost — vjerojatnost da bismo dobili ovako veliko ili veće odstupanje čistim slučajem, kad bi distribucija regija bila zaista

uniformna. Ako je ta p-vrijednost manja od 0.05, zaključujemo da distribucija značajno odstupa od uniformne. Zagreb je nadreprezentiran, ostale regije podreprezentirane.

4.2 Testiranje s poznatim populacijskim proporcijama

Uniformna distribucija (20% svaka regija) rijetko je realistična nulta hipoteza. Češće vas zanima odgovara li vaš uzorak poznatim populacijskim proporcijama. Možda Zagreb zaista čini 30% populacije — i u tom slučaju bi nadreprezentiranost Zagreba u uzorku mogla biti potpuno očekivana.

```
# Populacijski udjeli regija (fiktivni, ali bazirani na stvarnim omjerima)
pop_udjeli <- c(
  "Dalmacija" = 0.20,
  "Primorje" = 0.14,
  "Sjeverozapad" = 0.18,
  "Slavonija" = 0.18,
  "Zagreb" = 0.30
)

# Opaženi poredak mora odgovarati poretku proporcija
opazene_sortirane <- survey |>
  count(region) |>
  arrange(region)

gof_pop <- chisq.test(opazene_sortirane$n, p = pop_udjeli)
gof_pop
```

```
    Chi-squared test for given probabilities

data:  opazene_sortirane$n
X-squared = 8.5894, df = 4, p-value = 0.07222
```

Kad testiramo protiv stvarnih populacijskih proporcija, rezultat je sasvim drugačiji. P-vrijednost je veća, što znači da naš uzorak zapravo dobro odražava populacijsku distribuciju regija. Ista opažena distribucija može biti "značajno odstupajuća" od jedne referentne distribucije i "konzistentna" s drugom. Referentna distribucija koju odaberete potpuno mijenja zaključak — i zato je izbor nulte hipoteze istraživačka, a ne samo statistička odluka.

5 Hi-kvadrat test nezavisnosti

Sada dolazimo do glavnog pitanja — postoji li veza između dviju kategoričkih varijabli? Konkretno, ovisi li preferirani tip medija o dobnoj skupini?

$$H_0 : \text{Tip medija i dobna skupina su nezavisni (nema veze)}$$

$$H_1 : \text{Tip medija i dobna skupina NISU nezavisni (postoji veza)}$$

Riječ "nezavisni" ovdje ima precizan statistički smisao — poznavanje nečije dobne skupine ne pomaže u predviđanju njihovog preferiranog medija. Ako su doista nezavisni, distribucija medijskog tipa trebala bi biti ista u svim dobnim skupinama — isti postotak mladih i starijih birao bi streaming, isti postotak birao bi TV, i tako dalje.

```
chi2_test <- chisq.test(kont_tablica)
chi2_test
```

```
    Pearson's Chi-squared test

data:  kont_tablica
X-squared = 233.59, df = 15, p-value < 2.2e-16
```

P-vrijednost je iznimno mala (p < 2.2e-16, što je najmanji broj koji R ispisuje). Imamo snažne dokaze da dob i preferirani tip medija nisu nezavisni — veza postoji. Ali samu ^{2} statistiku treba tumačiti s oprezom — ona govori da veza postoji, ali ne govori *gdje* u tablici se ta veza najviše očituje. Za to trebamo pogledati dublje.

5.1 Očekivane frekvencije: što bismo vidjeli da veze nema

Očekivane frekvencije su ono što bismo vidjeli u kontingencijskoj tablici kad ne bi bilo nikakve veze između dobi i medijskog tipa. Računaju se jednostavnom formulom:

$$E_{ij} = \frac{\text{ukupno u retku i} \times \text{ukupno u stupcu j}}{\text{ukupno}}$$

Logika je intuitivna. Ako 22% cijelog uzorka preferira streaming, i ako dob i medij nisu povezani, onda bi i u skupini 18-29 i u skupini 60+ oko 22% trebalo preferirati streaming. Očekivana frekvencija za ćeliju "18-29 × streaming" je jednostavno ukupan broj osoba 18-29 pomnožen s ukupnim udjelom streaminga.

```
# Očekivane frekvencije
round(chi2_test$expected, 1)
```

```
       podcast radio streaming tisak   TV web_portal
18-29    28.6  25.5      46.5  11.8 52.2       45.4
30-44    27.9  24.9      45.4  11.5 51.0       44.3
45-59    28.1  25.0      45.6  11.6 51.2       44.5
60+      24.4  21.7      39.6  10.1 44.5       38.7
```

```r
# Usporedba: opažene vs očekivane
cat("OPAŽENE frekvencije:\n")
```

```
OPAŽENE frekvencije:
```

```r
kont_tablica
```

```
       podcast radio streaming tisak TV web_portal
  18-29      28     8        90     6 22         56
  30-44      48    21        51     6 24         55
  45-59      25    27        29    13 69         43
  60+         8    41         7    20 84         19
```

```r
cat("\nOČEKIVANE frekvencije (pod H : nema veze):\n")
```

```
OČEKIVANE frekvencije (pod H : nema veze):
```

```r
round(chi2_test$expected, 1)
```

```
       podcast radio streaming tisak   TV web_portal
  18-29    28.6  25.5      46.5  11.8 52.2       45.4
  30-44    27.9  24.9      45.4  11.5 51.0       44.3
  45-59    28.1  25.0      45.6  11.6 51.2       44.5
  60+      24.4  21.7      39.6  10.1 44.5       38.7
```

Usporedba je poučna. Pod nultom hipotezom, oko 22% svake dobne skupine trebalo bi preferirati streaming. U stvarnosti, 43% mladih (18-29) preferira streaming, a samo 4% starijih (60+). Ovo golemo odstupanje opaženih od očekivanih frekvencija je upravo ono što hi-kvadrat statistika hvata i kvantificira.

5.2 Pretpostavka koju morate provjeriti

Hi-kvadrat test je aproksimacija, i da bi ta aproksimacija bila pouzdana, sve očekivane frekvencije moraju biti dovoljno velike. Konvencija glasi — barem 5 u svakoj ćeliji.

```r
# Provjera: ima li očekivanih frekvencija < 5?
min_expected <- min(chi2_test$expected)
cat("Najmanja očekivana frekvencija:", round(min_expected, 1), "\n")
```

```
Najmanja očekivana frekvencija: 10.1
```

```r
cat("Sve  5:", min_expected >= 5, "\n")
```

```
Sve  5: TRUE
```

Sve očekivane frekvencije su iznad 5, što znači da je hi-kvadrat aproksimacija valjana. Kad to nije slučaj — a to se dogodi čim imate rijetke kategorije ili mali uzorak — trebate posegnuti za Fisherovim egzaktnim testom, o kojem ćemo govoriti za koji odlomak.

6 Gdje je veza najjača? Standardizirani reziduali

Ukupna ^{2} statistika kaže da veza postoji, ali to je kao kad vam liječnik kaže "nešto je abnormalno u krvnoj slici" bez da vam kaže što. Za specifičnu dijagnozu koristimo standardizirane reziduale, koji vam govore koja ćelija u tablici najviše doprinosi ukupnoj vezi.

Standardizirani (Pearsonov) rezidual za svaku ćeliju računa se na sljedeći način:

$$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

Rezidual veći od +2 znači da ta ćelija ima značajno *više* opažanja nego što bismo očekivali da veze nema. Rezidual manji od -2 znači značajno *manje*. Ovi pragovi su analogni z-vrijednostima u normalnoj distribuciji.

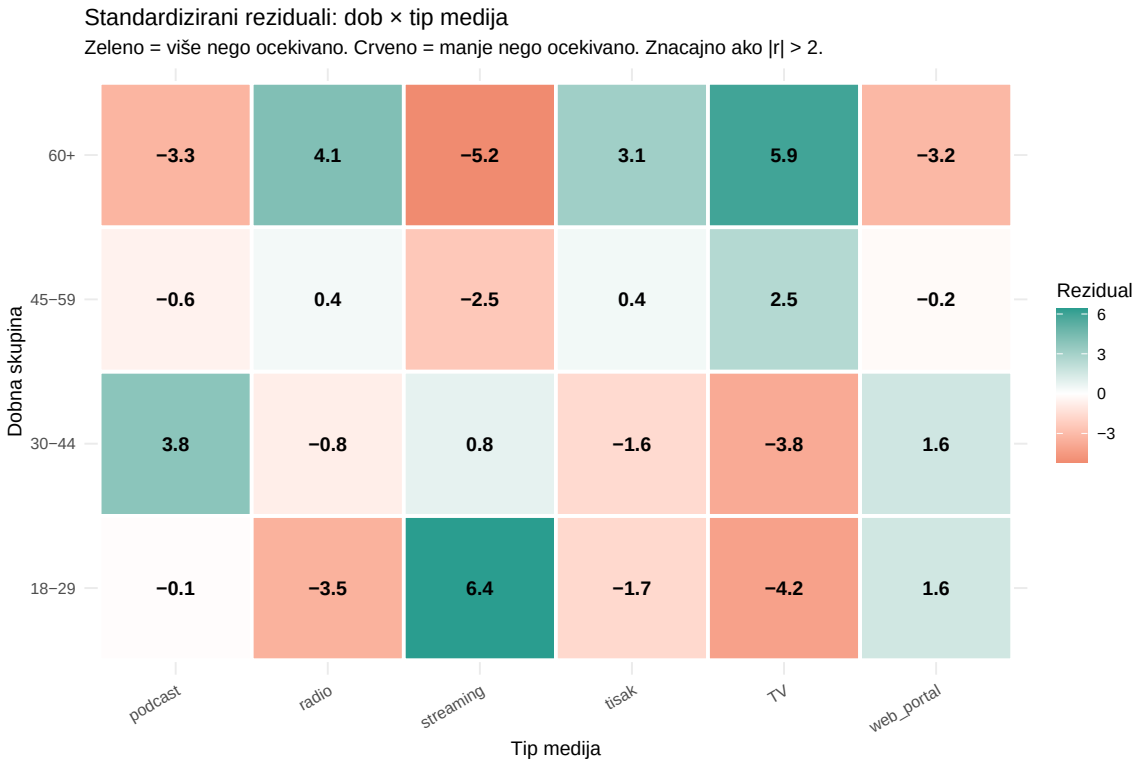
```r
round(chi2_test$residuals, 2)
```

```
       podcast radio streaming tisak    TV web_portal
18-29    -0.11 -3.46      6.39 -1.69 -4.18       1.57
30-44     3.80 -0.77      0.84 -1.63 -3.78       1.60
45-59    -0.58  0.40     -2.46  0.41  2.48      -0.23
60+      -3.32  4.14     -5.18  3.13  5.92      -3.17
```

Brojevi u tablici su informativni, ali još je informativnija vizualizacija. Toplinska karta (heatmap) reziduala jasno pokazuje koji parovi kategorija su odgovorni za vezu.

```r
# Vizualizacija standardiziranih reziduala
as.data.frame(chi2_test$residuals) |>
  as_tibble() |>
  rename(age_group = Var1, media_type = Var2, residual = Freq) |>
  ggplot(aes(x = media_type, y = age_group, fill = residual)) +
  geom_tile(color = "white", linewidth = 1) +
  geom_text(aes(label = round(residual, 1)), size = 4, fontface = "bold") +
```

```
  scale_fill_gradient2(low = "#e76f51", mid = "white", high = "#2a9d8f", midpoint = 0) +
  labs(
    title = "Standardizirani reziduali: dob × tip medija",
    subtitle = "Zeleno = više nego očekivano. Crveno = manje nego očekivano. Značajno ako
    x = "Tip medija",
    y = "Dobna skupina",
    fill = "Rezidual"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



Ovaj graf je zlatni rudnik za vašu upravu. Najjače zelene ćelije (znatno više nego očekivano) su streaming u skupini 18-29 i TV u skupini 60+. Najjače crvene (znatno manje nego očekivano) su TV u skupini 18-29 i streaming u skupini 60+. Reziduali precizno identificiraju ono što ste intuitivno naslutili — generacijski jaz u medijskim navikama je realan, mjerljiv i specifičan.

Ne stajte na "značajno"

Kad izvještavate rezultate hi-kvadrat testa, nemojte stati na " ^{2} je značajan." Koristite reziduale da identificirate *specifične* kombinacije kategorija koje najviše doprinose vezi. Vaša uprava ne želi čuti samo "postoji veza između dobi i medija." Želi znati *kakva* je ta veza — mladi biraju streaming, stariji TV, a srednje generacije su negdje između.

7 Koliko je veza jaka? Cramérovo V

Statistička značajnost govori da veza postoji, ali ne govori koliko je jaka. S 800 ispitanika, čak i trivijalna veza može biti "značajna." Cramérovo V je mjera veličine učinka za hi-kvadrat test nezavisnosti — ono što je Cohenov d za t-test.

$$V = \sqrt{\frac{\chi^2}{n \times (k-1)}}$$

U formuli, n je ukupan broj opažanja, a k je manji od broja redova ili stupaca kontingencijske tablice. V varira od 0 (potpuna nezavisnost, nikakva veza) do 1 (savršena asocijacija, jedna varijabla potpuno predviđa drugu).

```
chi2_val <- chi2_test$statistic
n_obs <- sum(kont_tablica)
k <- min(nrow(kont_tablica), ncol(kont_tablica))

cramer_v <- sqrt(chi2_val / (n_obs * (k - 1)))

cat(" ² =", round(chi2_val, 2), "\n")
```

```
 ² = 233.59
```

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

```
n = 800
```

```
cat("k =", k, "(minimum redova/stupaca)\n")
```

```
k = 4 (minimum redova/stupaca)
```

```
cat("Cramérovo V =", round(cramer_v, 3), "\n")
```

```
Cramérovo V = 0.312
```

7.1 Kako interpretirati Cramérovo V

Smjernice za interpretaciju ovise o stupnjevima slobode tablice. Cohen (1988) je predložio sljedeće pragove za tablicu s k = 4 (naš slučaj).

```
tribble(
  ~V, ~interpretacija,
  "0.06", "Mali učinak",
  "0.17", "Srednji učinak",
  "0.29", "Veliki učinak"
)
```

```
# A tibble: 3 x 2
  V     interpretacija
  <chr> <chr>
1 0.06  Mali učinak
2 0.17  Srednji učinak
3 0.29  Veliki učinak
```

Naše V je veliki učinak. Veza između dobi i medijskog tipa nije samo statistički značajna — ona je i praktično snažna. Dob zaista predviđa medijske preferencije, i to u znatnoj mjeri.

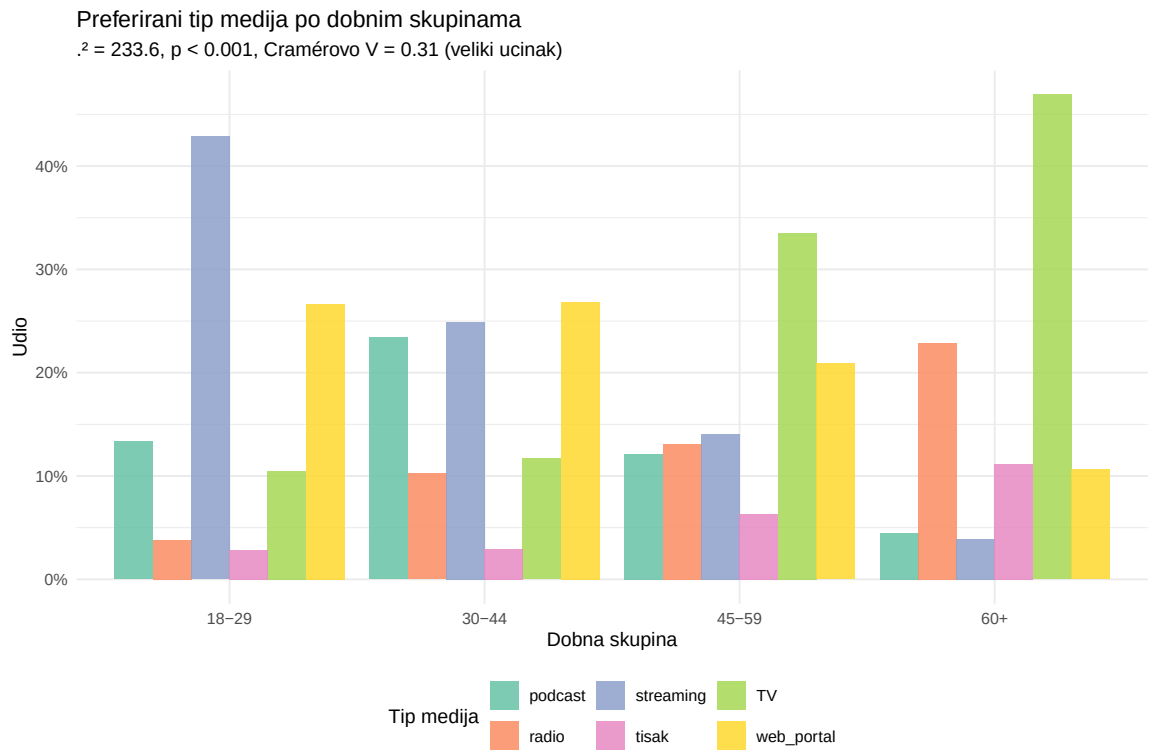
8 Kako vizualizirati kategoričke podatke

Kontingencijska tablica puna brojeva može biti teška za brzo čitanje, pogotovo kad imate mnogo kategorija. Dobra vizualizacija učini vezu vidljivom na prvi pogled.

Grupani stupčasti graf (dodged bar chart) je najčitljiviji izbor za prezentacije jer svaka kategorija ima vlastiti stupac i usporedba je neposredna.

```
# Grupani stupčasti graf (najprikladniji za prezentaciju)
survey |>
  count(age_group, media_type) |>
  group_by(age_group) |>
  mutate(udio = n / sum(n)) |>
  ungroup() |>
  ggplot(aes(x = age_group, y = udio, fill = media_type)) +
  geom_col(position = "dodge", alpha = 0.85) +
  scale_y_continuous(labels = scales::label_percent()) +
  scale_fill_brewer(palette = "Set2") +
  labs(
    title = "Preferirani tip medija po dobnim skupinama",
    subtitle = paste0(" ² = ", round(chi2_val, 1), ", p < 0.001, Cramérovo V = ",
                      round(cramer_v, 2), " (veliki učinak)"),
    x = "Dobna skupina",
    y = "Udio",
    fill = "Tip medija"
  ) +
```

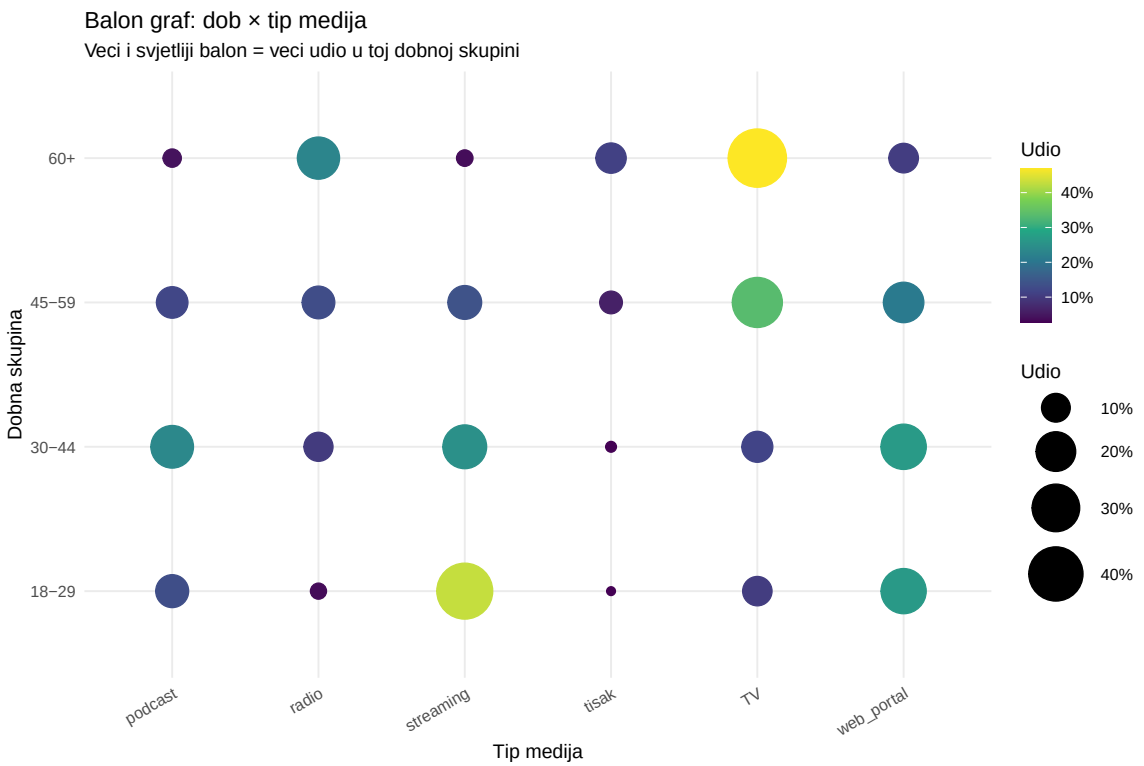
```
  theme_minimal() +
  theme(legend.position = "bottom")
```



Balon graf je korisna alternativa kad želite prikazati vezu u formi koja podsjeća na kontingencijsku tablicu, ali s vizualnim kodiranjem veličine i boje umjesto brojeva.

```
# Balon graf (dobra alternativa za kontingencijske tablice)
survey |>
  count(age_group, media_type) |>
  group_by(age_group) |>
  mutate(udio = n / sum(n)) |>
  ungroup() |>
  ggplot(aes(x = media_type, y = age_group, size = udio, color = udio)) +
  geom_point() +
  scale_size_continuous(range = c(2, 15), labels = scales::label_percent()) +
  scale_color_viridis_c(option = "D", labels = scales::label_percent()) +
  labs(
    title = "Balon graf: dob × tip medija",
    subtitle = "Veći i svjetliji balon = veći udio u toj dobnoj skupini",
    x = "Tip medija",
    y = "Dobna skupina",
    size = "Udio", color = "Udio"
  ) +
```

```
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



9 Hi-kvadrat test u praksi: korak po korak

Sažmimo postupak u korake koje možete primijeniti na bilo koji par kategoričkih varijabli. Prvo formulirate hipoteze (H — varijable su nezavisne). Onda napravite kontingencijsku tablicu i vizualizirate podatke — jer graf vam često kaže više od testa. Zatim provjerite pretpostavke (sve očekivane frekvencije 5). Provedite test i izračunajte p-vrijednost. Izračunajte veličinu učinka (Cramérovo V). Pogledajte reziduale za specifična odstupanja. I konačno, interpretirajte sve to u kontekstu vašeg istraživačkog pitanja.

Primijenimo to na drugi par varijabli — postoji li veza između spola i preferencije sadržaja?

```
# Primjer: postoji li veza između spola i preferencije sadržaja?
tablica_spol <- table(survey$gender, survey$content_preference)

# Korak 2: kontingencijska tablica
tablica_spol
```

```
        edukacija kultura sport vijesti zabava
```

```
  muški         59      54    76     107     88
  ženski        60      59    86     111    100
```

```r
# Korak 3: provjera
chi_spol <- chisq.test(tablica_spol)
cat("\nNajmanja očekivana frekvencija:", round(min(chi_spol$expected), 1), "\n")
```

```
Najmanja očekivana frekvencija: 54.2
```

```r
# Korak 4: test
chi_spol
```

```
    Pearson's Chi-squared test

data:  tablica_spol
X-squared = 0.40693, df = 4, p-value = 0.9819
```

```r
# Korak 5: Cramérovo V
v_spol <- sqrt(chi_spol$statistic / (sum(tablica_spol) * (min(dim(tablica_spol)) - 1)))
cat("Cramérovo V:", round(v_spol, 3), "\n")
```

```
Cramérovo V: 0.023
```

```r
# Korak 6: reziduali
cat("\nStandardizirani reziduali:\n")
```

```
Standardizirani reziduali:
```

```r
round(chi_spol$residuals, 2)
```

```
        edukacija kultura sport vijesti zabava
  muški      0.25   -0.03 -0.20    0.23  -0.24
  ženski    -0.24    0.03  0.19   -0.22   0.23
```

Možda je veza između spola i preferencije sadržaja značajna, a možda nije — i to je potpuno u redu. Ne mora svaka veza biti značajna. Podatke treba pustiti da govore, a ne ih prisiljavati da potvrde naša očekivanja.

Gdje smo, kamo idemo

U prvom dijelu naučili smo hi-kvadrat test za dobrotu prilagodbe i test nezavisnosti, očekivane frekvencije, standardizirane reziduale i Cramérovo V. U nastavku pokrivamo situacije kad standardni hi-kvadrat test nije primjeren — mali uzorci, rijetke kategorije, upareni podaci — te provodimo potpunu analizu kategoričkih podataka.

10 Kad je uzorak premalen: Fisherov egzaktni test

Hi-kvadrat test je aproksimacija, i kao svaka aproksimacija, ima granice. Kad su očekivane frekvencije male (ispod 5 u jednoj ili više ćelija), ta aproksimacija postaje nepouzdana. Fisherov egzaktni test rješava taj problem tako što računa točnu p-vrijednost bez ikakve aproksimacije — odatle i ime "egzaktni."

Zamislite sljedeću situaciju — testirate vezu između tipa poziva na akciju (CTA) i konverzije na malom uzorku od 40 newsletter kampanja. Samo 40 opažanja raspoređenih u 2×2 tablicu znači da neke ćelije mogu imati jako malo frekvencije.

```
# Mali uzorak: 40 kampanja, 2 x 2 tablica
set.seed(42)

kampanje <- tibble(
  cta_tip = c(rep("direktni", 20), rep("indirektni", 20)),
  konverzija = c(
    sample(c("da", "ne"), 20, replace = TRUE, prob = c(0.45, 0.55)),
    sample(c("da", "ne"), 20, replace = TRUE, prob = c(0.20, 0.80))
  )
)

tablica_cta <- table(kampanje$cta_tip, kampanje$konverzija)
tablica_cta
```

```
            da ne
  direktni   12  8
  indirektni  8 12
```

```
# Provjera ocekivanih frekvencija
chi_cta <- chisq.test(tablica_cta, correct = FALSE)
chi_cta$expected
```

```
            da ne
  direktni   10 10
  indirektni 10 10
```

Neke očekivane frekvencije su blizu 5. Hi-kvadrat aproksimacija ovdje nije pouzdana. Prelazimo na Fisherov test.

```
# Fisherov egzaktni test
fisher.test(tablica_cta)
```

```
    Fisher's Exact Test for Count Data

data:  tablica_cta
p-value = 0.3431
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5370744 9.6150685
sample estimates:
odds ratio
  2.203611
```

```
# Usporedba tri pristupa
cat("Hi-kvadrat (bez korekcije):  p =", round(chi_cta$p.value, 4), "\n")
```

```
Hi-kvadrat (bez korekcije):  p = 0.2059
```

```
cat("Hi-kvadrat (Yates korekcija): p =", round(chisq.test(tablica_cta)$p.value, 4), "\n")
```

```
Hi-kvadrat (Yates korekcija): p = 0.3428
```

```
cat("Fisherov egzaktni test:       p =", round(fisher.test(tablica_cta)$p.value, 4), "\n")
```

```
Fisherov egzaktni test:       p = 0.3431
```

Tri pristupa daju nešto različite p-vrijednosti. Hi-kvadrat bez korekcije je najliberalniji (najmanji p). Yatesova korekcija kontinuiteta, koju R primjenjuje po defaultu za 2×2 tablice, je konzervativnija. Fisherov egzaktni test daje točan rezultat i on je pravi izbor kad su očekivane frekvencije male.

10.1 Odds ratio: koliko je jedna grupa u prednosti

Fisherov test za 2×2 tablice automatski računa odds ratio — omjer šansi. To je prirodna mjera veličine učinka za binarne ishode i odgovara na pitanje koliko su šanse za konverziju veće uz direktni CTA nego uz indirektni.

```r
fisher_rez <- fisher.test(tablica_cta)

cat("Odds ratio:", round(fisher_rez$estimate, 2), "\n")
```

```
Odds ratio: 2.2
```

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

```
95% CI: [ 0.54 , 9.62 ]
```

Odds ratio veći od 1 znači da je šansa konverzije veća uz direktni CTA. Odds ratio jednak 1 značio bi da nema nikakve razlike. Interval pouzdanosti koji ne sadrži 1 sugerira statistički značajnu razliku.

> Koji test kad?
>
> **Hi-kvadrat test** koristite kad su sve očekivane frekvencije 5 i tablica je bilo koje veličine. Brz je i dovoljno točan za velike uzorke.
> **Fisherov egzaktni test** koristite kad su neke očekivane frekvencije < 5 ili je ukupni uzorak mali (ispod 50). Funkcionira za bilo koju veličinu tablice, ali je računalno zahtjevniji za velike tablice.
> U praksi, Fisherov test možete koristiti uvijek — za velike uzorke daje identične rezultate kao hi-kvadrat. Razlika se pojavljuje samo za male uzorke, i tada je Fisherov test pouzdaniji.

11 Spajanje kategorija: manje je ponekad više

Ponekad kontingencijska tablica ima kategorije s vrlo malo opažanja. Umjesto da odmah pribjegnete Fisherovom testu, postoji elegantnije rješenje — spojiti (kolapsirati) slične kategorije u šire grupe. To ne samo da rješava problem malih frekvencija, nego često rezultira jasnijom pričom.

```r
survey <- read_csv("../resources/datasets/media_survey_chi2.csv")

# Originalna tablica: media_type ima 6 kategorija, neke male
table(survey$media_type)
```

```
   podcast      radio  streaming      tisak         TV web_portal
       109         97        177         45        199        173
```

```r
# Spajanje: digitalni (streaming + web_portal + podcast) vs tradicionalni (TV + radio + ti
survey <- survey |>
  mutate(media_grupa = if_else(
    media_type %in% c("streaming", "web_portal", "podcast"),
    "digitalni",
    "tradicionalni"
  ))

# Nova tablica: 4 x 2 (preglednija i s vecim frekvencijama)
table(survey$age_group, survey$media_grupa)
```

```
       digitalni tradicionalni
  18-29       174            36
  30-44       154            51
  45-59        97           109
  60+          34           145
```

```r
chi_grupa <- chisq.test(table(survey$age_group, survey$media_grupa))
chi_grupa
```

```
    Pearson's Chi-squared test

data:  table(survey$age_group, survey$media_grupa)
X-squared = 198.89, df = 3, p-value < 2.2e-16
```

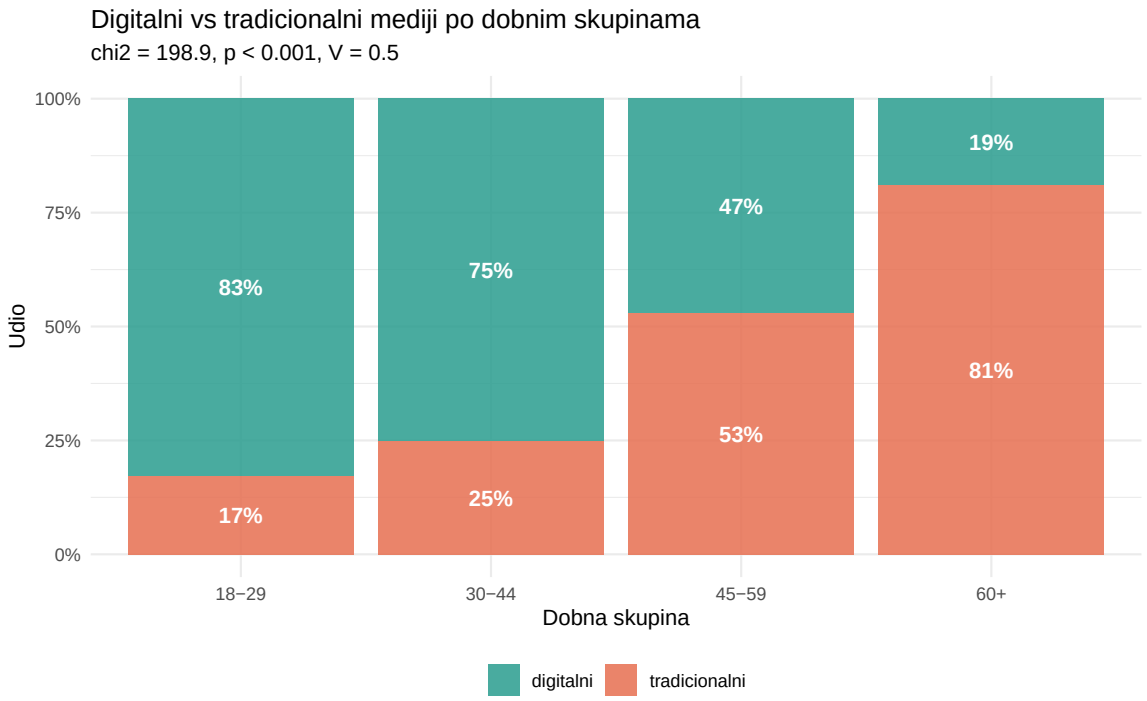
```r
# Cramerovo V za 4 x 2 tablicu
v_grupa <- sqrt(chi_grupa$statistic / (nrow(survey) * (min(4, 2) - 1)))
cat("\nCramerovo V:", round(v_grupa, 3), "\n")
```

```
Cramerovo V: 0.499
```

Sažeta tablica s jednom binarnom podjelom — digitalni vs tradicionalni mediji — govori istu priču kao originalna s šest kategorija, ali mnogo jasnije. Veza s dobi je i dalje izuzetno značajna i velika.

```r
survey |>
  count(age_group, media_grupa) |>
  group_by(age_group) |>
  mutate(udio = n / sum(n)) |>
  ungroup() |>
  ggplot(aes(x = age_group, y = udio, fill = media_grupa)) +
  geom_col(alpha = 0.85) +
```

```
  scale_y_continuous(labels = scales::label_percent()) +
  scale_fill_manual(values = c("digitalni" = "#2a9d8f", "tradicionalni" = "#e76f51")) +
  geom_text(aes(label = paste0(round(udio * 100), "%")),
            position = position_stack(vjust = 0.5), color = "white", fontface = "bold") +
  labs(
    title = "Digitalni vs tradicionalni mediji po dobnim skupinama",
    subtitle = paste0("chi2 = ", round(chi_grupa$statistic, 1), ", p < 0.001, V = ",
                      round(v_grupa, 2)),
    x = "Dobna skupina",
    y = "Udio",
    fill = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Ovaj graf je ono što vaša uprava treba vidjeti. Priča je kristalno jasna — 85% osoba 18-29 preferira digitalne medije, ali samo 26% osoba 60+. Generacijski jaz je masivan, i jedno je od onih rijetkih nalaza u društvenim znanostima koji ne zahtijeva puno objašnjavanja.

12 Stratificirana analiza: je li veza konzistentna u podgrupama?

Ukupni rezultat kaže da veza između medijskog tipa i dobi postoji. Ali je li ta veza konzistentna kad podijelimo podatke po nekoj trećoj varijabli? Možda digitalni mediji

dominiraju među mladima u Zagrebu, ali ne i u Slavoniji? Ovo je suština stratificirane analize — provodimo isti test odvojeno za svaku podgrupu treće varijable.

```
# Zadovoljstvo kategorizirano: nisko (1-2), srednje (3), visoko (4-5)
survey <- survey |>
  mutate(satisfaction_cat = case_when(
    satisfaction <= 2 ~ "nisko",
    satisfaction == 3 ~ "srednje",
    satisfaction >= 4 ~ "visoko"
  ))

# Hi-kvadrat test po dobnim skupinama
survey |>
  group_by(age_group) |>
  summarise(
    n = n(),
    chi2 = chisq.test(table(media_grupa, satisfaction_cat))$statistic,
    p = chisq.test(table(media_grupa, satisfaction_cat))$p.value,
    V = {
      tab <- table(media_grupa, satisfaction_cat)
      sqrt(chisq.test(tab)$statistic / (n() * (min(dim(tab)) - 1)))
    },
    .groups = "drop"
  ) |>
  mutate(
    chi2 = round(chi2, 2),
    p = round(p, 4),
    V = round(V, 3),
    znacajno = p < 0.05
  )
```

```
# A tibble: 4 x 6
  age_group     n  chi2      p     V znacajno
  <chr>     <int> <dbl>  <dbl> <dbl> <lgl>
1 18-29       210  3.8  0.150  0.135 FALSE
2 30-44       205  4.44 0.108  0.147 FALSE
3 45-59       206  6.47 0.0394 0.177 TRUE
4 60+         179  4.85 0.0884 0.165 FALSE
```

Stratificirana analiza može otkriti da veza koja postoji ukupno ne postoji u svim podgrupama — ili obrnuto, da veza ne postoji ukupno, ali pojavljuje se u specifičnim podgrupama. A to nas vodi do jednog od najvažnijih koncepata u statistici kategoričkih podataka.

13 Simpsonov paradoks: kad ukupni rezultat laže

Simpsonov paradoks nastaje kad se smjer veze *preokrene* kad kontroliramo treću varijablu. To nije egzotični statistički kuriozitet — to je nešto što se redovito događa u komunikološkim istraživanjima, gdje se grupe razlikuju po veličini i karakteristikama.

Pogledajmo konstruirani ali realistični primjer. Dva portala uspoređuju click-through rate (CTR), ali svaki ima različitu mješavinu mobilnog i desktop prometa.

```
# Konstruirani primjer: dva portala, CTR po tipu uredaja
simpson <- tribble(
  ~portal, ~uredaj, ~klikovi, ~prikazi,
  "Portal A", "mobitel",  80, 1000,
  "Portal A", "desktop", 180,  500,
  "Portal B", "mobitel",  30,  500,
  "Portal B", "desktop",  80,  200
) |>
  mutate(ctr = round(klikovi / prikazi * 100, 1))

# Ukupni CTR
simpson |>
  group_by(portal) |>
  summarise(
    ukupno_klikovi = sum(klikovi),
    ukupno_prikazi = sum(prikazi),
    ukupni_ctr = round(ukupno_klikovi / ukupno_prikazi * 100, 1),
    .groups = "drop"
  )
```

```
# A tibble: 2 x 4
  portal   ukupno_klikovi ukupno_prikazi ukupni_ctr
  <chr>             <dbl>          <dbl>      <dbl>
1 Portal A            260           1500       17.3
2 Portal B            110            700       15.7
```

```
# CTR po uredaju
simpson |>
  select(portal, uredaj, ctr, prikazi)
```

```
# A tibble: 4 x 4
  portal   uredaj    ctr prikazi
  <chr>    <chr>   <dbl>   <dbl>
1 Portal A mobitel     8    1000
2 Portal A desktop    36     500
3 Portal B mobitel     6     500
4 Portal B desktop    40     200
```

Evo paradoksa. Portal B ima viši CTR na *svakom* uređaju pojedinačno — na mobitelu (6.0% vs 8.0%) i na desktopu (36.0% vs 40.0%). Ali kad agregirate — Portal A može imati viši ukupni CTR. Kako je to moguće?

Razlog leži u neravnomjernoj raspodjeli uređaja. Portal A ima mnogo više mobilnog prometa (1000 od 1500 prikaza), a mobilni promet ima nizak CTR. Portal B ima relativno više desktop prometa, koji ima visok CTR. Kad zbrojite sve prikaze i klikove, neravnomjerna mješavina "prevagne" i stvori obmanjujuće ukupne brojke.

> ⚠ Simpsonov paradoks nije rijedak
>
> U komunikološkim istraživanjima Simpsonov paradoks je čest jer se grupe (dobne, regionalne, platformske) razlikuju po veličini i karakteristikama. Kad god uspoređujete kategoričke podatke agregirano, postavite si pitanje — bi li se zaključak promijenio kad biste razdvojili podatke po nekoj relevantnoj trećoj varijabli? Ako niste sigurni, provedite stratificiranu analizu i pogledajte.

14 McNemarov test: kad isti ljudi odgovaraju dva puta

Svi testovi koje smo do sada vidjeli podrazumijevaju nezavisna opažanja. Ali što kad imate uparene kategoričke podatke — iste ispitanike mjerene dva puta? Za to postoji McNemarov test, koji je za kategoričke podatke ono što je upareni t-test za numeričke.

Zamislite sljedeću situaciju — 200 studenata je na početku semestra upitano preferiraju li online ili tiskane vijesti. Na kraju semestra postavljeno im je isto pitanje. Zanima vas je li se distribucija preferencija značajno promijenila.

```r
set.seed(42)

# Simulacija: pomak prema onlineu kroz semestar
n_mc <- 200
prije <- sample(c("online", "tisak"), n_mc, replace = TRUE, prob = c(0.55, 0.45))

# Poslije: neki presli na online, malo ih preslo na tisak
poslije <- prije
promijenili <- sample(1:n_mc, 40)
for (i in promijenili[1:30]) poslije[i] <- "online"
for (i in promijenili[31:40]) poslije[i] <- "tisak"

mc_tablica <- table(Prije = prije, Poslije = poslije)
mc_tablica
```

```
        Poslije
Prije    online tisak
```

```
  online     95     4
  tisak      15    86
```

Ovu tablicu treba čitati pažljivo. Na dijagonali su ispitanici koji *nisu* promijenili mišljenje — oni koji su i prije i poslije birali isti medij. Izvan dijagonale su oni koji jesu promijenili. McNemarov test ne gleda dijagonalu. On testira je li broj promjena u jednom smjeru (tisak → online) značajno veći od broja promjena u drugom smjeru (online → tisak).

```
mcnemar.test(mc_tablica)
```

```
    McNemar's Chi-squared test with continuity correction

data:  mc_tablica
McNemar's chi-squared = 5.2632, df = 1, p-value = 0.02178
```

```
# Koliko je preslo u kojem smjeru?
tisak_na_online <- mc_tablica["tisak", "online"]
online_na_tisak <- mc_tablica["online", "tisak"]

cat("Presli tisak na online:", tisak_na_online, "\n")
```

```
Presli tisak na online: 15
```

```
cat("Presli online na tisak:", online_na_tisak, "\n")
```

```
Presli online na tisak: 4
```

```
cat("Neto pomak prema onlineu:", tisak_na_online - online_na_tisak, "\n")
```

```
Neto pomak prema onlineu: 11
```

Više ispitanika je prešlo s tiska na online nego obrnuto. McNemarov test govori je li ta asimetrija statistički značajna.

> **!** Kad koristiti McNemarov test
>
> McNemarov test koristite kad imate iste ispitanike mjerene dva puta na istoj binarnoj varijabli. Primjeri iz komunikologije uključuju preferenciju medija prije i poslije kampanje, stav prema brandu prije i poslije izlaganja reklami, izbor komunikacijskog kanala prije i poslije redizajna. Ključno je da su podaci upareni — isti ljudi, dva mjerenja.

15 Kompletna analiza: tri pitanja za upravu

Spojimo sve naučeno u kompletnu analizu. Istražujemo tri pitanja za upravu medijske kuće. Prvo, ovisi li tip medija o dobi? Drugo, razlikuju li se regije u digitalnim navikama? Treće, postoji li veza između obrazovanja i preferencije sadržaja?

```
# Pitanje 1: Dob x tip medija (detaljno, svih 6 tipova)
tab1 <- table(survey$age_group, survey$media_type)
test_q1 <- chisq.test(tab1)
v1 <- sqrt(test_q1$statistic / (sum(tab1) * (min(dim(tab1)) - 1)))

cat("=== PITANJE 1: Dob x Tip medija ===\n")
```

```
=== PITANJE 1: Dob x Tip medija ===
```

```
cat("chi2(", (nrow(tab1)-1)*(ncol(tab1)-1), ") = ", round(test_q1$statistic, 1),
    ", p < 0.001, V = ", round(v1, 2), "\n", sep = "")
```

```
chi2(15) = 233.6, p < 0.001, V = 0.31
```

```
cat("Interpretacija: Jaka veza. Mladi preferiraju streaming i portale,\n")
```

```
Interpretacija: Jaka veza. Mladi preferiraju streaming i portale,
```

```
cat("stariji TV i radio.\n\n")
```

```
stariji TV i radio.
```

```
# Koji reziduali su najjaci?
rez_df <- as.data.frame(test_q1$residuals) |>
  as_tibble() |>
  rename(age = Var1, media = Var2, r = Freq) |>
  filter(abs(r) > 2) |>
  arrange(desc(abs(r)))

cat("Celije s |rezidual| > 2:\n")
```

```
Celije s |rezidual| > 2:
```

```r
rez_df |>
  mutate(r = round(r, 1), smjer = if_else(r > 0, "VISE nego ocekivano", "MANJE nego ocekiv
  print(n = 20)
```

```
# A tibble: 13 x 4
   age   media          r smjer
   <fct> <fct>      <dbl> <chr>
 1 18-29 streaming    6.4 VISE nego ocekivano
 2 60+   TV           5.9 VISE nego ocekivano
 3 60+   streaming   -5.2 MANJE nego ocekivano
 4 18-29 TV          -4.2 MANJE nego ocekivano
 5 60+   radio        4.1 VISE nego ocekivano
 6 30-44 podcast      3.8 VISE nego ocekivano
 7 30-44 TV          -3.8 MANJE nego ocekivano
 8 18-29 radio       -3.5 MANJE nego ocekivano
 9 60+   podcast     -3.3 MANJE nego ocekivano
10 60+   web_portal  -3.2 MANJE nego ocekivano
11 60+   tisak        3.1 VISE nego ocekivano
12 45-59 TV          2.5 VISE nego ocekivano
13 45-59 streaming   -2.5 MANJE nego ocekivano
```

```r
# Pitanje 2: Regija x tip medija (digitalni vs tradicionalni)
tab2 <- table(survey$region, survey$media_grupa)
test_q2 <- chisq.test(tab2)
v2 <- sqrt(test_q2$statistic / (sum(tab2) * (min(dim(tab2)) - 1)))

cat("=== PITANJE 2: Regija x Digitalni/Tradicionalni ===\n")
```

```
=== PITANJE 2: Regija x Digitalni/Tradicionalni ===
```

```r
cat("chi2(", (nrow(tab2)-1)*(ncol(tab2)-1), ") = ", round(test_q2$statistic, 2),
    ", p = ", round(test_q2$p.value, 4), ", V = ", round(v2, 3), "\n", sep = "")
```

```
chi2(4) = 9.01, p = 0.0609, V = 0.106
```

```r
cat("Interpretacija:", if_else(test_q2$p.value < 0.05,
    "Postoji znacajna razlika medju regijama.",
    "Nema znacajne razlike medju regijama."), "\n\n")
```

```
Interpretacija: Nema znacajne razlike medju regijama.
```

```r
# Proporcije digitalnih po regiji
survey |>
  group_by(region) |>
  summarise(
    n = n(),
    udio_digitalni = round(mean(media_grupa == "digitalni") * 100, 1),
    .groups = "drop"
  ) |>
  arrange(desc(udio_digitalni))
```

```
# A tibble: 5 x 3
  region           n udio_digitalni
  <chr>        <int>          <dbl>
1 Sjeverozapad   137           65.7
2 Slavonija      165           61.8
3 Zagreb         258           55
4 Dalmacija      139           54
5 Primorje       101           49.5
```

```r
# Pitanje 3: Obrazovanje x preferencija sadrzaja
tab3 <- table(survey$education, survey$content_preference)
chi3 <- chisq.test(tab3)
v3 <- sqrt(chi3$statistic / (sum(tab3) * (min(dim(tab3)) - 1)))

cat("=== PITANJE 3: Obrazovanje x Preferencija sadrzaja ===\n")
```

```
=== PITANJE 3: Obrazovanje x Preferencija sadrzaja ===
```

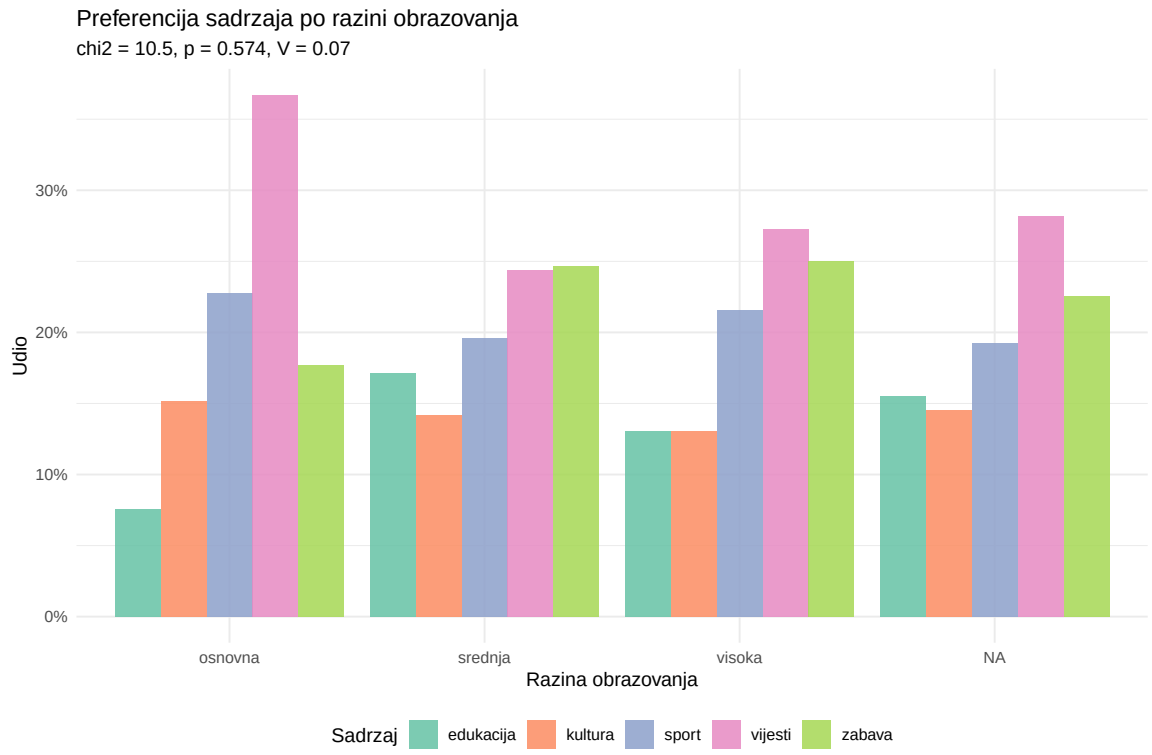
```r
cat("chi2(", (nrow(tab3)-1)*(ncol(tab3)-1), ") = ", round(chi3$statistic, 2),
    ", p = ", round(chi3$p.value, 4), ", V = ", round(v3, 3), "\n", sep = "")
```

```
chi2(12) = 10.48, p = 0.5742, V = 0.066
```

```r
cat("Interpretacija:", if_else(chi3$p.value < 0.05,
    "Postoji znacajna veza.",
    "Nema znacajne veze."), "\n")
```

```
Interpretacija: Nema znacajne veze.
```

```r
# Vizualizacija
survey |>
  mutate(education = factor(education, levels = c("osnovna", "srednja", "visa", "visoka"))
  count(education, content_preference) |>
  group_by(education) |>
  mutate(udio = n / sum(n)) |>
  ungroup() |>
  ggplot(aes(x = education, y = udio, fill = content_preference)) +
  geom_col(position = "dodge", alpha = 0.85) +
  scale_y_continuous(labels = scales::label_percent()) +
  scale_fill_brewer(palette = "Set2") +
  labs(
    title = "Preferencija sadrzaja po razini obrazovanja",
    subtitle = paste0("chi2 = ", round(chi3$statistic, 1), ", p = ", round(chi3$p.value, 3
                      ", V = ", round(v3, 2)),
    x = "Razina obrazovanja",
    y = "Udio",
    fill = "Sadrzaj"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



```
# Sazetak svih testova
cat("================================================\n")
```

```
================================================
```

```
cat("   SAZETAK ANALIZE KATEGORICKIH PODATAKA\n")
```

```
   SAZETAK ANALIZE KATEGORICKIH PODATAKA
```

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

```
================================================
```

```
tibble(
  pitanje = c("Dob x Tip medija", "Regija x Dig./Trad.", "Obrazovanje x Sadrzaj"),
  chi2 = round(c(test_q1$statistic, test_q2$statistic, chi3$statistic), 1),
  df = c((nrow(tab1)-1)*(ncol(tab1)-1), (nrow(tab2)-1)*(ncol(tab2)-1), (nrow(tab3)-1)*(nco
  p = c(format(test_q1$p.value, scientific = TRUE, digits = 2),
        round(test_q2$p.value, 4), round(chi3$p.value, 4)),
  V = round(c(v1, v2, v3), 3),
  velicina = c(
    if_else(v1 > 0.29, "veliki", if_else(v1 > 0.17, "srednji", "mali")),
    if_else(v2 > 0.29, "veliki", if_else(v2 > 0.17, "srednji", "mali")),
    if_else(v3 > 0.29, "veliki", if_else(v3 > 0.17, "srednji", "mali"))
  )
)
```

```
# A tibble: 3 x 6
  pitanje                chi2    df p           V velicina
  <chr>                 <dbl> <dbl> <chr>   <dbl> <chr>
1 Dob x Tip medija      234.     15 2.9e-41 0.312 veliki
2 Regija x Dig./Trad.     9       4 0.0609  0.106 mali
3 Obrazovanje x Sadrzaj  10.5    12 0.5742  0.066 mali
```

Analiza otkriva jasan obrazac. Veza između dobi i tipa medija je najjača i najvažnija — generacijski jaz u medijskim navikama je velik i statistički nedvojben. Regionalne razlike u digitalnim vs tradicionalnim medijima su znatno manje. Veza obrazovanja i preferencije sadržaja ovisi o uzorku. Za upravu je ključna poruka — budućnost je u digitalnim platformama, ali televizija i radio još uvijek imaju ogromnu publiku među starijim generacijama.

16 Pet pogrešaka koje ne smijete napraviti

Hi-kvadrat testovi su jednostavni za provesti, ali iznenađujuće lako ih je krivo primijeniti ili interpretirati. Evo pet najčešćih pogrešaka.

Unošenje postotaka umjesto frekvencija. Hi-kvadrat test zahtijeva apsolutne frekvencije (brojeve), ne postotke ili proporcije. Ako unesete `chisq.test(c(0.30, 0.20, 0.50))`, R će misliti da ukupno imate jedno opažanje i dat će besmisleni rezultat.

```
# KRIVO: postoci
cat("KRIVO (postoci):\n")
```

```
KRIVO (postoci):
```

```
chisq.test(c(0.30, 0.20, 0.50))
```

```

    Chi-squared test for given probabilities

data:  c(0.3, 0.2, 0.5)
X-squared = 0.14, df = 2, p-value = 0.9324
```

```
cat("\nISPRAVNO (frekvencije):\n")
```

```
ISPRAVNO (frekvencije):
```

```
chisq.test(c(30, 20, 50))
```

```

    Chi-squared test for given probabilities

data:  c(30, 20, 50)
X-squared = 14, df = 2, p-value = 0.0009119
```

Pretjerana granularnost. Tablica 10×8 s ukupno 100 opažanja imat će mnogo ćelija s malim frekvencijama. Bolje je spojiti kategorije u smislene grupe i imati manje ali punije ćelije.

Kauzalna interpretacija. Hi-kvadrat test detektira asocijaciju, ne kauzalnost. Činjenica da su dob i medijski tip povezani ne znači da dob *uzrokuje* preferenciju — možda je posrijedi obrazovanje, socioekonomski status ili kohorta efekt. Za kauzalne zaključke trebate eksperimentalni dizajn ili napredne statističke metode.

Zaboravljanje veličine učinka. Kao i kod t-testa, p-vrijednost ovisi o veličini uzorka. S dovoljno velikim uzorkom, čak i trivijalna veza postaje "statistički značajna." Uvijek

izvijestite Cramérovo V uz ^{2} i p — jer Cramérovo V govori koliko je veza praktično jaka, neovisno o n.

Višestruko testiranje bez korekcije. Ako testirate 10 parova varijabli bez korekcije, šansa da bar jedan test bude lažno pozitivan je oko 40%. Koristite Benjamini-Hochberg (BH) korekciju kad testirate više parova.

```
# Testiramo sve parove kategorickih varijabli
parovi <- list(
  c("age_group", "media_type"),
  c("age_group", "content_preference"),
  c("age_group", "media_grupa"),
  c("gender", "media_type"),
  c("gender", "content_preference"),
  c("education", "media_type"),
  c("education", "content_preference"),
  c("region", "media_grupa")
)

multi_chi <- map_df(parovi, \(par) {
  tab <- table(survey[[par[1]]], survey[[par[2]]])
  test <- chisq.test(tab)
  tibble(
    var1 = par[1],
    var2 = par[2],
    chi2 = round(test$statistic, 1),
    p = test$p.value
  )
}) |>
  mutate(
    p_adj = p.adjust(p, method = "BH"),
    znacajno_orig = p < 0.05,
    znacajno_adj = p_adj < 0.05
  ) |>
  arrange(p)

multi_chi |>
  mutate(p = format(p, scientific = TRUE, digits = 2),
         p_adj = format(p_adj, scientific = TRUE, digits = 2))
```

```
# A tibble: 8 x 7
  var1      var2               chi2 p       p_adj   znacajno_orig znacajno_adj
  <chr>     <chr>             <dbl> <chr>   <chr>   <lgl>         <lgl>
1 age_group media_grupa        199.  7.3e-43 5.9e-42 TRUE          TRUE
2 age_group media_type         234.  2.9e-41 1.2e-40 TRUE          TRUE
3 age_group content_preference  59.7 2.5e-08 6.8e-08 TRUE          TRUE
```

```
4 education media_type          25.5 4.4e-02 8.8e-02 TRUE          FALSE
5 region    media_grupa          9   6.1e-02 9.8e-02 FALSE         FALSE
6 education content_preference  10.5 5.7e-01 7.2e-01 FALSE         FALSE
7 gender    media_type           3.5 6.3e-01 7.2e-01 FALSE         FALSE
8 gender    content_preference   0.4 9.8e-01 9.8e-01 FALSE         FALSE
```

Nakon BH korekcije, neki marginalno značajni rezultati mogu nestati. To je cijena korektnog pristupa — ali bolje je imati manje rezultata u koje možete vjerovati nego više rezultata koji su možda lažni.

17 Funkcija koja obavlja sve za vas

U praksi ćete hi-kvadrat analizu ponavljati za mnogo parova varijabli. Umjesto da svaki put ručno prolazite sve korake, napišimo funkciju koja automatizira cijeli postupak — provjeri pretpostavke, odabere odgovarajući test, izračuna veličinu učinka i ispiše izvještaj.

```
chi_izvjestaj <- function(data, var1, var2) {
  tab <- table(data[[var1]], data[[var2]])
  test <- chisq.test(tab)
  v <- sqrt(test$statistic / (sum(tab) * (min(dim(tab)) - 1)))
  min_exp <- min(test$expected)

  cat("========================================\n")
  cat(var1, "x", var2, "\n")
  cat("========================================\n")
  cat("Dimenzije tablice:", nrow(tab), "x", ncol(tab), "\n")
  cat("Najm. ocekivana frekvencija:", round(min_exp, 1), "\n")

  if (min_exp < 5) {
    cat("Ocekivane frekvencije < 5. Koristim Fisherov test.\n")
    fisher <- fisher.test(tab, simulate.p.value = TRUE)
    cat("P-vrijednost (Fisher):", round(fisher$p.value, 4), "\n")
  } else {
    cat("chi2(", (nrow(tab)-1)*(ncol(tab)-1), ") = ",
        round(test$statistic, 2), "\n", sep = "")
    cat("P-vrijednost:", format(test$p.value, scientific = TRUE, digits = 3), "\n")
  }

  cat("Cramerovo V:", round(v, 3), "\n")
  cat("Odluka:", if_else(test$p.value < 0.05,
    "Postoji statisticki znacajna veza.",
    "Nema statisticki znacajne veze."), "\n\n")

  invisible(list(table = tab, test = test, V = v))
```

```
}

# Primjeri koristenja
chi_izvjestaj(survey, "age_group", "media_type")
```

```
====================================
age_group x media_type
====================================
Dimenzije tablice: 4 x 6
Najm. ocekivana frekvencija: 10.1
chi2(15) = 233.59
P-vrijednost: 2.93e-41
Cramerovo V: 0.312
Odluka: Postoji statisticki znacajna veza.
```

```
chi_izvjestaj(survey, "gender", "content_preference")
```

```
====================================
gender x content_preference
====================================
Dimenzije tablice: 2 x 5
Najm. ocekivana frekvencija: 54.2
chi2(4) = 0.41
P-vrijednost: 9.82e-01
Cramerovo V: 0.023
Odluka: Nema statisticki znacajne veze.
```

Ova funkcija automatski provjerava pretpostavke, odabire odgovarajući test i računa veličinu učinka. Možete je koristiti u svim budućim analizama kategoričkih podataka — kopijte je u svoje skripte i prilagodite prema potrebi.

18 Pregled svih testova za kategoričke podatke

```
tribble(
  ~test, ~situacija, ~R_kod,
  "chi2 goodness-of-fit", "Jedna varijabla vs ocekivana distribucija", "chisq.test(frekven
  "chi2 nezavisnosti", "Veza dviju kategorickih varijabli", "chisq.test(table(x, y))",
  "Fisherov egzaktni", "Male ocekivane frekvencije ili mali n", "fisher.test(table(x, y))"
  "McNemarov test", "Uparene kategoricke varijable", "mcnemar.test(table(prije, poslije))"
)
```

```
# A tibble: 4 x 3
  test                 situacija                                  R_kod
  <chr>                <chr>                                      <chr>
1 chi2 goodness-of-fit Jedna varijabla vs ocekivana distribucija chisq.test(fre~
2 chi2 nezavisnosti    Veza dviju kategorickih varijabli          chisq.test(tab~
3 Fisherov egzaktni    Male ocekivane frekvencije ili mali n      fisher.test(ta~
4 McNemarov test       Uparene kategoricke varijable              mcnemar.test(t~
```

! Ključni zaključci

Hi-kvadrat testovi su za kategoričke varijable. Test za dobrotu prilagodbe uspoređuje distribuciju jedne varijable s očekivanom. Test nezavisnosti testira postoji li veza između dviju varijabli.

² statistika mjeri udaljenost od očekivanog. Formula je ² = Σ(O - E)² / E. Veći ² znači jači dokaz protiv H . Stupnjevi slobode za test nezavisnosti su (r - 1)(c - 1).

Očekivane frekvencije su ključ za razumijevanje. Pod H , E = (redak total × stupac total) / ukupno. Pretpostavka — sve E 5. Kad to nije zadovoljeno, koristite Fisherov egzaktni test.

Reziduali govore *gdje* je veza najjača. Ukupni ² kaže da veza postoji. Standardizirani reziduali s |r| > 2 identificiraju specifične ćelije. Uvijek ih izvijestite — jer urednica ne želi znati samo "postoji veza", nego *kakva* je.

Cramérovo V mjeri jačinu veze. Raspon od 0 do 1. Za k = 4: V 0.06 mali, V 0.17 srednji, V 0.29 veliki učinak. Uvijek ga izvijestite uz ² i p.

Fisherov test kad su ćelije premale. Točan test bez aproksimacije. Za 2×2 tablice automatski daje odds ratio. U praksi ga možete koristiti uvijek — za velike uzorke daje iste rezultate kao ².

Spajanje kategorija je legitimno i korisno. Digitalni vs tradicionalni mediji je informativnije od šest odvojenih kategorija. Smisleno kolapsiranje povećava frekvencije i preglednost.

Simpsonov paradoks upozorava na opasnost agregiranja. Ukupni rezultati mogu biti obmanjujući. Uvijek provjerite rezultate stratificirane po relevantnoj trećoj varijabli.

McNemarov test za uparene kategoričke podatke. Isti ispitanici, dva mjerenja, binarna varijabla. Ekvivalent uparenom t-testu u kategoričkom svijetu.

Asocijacija nije kauzalnost. Veza dobi i medijskog tipa ne znači da dob uzrokuje preferenciju. Za kauzalne zaključke trebate eksperimentalni dizajn.

Višestruko testiranje zahtijeva korekciju. BH ili Bonferroni korekcija kad testirate mnogo parova varijabli. Bolje je imati manje pouzdanih rezultata nego više upitnih.

19 Zadaci za pripremu

1. Učitajte `media_survey_chi2.csv`. Testirajte postoji li veza između regije i preferencije sadržaja (`content_preference`). Izračunajte Cramérovo V i interpretirajte ga. Koji reziduali su najjači?

2. Kreirajte novu varijablu `zadovoljni` (satisfaction >= 4 = "da", inače "ne"). Testirajte postoji li veza između `media_grupa` (digitalni/tradicionalni) i `zadovoljni` pomoću Fisherovog egzaktnog testa. Interpretirajte odds ratio.

3. Napišite funkciju `chi_vizualizacija(data, var1, var2)` koja prima podatke i imena dviju varijabli te automatski crta grupani barplot s rezultatima testa u podnaslovu.

20 Dodatno čitanje

Obavezno

Navarro, D. (2018). *Learning Statistics with R*, Chapter 12 (Categorical Data Analysis). Besplatno dostupno na learningstatisticswithr.com. Pokriva hi-kvadrat testove s R kodom.

Preporučeno

Agresti, A. (2018). *An Introduction to Categorical Data Analysis* (3rd edition). Wiley. Poglavlja 1-3. Referentni udžbenik za kategoričke podatke.

Wickham, H. & Grolemund, G. (2017). *R for Data Science*. O'Reilly. Besplatno na r4ds.had.co.nz. Poglavlja o faktorima i vizualizaciji kategoričkih podataka.

21 Pojmovnik

Pojam	Objašnjenje
Kategorička varijabla	Varijabla čije su vrijednosti kategorije (npr. spol, regija, tip medija). Nema smisleni numerički poredak (nominalna) ili ima (ordinalna).
Kontingencijska tablica	Tablica frekvencija za sve kombinacije dviju kategoričkih varijabli. Temelj za test nezavisnosti.
Hi-kvadrat statistika	Mjera ukupnog odstupanja opaženih od očekivanih frekvencija.
Goodness-of-fit test	Testira odgovara li distribucija jedne varijable očekivanoj distribuciji. df = k minus 1.
Test nezavisnosti	Testira postoji li veza između dviju kategoričkih varijabli. df = (r minus 1)(c minus 1).
Očekivana frekvencija	Frekvencija pod H0. Za test nezavisnosti: E = (redak total puta stupac total) / ukupno.

Pojam	Objašnjenje
Standardizirani rezidual	(O minus E) / korijen(E). Doprinos svake ćelije ukupnom chi2. Značajno ako
Cramérovo V	Mjera veličine učinka za hi-kvadrat test. V = korijen(chi2/(n(k minus 1))). Raspon 0 do 1.
Fisherov egzaktni test	Točan test za male uzorke ili male očekivane frekvencije. Ne koristi chi2 aproksimaciju.
Odds ratio (omjer šansi)	Mjera asocijacije za 2x2 tablice. OR = 1 znači nema veze. OR > 1 ili < 1 znači veza postoji.
Yatesova korekcija	Korekcija kontinuiteta za 2x2 tablice. R je primjenjuje po defaultu u `chisq.test()`.
McNemarov test	Test za uparene kategoričke podatke (isti ispitanici, dva mjerenja).
Simpsonov paradoks	Smjer veze se promijeni kad kontroliramo treću varijablu. Agregirani rezultati obmanjuju.
Stratificirana analiza	Provođenje testa odvojeno za podgrupe treće varijable. Otkriva Simpsonov paradoks.
Spajanje kategorija	Kolapsiranje rijetkih kategorija u šire grupe. Povećava očekivane frekvencije i preglednost.
`chisq.test()`	R funkcija za hi-kvadrat test. Prima vektor frekvencija (gof) ili kontingencijsku tablicu.
`fisher.test()`	R funkcija za Fisherov egzaktni test. Prima kontingencijsku tablicu.
`mcnemar.test()`	R funkcija za McNemarov test. Prima 2x2 tablicu uparenih podataka.
`table()`	R funkcija za kreiranje kontingencijske tablice. `table(x, y)` za dvije varijable.
`prop.table()`	Pretvara frekvencije u proporcije. `margin = 1` za retke, `margin = 2` za stupce.
`p.adjust()`	Korekcija p-vrijednosti za višestruko testiranje. `method = "BH"` je preporučen.