Statistički metodi

Praktični pristup u R-u

Aleksandar Tomašević

05.02.25

Sadržaj

# Osnovne informacije

Dobro došli na *HTML* verziju udžbenika *Statistički metodi: Praktični pristup u R-u*.

Meni sa desne strane vam omogućava da brzo pređete na željeno poglavlje. Ukoliko koristite mobilni uređaj, meni možete otvoriti klikom na gornji levi ugao ekrana.

Udžbenik će zauvek biti dostupan besplatno i isključivo u elektronskom obliku.

**PDF verzija** udžbenika je dostupna u okviru Digitalne biblioteke Filozofskog fakulteta: <https://digitalna.ff.uns.ac.rs/sadrzaj/2025/978-86-6065-902-8>.

Ukoliko imate pitanja, nedoumice ili sugestije vezane za formu ili sadržaj ovog udžbenika, slobodno me [kontaktirajte](mailto:atomashevic@ff.uns.ac.rs). Više informacija o meni možete pronaći na [www.atomasevic.com](https://www.atomasevic.com) (engleski) ili na sajtu [Filozofskog fakulteta u Novom Sadu](https://www.ff.uns.ac.rs/sr/fakultet/odseci/sociologija/zaposleni/aleksandar-tomasevic) (srpski).

## Impresum

ISBN: 978-86-6065-902-8

Recezenti: prof. dr Valentina Sokolovska, prof. dr Tanja Jevremov

Izdavač: Filozofski fakultet u Novom Sadu, 2025

Digitalna biblioteka Filozofskog fakulteta: <https://digitalna.ff.uns.ac.rs/sadrzaj/2025/978-86-6065-902-8>

## Citiranje

Tomašević, A. (2025). *Statistički metodi: Praktični pristup u R-u*. Filozofski fakultet u Novom Sadu.

@BOOK{tomasevic2025statis,  
 title = "Statistički metodi: Praktični pristup u R-u",  
 author = "Tomašević, Aleksandar",  
 publisher = "Filozofski fakultet u Novom Sadu",  
 address = "Novi Sad",  
 year = 2025,  
 isbn = 9788660659028  
}

# 1. Uvod

## 1.1 Kome je ovaj udžbenik namenjen?

Ovaj udžbenik je namenjen svima koji su tokom studija ili profesionalnog usavršavanja prepoznali potrebu za primenom statističkih metoda. Optimalno predznanje za praćenje ovog udžbenika obuhvata poznavanje osnovnih pojmova deskriptivne statistike. Međutim, to nije strogo neophodno - u narednom poglavlju ([Odeljak 2.3](#sec-ds)) ćemo obnoviti ključne koncepte deskriptivne statistike (aritmetičku sredinu, standardnu devijaciju) dok se istovremeno upoznajemo sa programskim jezikom R. Ako ste ranije pohađali bilo kakav osnovni kurs statistike, bilo na fakultetu ili putem interneta, verovatno ste se već susreli sa opisivanjem i tumačenjem podataka. To je prvi korak ka kvantitativnom sagledavanju problema u vašoj disciplini. Ovaj udžbenik vas vodi dalje i uvodi u svet **statističkog zaključivanja**.

A šta je zapravo statističko zaključivanje? U osnovi, to je primena statističkih metoda za rešavanje konkretnih problema. Međutim, važno je napomenuti da retko dobijamo jednoznačne odgovore primenom samo jedne metode. Upravo zato je cilj ovog udžbenika da vas vodi kroz osnovne korake ovog procesa, koristeći praktičan pristup.

## 1.2 Šta je praktičan pristup?

Praktičan pristup podrazumeva da se **svi koraci zaključivanja prevode u računarski kod**. Ovaj pristup se razlikuje od *point and click* principa koji je karakterističan za statističke programe poput **SPSS**-a ili **JASP**-a. U tim programima korisnik najčešće definiše osnovne parametre analize i odmah dobija rezultate, dok su međukoraci sakriveni od pogleda.

Koja je mana ovakvog pristupa? Zamislimo analogiju sa učenjem klavira. Kurs bi počeo solfeđom i učenjem notnog zapisa, nastavio bi se vežbama prstiju, i tu bi se zaustavio. Kada dođe trenutak za nastup, seli biste za sintisajzer, [izabrali jednu od predefinisanih kompozicija i pritisnuli *play*](https://www.youtube.com/watch?v=YTX-4mqITEA). Verovatno biste razvili dobro poznavanje kompozicije i istančan sluh za prepoznavanje tonova i ritma, ali da li biste zaista savladali sviranje instrumenta? Odgovor je jasan - ne biste.

Tradicionalna nastava statistike često liči na ovaj scenario. Obično počinjemo od formalizma, savladavajući statističku notaciju (kao notni zapis), zatim prelazimo na teoriju verovatnoće (kao solfeđo), nakon toga na analitičke postupke sprovođenja statističkih metoda (kao vežbe prstiju) i tu zastajemo. Kada dođe trenutak za praktični rad, otvaramo SPSS (naš statistički „sintisajzer“), biramo analizu koju želimo da primenimo na određenom skupu podataka (našu „kompoziciju“), podešavamo parametre analize (kao „tempo“ i „tonalitet“) i brzo dobijamo rezultate.

Spoj teorije i prakse u takvom pristupu predstavlja značajan izazov. Studenti često usmeravaju pažnju na delove kursa koji su im neophodni za savladavanje neposrednih prepreka - kolokvijuma ili ispita. Razumevanje logike i principa statističkog zaključivanja u velikoj meri zavisi od sposobnosti nastavnika da kreira kurs koji efikasno povezuje teoriju i praksu, kako kroz nastavu, tako i kroz različite oblike provere znanja.

U ovom udžbeniku prevodićemo logičke korake statističkog zaključivanja u računarski kod koristeći programski jezik R. Fokusiraćemo se na suštinski kod, koristeći samo ono što je neophodno za dolazak do zaključka. Važna napomena - prethodno znanje R-a nije preduslov za praćenje ovog udžbenika, niti je cilj da kroz njega detaljno savladate R. Statističko zaključivanje i programiranje su dve odvojene veštine, svaka kompleksna na svoj način, a naš fokus je na prvoj.

Nećemo se upuštati u izradu složenih vizuelizacija, obradu i transformaciju podataka, optimizaciju koda ili napredne statističke modele. Naš glavni cilj je da kroz praktičnu primenu koda razvijemo intuiciju o statističkom zaključivanju i učinimo proces donošenja zaključaka jasnim i pristupačnim. Ako savladate materijal iz ovog udžbenika, nećete steći sve veštine potrebne za naprednu analizu podataka u R-u. Nećete upoznati nijedan od ključnih paketa u R-u koji se koriste za obradu, analizu i statističko modelovanje (poput tidyverse ili easystats). Naš cilj je konkretniji i jednostavniji: osnove statističkog zaključivanja u R-u, korak po korak, linija po linija koda, bez dodatnih paketa.

## 1.3 Zašto je sve ovo važno?

Ako ste studentkinja ili student društvenih nauka, tokom studija ćete primetiti da su ove discipline (izuzev humanističkih) dominantno kvantitativno orijentisane. To znači da je za praćenje aktuelnih istraživanja, razumevanje rezultata i interpretaciju zaključaka neophodno poznavanje statističkih metoda i osnovnih principa statističkog zaključivanja. Čak i ako nikada ne sprovedete sopstveno kvantitativno istraživanje, vrlo verovatno ćete koristiti rezultate tuđih istraživanja. Međutim, sami rezultati imaju malu vrednost ako ne razumete principe i logiku kojima se od sirovih podataka dolazi do zaključaka.

Prevođenje logičkih koraka u računarski kod pri rezonovanju i donošenju zaključaka predstavlja jednu od ključnih veština za profesionalni razvoj u 21. veku. Tržište rada sve više vrednuje sposobnost kvantitativnog razmišljanja i rada sa podacima.

Za studente koji pohađaju kurs **Statistički metodi** na Filozofskom fakultetu, ovo je tek početak. Kroz kurseve **Multivarijantna analiza** i **Uvod u mrežnu analizu** imaćete priliku da unapredite svoje znanje statistike i programskog jezika *R*. Ovi kursevi nude praktična znanja koja su direktno primenjiva u različitim profesionalnim okruženjima.

## 1.4 Zadaci i interaktivni kod

Na kraju svakog poglavlja nalazi se odeljak „Zadaci“ koji vam omogućava praktičnu primenu naučenog gradiva. Osnovni zadaci traže reprodukciju analiza iz poglavlja na novim skupovima podataka, uz pitanja koja će vas podstaći na dublje razumevanje sprovedenih koraka. Složeniji zadaci, označeni sa jednom (\*) ili dve (\*\*) zvezdice, zahtevaju modifikaciju postojećih analiza ili primenu naprednih metoda koje prevazilaze okvire poglavlja.

Pri prvom čitanju udžbenika, preporučujem da uradite samo prvi zadatak i nastavite dalje. Nakon što savladate celokupno gradivo, možete se vratiti zahtevnijim zadacima.

Ukoliko primetite grešku u nekom zadatku ili vam je potrebno pojašnjenje određenog problema, slobodno me kontaktirajte.

Za rešavanje zadataka na raspolaganju su vam dve opcije.

1. Koristite interaktivnu konzolu koja se nalazi na kraju svakog poglavlja. Konzola vam omogućava direktno izvršavanje R koda i trenutni prikaz rezultata. Eksperimentišite sa kodom, menjajte parametre i posmatrajte kako te izmene utiču na rezultate. Zapamtite - konzola ne pamti vaše promene nakon zatvaranja stranice, ali kod uvek možete sačuvati kao skriptu na vašem računaru.
2. Instalirajte R i RStudio na svom računaru. Detaljno uputstvo za instalaciju nalazi se na [ovoj stranici](https://rstudio-education.github.io/hopr/starting.html). Ovaj pristup vam daje potpunu kontrolu nad svojim kodom i rezultatima, omogućavajući rad sa zadacima i podacima i van okvira ovog udžbenika.

## 1.5 Organizacija udžbenika

U [drugom poglavlju](R.qmd) upoznajemo se sa programskim jezikom R. Kroz obnavljanje osnovnih pojmova deskriptivne statistike, naučićemo kako da koristimo R u praksi.

[Treće poglavlje](sz.qmd) nas vodi kroz srž statističkog zaključivanja - kako iz uzorka izvoditi validne zaključke o populaciji.

[Četvrto poglavlje](verovatnoca.qmd) postavlja temelje teorije verovatnoće, dok istovremeno produbljujemo naše poznavanje R-a.

[Peto poglavlje](normalna.qmd) nam predstavlja normalnu distribuciju - statistički alat koji je ključan za razumevanje većine statističkih metoda.

U [šestom poglavlju](ztest.qmd) rešavamo konkretan istraživački problem od početka do kraja. Ovde se upoznajemo sa testiranjem hipoteza, intervalima poverenja i konceptom statističke značajnosti.

[Sedmo poglavlje](ttest.qmd) nas uvodi u t-test - osnovni statistički metod za poređenje aritmetičkih sredina dva nezavisna uzorka.

[Osmo poglavlje](regresija.qmd) predstavlja linearnu regresiju u svojoj najjednostavnijoj formi - moćan alat za modelovanje odnosa između varijabli.

U [devetom poglavlju](korelacija.qmd) istražujemo korelaciju i linearnu zavisnost između kvantitativnih varijabli, nadovezujući se na koncepte iz prethodnog poglavlja.

[Deseto poglavlje](anova.qmd) proširuje logiku t-testa i uvodi analizu varijanse (ANOVA). Ovaj metod nam omogućava poređenje više od dva nezavisna uzorka. Takođe, otkrivamo neočekivanu vezu između ANOVA-e i regresije.

U [jedanaestom poglavlju](olm.qmd) dublje istražujemo ovu vezu i otkrivamo zašto naslov ovog udžbenika nije sasvim precizan.

## 1.6 Tehničke napomene

Za najbolje iskustvo učenja, preporučujemo korišćenje [*web* verzije](https://sm.atomasevic.com) umesto PDF formata. Iako oba formata sadrže identičan sadržaj, HTML verzija nudi nekoliko ključnih prednosti - jednostavno kopiranje R koda, rad sa interaktivnom konzolom i efikasno pretraživanje sadržaja. PDF verzija služi kao praktična alternativa za čitanje bez pristupa internetu ili za štampanje materijala. Imajte na umu da je internet veza neophodna za preuzimanje podataka i rad sa interaktivnim konzolama.

Ovaj udžbenik je razvijen pomoću Quarto (Allaire i ostali, 2022) sistema koji omogućava integraciju teksta i koda u jedinstven dokument. Za implementaciju interaktivnog koda korišćen je webR (Stagg i ostali, 2023) sistem.

Svaka statistička analiza u udžbeniku praćena je odgovarajućim računarskim kodom koji možete direktno primeniti u vašem R okruženju. Izuzetak čine složeniji grafikoni čiji bi duži kod narušio čitljivost teksta. Kompletan izvorni kod za sve grafikone dostupan je u javnom repozitorijumu udžbenika.

Radi doslednosti između rezultata u tekstu i onih koje ćete dobiti u R konzoli, koristimo decimalni separator . umesto uobičajenog ,. Moguća su i druga manja odstupanja od jezičkih konvencija zbog usklađivanja sa standardnom statističkom i tehničkom notacijom u R-u. Ovaj udžbenik je tehnički dokument koji je prvenstveno namenjen učenju i primeni statističkih metoda u R-u. Sve greške i propusti su isključivo odgovornost autora.

# 2. Upoznavanje sa R-om

## 2.1 Šta je R?

R (R Core Team, 2024) je veoma koristan programski jezik primenljiv u brojnim oblastima. Ovaj svestrani alat služi za razvoj aplikacija, analizu podataka, statističke proračune, vizualizaciju informacija, mašinsko učenje, obradu slika i teksta, kao i za web razvoj. Ključna prednost R-a je njegova dostupnost - besplatan je i otvorenog koda, što znači da ga svako može preuzeti, koristiti, modifikovati i deliti bez ograničenja.

U okviru našeg kursa, R ćemo koristiti na tri osnovna načina:

1. za izračunavanje vrednosti,
2. za kreiranje i čuvanje **objekata**,
3. za primenu **funkcija**.

Praktično gledano, naš rad će se sastojati od zadavanja preciznih instrukcija R-u za obradu i analizu podataka. Pisanje R koda je proces kreiranja niza jasnih uputstava koje R treba da izvrši. Kada se povežu na smislen način, ove instrukcije predstavljaju logičku implementaciju statističkih metoda koje želimo da primenimo.

|  |  |
| --- | --- |
|  | **Osnovni R**  Mnogi udžbenici statistike koji koriste R oslanjaju se na **tidyverse** i **ggplot2** softverske pakete. Ovi paketi su izuzetno korisni i značajno pojednostavljuju rad sa podacima. Ipak, u ovom udžbeniku fokusiraćemo se isključivo na osnovni (eng. *base*) R, bez dodatnih paketa. Kroz primenu elementarnih funkcija za obradu i vizualizaciju podataka, savladaćemo sve statističke metode koje su predmet našeg kursa.  Postoji i dodatna prednost ovog pristupa - sintaksa i logika osnovnog R-a je pristupačnija i sličnija drugim programskim jezicima poput Python-a. Kada savladate osnovni R, prelazak na druge programske jezike biće znatno jednostavniji. |

R je napredan alat, ali njegov osnovni princip rada nije mnogo drugačiji od digitalnog kalkulatora. Ključna razlika je u tome što R može da obradi značajno veće količine podataka i izvrši kompleksnije operacije, što ga čini izuzetno efikasnim alatom za statističku analizu.

|  |
| --- |
| Slika 2.1: Prikaz Google kalkulatora |

Na slici vidimo da interfejs osnovnog Google kalkulatora sadrži:

* tastere za unos brojeva,
* tastere za izbor aritmetičkih operacija,
* tastere za izbor **matematičkih funkcija**.

Rad sa kalkulatorom prati jasan obrazac - unosimo brojeve, biramo operaciju i dobijamo rezultat. Na primer, za izračunavanje kvadratnog korena broja 25, unosimo 25, pritiskamo taster za kvadratni koren i očitavamo rezultat.

R funkcioniše po sličnom principu, ali sa znatno većim mogućnostima. U R-u prvo pripremamo podatke - bilo direktnim unosom ili učitavanjem iz eksternih izvora poput Excel tabela. Nakon toga, primenjujemo odgovarajuće funkcije za obradu tih podataka.

Evo konkretnog primera: recimo da znamo da je varijansa neke varijable 198 i želimo izračunati standardnu devijaciju. U nastavku je prikazan kod koji to radi. Tokom čitanja udžbenika često ćemo analizirati kodove koji se koriste za različite proračune i vizualizacije.

varijansa <- 198  
standardna\_devijacija <- sqrt(varijansa)  
standardna\_devijacija

Line 1

Unosimo vrednost 198. Operator <- dodeljuje ovu vrednost objektu varijansa. Sam izgled ovog operatora ilustruje proces - poput strelice koja „ubacuje“ vrednost u objekat.

Line 2

Izračunavamo standardnu devijaciju. Funkcija sqrt() računa kvadratni koren prosleđene vrednosti. Matematički, to je direktna implementacija formule koju smo ranije videli.

Line 3

Prikazujemo rezultat. R nam daje vrednost čim navedemo ime objekta ili funkcije - jednostavno i efikasno.

[1] 14.07125

Nakon izvršavanja, R prikazuje vrednost standardne devijacije od približno 14.07.

R ima dve značajne prednosti u odnosu na klasični kalkulator:

1. Transparentnost procesa: Svaki korak je precizno dokumentovan. Imena varijabli jasno označavaju **svrhu** naših postupaka, što čini ceo proces razumljivijim od običnog računanja.
2. Strukturiranost: Analiza je podeljena na jasne, logične korake. Možemo pratiti kako se svaki korak nadovezuje na prethodni, gradeći celovitu sliku analize.

Ovaj primer pokazuje tri osnovna načina korišćenja R-a. Prvo, dodeljujemo vrednost objektu varijansa. U R-u, objekti deluju kao virtuelni „kontejneri“ za podatke i vrednosti. Konkretno, objekat varijansa prima vrednost 198.

Opšta sintaksa za dodelu vrednosti objektu u R-u je:

ime\_objekta <- vrednost

Objekat u R-u može sadržati različite tipove podataka - od pojedinačnih vrednosti do nizova, drugih objekata ili rezultata funkcija. Dobar primer je linija koda standardna\_devijacija <- sqrt(varijansa). Ovde koristimo funkciju sqrt za računanje kvadratnog korena, što nas dovodi do važnog koncepta - funkcija.

Funkcije su osnovni gradivni blokovi R-a. To su predefinisane celine koda koje efikasno izvršavaju specifične operacije. Na primer, funkcija sqrt je dizajnirana sa jednim ciljem - da izračuna kvadratni koren broja. Kada radimo u R-u, funkcije pozivamo koristeći sledeću sintaksu:

ime\_funkcije(argument1, argument2, ...)

Argumenti funkcije predstavljaju ulazne podatke koje funkcija koristi za izvršavanje operacija. To su informacije koje funkcija zahteva kako bi obavila svoj zadatak. Na primer, funkcija sqrt() koristi objekat varijansa kao argument za izračunavanje kvadratnog korena. Rezultat funkcije - vrednost koja nastaje nakon obrade ulaznih podataka - nazivamo izlazom. U našem primeru, taj izlaz (rezultat funkcije sqrt()) čuvamo u objektu standardna\_devijacija.

Kada niz ovakvih naredbi sačuvamo kao tekstualni fajl, dobijamo ono što u programiranju zovemo skriptom. Skripta je, u suštini, skup preciznih instrukcija koje računar izvršava jednu za drugom. Iako se u ovom udžbeniku nećemo detaljno baviti konceptom skripti, koristićemo ih kao efikasan način za čuvanje i ponovno korišćenje koda. Ovakav pristup nam omogućava da lako ponavljamo analize ili ih prilagođavamo novim podacima.

Svaki korak analize u ovom udžbeniku praćen je R kodom koji možete pregledati. Jedini izuzetak predstavljaju složeni grafikoni, za čije iscrtavanje je potrebno više desetina linija koda. Taj kod namerno izostavljamo kako bismo održali fokus na suštini gradiva. Sve vizualizacije u ovom udžbeniku kreirane su u osnovnom R-u, koristeći isključivo standardne funkcije za iscrtavanje grafičkih elemenata.

Ovakav pristup nam omogućava da jasno vidimo vezu između teorije i prakse, bez nepotrebnih tehničkih detalja koji bi mogli da zamagle suštinu. Ideja je jednostavna - najpre razumemo koncept, zatim vidimo kako se on implementira u kodu, i na kraju posmatramo rezultat. Ovaj metodički pristup pokazao se jako efikasnim u savladavanju statističkih koncepata.

## 2.2 Kako se koristi R?

R je programski jezik koji najbolje funkcioniše u **interaktivnom režimu**. Ovo znači da možete unositi naredbe direktno u **konzolu** i odmah dobiti rezultate - kao da vodite dijalog sa računarom. Ovakav pristup čini R izuzetno praktičnim za statističku analizu i eksperimentisanje sa podacima.

Za rad sa R-om na vašem računaru potrebno je instalirati R i RStudio. Ako preferirate rad u web okruženju, posetite interaktivnu verziju ovog udžbenika na <https://sm.atomasevic.com>, gde možete koristiti R direktno u okviru dokumenta.

Najbolje ćemo pokazati kako R funkcioniše tako što ćemo zajedno da osvežimo znanje o osnovnim konceptima deskriptivne statistike.

## 2.3 Deskriptivna statistika

U našem istraživanju učestvovalo je ispitanika. Od svakog smo prikupili podatak o njihovim mesečnim primanjima. Varijabla predstavlja te odgovore, tako da vrednosti čine naš skup podataka.

Prva mera koju ćemo izračunati je **aritmetička sredina uzorka** - jednostavno rečeno, prosek primanja svih ispitanika. Formula za njeno izračunavanje izgleda ovako:

Da vidimo kako ovu formulu možemo implementirati u R-u. Naši podaci su smešteni u **CSV fajl** koji možemo preuzeti sa interneta.

|  |  |
| --- | --- |
|  | **CSV**  CSV format je jedan od najčešće korišćenih formata za skladištenje podataka. Naziv CSV potiče od engleskog izraza *Comma Separated Values*, što označava vrednosti razdvojene zarezima. Zamislite jednostavnu tabelu gde su podaci u kolonama razdvojeni zarezima umesto vertikalnih linija. Evo primera kako bi izgledali osnovni podaci o starosti i polu ispitanika u CSV formatu:  starost,pol 25,m 30,z 22,m  Format CSV je elegantan u svojoj jednostavnosti - pristupačan je i za računare i za ljude. U suštini, to je običan tekstualni fajl koji organizuje podatke u redove, gde su vrednosti razdvojene zarezima. Uzmimo kao primer fajl sa dve varijable - starost i pol. Svaki red predstavlja jednog ispitanika, a njihovi podaci su razdvojeni zarezima. U našem primeru imamo tri ispitanika: prvi ima 25 godina i muškog je pola, drugi 30 godina i ženskog pola, a treći 22 godine i muškog pola. Ova struktura nam omogućava intuitivno razumevanje podataka, što je ključno za efikasnu analizu.  CSV fajl koji ćemo koristiti u narednom primeru izgleda ovako:  "primanja" 340 1069 522 ...  Ovaj fajl je jednostavan, ali efikasan. Sadrži jednu varijablu - primanja. Svaki red nakon zaglavlja predstavlja podatak o primanjima jednog ispitanika. Struktura je precizna i jasna: sa 350 ispitanika, fajl ima tačno 351 red (prvi red sadrži naziv varijable). Ova elegantna organizacija podataka nam omogućava direktnu i preciznu analizu primanja naših ispitanika. |

Prvo ćemo učitati podatke iz CSV fajla u R.

fajl <- "https://gist.githubusercontent.com/atomashevic/492eaf9250b68a35557f224b20e8b310/raw/94329ddd24964e6a9fe0d392a6482d556aa2b54b/primanja.csv"  
podaci <- read.csv(fajl)  
head(podaci)

Line 1

Učitavanje podataka. Na početku definišemo putanju do fajla - ovo je URL adresa koja vodi do našeg fajla sa podacima. Ovaj pristup nam omogućava jednostavan pristup podacima bez lokalne kopije.

Line 2

Podatke učitavamo u objekat podaci. Funkcija read.csv je elegantno rešenje za učitavanje podataka iz CSV fajla. Ona efikasno obrađuje podatke koristeći putanju definisanu u prethodnom koraku.

Line 3

Primenjujemo funkciju head koja prikazuje prvih nekoliko redova učitanih podataka. Ova funkcija je izuzetno praktična kada radimo sa velikim skupovima podataka - umesto pregledanja cele tabele, dobijamo brz uvid u strukturu i sadržaj podataka. To nam omogućava da odmah procenimo da li su podaci ispravno učitani i spremni za analizu.

primanja  
1 340  
2 1069  
3 522  
4 356  
5 827  
6 340

|  |  |
| --- | --- |
|  | **Skupovi ili tabele podataka**  U našem radu često ćemo koristiti izraze *skup podataka* i *tabela podataka*. Iako deluju slično, između njih postoji suptilna razlika. U R-u podaci se čuvaju u objektima koje nazivamo *data frame*, što na srpski prevodimo kao *tabela podataka*. S druge strane, *skup podataka* (eng. *data set*) odnosi se na sam fajl sa podacima koji može biti sačuvan na različitim lokacijama - serveru, repozitorijumu, hard disku ili drugde.  Oba pojma dele istu osnovnu strukturu i organizaciju podataka. Zamislite ih kao tablicu sa dve dimenzije: **opservacije** i **varijable**. Opservacije predstavljaju redove, a varijable kolone u toj tablici.  Kada govorimo o jednoj opservaciji, mislimo na sve informacije koje smo prikupili o jednom konkretnom „subjektu“ našeg istraživanja. To može biti osoba (ispitanik), ali i nešto sasvim drugo - firma, region, pa čak i država. Jedinica opservacije nam govori šta tačno svaka opservacija predstavlja u našem skupu podataka.  Opservacije označavamo rednim brojevima - prva opservacija nalazi se u prvom redu tabele podataka, druga u drugom, i tako redom. Ovakva struktura nam omogućava precizno praćenje i analizu podataka za svakog pojedinačnog subjekta našeg istraživanja.  podaci[1, ]  Line 1  Prikaz prve opservacije iz tabele podataka.  [1] 340  Kolone skupa podataka nazivamo varijablama. Svaka varijabla predstavlja jednu karakteristiku koja se menja između opservacija. Uzmimo za primer varijablu „pol“ - ona beleži ovu karakteristiku za svakog ispitanika u našem skupu. Termin „promenljiva“ nije slučajan - koristi se upravo zato što vrednosti variraju od ispitanika do ispitanika. Razmislite na trenutak: kada bi svi ispitanici bili istog pola, ta varijabla ne bi nosila nikakvu korisnu informaciju.  Ova struktura podataka je temelj našeg rada u R-u, i važno je da je dobro razumemo pre nego što krenemo dalje. |

Kada R učita CSV fajl, kreira tabelu podataka (eng. *data frame*). Naša tabela sadrži jednu kolonu - primanja, koja beleži mesečna primanja svakog ispitanika. Svaki red predstavlja podatak o jednom ispitaniku.

Sa podacima spremnim za analizu, možemo izračunati aritmetičku sredinu uzorka. Pogledajmo matematičku formulaciju:

Za direktnu primenu ove formule u R-u, potrebna su nam dva osnovna elementa:

1. Zbir svih vrednosti u uzorku: sum(podaci$primanja)
2. Broj ispitanika: nrow(podaci)

Ova dva elementa čine srž našeg izračunavanja. Pogledajmo kako ih implementiramo u R-u.

AS <- sum(podaci$primanja) / nrow(podaci)  
round(AS,2)

Line 1

Proračun aritmetičke sredine. Funkcija sum izračunava zbir svih vrednosti u vektoru, dok nrow određuje broj redova u tabeli.

Line 2

Zaokruživanje rezultata na dve decimale korišćenjem funkcije round.

[1] 799.8

Izvršavanjem ovog koda dobijamo aritmetičku sredinu uzorka od približno 800. Vredan je pomena i elegantniji pristup - funkcija mean(). Isprobajte je i uporedite rezultate. Iako je mean() računski efikasnija, naš prvobitni pristup nam omogućava direktan uvid u logiku izračunavanja.

Nastavimo sa dubljom analizom. Pored srednje vrednosti, ključno je razumeti i **raspršenost** podataka u uzorku. Zamislite dva uzorka koja dele istu aritmetičku sredinu, ali imaju različitu raspršenost - takva situacija nam otkriva suštinske karakteristike podataka koje proučavamo.

Ilustrujmo ovo na konkretnom primeru. Pogledajmo ispitanika broj 42, čija su primanja . Možemo precizno izmeriti koliko njegova primanja odstupaju od proseka uzorka. Ovaj proračun nam otkriva individualnu varijaciju unutar našeg skupa podataka.

ispitanik <- podaci$primanja[42]  
udaljenost <- ispitanik - AS  
udaljenost

Line 1

Izdvajamo primanja ispitanika pod rednim brojem 42.

Line 2

Računamo razliku između njegovih primanja i proseka uzorka.

Line 3

Prikazujemo rezultat.

[1] -459.8

Analiza pokazuje da su primanja ovog ispitanika za približno 460 evra niža od proseka uzorka. Razmotrimo ovo u širem kontekstu. Uzmimo za primer skandinavsku zemlju gde prosečna primanja iznose 5460 evra. Tamo bi ispitanik sa primanjima od 5000 evra takođe bio 460 evra ispod proseka. Međutim, isto odstupanje od proseka ima sasvim različito značenje u društvu gde je prosek 5460 evra, u odnosu na društvo gde je prosek 800 evra.

Ova situacija nam jasno pokazuje zašto ne možemo posmatrati odstupanja od proseka izolovano. Potrebno je da odredimo prosečno odstupanje od proseka, koje će nam poslužiti kao referentna tačka za analizu svih pojedinačnih odstupanja, uključujući i našeg ispitanika broj 42.

Usled jedinstvenih karakteristika aritmetičke sredine, najprecizniji način merenja prosečnog odstupanja je kroz kvadriranje individualnih odstupanja. Ovim pristupom dolazimo do varijanse, čiji kvadratni koren predstavlja standardnu devijaciju.

Varijansu uzorka računamo prema sledećoj formuli:

|  |  |
| --- | --- |
|  | **Zašto je varijansa ?**  Tokom kursa primetićete važnu konvenciju - vrednosti koje izračunavamo za **uzorak**, poput aritmetičke sredine, varijanse, standardne devijacije i relativne frekvencije, označavamo **latiničnim slovima** (). Ove statistike uzorka nisu samo brojevi - one nam služe kao prozor u nepoznate parametre cele populacije. R nam omogućava da izračunamo ove vrednosti i iskoristimo ih za razumevanje šire slike.  Kada govorimo o **populaciji**, koristimo **mala grčka slova** () za označavanje parametara. Ovi parametri predstavljaju stvarne vrednosti karakteristika populacije koje ne možemo direktno izmeriti. Ipak, možemo ih proceniti koristeći podatke iz našeg uzorka. Upravo ovaj proces - procena parametara populacije na osnovu uzorka - čini srž statističkog zaključivanja. To je kao da sastavljamo slagalicu cele slike koristeći samo jedan njen deo. |

Podelićemo proces izračunavanja varijanse na četiri jasna koraka:

1. Prvo računamo individualna odstupanja od proseka: R podaci$primanja - AS. Ovo nam pokazuje koliko svaka vrednost odstupa od središta distribucije.
2. Zatim kvadriramo ta odstupanja: R (podaci$primanja - AS)^2. Ovaj korak je elegantan - eliminišemo negativne vrednosti i naglašavamo veća odstupanja.
3. Potom sabiramo sve kvadrirane vrednosti: R sum((podaci$primanja - AS)^2). Ova suma nam daje ukupnu meru raspršenosti.
4. Na kraju, delimo zbir sa n-1: R sum((podaci$primanja - AS)^2) / (nrow(podaci) - 1). Ovim dobijamo prosečno kvadrirano odstupanje.

Svaki korak vodi ka dubljem razumevanju strukture naših podataka. Ova postupnost nije slučajna - ona nam omogućava da vidimo kako se matematički koncepti pretvaraju u konkretne mere varijabilnosti.

|  |  |
| --- | --- |
|  | **Vektori u R-u**  Pogledajmo detaljnije komandu podaci$primanja - AS. Šta tačno tražimo od R-a? Želimo da **za svakog ispitanika** izračunamo odstupanje njegovih primanja od proseka uzorka. Objekat AS je aritmetička sredina uzorka - jedan broj. Nasuprot tome, podaci$primanja je **vektor**, struktura koja sadrži niz brojeva (u našem slučaju, primanja svih ispitanika).  Kada od vektora oduzmemo skalar (pojedinačnu vrednost), R primenjuje tu operaciju na svaki element vektora. Rezultat je nov vektor koji sadrži individualna odstupanja svakog elementa od oduzete vrednosti. Ova operacija nam efikasno pokazuje kako su primanja raspoređena u odnosu na prosek. |

Hajde da ove korake pretočimo u R kod. Pogledajte kako to izgleda:

varijansa <- sum((podaci$primanja - AS)^2) /  
 (nrow(podaci) - 1)  
round(varijansa,2)

Line 2

Ovde izračunavamo varijansu. Funkcija sum sabira sve vrednosti u vektoru, dok nrow određuje broj redova u tabeli.

Line 3

Rezultat zaokružujemo na dve decimale koristeći funkciju round. Ovo čini numeričke vrednosti preglednijim.

[1] 158021.6

R nam nudi elegantniji pristup izračunavanju varijanse. Funkcija var() nam omogućava da ovaj proračun izvršimo u jednom koraku. Pogledajmo kako:

varijansa <- var(podaci$primanja)  
round(varijansa,2)

Line 1

Funkcija var() elegantno računa varijansu - ona obrađuje vektor i vraća nam tačnu vrednost.

Line 2

Funkcija round zaokružuje rezultat na dve decimale, čineći broj preglednijim za analizu.

[1] 158021.6

Šta nam govori dobijeni rezultat? Varijansa, iako matematički precizna, nije intuitivna za tumačenje. Razlog je jednostavan - nije izražena u istim jedinicama kao naši izvorni podaci. Srećom, postoji elegantan način da ovo prevaziđemo. Transformisaćemo je u meru koja je prirodnija za razumevanje - **standardnu devijaciju**.

Postupak je jednostavan: izračunavamo kvadratni koren varijanse. Ovaj naizgled jednostavan matematički korak nam otvara put ka jasnijem razumevanju raspršenosti podataka. Standardna devijacija, za razliku od varijanse, izražena je u istim jedinicama kao i originalni podaci, što nam omogućava direktniju interpretaciju rezultata.

standardna\_devijacija <- sqrt(varijansa)  
round(standardna\_devijacija,2)

Line 1

Izračunavanje standardne devijacije koristeći funkciju sqrt koja računa kvadratni koren argumenta.

Line 2

Zaokruživanje rezultata na dve decimale.

[1] 397.52

Dobijeni rezultat od približno 400 evra predstavlja standardnu devijaciju uzorka. Šta nam ova vrednost govori? U osnovi, ona ukazuje da primanja tipičnog ispitanika variraju oko 400 evra u odnosu na prosečnu vrednost. Ova mera raspršenosti daje nam precizniju sliku o tome kako su primanja distribuirana oko prosečne vrednosti od 800 evra.

Standardna devijacija nam omogućava sledeći logičan korak - **standardizaciju** odstupanja od proseka. Kroz ovaj postupak pretvaramo pojedinačna odstupanja iz originalnih jedinica mere u univerzalne, uporedive vrednosti. Tako dobijamo mogućnost da tumačimo odstupanja nezavisno od proseka uzorka ili korišćenih mernih jedinica.

Razmotrimo konkretan primer: ispitanik 42 ima primanja koja su 460 evra ispod proseka. Umesto da ovu razliku posmatramo u apsolutnim vrednostima, možemo je standardizovati i izraziti kroz broj standardnih devijacija. Rezultat ovog postupka nazivamo **Z-skor**.

U R-u to izgleda ovako:

Z <- (ispitanik - AS) / standardna\_devijacija  
round(Z,2)

Line 1

Izračunavanje Z-skora prati preciznu formulu.

[1] -1.16

Dobijeni rezultat od -1,16 pokazuje da se ispitanik nalazi 1,16 standardnih devijacija ispod proseka. Ovaj broj ima značenje samo u okviru šire **statističke distribucije**. Detaljniju interpretaciju Z-skorova i elegantno **pravilo tri sigme** obradićemo u poglavlju o [normalnoj distribuciji](normalna.qmd).

Prirodan sledeći korak je **standardizacija** celokupnog uzorka. Kroz ovu transformaciju stvaramo novu varijablu koja ima aritmetičku sredinu 0 i standardnu devijaciju 1.

Da bismo potvrdili ovu transformaciju, kreiraćemo varijablu Zi koja sadrži standardizovane vrednosti primanja. Ova vektorska operacija transformiše svaku vrednost varijable primanja u standardizovanu vrednost. Proces je jednostavan ali moćan - od svake vrednosti oduzimamo prosečnu vrednost i delimo rezultat standardnom devijacijom. Time dobijamo novi vektor Zi sa standardizovanim vrednostima koje nam omogućavaju direktno poređenje različitih opservacija.

Zi <- (podaci$primanja - AS) / standardna\_devijacija  
head(Zi)

Line 1

Kreiramo vektor standardizovanih vrednosti primanja.

Line 2

Prikazujemo nekoliko prvih vrednosti radi ilustracije.

[1] -1.15667340 0.67719983 -0.69883400 -1.11642378 0.06842435 -1.15667340

Prirodan sledeći korak je provera ispravnosti naše standardizacije. Teorijski očekujemo da standardizovane vrednosti imaju aritmetičku sredinu 0 i standardnu devijaciju 1. Proverimo da li je to zaista tako. Za jasan i pregledan prikaz rezultata koristićemo funkciju cat.

Z\_AS <- mean(Zi)  
  
Z\_SD <- sd(Zi)  
  
cat("Aritmetička sredina standardizovanih vrednosti:", round(Z\_AS,2), "\n")  
  
cat("Standardna devijacija standardizovanih vrednosti:", round(Z\_SD,2))

Line 1

Izračunavamo aritmetičku sredinu standardizovanih vrednosti.

Line 3

Računamo standardnu devijaciju standardizovanih vrednosti.

Line 5

Prikazujemo aritmetičku sredinu zaokruženu na dve decimale.

Line 7

Ispisujemo standardnu devijaciju zaokruženu na dve decimale.

Aritmetička sredina standardizovanih vrednosti: 0   
Standardna devijacija standardizovanih vrednosti: 1

Z-skorovi nam daju moćan alat za interpretaciju vrednosti, nezavisno od jedinice mere ili apsolutne veličine. Njihova ključna prednost je mogućnost direktnog poređenja vrednosti iz različitih uzoraka. Bilo da analiziramo visinu ispitanika ili primanja u različitim valutama, Z-skorovi nam omogućavaju precizno određivanje relativne pozicije svake vrednosti. Vrednost Z-skora blizu 0 ukazuje na malu udaljenost od proseka, dok pozitivan Z-skor jasno označava vrednost iznad proseka uzorka.

Ovim završavamo uvod u R i osnovne koncepte deskriptivne statistike. Do sada smo se fokusirali na kvantitativne varijable, ali u narednim poglavljima detaljno obrađujemo i kvalitativne varijable, kao i njihovu praktičnu primenu u R-u. Pre nego što nastavite dalje, preporučujem da ponovite ključne koncepte deskriptivne statistike sažete u 10 tačaka iznad. Za efikasno utvrđivanje znanja, najbolje je da rešite zadatke na kraju poglavlja - oni će vam pomoći da izgradite čvrste temelje za dalji rad u R-u.

|  |  |
| --- | --- |
|  | **Tipovi varijabli**  U prethodnom poglavlju upoznali ste se sa konceptom mernih skala (nominalna, ordinalna, intervalna i racio). Merna skala je koncept koji se prvenstveno odnosi na dizajn istraživanja i psihometrijske karakteristike mernog instrumenta. Za naše potrebe, napravićemo jednostavnu ali efikasnu podelu na **numeričke i tekstualne** varijable.  Ova podela je istovremeno intuitivna i praktična - tekstualne varijable sadrže vrednosti zapisane kao niz alfanumeričkih znakova (slova „A“-„Ž“, cifre „0“-„9“, kao i specijalne znakove poput „!“, „#“, „?“), dok numeričke varijable sadrže vrednosti koje R direktno prepoznaje i obrađuje kao brojeve.  Logično pitanje koje se nameće je: „Zar nisu i brojevi alfanumerički znakovi?“ Suština je u načinu na koji R interpretira te vrednosti. Na primer, 999 R tumači kao broj s kojim može da vrši matematičke operacije, dok "999" tretira kao običan tekst. Sve što se nalazi između navodnika (") R interpretira kao tekst, bez obzira na sadržaj.  U okviru obe kategorije varijabli postoji posebna i izuzetno korisna podgrupa - **binarne varijable**. One su specifične jer imaju tačno dve moguće vrednosti.  Kod tekstualnih varijabli, klasičan primer je varijabla pol sa dve opcije: "Muški" ili "Ženski". Savremena istraživanja često zahtevaju širi spektar opcija, ali to prevazilazi okvire našeg trenutnog razmatranja.  Kada govorimo o numeričkim varijablama, posebnu pažnju zaslužuju binarne ili „dummy“ varijable. Ove indikatorske varijable tipično uzimaju vrednosti 0 ili 1. Uzmimo za primer varijablu zaposlen: vrednost 1 označava zaposlenu osobu, dok 0 označava nezaposlenu. Ovakva varijabla elegantno ukazuje na prisustvo ili odsustvo određene karakteristike kod svakog ispitanika u studiji. |

## 2.4 Pregled deskriptivne statistike

Suština deskriptivne statistike, koja je ključna za razumevanje ostatka knjige, može se svesti na 10 osnovnih tačaka:

1. Analiza pojedinačne varijable obuhvata dve ključne komponente: centralnu tendenciju i varijabilitet - odnosno kako se vrednosti grupišu oko centralne mere.
2. Tri fundamentalne centralne mere koje koristimo su modus, medijana i aritmetička sredina.
3. Modus predstavlja vrednost koja se najčešće pojavljuje u skupu podataka.
4. Medijana je vrednost koja deli uređeni niz podataka na dva jednaka dela - elegantno rešenje za pronalaženje sredine skupa.
5. Aritmetička sredina je računski izvedena mera centralne tendencije i predstavlja prvi centralni moment varijable - preciznije, prosek svih vrednosti.
6. Za merenje varijabiliteta koristimo dva moćna alata: varijansu i standardnu devijaciju.
7. Varijansa, kao drugi centralni moment varijable, predstavlja prosečno kvadrirano odstupanje od aritmetičke sredine.
8. Standardna devijacija, kao kvadratni koren varijanse, daje nam intuitivniju meru raspršenosti u originalnim jedinicama mere.
9. Standardizacija transformiše originalnu varijablu X u Z-skor kroz formulu . Na primer, u situaciji gde je prosek ispita 10 poena sa standardnom devijacijom 3, student koji je osvojio 16 poena ima Z-skor 2, što znači da je njegov rezultat dve standardne devijacije iznad proseka.
10. Standardizovana varijabla ima dve karakteristične osobine: aritmetičku sredinu 0 i standardnu devijaciju 1, što matematički zapisujemo kao , .

Ovi koncepti predstavljaju temelj za razumevanje složenijih statističkih analiza koje slede.

## 2.5 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1**  Uporedite ispitanike 42 i 100 iz našeg skupa podataka. Analizirajte kako se njihova primanja odnose prema proseku uzorka. Fokusirajte se na dve ključne dimenzije - intenzitet i smer odstupanja od prosečne vrednosti. Na ovaj način ćete direktno primeniti koncepte deskriptivne statistike koje smo obradili. |

|  |  |
| --- | --- |
|  | **Zadatak 2**  Razmotrimo ponovo primer skandinavske zemlje. U toj zemlji prosečna plata iznosi 5460 evra, sa standardnom devijacijom od 870 evra. Naš zadatak je da odredimo poziciju osobe čija primanja iznose 5000 evra u odnosu na prosek - preciznije, koliko standardnih devijacija je ta osoba udaljena od proseka? Nakon toga, uporedite relativnu poziciju ovog hipotetičkog ispitanika sa pozicijama ispitanika 42 i 100 iz našeg uzorka. Obratite pažnju na način na koji se svako od njih pozicionira u odnosu na prosek svog uzorka i razmislite šta nam to govori o distribuciji primanja u ove dve različite populacije. |

|  |  |
| --- | --- |
|  | **Zadatak 3 \***  Koristeći isključivo funkcije koje smo obradili u ovom poglavlju, dokažite da je rezultat izraza podaci$primanja - AS zaista vektor. Za rešenje razmotrite osnovne karakteristike vektora i načine na koje ih možemo proveriti programski. |

|  |  |
| --- | --- |
|  | **Zadatak 4 \*\***  Pronađite u našem uzorku ispitanika čija je pozicija u odnosu na prosek najbliža poziciji hipotetičkog ispitanika iz skandinavske zemlje (iz Zadatka 2).  Za pretragu vektora podaci$primanja i izdvajanje rednih brojeva opservacija koje se nalaze između vrednosti i , možete koristiti sledeću funkciju:  which(podaci$primanja > a & podaci$primanja < b)  Iskoristite ovu funkciju da rešite zadatak. |

# 3. Uzorci i statističko zaključivanje

## 3.1 Populacija i uzorak

Započnimo naše putovanje kroz statistiku upoznavanjem dva osnovna pojma: **populacije** i **uzorka**. Pre svega, važno je razjasniti čestu zabludu - populacija u statistici nema veze s brojem stanovnika neke zemlje. Ovi pojmovi su direktno povezani sa specifičnim istraživačkim problemom koji proučavamo.

Populacija predstavlja skup svih entiteta koji su predmet našeg istraživanja. To mogu biti ljudi, ali i mnogo toga drugog. Evo nekoliko primera:

1. Pri istraživanju stavova srednjoškolaca u Srbiji, populacija obuhvata **sve** srednjoškolce u zemlji.
2. Kod proučavanja kriminalnih aktivnosti među penzionerima u Vojvodini, populacija je (nadamo se) manja grupa - svi penzioneri u Vojvodini koji se bave kriminalnim radnjama.
3. Populacija se ne mora sastojati samo od ljudi. To mogu biti firme, institucije, Instagram profili, sajtovi, političke partije, pa čak i čitave države.

Suštinski je bitno shvatiti da populacija predstavlja celokupan skup entiteta na koje se odnosi naše istraživačko pitanje. Bez obzira na veličinu ili prirodu, populacija je uvek određena našim istraživačkim fokusom.

Drugim rečima, populacija je skup svih entiteta koji su u središtu našeg istraživačkog poduhvata. Zanimljivo je da je populacija gotovo uvek **nedostupna** za potpuno istraživanje. Bez obzira na to koliko je populacija velika ili mala, praktično je nemoguće istraživanjem obuhvatiti baš svakog njen element.

I upravo tu dolazimo do suštine **statističkog zaključivanja**: imamo u vidu određenu populaciju, ali pošto je ne možemo u potpunosti istražiti, usmeravamo se na **uzorak** - jedan mali deo populacije koji možemo detaljno proučiti. Zamislite sledeću situaciju: pokušavate da procenite ukus velikog lonca gulaša probajući samo jednu kašiku. Pri tome je veličina te „kašike“ često ograničena praktičnim uslovima, poput vremena i raspoloživog budžeta.

Kako dobijamo podatke u uzorku? To je oblast metodologije prikupljanja podataka i time se nećemo detaljno baviti. Za nas je ključan koncept **prostog slučajnog uzorka** - uzorka koji daje svakom članu populacije jednaku verovatnoću da bude izabran. Proces uzorkovanja funkcioniše kao lutrija, gde mi kao istraživači nemamo uticaj na to ko će biti „izvučen“. Iz perspektive statistike, uzorak posmatramo kao loto bubanj iz kojeg nasumično izvlačimo kuglice (jedinice uzorka).

|  |  |
| --- | --- |
|  | **„Reprezentativan“ uzorak**  U društvenim naukama često se koristi termin „reprezentativan“ uzorak. Značenje ovog termina je, međutim, nejasno i varira među istraživačima. Metodološki udžbenici sociologije nude drugačije objašnjenje ovog termina u poređenju sa udžbenicima psihologije ili statistike. U ovom udžbeniku taj pojam nećemo koristiti.  Uobičajeno shvatanje reprezentativnosti podrazumeva da distribucija različitih obeležja (pol, klasa, obrazovanje, starost, zarade) u uzorku mora odgovarati distribuciji tih obeležja u populaciji. Na primer, ako u Srbiji ima 21% visokoobrazovanih, 45% srednje obrazovanih i 34% osoba sa osnovnim obrazovanjem, očekivali bismo slične proporcije u našem uzorku. Mnogi metodološki priručnici tvrde da je ovako definisana reprezentativnost neophodan preduslov za dalje analize.  Međutim, uzorkovanje nije tema ovog udžbenika i nećemo se zadržavati na konceptima čije je značenje sporno. Umesto toga, usredsredićemo se na podatke dobijene prostim slučajnim uzorkom. U narednim poglavljima ćemo videti kako takve podatke možemo koristiti za odgovaranje na naša istraživačka pitanja.  Neki autori tvrde da je sam koncept reprezentativnosti logički paradoksalan i samim tim neupotrebljiv (Stuart, 1987). Razmislimo: da bismo utvrdili da li uzorak verno predstavlja populaciju, morali bismo imati detaljno znanje o samoj populaciji. Ali kada bismo zaista posedovali takvo znanje, uzorak nam ne bi ni bio potreban. Upravo u ovoj cirkularnosti leži suštinski problem koncepta reprezentativnosti.  Važnije je doći do ispravnog zaključka tako da distribucija ishoda u uzorku odgovara distribuciji ishoda u populaciji. Na primer, kada na osnovu podataka iz uzorka utvrdimo da muškarci u Srbiji imaju viša primanja u odnosu na žene, želimo biti sigurni da te razlike postoje i u populaciji. Ovo zapravo nije povezano s tim da li vaš uzorak sadrži 39%, 49% ili 59% muškaraca. |

Po završetku istraživanja na uzorku, prikupljeni podaci se organizuju u skupove ili matrice podataka. To možete zamisliti kao veliku tabelu gde svaki red predstavlja jednu jedinicu uzorka (obično jednog ispitanika), dok svaka kolona predstavlja jednu varijablu. Varijable mogu sadržati različite vrste podataka - brojeve, kategorije ili tekst - koji pokazuju rezultate merenja na tom uzorku. Na primer, to mogu biti odgovori ispitanika iz ankete.

Kada imamo ovakav skup podataka, možemo ga analizirati koristeći alate deskriptivne statistike. Ovi alati nam omogućavaju da:

1. Opišemo centralne tendencije varijabli (oko kojih vrednosti se grupišu podaci),
2. Istražimo varijabilitet (koliko podaci odstupaju od centralnih vrednosti),
3. Predstavimo podatke kroz grafikone,
4. Standardizujemo varijable (što olakšava poređenja).

Ovi postupci nam pomažu da razumemo suštinske obrasce u našim podacima i daju nam prvi uvid u karakteristike uzorka koji proučavamo.

Šta se desi u uzorku, ostaje u uzorku.

Sve što otkrijemo na nivou uzorka ostaje vezano za taj uzorak. Bez obzira na njegov kvalitet, uzorak je samo mali prozor u populaciju - ništa više od toga. Ne možemo jednostavno preslikati opis uzorka na celu populaciju. Centralni problem statističkog zaključivanja jeste upravo ovo „premošćavanje“ između uzorka i populacije. Moramo ovo držati na umu: uzorak i populacija nisu isto. Uzorak predstavlja samo fragment celine i daje nam nepotpune informacije o populaciji. Ali nema razloga za brigu - kao što ćete uskoro videti, ljudi su prilično dobri u donošenju odluka na osnovu nepotpunih informacija, a statistika nam pomaže da to radimo na sistematičan i objektivan način.

|  |  |
| --- | --- |
|  | **Grčke babe i latinične žabe**  U statistici koristimo dve vrste matematičke notacije:   1. **Latinična slova** označavaju vrednosti kao što su aritmetička sredina (), varijansa (), standardna devijacija () ili relativna frekvencija (). Ove vrednosti nazivamo **statistike uzorka** jer ih možemo direktno izračunati iz podataka kojima raspolažemo. 2. **Mala grčka slova** () predstavljaju parametre populacije. Ovi parametri su nam nedostupni - znamo da postoje, ali ih nikada ne možemo precizno izračunati. Možemo ih samo približno **oceniti**, što predstavlja jedan od temeljnih ciljeva statističke analize.   Razlika između ove dve kategorije mora biti jasna. Moramo razlikovati ono što znamo (uzorak) od onoga što želimo saznati (populacija).  Na primer, kada napišemo , to nam govori da je aritmetička sredina uzorka 10. Ovo je konkretna statistika koju smo izračunali iz podataka kojima raspolažemo.  Kada napišemo , postavljamo pretpostavku o aritmetičkoj sredini populacije. Iako ne možemo biti potpuno sigurni u tačnost ove vrednosti, ovakve pretpostavke služe kao polazne tačke za dalju analizu. One nam daju okvir za razmišljanje o populaciji, uprkos tome što nemamo potpune informacije o njoj.  Ova dva načina zapisivanja ( i ) imaju suštinski različita značenja. Pri interpretaciji statističkih izraza, precizna notacija je ključna za razumevanje šta tačno merimo ili pretpostavljamo. |

## 3.2 Mali i veliki svetovi

Priča o Kolumbovom „otkriću“ Amerike i planu da stigne do Indije ploveći preko Atlantika savršeno ilustruje odnos između modela i stvarnosti. Kao i većina obrazovanih ljudi svog vremena, Kolumbo je bio svestan sferne prirode Zemlje. Međutim, njegova procena veličine planete značajno je odstupala od stvarnosti.

Razmotrite sledeće: Kolumbo je odbacio Aristotelov proračun prečnika Zemlje iz 2. veka pre nove ere (koji je, uzgred, bio izuzetno precizan). Umesto toga, priklonio se novijim, ali netačnim procenama koje su sugerisale da je Zemlja približno 3/4 stvarne veličine.

Ova greška u modelu dovela je do značajnog odstupanja u procenama. Prema Kolumbovim kalkulacijama, razdaljina između Kanarskih ostrva i Japana iznosila je oko 2400 nautičkih milja. Stvarna razdaljina je četiri puta veća. Ova razlika između modela i stvarnosti direktno je uticala na planiranje njegovog putovanja.

Svoj **mali svet** - skup pretpostavki o planeti Zemlji - Kolumbo je izgradio na temelju ograničenih podataka kojima je raspolagao. Na osnovu tih pretpostavki, kreirao je mape i planirao putovanje. U tom istorijskom trenutku, niko nije mogao imati preciznu predstavu o stvarnoj veličini i geometriji naše planete. Prema Kolumbovom modelu sveta, plovidba zapadno preko Atlantika trebalo je da relativno brzo dovede do obala Japana, a zatim i do Indije. Nakon duže plovidbe nego što je očekivao, kad je konačno ugledao kopno, Kolumbo je bio ubeđen da je ostrvo pred njim deo arhipelaga Istočne Indije. Drugim rečima, čvrsto je verovao u ispravnost svog modela i smatrao je da je, prateći svoje proračune, stigao na željeno odredište. Međutim, **veliki svet** - objektivna stvarnost - pokazao se fundamentalno drugačijim od njegovih pretpostavki.

**Mali svet** predstavlja ono čime mi, kao istraživači, raspolažemo - naše ideje, hipoteze i nepotpune informacije. Baš kao što je Kolumbo imao svoje pretpostavke, naš zadatak je da proverimo validnost tih ideja i hipoteza suočavajući ih sa stvarnošću (**velikim svetom**). Ovaj proces nije ni jednostavan ni očigledan. Uzmimo za primer Kolumba, koji veoma dugo nije prihvatio činjenicu da je otkrio novi kontinent umesto zapadnog puta do Indije. Njegova priča nam pokazuje koliko snažno možemo biti vezani za svoje pretpostavke, čak i kad se suočimo sa dokazima koji ih osporavaju.

Naše putovanje počinje formulisanjem **malih svetova**, odnosno **statističkih modela**. Ovi modeli, poput Kolumbovih mapa, predstavljaju pojednostavljenu sliku stvarnosti. Daju nam precizan opis malog sveta - društvenih fenomena koji su u središtu našeg istraživačkog problema. Razmotrimo primer istraživanja ekonomskih posledica rodnih nejednakosti u Srbiji. Možemo postaviti dva jednostavna modela koji opisuju dve različite verzije našeg društva.

**Model 1:** Zarada zaposlenih u Srbiji zavisi od njihovog pola/roda. Muškarci u Srbiji u proseku imaju više primanja od žena.

**Model 2:** Zarada zaposlenih u Srbiji ne zavisi od njihovog pola/roda. Muškarci i žene u Srbiji u imaju jednaka prosečna primanja.

Model se sastoji od pretpostavki koje predstavljaju uprošćavanje stvarnosti, odosno opis malog sveta. Uprošćavanje u ovom primeru podrazumeva i da ne razmatramo druge *faktore* za koje smo prilično sigurni da mogu uticati na zaradu: radno iskustvo, obrazovanje, itd. Ova dva modela nam, međutim, pričaju veoma različite priče. Prvi oslikava Srbiju kao društvo u kojem postoje značajne rodne razlike u ličnim primanjima, dok drugi predstavlja sliku egalitarnog društva bez rodnih nejednakosti u sferi zarada. Svaki od ovih modela nosi sa sobom različite implikacije za razumevanje društvene stvarnosti u Srbiji.

Jasno je da oba modela ne mogu istovremeno biti tačna. Centralni cilj ovog udžbenika je da savladate metode koje će vam pomoći da odredite koji od predloženih modela (bilo da ih je dva ili više) najtačnije opisuje stvarnost, odnosno koji najbolje aproksimira „veliki svet“. Ove metode će vam omogućiti da na sistematičan i objektivan način procenite koji model najvernije odgovara podacima i realnosti koju nastojimo da razumemo.

### 3.2.1 Statistički modeli

Sve što smo rekli o modelima primenjuje se na matematičke, biološke, medicinske i druge naučne modele koji se zasnivaju na povezanom skupu pretpostavki o istraživanom problemu. **Statistički modeli** se, međutim, izdvajaju posebnim načinom na koji predstavljaju i zapisuju ideje i pretpostavke. Neke elemente deskriptivne statistike već smo videli u opisu dva modela, na primer kroz koncept **jednakih prosečnih primanja**.

Šta tačno znači tvrdnja „muškarci i žene u Srbiji imaju jednaka prosečna primanja“? Pre svega, govorimo o dve populacije - populaciji muškaraca i populaciji žena u Srbiji. Kada bismo imali podatke o primanjima svakog muškarca u Srbiji, mogli bismo izračunati njihovu aritmetičku sredinu. Pošto je to praktično nemoguće, ta aritmetička sredina (prosečna primanja muškaraca) postaje **parametar** koji označavamo sa . Analogno tome, za populaciju žena imamo parametar . Model 2, koji pretpostavlja jednaka prosečna primanja, matematički zapisujemo kao .

Da bismo proverili istinitost ove pretpostavke, potrebni su nam podaci - konkretna merenja i zabeležene vrednosti na osnovu kojih ćemo utvrditi da li je pretpostavka tačna ili ne. Međutim, kao što smo već naglasili, ti podaci su nam dostupni samo na nivou **uzorka**. Iz njih možemo izračunati i .

Suština je da ne pomešamo ono što znamo o uzorku s onim što želimo saznati o populaciji (setite se naših grčkih baba i latiničnih žaba). Potreban nam je pouzdan **metod** koji će nam omogućiti da na osnovu podataka iz uzorka donesemo zaključke o populaciji.

U narednim poglavljima upoznaćete se s dva ključna pristupa: **statističkim testovima** i **statističkim ocenjivanjem**. Obe grupe metoda počivaju na ideji koja nije odmah očigledna i često zbunjuje početnike. Centralni koncept je sledeći: statistički metodi upoređuju statistike uzorka s teorijskim vrednostima svih mogućih uzoraka koje bismo mogli očekivati prema pretpostavkama našeg statističkog modela. Zvuči složeno? Ne brinite - razložićemo ovaj proces na jednostavne korake.

## 3.3 Testiranje Kolumbovog modela

Slučaj Kolumbovog putovanja i otkrića Amerike savršeno ilustruje odnos između podataka i hipoteza. On je posmatrao novootkrivenu teritoriju isključivo kroz prizmu svoje pretpostavke - unutar svog malog sveta gde je navodno stigao do Indije. Svoja zapažanja je interpretirao u skladu s tim modelom.

Statističko zaključivanje se bazira na direktnom suočavanju podataka i modela. Cilj je utvrditi da li realnost (izmereni podaci) potvrđuje ili opovrgava naš model. Kolumbo nije bio spreman da prihvati grešku u svojoj pretpostavci, ali su kasnija istraživanja nedvosmisleno pokazala da podaci ne podržavaju njegovu teoriju - otkriven je novi kontinent, a ne zapadni put do Indije. Da bismo izbegli takvu subjektivnost u zaključivanju, statistički metodi nude precizna pravila koja nam govore kada treba odbaciti model i prihvatiti da podaci ne podržavaju naše pretpostavke.

Statistički testovi i statističko ocenjivanje predstavljaju moćne alate koji koriste podatke iz uzorka da bi doneli zaključke o populaciji. Podaci koje dobijamo iz uzorka nisu cilj sami po sebi - oni su početna tačka u procesu donošenja zaključaka. Koristimo ih da napravimo sledeći korak u nastojanju da odgovorimo na naša ključna pitanja. Najčešće postavljamo pitanja sa ciljem da neki fenomen predvidimo ili objasnimo.

Suština ovakvog pristupa leži u poređenju informacija dobijenih iz statistika s našim pretpostavkama i modelima o populaciji. Kada se naše pretpostavke dobro poklapaju s rezultatima iz uzorka, možemo formulisati odgovore na istraživačka pitanja u skladu s tim pretpostavkama. Međutim, ta sigurnost nikada nije apsolutna zbog prirode informacija dobijenih iz uzorka, odnosno zbog prisustva **statističke greške**. S druge strane, ako postoji značajno neslaganje između naših pretpostavki i informacija iz uzorka, odbacićemo te pretpostavke - ponovo, uz određeni stepen (ne)sigurnosti.

Zbog toga, ne polazimo od apsolutne sigurnosti ili neizvesnosti u procesu donošenja zaključaka. Naš zadatak je da se suočimo sa neizvesnošću, nepotpunim informacijama i rizikom pravljenja grešaka. U nauci, jedino statistika i statističko zaključivanje nude sistemski pristup rešavanju ovog izazova.

Suština ovog procesa zaključivanja leži u preciznom merenju koliko se naše pretpostavke (model, mali svet) poklapaju ili razilaze s onim što smo dobili iz uzorka. Za to koristimo **statističke distribucije**. Baš kao što smo ranije izračunali odstupanje zarade pojedinca od prosečne zarade u uzorku ([Odeljak 2.3](#sec-ds)), možemo izmeriti i koliko naše pretpostavke „odstupaju“ od podataka. Primenom jasno definisanih pravila procenjujemo da li je to odstupanje malo (što ukazuje na slaganje pretpostavke s podacima) ili veliko (što sugeriše neslaganje). Ovaj pristup nam pruža objektivan metod za procenu koliko naše ideje odgovaraju stvarnosti koju istražujemo.

Ključni izazov u savladavanju statističkih metoda je razumevanje jezika i logike **statističkih distribucija**, što neizbežno uključuje i ovladavanje osnovnim principima **verovatnoće**. U narednom poglavlju započinjemo upravo s tim konceptima.

|  |  |
| --- | --- |
|  | **Vremenska prognoza i kišobran**  U statistici, odnos između pretpostavke i podataka uvek se izražava kroz verovatnoću. Često analiziramo i pojam rizika, posebno rizik donošenja pogrešne odluke.  Ove rizike procenjujemo svakodnevno. Zamislite situaciju: vaša aplikacija za vremensku prognozu pokazuje 90% verovatnoće za kišu sutra, baš kada treba da krenete na fakultet. Ako odlučite da ponesete kišobran, prihvatate 10% rizik da ćete ga nositi bez potrebe. Većina ljudi će prihvatiti ovaj rizik jer je kiša gotovo izvesna. Međutim, ako je verovatnoća kiše samo 5%, većina će ostaviti kišobran kod kuće - rizik od pokisnuća je minimalan.  Najzanimljiviji slučajevi su oni granični, na primer kada je verovatnoća kiše 40% ili 50%. U tim situacijama dolaze do izražaja individualne razlike u proceni rizika. Da bismo u statistici izbegli subjektivnost, fokusiramo se na ekstremne vrednosti ili „repove“ statističkih distribucija, analizirajući situacije gde se rizici kreću između 1% i 10%. |

## 3.4 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1**  Zamislite da ste deo tima koji proučava vršnjačko nasilje u školama u Srbiji. Pred vama je ograničena evidencija broja prijavljenih slučajeva nasilja, a vaš tim polazi od pretpostavke da je u poslednjih 10 godina došlo do porasta nasilja u školama.  Razmotrite, šta u ovom slučaju predstavlja mali svet, a šta veliki svet sa kojim se suočavate? |

|  |  |
| --- | --- |
|  | **Zadatak 2**  Napišite dva modela koji bi bili adekvatni za istraživanje razlika u uspehu učenika gimnazija i stručnih škola na prijemnim ispitima. |

|  |  |
| --- | --- |
|  | **Zadatak 3 \***  Napišite tri modela koji bi mogli objasniti odnos između potrošačke korpe i minimalne zarade u Srbiji. Obavezno definišite svoje pretpostavke jasnim, nedvosmislenim jezikom. |

# 4. Verovatnoća

Teorija verovatnoće je matematička disciplina koja nam omogućava da kvantifikujemo neizvesnost i donosimo zaključke u situacijama gde imamo ograničene informacije o događajima i njihovim ishodima.

Ova definicija može delovati apstraktno, ali teorija verovatnoće je zapravo veoma praktičan alat koji prožima sve naučne discipline. Nije slučajno što se često naziva „logikom nauke“ (Jaynes, 2003). Da bismo razumeli njenu moć i praktičnu vrednost, počećemo od osnovnih koncepata i videti kako se oni primenjuju kroz praktične primere u programskom jeziku R.

## 4.1 Događaji i ishodi

Osnovni pojam teorije verovatnoće je **eksperiment**. Eksperiment u ovom kontekstu predstavlja nešto sasvim drugačije od eksperimenta u fizici ili hemiji. Eksperiment je bilo koja situacija u kojoj postoji neizvesnost o tome šta će se desiti.

Uzmimo za primer bacanje kockice. To je idealan primer eksperimenta jer ne možemo sa sigurnošću predvideti ishod. Kada bacimo kockicu, može se **dogoditi** jedan od šest mogućih ishoda. **Događaj** predstavlja konkretnu realizaciju eksperimenta, tokom koje se ostvario jedan od tih mogućih ishoda.

Da pojednostavimo ovaj koncept. Kada bacite kockicu i ona pokaže broj 3, „bacanje kockice“ je naš *eksperiment*, „kockica je pokazala 3“ je *događaj* koji se odigrao, a sam broj 3 je *ishod* tog eksperimenta. Ovo razlikovanje između eksperimenta, događaja i ishoda je ključno za razumevanje teorije verovatnoće.

## 4.2 Verovatnoća

Verovatnoća je mera izvesnosti događaja - broj između 0 i 1 koji nam govori koliko je verovatno da će se određeni događaj ostvariti. Vrednost 0 označava nemoguć događaj, dok vrednost 1 označava izvestan događaj koji će se sigurno desiti.

Za računanje verovatnoće neophodno je razumeti strukturu eksperimenta: koliko mogućih ishoda postoji i kako se događaj koji nas zanima uklapa u tu strukturu.

Uzmimo praktičan primer: želimo izračunati verovatnoću da kockica pokaže paran broj. Ovaj događaj se može ostvariti kada kockica pokaže 2, 4 ili 6. Bacanje kockice ima ukupno 6 mogućih ishoda, pri čemu je svaki ishod jednako verovatan. Verovatnoća ovog događaja je stoga 3/6, odnosno 0.5.

Princip izračunavanja je jasan - broj povoljnih ishoda delimo sa ukupnim brojem mogućih ishoda. U našem primeru, događaj se može ostvariti na 3 načina, dok je ukupan broj ishoda 6. Količnik 3/6 = 0.5 nam govori da postoji 50% šanse da kockica pokaže paran broj.

## 4.3 Uslovna verovatnoća

Uslovna verovatnoća predstavlja verovatnoću da će se jedan događaj desiti pod uslovom da se drugi događaj već dogodio. Razmotrimo sledeću situaciju: bacili smo kockicu i znamo da je pao paran broj. Kolika je verovatnoća da je taj broj baš 2?

Princip izračunavanja uslovne verovatnoće je prirodan: prebrojavamo povoljne ishode i delimo ih ukupnim brojem mogućih ishoda, ali sada uzimamo u obzir samo one ishode koji zadovoljavaju postavljeni uslov. U našem primeru, paran broj na kockici može biti 2, 4 ili 6 (ukupno 3 mogućnosti), a broj 2 se pojavljuje samo jednom. Stoga je verovatnoća da smo dobili 2, pod uslovom da znamo da je broj paran, jednaka 1/3.

U matematičkoj notaciji, ako događaj dobijanja parnog broja označimo sa , a događaj dobijanja broja 2 sa , uslovnu verovatnoću zapisujemo kao . Ovde predstavlja dodatnu informaciju koja modifikuje našu procenu verovatnoće događaja .

Uporedimo ovo sa neuslovljenom verovatnoćom. Verovatnoća da pri proizvoljnom bacanju kockice dobijemo 2 je . Verovatnoća dobijanja parnog broja je . Međutim, verovatnoća dobijanja 2 pod uslovom da je broj paran iznosi . Razlika između i ilustruje kako dodatna informacija o događaju menja našu procenu verovatnoće događaja .

## 4.4 Verovatnoća u R-u

Pogledajmo kako ovo funkcioniše na praktičnom primeru. U CSV datoteci kandidati.csv imamo podatke o glasanju građana jednog grada u Srbiji na izborima za gradonačelnika. U uzorku od 400 građana, pratili smo glasove za tri kandidata: Anu, Bogdana i Veljka. Za svakog ispitanika smo takođe zabeležili da li je mlad (do 35 godina) ili ne. Učitajmo ove podatke u R i analizirajmo ih.

podaci <- read.csv("https://sm.atomasevic.com/data/kandidati.csv")  
head(podaci)

kandidat mladi  
1 Ana 0  
2 Bogdan 1  
3 Bogdan 0  
4 Ana 0  
5 Ana 0  
6 Ana 0

Varijabla podaci$kandidat je tekstualnog tipa, dok je podaci$mladi numerička binarna (ili *dummy*) varijabla.

Da bismo razumeli ove podatke kroz prizmu teorije verovatnoće, posmatrajmo čin glasanja kao eksperiment. Kada građanin pristupi glasačkom mestu i bira jednog od kandidata, to za nas predstavlja realizaciju slučajnog eksperimenta. Mogući ishodi su tri kandidata, a ostvareni događaj je konkretan izbor glasača.

Prvo ćemo analizirati raspodelu glasova među kandidatima koristeći funkciju table().

table(podaci$kandidat)

Ana Bogdan Veljko   
 228 152 20

Pogledajmo dobijenu tabelu frekvencija. Ana je osvojila najveći broj glasova u uzorku, zatim sledi Bogdan, dok je Veljko dobio najmanje glasova. Koristeći funkciju nrow(podaci) možemo dobiti ukupan broj opservacija u uzorku. Izračunajmo sada verovatnoću da je nasumično izabrani građanin glasao za Anu.

Ovde je verovatnoća da je nasumično izabrani građanin glasao za Anu, gde predstavlja broj građana koji su glasali za Anu, a označava ukupan broj građana u uzorku.

n <- nrow(podaci)  
tabela <- table(podaci$kandidat)  
f\_ana <- tabela["Ana"]  
p\_ana <- f\_ana / n  
p\_ana

Line 2

Kreiramo tabelu frekvencija za varijablu kandidat

Line 3

Iz tabele izdvajamo broj glasova za Anu

Line 4

Verovatnoću računamo kao odnos broja glasova za Anu i ukupnog broja opservacija

Line 5

Prikazujemo izračunatu verovatnoću

Ana   
0.57

Verovatnoća da će nasumično izabrani ispitanik glasati za Anu iznosi 0.57, odnosno 57%. Postoji jednostavniji način da dobijemo iste rezultate za sve kandidate. Umesto kreiranja tabele frekvencija, možemo koristiti funkciju prop.table() koja direktno računa proporcije. U ovom kontekstu, te proporcije predstavljaju verovatnoće glasanja za svakog kandidata.

prop.table(table(podaci$kandidat))

Line 1

Funkcija prop.table() prima tabelu frekvencija i računa proporcije. Ovo je prvi put da imamo da je input funkcije ouput prethodne funkcije. R će prvo napraviti tabelu frekvencija, a zatim ubaciti tu tabelu kao input za prop.table() funckiju.

Ana Bogdan Veljko   
 0.57 0.38 0.05

Ova funkcija nam daje proporcije za sva tri kandidata. Kada ih saberemo, dobijamo 1, što je logično jer je suma verovatnoća svih mogućih ishoda eksperimenta jednaka 1.

Razmotrimo sada uslovnu verovatnoću u praksi. Pretpostavimo da želimo izračunati verovatnoću da je nasumično izabrani građanin glasao za Anu, uz uslov da znamo da je osoba mlada. Preciznije, ako je poznato da je reč o mladom biraču, kolika je verovatnoća da je glas dat Ani?

Ovo možemo izračunati na dva načina. Prvi pristup je da izdvojimo podskup mladih ispitanika iz uzorka, a zatim izračunamo proporciju glasova za Anu u tom podskupu.

podaci\_mladi <- podaci[podaci$mladi == 1, ]  
prop.table(table(podaci\_mladi$kandidat))

Line 1

Kreiramo novi objekat podaci\_mladi koji sadrži samo one ispitanike iz uzorka koji su mladi.

Line 2

Izračunavamo proporcije za Anu u tom manjem uzorku.

Ana Bogdan Veljko   
0.34375 0.28125 0.37500

|  |  |
| --- | --- |
|  | **Brojanje u R-u**  Za izračunavanje verovatnoće, potrebno je prebrojati različite načine na koje se događaj može ostvariti. Pogledajmo kako se ovo prebrojavanje izvodi u R-u.  Da bismo ilustrovali ovaj koncept, prebrojićemo mlade ispitanike u našem uzorku. Počnimo od varijable podaci$mladi.  head(podaci$mladi)  [1] 0 1 0 0 0 0  Kada govorimo o mladim ispitanicima koji su označeni indikatorom podaci$mladi == 1, važno je obratiti pažnju na dvostruki znak jednakosti ==. U R-u je ovo operator za testiranje jednakosti. Kada napišemo podaci$mladi == 1, zapravo tražimo od R-a da za svaku opservaciju u skupu podataka proveri da li je vrednost varijable mladi jednaka 1. Pogledajmo šta dobijamo kada ovo izvršimo:  podaci$mladi == 1  Line 1  Kada primenimo operator == na vektor podaci$mladi, R proverava svaki element vektora i vraća logički vektor iste dužine. Elementi ovog vektora su TRUE za mlade ispitanike i FALSE za ostale.  [1] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  [25] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE  [49] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  [73] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  [85] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  [97] FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [121] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [145] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE [157] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE [205] FALSE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [217] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [229] FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE [241] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE [253] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [265] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE [277] FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE [301] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE [313] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE [325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE [349] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [361] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [373] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [385] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE [397] FALSE FALSE FALSE FALSE |

Podaci otkrivaju zanimljiv obrazac. Dok u celom uzorku Ana ima podršku od 57%, među mladim biračima ta podrška pada na 34%. Istovremeno, Veljkova podrška pokazuje dramatičan rast - sa 5% u celom uzorku na 37.5% među mladima. Ovo je jasan indikator da Veljko uživa značajno veću popularnost među mladom populacijom nego što to opšti podaci sugerišu.

Da bismo bolje razumeli odnos između starosne grupe i izbornih preferencija, kreiraćemo unakrsnu tabelu (kros-tabulaciju) za varijable mladi i kandidat. Ova tabela će nam precizno prikazati kako se glasovi mladih i starijih građana raspodeljuju među kandidatima - Anom, Bogdanom i Veljkom.

table(podaci$mladi, podaci$kandidat)

Line 1

Kreiramo unakrsnu tabelu za varijable mladi i kandidat. Funkcija table() ovde prima dva argumenta, što nam omogućava da analiziramo odnos između dve varijable. Za razliku od jednostavne tabele frekvencija koja nastaje sa jednim argumentom, dva argumenta generišu matricu koja prikazuje presek ovih varijabli.

Ana Bogdan Veljko  
 0 217 143 8  
 1 11 9 12

Ova tabela nam daje važan uvid u podatke: u našem uzorku imamo 32 mlada ispitanika. Jasno možemo videti raspodelu glasova starijih ispitanika (prvi red tabele) i mladih (drugi red). Međutim, sama tabela ne govori direktno o verovatnoćama. Da bismo izračunali te verovatnoće, potrebno je da konvertujemo frekvencije iz tabele u proporcije.

prop.table(table(podaci$mladi, podaci$kandidat), margin = 1)

Line 1

Kreiramo kros-tabulaciju kao input za funkciju prop.table(). Parametar margin=1 određuje da li se proporcije računaju po redovima (1) ili kolonama (2).

Ana Bogdan Veljko  
 0 0.58967391 0.38858696 0.02173913  
 1 0.34375000 0.28125000 0.37500000

Ovaj pristup daje iste rezultate kao prethodni, ali uz značajnu prednost - omogućava nam direktno poređenje verovatnoća za svakog kandidata u zavisnosti od starosne grupe. Podaci otkrivaju jasan obrazac: Ana dominira među starijim ispitanicima, dok je situacija među mladima bitno drugačija. U mlađoj populaciji, Ana i Veljko imaju gotovo identičnu podršku, što ukazuje na fundamentalno drugačiju strukturu glasačkih preferencija u ovoj demografskoj grupi.

## 4.5 Distribucije verovatnoće

Nejednakosti koje smo uočili u prethodnom primeru pokazuju da nisu svi ishodi jednako verovatni. Veća je verovatnoća da će mlad birač glasati za Veljka nego stariji. Kada smo analizirali bacanje kockice, svaki ishod je bio jednako verovatan - verovatnoća dobijanja broja 1 jednaka je verovatnoći dobijanja brojeva 2, 3, 4, 5 ili 6.

Statistički opis jednakosti ili nejednakosti verovatnoća ishoda eksperimenta predstavljen je **distribucijom verovatnoće**. Distribucije verovatnoće možemo opisati ili matematičkom formulom koja definiše pravilnost pojavljivanja ishoda, ili grafički, linijama koje prikazuju verovatnoće svakog mogućeg ishoda.

Najjednostavnija distribucija verovatnoće je **diskretna uniformna distribucija**, koju smo već sreli pri analizi bacanja kockice. Možemo je zapisati kao:

gde je događaj, a broj mogućih ishoda.

|  |  |
| --- | --- |
|  | **Zašto diskretna?**  Distribucije verovatnoće delimo na diskretne i neprekidne (kontinuirane). Ova podela proističe iz prirode ishoda koje opisuju. Diskretne distribucije opisuju ishode koji se mogu prebrojati, dok neprekidne opisuju ishode koji se ne mogu prebrojati. Razjasnimo ovo kroz primere.  Ključno pitanje koje pomaže u razlikovanju ove dve vrste distribucija je: „Koliko mogućih ishoda postoji između vrednosti a i b?“  Uzmimo bacanje kockice kao primer diskretne distribucije. Između 3 i 6 imamo tačno 2 moguća ishoda (4 i 5). Ili ako posmatramo starost ljudi u godinama, između 25 i 35 godina možemo precizno nabrojati sve vrednosti (26, 27, …, 34).  Nasuprot tome, merenje visine ljudi predstavlja primer neprekidne distribucije. Između 160 cm i 180 cm teoretski postoji beskonačno mnogo mogućih vrednosti. Čak i u najmanjem intervalu, recimo između 160 cm i 160,0000001 cm, matematički gledano postoji beskonačno mnogo mogućih vrednosti. |

Hajde da nacrtamo grafikon ove distribucije. Prvo ćemo da definišemo ishode i njihove verovatnoće.

k <- 6  
x <- 1:k  
y <- rep(1/k, k)

Line 1

Definišemo broj mogućih ishoda

Line 2

Kreiramo vektor ishoda, od 1 do k

Line 3

Kreiramo vektor verovatnoća, gde je svaka verovatnoća jednaka 1/k

Sada možemo grafički predstaviti ove rezultate iscrtavajući vertikalnu liniju iznad svakog ishoda.

par(family = "Jost")  
  
plot(x, y,  
 type = "h", lwd = 3, col = "#5C88DAFF",  
 main = "Diskretna uniformna distribucija",  
 xlab = "Ishod", ylab = "Verovatnoća")  
  
points(x, y, pch = 16, col = "#CC0C00FF", cex = 1.5)

Line 3

x i y predstavljaju vektore ishoda (x-osa) i njihovih verovatnoća (y-osa)

Line 4

type = "h" kreira vertikalne linije. Parametri lwd=3 i col="#5C88DAFF" definišu debljinu i boju linija

Lines 5-6

main, xlab i ylab definišu naslov grafikona i oznake osa

Line 8

points() dodaje tačke na već iscrtani grafikon

|  |
| --- |
| Slika 4.1: Diskretna uniformna distribucija |

Grafikon precizno prikazuje jednaku raspodelu verovatnoće za sve ishode, gde verovatnoća svakog ishoda iznosi približno 0.17. Kroz ovakav vizuelni prikaz jasno se uočava osnovna karakteristika diskretne uniformne distribucije - svi ishodi imaju identičnu verovatnoću pojavljivanja.

Razmotrimo sada složeniji scenario koji uključuje bacanje dve kockice. Posebno nas zanima distribucija zbira brojeva koji se pojavljuju na kockicama.

Matematički gledano, prva kockica (A) i druga kockica (B) svaka imaju po 6 mogućih ishoda. Ukupan broj mogućih kombinacija je njihov proizvod: . Ova naizgled jednostavna situacija otvara put ka dubljoj analizi distribucije verovatnoća.

A <- 1:6  
B <- 1:6  
  
X <- expand.grid(A, B)  
colnames(X) <- c("A", "B")  
X$suma <- X$A + X$B  
  
head(X)

Lines 1-2

Kreiramo vektore A i B koji predstavljaju ishode prve i druge kockice.

Line 4

Funkcija expand.grid() kreira sve moguce kombinacije vrednosti iz A i B.

Line 5

Postavljamo imena kolona na „A“ i „B“.

Line 6

Dodajemo novu kolonu suma koja predstavlja zbir brojeva na kockicama u svim kombinacijama.

A B suma  
1 1 1 2  
2 2 1 3  
3 3 1 4  
4 4 1 5  
5 5 1 6  
6 6 1 7

U prvih nekoliko redova vidimo kombinacije i njihove zbirove, ali da bismo sagledali celokupnu sliku svih mogućih zbirova i njihovih frekvencija, koristimo funkciju table().

table(X$suma)

2 3 4 5 6 7 8 9 10 11 12   
 1 2 3 4 5 6 5 4 3 2 1

Ova tabela prikazuje distribuciju koju intuitivno razumemo kroz iskustvo sa bacanjem kockica. Zbir 12 je redak događaj koji zahteva specifičnu kombinaciju bacanja, dok je zbir 2 jednako redak na suprotnom kraju distribucije. Zbir 1 je nemoguć događaj, kao i svaki zbir veći od 12. Zbirovi između 5 i 9 se javljaju najčešće, što je direktna posledica većeg broja mogućih kombinacija koje vode do ovih ishoda.

Da bismo kvantifikovali ove verovatnoće preciznije, koristimo funkciju prop.table().

prop.table(table(X$suma))

Line 1

Funkcija prop.table() prima tabelu frekvencija i računa proporcije.

2 3 4 5 6 7 8   
0.02777778 0.05555556 0.08333333 0.11111111 0.13888889 0.16666667 0.13888889   
 9 10 11 12   
0.11111111 0.08333333 0.05555556 0.02777778

Prikažimo ovo grafički koristeći sličan pristup kao kod uniformne distribucije.

par(family = "Jost")  
  
plot(table(X$suma),  
 type = "h", lwd = 3, col = "#5C88DAFF",  
 main = "Distribucija zbirova dve kockice",  
 xlab = "Zbir", ylab = "Frekvencija")

Line 3

table(X$suma) daje nam tabelu frekvencija za zbirove.

Line 4

type = "h" govori R-u da nacrta histogram. lwd=3 i col="#5C88DAFF" su opcije koje određuju debljinu i boju linija.

Lines 5-6

main i xlab i ylab su opcije koje određuju naslov grafikona i imena osa.

|  |
| --- |
| Slika 4.2: Distribucija zbirova dve kockice |

Ovaj grafikon ilustruje distribuciju verovatnoća za različite zbirove. Najviši vrh grafikona predstavlja najčešći ishod, dok najniže tačke ukazuju na najmanje verovatne rezultate. Pred nama je primer diskretne distribucije verovatnoće koja odstupa od uniformne, s obzirom na to da se verovatnoće pojedinih zbirova međusobno razlikuju.

Razmotrimo sada kako bi izgledala distribucija kada bacamo četiri kockice. Ovaj primer će nam pomoći da bolje razumemo kako se distribucija menja sa povećanjem broja slučajnih događaja.

A <- 1:6  
B <- 1:6  
C <- 1:6  
D <- 1:6  
  
X <- expand.grid(A, B, C, D)  
colnames(X) <- c("A", "B", "C", "D")  
X$suma <- X$A + X$B + X$C + X$D  
  
par(family = "Jost")  
  
plot(table(X$suma),  
 type = "h", lwd = 3, col = "#5C88DAFF",  
 main = "Distribucija zbirova 4 kockice",  
 xlab = "Zbir", ylab = "Frekvencija")

Lines 1-4

Kreiramo vektore A, B, C i D koji predstavljaju ishode prve, druge, treće i četvrte kockice.

Line 7

Postavljamo imena kolona na „A“, „B“, „C“ i „D“.

Line 8

Dodajemo novu kolonu suma koja predstavlja zbir brojeva na kockicama u svim kombinacijama.

Line 15

Kreiramo grafikon koji prikazuje distribuciju zbirova.

|  |
| --- |
| Slika 4.3: Distribucija zbirova 4 kockice |

Grafikon koji dobijamo pokazuje značajno drugačiju distribuciju, ali zadržava prepoznatljiv oblik. Dobijanje zbira 24 predstavlja izuzetno redak događaj, dok se najveća verovatnoća koncentrisuje oko zbirova između 11 i 17.

Ova dva grafikona otkrivaju fundamentalni obrazac u teoriji verovatnoće. Ishodi sa najvećom verovatnoćom grupišu se oko centralne vrednosti distribucije, odnosno oko proseka. Sa udaljavanjem od proseka, bilo prema većim ili manjim vrednostima, frekvencija i verovatnoća dosledno opadaju. Ekstremne vrednosti, one najudaljenije od proseka, pojavljuju se sa najmanjom verovatnoćom.

Ovo nas dovodi do ključnog pitanja: koja matematička pravilnost opisuje ovo opadanje verovatnoće? Koliki je tačno odnos između verovatnoće dobijanja zbira 8 i zbira 10? Ovakve pravilnosti u distribuciji verovatnoća opisuju se pomoću složenijih matematičkih modela, među kojima je **normalna distribucija** najvažnija i najšire primenjena. U sledećem poglavlju detaljno ćemo istražiti osobine i primene ove distribucije.

## 4.6 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1**  Koristeći R, izračunajte sledeće uslovne verovatnoće:   1. - verovatnoću da je ispitanik glasao za Anu, ako znamo da je mlad 2. - verovatnoću da je ispitanik glasao za Bogdana, ako znamo da nije mlad 3. - verovatnoću da je ispitanik glasao za Bogdana, ako znamo da je mlad   Interpretirajte dobijene rezultate i objasnite šta nam oni govore o odnosu između glasačkih preferencija i starosnih grupa. |

|  |  |
| --- | --- |
|  | **Zadatak 2**  Izračunajmo sledeće uslovne verovatnoće koristeći R:   1. - verovatnoću da je ispitanik mlad, ako znamo da je glasao za Anu 2. - verovatnoću da je ispitanik mlad, ako znamo da je glasao za Veljka   Ove verovatnoće će nam dati jasniju sliku o starosnoj strukturi biračkog tela svakog kandidata. Uporedite dobijene rezultate i razmotrite šta nam oni govore o odnosu između kandidata i demografske strukture njihovih birača. |

|  |  |
| --- | --- |
|  | **Zadatak 3 \***  Kreirajte grafikon distribucije verovatnoća za scenario bacanja 10 kockica i analizirajte raspodelu njihovih zbirova. |

|  |  |
| --- | --- |
|  | **Zadatak 4 \*\***  Analizirajmo podatke primanja.csv. Potrebno je izračunati verovatnoću da mesečna primanja ispitanika iz uzorka padaju u opseg između 1200 i 1300 evra. Ovo je praktičan primer za razumevanje diskretnih distribucija verovatnoće u realnom kontekstu. |

|  |  |
| --- | --- |
|  | **Zadatak 5 \*\***  Koristeći podatke primanja.csv, odredite verovatnoću da ispitanik ima mesečna primanja koja odstupaju za više od 3 standardne devijacije iznad aritmetičke sredine (odnosno, Z-skor veći od 3).  Ovaj zadatak nam pomaže da razumemo kako identifikovati statističke ekstreme u realnim podacima. |

# 5. Normalna distribucija

Za razliku od prethodnog poglavlja, ovde ćemo se baviti neprekidnim, odnosno kontinuiranim varijablama. Kao primer uzećemo visinu izraženu u centrimetrima, koja može uzeti bilo koju vrednost iz određenog intervala.

Kod ovakvih varijabli, verovatnoća pojedinačne vrednosti je praktično nula. Zašto? Zbog prirode neprekidnog intervala - verovatnoća da ćemo u populaciji pronaći osobu visoku tačno 183,9088005321cm je zanemarljiva. Umesto toga, kod neprekidnih varijabli postavljamo praktičnija pitanja, poput „kolika je verovatnoća da će nasumično odabrani ispitanik biti visok između 180cm i 185cm?“.

Hajde da vidimo kako ovo funkcioniše na stvarnim podacima o visini ispitanika.

podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/5f08daba08a4de48265c6ebf6da693dd/raw/be4b8f8c14a0acb57357a0f40e3b883419f80755/visine.csv")  
  
head(podaci)  
  
cat("Broj ispitanika je: ", nrow(podaci), "\n")

Line 1

Učitavanje podataka o visinama ispitanika

Line 3

Prikaz prvih nekoliko redova tabele

Line 5

Ispis broja ispitanika u tabeli

pol visina  
1 M 176.7964  
2 M 179.2737  
3 M 192.6903  
4 M 181.5288  
5 M 181.9697  
6 M 193.8630  
Broj ispitanika je: 400

U skupu podataka imamo 400 ispitanika sa precizno izmerenim visinama (do četiri decimale). Uz visinu, za svakog ispitanika imamo i podatak o polu.

Izračunajmo verovatnoću pronalaska osobe visine između 180cm i 185cm u našem uzorku.

podskup <- podaci[podaci$visina > 180 & podaci$visina < 185,]  
  
cat("Broj ispitanika visine između 180cm i 185cm je: ", nrow(podskup), "\n")  
  
verovatnoca <- nrow(podskup) / nrow(podaci)  
  
  
cat("Verovatnoća da ćemo u uzorku pronaći neku osobu visine između 180cm i 185cm je: ", verovatnoca, "\n")

Line 1

Izdvajanje podskupa ispitanika visine između 180cm i 185cm

Line 5

Izračunavanje verovatnoće kao odnos broja ispitanika u podskupu i ukupnog broja ispitanika

Broj ispitanika visine između 180cm i 185cm je: 45   
Verovatnoća da ćemo u uzorku pronaći neku osobu visine između 180cm i 185cm je: 0.1125

Vidimo da ta verovatnoća iznosi oko 11%. Za razliku od diskretnih varijabli, kod kontinuiranih varijabli nemamo mogućnost da jednostavno izračunamo verovatnoću za sve moguće ishode. Potrebno je da ih grupišemo u intervale (na primer, između 180cm i 185cm) i izračunamo frekvencije tih intervala.

Za vizuelni prikaz frekvencija intervala koristimo histogram, koji možemo kreirati pomoću funkcije hist u R-u.

par(family = "Jost")  
  
hist(podaci$visina,  
 breaks = 10,  
 main = "Histogram visina", xlab = "Visina", ylab = "Frekvencija", col = "#5C88DAFF", lwd = 3)

Line 3

podaci$visina je vektor podataka o visinama, osnovni argument funkcije hist

Line 4

breaks argument govori R-u na koliko malih intervala treba podeliti sve vrednosti varijable podaci$visina

|  |
| --- |
| Slika 5.1: Histogram visina |

Na histogramu uočavamo raspored visina u populaciji koji podseća na distribuciju zbirova bacanja kockica iz prethodnog poglavlja. Za jasniju interpretaciju histograma, pogledajmo osnovne mere deskriptivne statistike.

as\_visina <- mean(podaci$visina)  
med\_visina <- median(podaci$visina)  
sd\_visina <- sd(podaci$visina)  
  
cat("Aritmetička sredina visina je: ", as\_visina, "\n")

Aritmetička sredina visina je: 172.3499

cat("Medijana visina je: ", med\_visina, "\n")

Medijana visina je: 171.3958

cat("Standardna devijacija visina je: ", sd\_visina, "\n")

Standardna devijacija visina je: 9.352587

Histogram otkriva važan obrazac: većina visina grupiše se oko proseka (približno 172cm), a verovatnoća pojave ekstremnih vrednosti, poput 150cm ili 200cm, značajno je manja. Jasno je vidljivo da verovatnoća pronalaska osobe čija visina odstupa od proseka opada proporcionalno sa tim odstupanjem.

Ovakva pravilnost u opadanju verovatnoće nije proizvoljna. Da bismo preciznije analizirali ovaj obrazac i bolje razumeli njegovu prirodu, korisno je predstaviti distribuciju visina kontinuiranom linijom.

par(family = "Jost")  
  
plot(density(podaci$visina),  
 main = "Distribucija visina",  
 xlab = "Visina", ylab = "Gustina",  
 lwd = 3,  
 xlim = c(120, 220),  
 col = "#5C88DAFF")

Line 3

density funkcija računa gustinu verovatnoće za svaku vrednost u našem skupu podataka

Line 6

lwd argument određuje debljinu iscrtane linije

Line 8

xlim argument određuje granice grafikona na x-osi

|  |
| --- |
| Slika 5.2: Distribucija visina |

Na grafikonu vidimo **liniju (krivu) distribucije gustine** koja pokazuje verovatnoću pojave različitih visina. Ne ulazeći duboko u matematičku definiciju gustine, možemo je posmatrati kao meru verovatnoće pojave određene visine, normalizovanu tako da je ukupna površina ispod krive jednaka 1. Grafikon ilustruje kako je verovatnoća najveća oko prosečne vrednosti (približno 172cm) i postepeno opada sa udaljavanjem od nje. Nagib krive precizno opisuje brzinu opadanja verovatnoće s odstupanjem od proseka.

Ova kriva pokazuje karakterističan oblik koji odgovara najvažnijoj teorijskoj distribuciji verovatnoće – **normalnoj distribuciji**. Poznata je i kao Gausova kriva, po **Karlu Fridrihu Gausu**, koji je do nje došao proučavajući obrasce grešaka u astronomskim merenjima (Gauss, 1823).

Pokušajmo da razumemo kako je Gaus otkrio normalnu distribuciju. Dok je noć za noći merio pozicije zvezda na nebu, uočio je nešto neočekivano: većina merenja se grupisala oko jedne centralne vrednosti, a greške su se javljale simetrično sa obe strane. Drugim rečima, verovatnoća da će izmerena pozicija zvezde odstupati za određenu vrednost iznad stvarne pozicije bila je jednaka verovatnoći da će odstupati za istu vrednost ispod.

Ova simetrična priroda grešaka imala je značajne praktične implikacije. Ponavljanjem merenja, pozitivne i negativne greške su se međusobno potirale, što je vodilo ka sve preciznijoj proceni stvarne pozicije zvezde. Naravno, drastične greške u merenju bile su retke i obično su bile posledica tehničkih problema ili nepovoljnih atmosferskih uslova. Upravo ovakav obrazac merenja i grešaka postavio je temelje onome što danas poznajemo kao normalnu distribuciju.

Da bismo konstruisali krivu normalne distribucije za naše podatke o visini, potrebna su nam dva ključna parametra: aritmetička sredina i standardna devijacija. Aritmetička sredina definiše centar distribucije - tačku maksimalne verovatnoće koja predstavlja vrh krive. Ona služi kao referentna tačka koja pozicionira distribuciju na x-osi.

Standardna devijacija ima ključnu ulogu u oblikovanju normalne distribucije. Ona definiše koliko su podaci koncentrisani oko aritmetičke sredine, odnosno koliko brzo opada verovatnoća sa udaljavanjem od centra distribucije. Kada je standardna devijacija manja, distribucija je uža i više zašiljena - podaci su zbijeniji oko proseka. Nasuprot tome, veća standardna devijacija rezultira širom, razvučenijom distribucijom gde su podaci raspoređeni na većem intervalu vrednosti.

Da bismo bolje razumeli kako ova dva parametra oblikuju normalnu distribuciju naših podataka o visini, prikažimo je grafički.

par(family = "Jost")  
  
curve(dnorm(x, as\_visina, sd\_visina),  
 from = 120, to = 220,  
 main = "Normalna distribucija visina",  
 xlab = "Visina", ylab = "Gustina",  
 lwd = 3,  
 col = "#5C88DAFF")

Line 3

dnorm (density-normal) je funkcija koja računa vrednost gustine normalne distribucije za datu vrednost AS i SD

Line 4

from i to argumenti su granice grafikona na x-osi

|  |
| --- |
| Slika 5.3: Normalna distribucija visina |

Funkcija density računa gustinu distribucije na osnovu **empirijskih podataka**. Ona procenjuje verovatnoću ishoda koristeći stvarne opservacije. Nasuprot tome, dnorm izračunava gustinu normalne distribucije pomoću **teorijskih parametara**, pokazujući nam idealan raspored podataka prema zadatim vrednostima.

Kada uporedimo krivu normalne distribucije sa našim histogramom, vidimo da ona ima sličan, ali pravilniji oblik i izraženiju strmost. Ova strmost nam govori da verovatnoća pronalaska ekstremnih vrednosti (recimo, visina ispod 140cm ili iznad 210cm) brzo opada kako se udaljavamo od centra distribucije. Oblik i širina normalne distribucije direktno zavise od njene standardne devijacije.

Da bismo jasnije videli kako standardna devijacija oblikuje distribuciju, pogledajmo kako bi izgledala kada bismo je udvostručili.

par(family = "Jost")  
  
curve(dnorm(x, as\_visina, 1.5\*sd\_visina),  
 from = 120, to = 220,  
 main = "Normalna distribucija visina",  
 xlab = "Visina", ylab = "Gustina",  
 lwd = 3,  
 col = "#5C88DAFF")

Line 3

1.5\*sd\_visina - povećava standardnu devijaciju za 50%

Line 4

from i to argumenti su granice grafikona na x-osi

|  |
| --- |
| Slika 5.4: Normalna distribucija visina sa povećanom standardnom devijacijom |

Sada uočavamo da je distribucija manje strma i šira. Kao posledica toga, ekstremnije vrednosti, poput visine od 140cm, postaju verovatnije nego na prethodnom grafikonu. Ova promena ilustruje kako povećanje standardne devijacije utiče na oblik normalne distribucije, čineći je „razvučenijom“ i manje koncentrisanom oko srednje vrednosti.

|  |  |
| --- | --- |
|  | **Standardna devijacija i preciznost**  Od svih koncepata deskriptivne statistike, standardna devijacija je često najteža za intuitivno razumevanje. Interesantno je da Gaus, prilikom formulisanja matematičkog modela normalne distribucije, nije koristio standardnu devijaciju. Umesto toga, uveo je koncept **preciznosti**.  Razmislimo o tome: Gaus je svoj rad fokusirao na analizu grešaka merenja. U tom okviru, veća preciznost znači strmiju i užu normalnu distribuciju, jer je manja verovatnoća značajnog odstupanja od tačne vrednosti. Suprotno tome, manja preciznost daje širu distribuciju, što odražava veću verovatnoću različitih grešaka merenja.  U matematičkom smislu, preciznost je definisana kao , što direktno pokazuje inverzni odnos sa varijansom.  Hajde da izračunamo preciznost za oba prethodna primera i vidimo šta nam brojevi govore.  preciznost1 = 1/(sd\_visina^2) preciznost2 = 1/(1.5\*sd\_visina)^2  cat("Preciznost u prvom primeru je: ", preciznost1, "\n")  Preciznost u prvom primeru je: 0.01143237  cat("Preciznost u drugom primeru je:", preciznost2, "\n")  Preciznost u drugom primeru je: 0.005081055  Vidimo da je preciznost u prvom slučaju 0.01, a u drugom 0.005, što pokazuje da se preciznost dvostruko smanjila u drugom primeru.  Iako nećemo dalje koristiti koncept preciznosti u ovom kursu, bitno je razumeti da standardna devijacija predstavlja recipročnu vrednost preciznosti. Ovaj odnos nam pruža jasan uvid - povećanje standardne devijacije direktno znači smanjenje preciznosti merenja, i obrnuto. |

## 5.1 Zašto je normalna distribucija bitna?

Vratimo se na Gausov primer merenja greške pri posmatranju astronomskih objekata. Ako znamo da greška ima normalnu distribuciju, možemo izračunati verovatnoću pojave **značajne greške**. Šta to konkretno znači? To znači da možemo precizno odrediti verovatnoću da će se tokom istraživanja pojaviti greška koja može bitno uticati na naše zaključke. Drugim rečima, možemo kvantifikovati rizik da naši rezultati budu pogrešni.

Zar ne bismo želeli da taj rizik eliminišemo? U nauci to nije moguće zbog inherentnih ograničenja mernih instrumenata i činjenice da radimo sa uzorcima, a ne sa celom populacijom. Rizik greške je neizbežan, ali normalna distribucija nam pruža moćan alat za njegovu preciznu kvantifikaciju i opis.

Matematički gledano, normalna distribucija (kao i svaka druga distribucija verovatnoće) je **funkcija**. Kada u nju uvrstimo neku vrednost (u našem slučaju aritmetičku sredinu uzorka), ona nam vraća verovatnoću pojavljivanja te vrednosti u populaciji koja je definisana parametrima - aritmetičkom sredinom i varijansom.

Konkretnije, normalna distribucija nam daje - verovatnoću da ćemo iz populacije sa aritmetičkom sredinom i standardnom devijacijom dobiti uzorak čija je aritmetička sredina . Elegantno i precizno, zar ne?

|  |  |
| --- | --- |
|  | Pažnja, uznemirujući sadržaj! |

Normalna distribucija je vizuelno jednostavna, ali matematički složena funkcija.

Međutim, njen suštinski važan aspekt, koji joj daje prepoznatljivi oblik zvona je:

Što je opservacija dalja od aritmetičke sredine u odnosu na standardnu devijaciju, to je manja verovatnoća njenog pojavljivanja. Ovaj odnos nije linearan - grafikon normalne distribucije nije nacrtan pravim linijama već eksponencijalnom krivom, gde pad verovatnoće određuje kvadrat odstupanja i standardna devijacija.

Ovo merenje odstupanja od aritmetičke sredine pomoću standardne devijacije nije nam nepoznato. Sreli smo ga već kod normalizacije podataka i formiranja Z-skorova.

Primenimo sada naše prethodno znanje i pogledajmo kako izgleda normalna distribucija **standardizovanih visina**.

par(family = "Jost")  
  
Z <- (podaci$visina - as\_visina) / sd\_visina  
hist(Z, breaks = 10, main = "Histogram standardizovanih visina", xlab = "Z-skor", ylab = "Frekvencija", col = "#5C88DAFF", lwd = 3)  
  
cat("Aritmetička sredina standardizovanih visina je: ", round(mean(Z),2), "\n")  
cat("Standardna devijacija standardizovanih visina je: ", round(sd(Z),2), "\n")

Line 3

Standardizacija visina računanjem Z-skorova

Line 4

Crtanje histograma standardizovanih visina

Aritmetička sredina standardizovanih visina je: 0   
Standardna devijacija standardizovanih visina je: 1

|  |
| --- |
| Slika 5.5: Normalna distribucija standardizovanih visina |

Aritmetička sredina standardizovanih visina je 0, a standardna devijacija 1. Ovo nije slučajnost - to je direktna posledica matematičke definicije standardizacije, gde od svake vrednosti oduzimamo aritmetičku sredinu i delimo je standardnom devijacijom. Kada pogledamo detaljnije, uočavamo da raspored podataka nije savršeno simetričan i da je blago pomeren ulevo, sa centrom između 0 i -1. Ovo je sasvim očekivano - empirijske distribucije retko kada pokazuju matematički savršen normalan oblik, ali normalna distribucija i dalje ostaje izvanredan model za opisivanje prirodne tendencije podataka da se grupišu oko centralne vrednosti, sa postepenim opadanjem verovatnoće prema krajevima distribucije.

Izračunajmo sada konkretnu verovatnoću da se ispitanik nalazi u intervalu od jedne standardne devijacije oko aritmetičke sredine. To možemo uraditi na sledeći način:

mean(Z > -1 & Z < 1)

Line 1

Izračunavanje verovatnoće da je Z-skor između -1 i 1

[1] 0.6775

Ovaj izraz u R-u može delovati kompleksno: tražimo verovatnoću, a koristimo aritmetičku sredinu? Šta se zapravo dešava? R prolazi kroz sve Z-skorove i proverava koji su između -1 i 1. Svaki skor koji zadovoljava ovaj uslov označava sa 1, a ostale sa 0. Zatim deli zbir jedinica sa ukupnim brojem opservacija. Rezultat je upravo ono što nam treba - verovatnoća da opservacija pripada zadatom intervalu, izračunata kao relativna frekvencija.

Dakle, verovatnoća da visina nekog ispitanika odstupa najviše jednu standardnu devijaciju od aritmetičke sredine iznosi 67.7%. To znači da se oko dve trećine svih ispitanika nalazi unutar jedne standardne devijacije od proseka. Ovaj raspon definiše tipične visine u našem uzorku.

Šta se dešava kada proširimo „komšiluk“ na 2 standardne devijacije?

[1] 0.96

1. Izračunavanje verovatnoće da je Z-skor između -2 i 2

Vidimo da je to približno 96% svih vrednosti. Samo 4% opservacija je izvan tog intervala.

Šta se dešava kada proširimo „komšiluk“ na 3 standardne devijacije?

[1] 1

1. Izračunavanje verovatnoće da je Z-skor između -3 i 3

Sve vrednosti (100%) nalaze se unutar 3 standardne devijacije od aritmetičke sredine. Nema opservacija koje bi bile udaljenije od tog intervala.

Standardizovana udaljenost od aritmetičke sredine nam omogućava da precizno odredimo šta je **tipično** za uzorak, a šta možemo smatrati retkim ili **ekstremnim**.

Kada se vratimo na Gausov rad o greškama merenja, ovo ima praktičnu primenu - 96% svih izvršenih merenja imaće grešku manju od 2 standardne devijacije. Poznavanjem standardne devijacije merenja procenjujemo da li je prihvatljivo da 96% grešaka bude u tom rasponu. Pri tome prihvatamo činjenicu da će 4% grešaka biti veće od 2 standardne devijacije, što je podsetnik da su retke, ali značajne greške neizbežne.

Jedna od elegantnih karakteristika normalne distribucije je da kod podataka koji je prate, raspored Z-skorova sledi jednostavno numeričko pravilo: 68-95-99.7.

* 68% svih Z-skorova nalazi se u intervalu od -1 do 1
* 95% svih Z-skorova je u rasponu od -2 do 2
* 99.7% svih Z-skorova obuhvaćeno je intervalom od -3 do 3

par(family = "Jost")  
z <- seq(-4, 4, length=1000)  
  
density <- dnorm(z)  
  
plot(z, density, type="l", lwd=3, col= "#5C88DAFF",  
 main="Pravilo 3 sigme",  
 xlab="Z-skor", ylab="Gustina")  
  
polygon(c(-1, seq(-1, 1, length=100), 1), c(0, dnorm(seq(-1, 1, length=100)), 0), col=rgb(0, 0, 1, 0.2))  
polygon(c(-2, seq(-2, 2, length=100), 2), c(0, dnorm(seq(-2, 2, length=100)), 0), col=rgb(0, 0, 1, 0.2))  
polygon(c(-3, seq(-3, 3, length=100), 3), c(0, dnorm(seq(-3, 3, length=100)), 0), col=rgb(0, 0, 1, 0.2))  
  
abline(v=c(-1, 1), col="#CC0C00FF", lty=2, lwd=2)  
abline(v=c(-2, 2), col="#00AF66FF", lty=2, lwd=2)  
abline(v=c(-3, 3), col="black", lty=2, lwd=2)  
  
text(0, 0.35, "68%", col="#CC0C00FF", font=2)  
text(0, 0.15, "95%", col="#00AF66FF", font = 2)  
text(0, 0.05, "99.7%", col="black", font = 2)  
  
legend("topright", legend=c("1 SD", "2 SD", "3 SD"),  
 col=c("#CC0C00FF","#00AF66FF", "black"), lty=2, lwd=3, bty="n")

Line 2

seq(-4, 4, length=1000) generiše 1000 vrednosti između -4 i 4

Line 4

Kreiranje normalne distribucije vrednosti između -4 i 4

Line 8

Crtannje krive normalne distribucije

Line 12

Dodajemo osenčenje/obojene oblasti koje odgovaraju intervalima od -1 do 1, -2 do 2 i -3 do 3

Lines 14-16

Dodajemo vertikalne linije koje označavaju granice intervala

Line 23

Dodajemo legendu koja objašnjava boje i linije na grafikonu

|  |
| --- |
| Slika 5.6: Pravilo 3 sigme |

Ovo pravilo poznaje se kao **empirijsko pravilo** ili **pravilo tri sigme**. Iako nije univerzalno primenljivo, daje nam moćan alat za procenu verovatnoće odstupanja opservacija od aritmetičke sredine.

Distribucija Z-skorova koju ovde opisujemo naziva se Z-distribucija ili **standardizovana normalna distribucija**. Njena široka primena dolazi iz činjenice da x-osa nije vezana za konkretne jedinice mere ili veličine opservacija sa kojima radimo. Ovo svojstvo omogućava nam da direktno poredimo različite skupove podataka.

No, pravi potencijal Z-distribucije u statističkom zaključivanju otkriva se kada umesto distribucije podataka razmotrimo **distribuciju aritmetičkih sredina uzoraka**. Ovo je složen koncept, ali možemo ga ilustrovati kroz jedan hipotetički primer.

## 5.2 Šta bi bilo kada bi bilo?

Vratimo se na primer sa primanjima građana. Radili smo sa podacima o primanjima ispitanika, ali hajde da se vratimo u domen **malog sveta** i razmotrimo kakva bi mogla biti primanja građana Srbije. Nemamo konkretne podatke pred sobom i krećemo se u domenu pretpostavki. Recimo da smo optimistični i pretpostavljamo da su **prosečna primanja građana Srbije** 1000€. Ne govorimo o uzorku, već o populaciji - celokupnom radno aktivnom stanovništvu.

Kako formalno zapisujemo ovu pretpostavku?

Na levoj strani je nepoznati **parametar** populacije, odnosno aritmetička sredina varijable na nivou populacije (prosečna primanja građana Srbije). Na desnoj strani je naša optimistična pretpostavka. Budući da su parametri uvek nepoznati, ovaj znak jednakosti izražava pretpostavku o mogućoj vrednosti parametra .

U takvom scenariju, šta možemo očekivati ako sprovedemo anketno istraživanje na uzorku i pitamo ispitanike o njihovim primanjima? Slično Gausovom iskustvu, ne možemo očekivati da će aritmetička sredina uzorka biti precizno 1000€. Razlog je jednostavan - svaki uzorak, bez obzira na njegov kvalitet, predstavlja samo segment populacije i nosi nepotpune informacije o njoj.

Preciznije, govorimo o **varijabilitetu uzorka**. Zamislite uzorak kao rezultat izvlačenja loto loptica - možda nam se u uzorku nađu pretežno siromašniji građani, ili nam se nekoliko milionera slučajno pojavi u podacima. Nemamo kontrolu nad ovim procesom, i sasvim je normalno očekivati da prosek našeg uzorka neće biti identičan stvarnom proseku populacije. Ako bismo ovo preveli na jezik prirodnih nauka, rekli bismo da prosečna primanja koja dobijemo u uzorku sadrže **grešku merenja** u odnosu na prosek populacije.

Kako analizirati ove greške? Kako proceniti da li smo u uzorku pronašli mala ili velika odstupanja od proseka populacije? Odgovor leži u specifičnoj vrsti Z-distribucije koja se zove **distribucija aritmetičkih sredina uzoraka**.

## 5.3 Distribucija aritmetičkih sredina uzoraka

Sagledajmo situaciju: zamislili smo populaciju sa aritmetičkom sredinom od 1000€, odnosno . Kada iz te populacije (npr. radno stanovništvo Srbije) uzmemo uzorak od 60 ispitanika i pitamo ih o njihovim primanjima, izračunavamo aritmetičku sredinu uzorka .

Simulirajmo ovaj proces. U ovom primeru, pretpostavićemo da je standardna devijacija populacije . U realnim istraživanjima ova vrednost nam nije poznata, ali nam je potrebna za simulaciju.

set.seed(12345)  
N = 60  
mi = 1000  
sigma = 600  
  
X = rnorm(N,mi,sigma)  
AS = mean(X)  
  
cat("Aritmetička sredina uzorka je: ", AS)

Line 1

set.seed(12345) - omogućava nam da reprodukujemo isti rezultat u slučaju da želimo da ponovimo eksperiment.

Line 6

rnorm(N,mi,sigma) - generiše nam 100 slučajnih brojeva iz normalne distribucije sa aritmetičkom sredinom 1000 i standardnom devijacijom 400.

Aritmetička sredina uzorka je: 1117.109

Dobijamo , aritmetičku sredinu uzorka koja je veća od aritmetičke sredine populacije. Zbog čega? Zbog greške merenja aritmetičke sredine na osnovu uzorka. Budući da uzorak ne sadrži sve informacije o populaciji, broj koji dobijemo ne može potpuno tačno opisati aritmetičku sredinu populacije, već neminovno dolazi do odstupanja.

Hajde da izračunamo to odstupanje:

odstupanje = AS - mi  
cat("Odstupanje je: ", odstupanje)

Odstupanje je: 117.1094

Da li je ovo odstupanje značajno? U ovom trenutku ne možemo dati precizan odgovor. Simulirajmo ovaj proces 100 puta da bismo razumeli distribuciju mogućih aritmetičkih sredina uzoraka koje bismo mogli dobiti u realnom istraživanju sa istim parametrima populacije.

1. k = 100 - broj uzoraka koje želimo da simuliramo
2. sredine\_uzoraka = c() - pravimo prazan vektor koji će sadržati sredine uzoraka
3. for (i in 1:k) - petlja koja će se izvršiti 100 puta. U programiranju, petlje sadrže niz instrukcija koje se izvršavaju više put. U ovom slučaju izvršićemo tri naredbe 100 puta.
4. uzorak = rnorm(N,mi,sigma) - pravimo uzorak od 100 slučajnih brojeva iz normalne distribucije sa aritmetičkom sredinom 1000 i standardnom devijacijom 400. Potom računamo aritmetičku sredinu tog uzorka.
5. sredine\_uzoraka = c(sredine\_uzoraka,AS) - pridružujemo sredinu uzorka vektoru sredine\_uzoraka. Funkcija c() spaja dve vrednosti u jednu.

Hajde da vidimo nekoliko vrednosti sredine\_uzoraka.

head(sredine\_uzoraka)

[1] 1138.2487 984.8782 1057.9348 946.0477 1135.9611 1038.9922

Očigledno, sredine uzoraka su različite. Da bi ih bolje uporedili, možemo da ih prikažemo histogramom.

par(family = "Jost")  
hist(sredine\_uzoraka,  
 ylab='Frekvencija',  
 xlab='Aritmetička sredina uzorka',  
 main='100 uzoraka',  
 col="#5C88DAFF",  
 lwd=3)

|  |
| --- |
| Slika 5.7: Histogram aritmetičkih sredina uzoraka |

Histogram jasno pokazuje približno normalnu distribuciju. Najveća koncentracija aritmetičkih sredina uzoraka nalazi se oko vrednosti 1000, dok njihova učestalost postepeno opada ka krajevima distribucije. Tipičan uzorak će dati rezultat blizak aritmetičkoj sredini populacije, mada su moguća i značajnija odstupanja.

Važno je naglasiti da je ovo neprekidna (kontinuirana) distribucija, što znači da posmatramo intervale vrednosti, jer je verovatnoća pojave bilo koje pojedinačne vrednosti jednaka nuli.

Otkriće da aritmetičke sredine uzoraka formiraju normalnu distribuciju predstavlja temelj statističkog zaključivanja. U teoriji verovatnoće, ovo fundamentalno svojstvo poznato je kao centralna granična teorema.

|  |  |
| --- | --- |
|  | **Šta je centralna granična teorema?**  Ova teorema je kamen temeljac matematičke statistike. Iako nećemo ulaziti u njen formalni matematički zapis, možemo je objasniti jednostavno i direktno.  Počnimo ovako: imamo nezavisnih slučajnih uzoraka veličine iz iste populacije. Ta populacija ima aritmetičku sredinu i konačnu varijansu . Sa označavamo aritmetičku sredinu bilo kog od tih uzoraka.  Centralna granična teorema je jednostavna ali moćna: kad raste, distribucija aritmetičkih sredina teži ka standardizovanoj normalnoj raspodeli. Matematički to znači da standardizovana vrednost svake aritmetičke sredine prati normalnu distribuciju kada teži beskonačnosti.  Šta ovo znači u praksi? Na velikim uzorcima ne vidimo ekstremna odstupanja aritmetičkih sredina. Odstupanja se grupišu oko nule i prate predvidljiv obrazac koji definišu zakoni verovatnoće. To nam omogućava da preciznije procenimo parametre populacije i radimo pouzdanije statističke analize.  Praktična primena ove teoreme je direktna i suštinska. Ona nam pruža matematički okvir za razumevanje veze između našeg uzorka i populacije koju istražujemo. To je osnovni alat statističkog zaključivanja koji nam omogućava da precizno kvantifikujemo neizvesnost naših procena. |

Zahvaljujući ovoj teoremi možemo izračunati teorijsku distribuciju aritmetičkih sredina uzoraka iz određene populacije. Za to nam je neophodna informacija o aritmetičkoj sredini populacije, ali i još jedan ključni element. Taj element koji nedostaje naziva se **standardna greška aritmetičke sredine uzorka**.

|  |  |
| --- | --- |
|  | **Zakon velikih brojeva**  Zakon velikih brojeva je jedan od najčešće pogrešno interpretiranih koncepata u statistici. Kada čujete da će se nešto „sigurno desiti“ ili da nešto možemo „100% očekivati“ na veoma velikom uzorku „jer važi zakon velikih brojeva“, znajte da je to potpuno pogrešno tumačenje i da osoba koja to govori ne razume osnovne principe statistike.  Zakon velikih brojeva je direktna posledica centralne granične teoreme. On nam govori da se na velikom uzorku aritmetička sredina distribucije približava aritmetičkoj sredini populacije. Preciznije, ako uzmemo sve više i više uzoraka, **aritmetička sredina aritmetičkih sredina tih uzoraka** će konvergirati ka aritmetičkoj sredini populacije.  Zašto je to tako? Mehanizam je jednostavan - kada izvlačimo veliki broj uzoraka iz populacije, individualna odstupanja aritmetičkih sredina od prave vrednosti se međusobno poništavaju. Konačan prosek tih aritmetičkih sredina prirodno teži ka aritmetičkoj sredini populacije.  Ključno je razumeti da to ne znači da će bilo koji pojedinačni uzorak biti savršeno tačan ili imati nulto odstupanje od prave vrednosti. To je matematički nemoguće - nijedan uzorak, bez obzira na broj opservacija i veličinu, ne može dati apsolutno preciznu vrednost aritmetičke sredine populacije. Imajte ovo na umu kada se susretnete sa pozivanjem na zakon velikih brojeva. |

## 5.4 Standardna greška

Standardna greška je standardna devijacija distribucije aritmetičkih sredina uzoraka. Ona pokazuje koliko varijacija u proseku možemo očekivati kada računamo aritmetičke sredine na uzorcima jednake veličine koje uzimamo iz iste populacije.

Drugim rečima, ona predstavlja grešku merenja kada procenjujemo aritmetičku sredinu uzorka. Standardna greška nam otkriva preciznost našeg izračunavanja aritmetičke sredine uzorka. Ovaj pojam smo sreli kod centralne granične teoreme i izražava se sledećom formulom:

Kao što vidite standardna greška zavisi od dva faktora:

1. Standardne devijacije varijable na nivou populacije
2. Veličine uzorka

Standardnu devijaciju populacije ne možemo kontrolisati - ona je deo **velikog sveta**. Na primer, da li su zarade u društvu relativno ujednačene (mali varijabilitet) ili postoje ekstremne razlike (veliki varijabilitet), to ne zavisi od nas već od karakteristika društva koje proučavamo. Slično tome, raspon visina u populaciji je biološka činjenica na koju ne možemo uticati.

Ono što je pod našom kontrolom jeste veličina uzorka. Veći uzorak znači manju standardnu grešku. Međutim, važno je razumeti da ovaj odnos nije linearan - ne možete jednostavnim udvostručavanjem uzorka prepoloviti grešku merenja. Matematički gledano, standardna greška teži nuli tek kada veličina uzorka dostigne veličinu populacije. Ovo je logično jer u tom slučaju ne bismo više radili sa uzorkom, već bismo imali potpune informacije o celoj populaciji.

Standardna greška definiše oblik distribucije aritmetičkih sredina uzoraka. Što je manja standardna greška, to je distribucija sredina uzoraka uža i viša, što direktno ukazuje na veću preciznost u proceni aritmetičke sredine uzorka.

Međutim, postavlja se praktično pitanje: kako izračunati standardnu grešku kada ne znamo standardnu devijaciju (tj. varijansu) populacije? U odsustvu informacija o populaciji, logičan pristup je da koristimo standardnu devijaciju uzorka kao najbolju dostupnu ocenu standardne devijacije populacije.

Ova formula je u stvari ocena standardne devijacije populacije i ona je u skladu sa centralnom graničnom teoremom.

|  |  |
| --- | --- |
|  | **Zašto grešimo?**  U statistici pojam greške nije indikator neispravnosti ili propusta. Greška predstavlja prirodne varijacije koje proizilaze iz procesa slučajnog uzorkovanja. Svi oblici neizvesnosti i nepotpunih informacija inherentnih uzorku čine statističku grešku. Možemo je posmatrati kao meru nepreciznosti koja neizbežno prati svako istraživanje zasnovano na uzorcima. |

Vratimo se na konkretan primer sa primanjima građana. Definisali smo da je aritmetička sredina populacije 1000€, a standardna devijacija 600€. Radimo sa uzorkom od 60 ispitanika. Potrebno je da vizuelno predstavimo distribuciju aritmetičkih sredina uzoraka. Centar ove distribucije je na 1000€, a njena standardna devijacija je zapravo standardna greška. Prvi korak je izračunavanje te standardne greške.

sigma = 600  
n = 60  
s\_AS = sigma / sqrt(n) #<1>\  
cat("Standardna greška je: ", s\_AS)

Standardna greška je: 77.45967

1. sigma / sqrt(n) - izračunava se kao količnik standardne devijacije populacije i kvadratnog korena veličine uzorka.

Dakle, preciznost sa kojom možemo utvrditi koliko iznosi *aritmetička sredina* uzorka je 77.46€. Sada ćemo da iskoristimo ovu vrednost da nacrtamo liniju distribucije aritmetičkih sredina uzoraka.

par(family = "Jost")  
  
curve(dnorm(x,mi,s\_AS),  
 from=mi-5\*s\_AS,  
 to=mi+5\*s\_AS,  
 lwd=3,  
 col = "#5C88DAFF",  
 ylab='Gustina verovatnoće',  
 xlab='Aritmetička sredina uzorka',  
 main='Distribucija aritmetičkih sredina uzoraka')

Line 3

Teorijska normalna distribucija čija je aritmetička sredina 1000 i standardna devijacija 40.

|  |
| --- |
| Slika 5.8: Distribucija aritmetičkih sredina uzoraka |

Ova normalna distribucija pokazuje šta možemo očekivati kada sprovodimo istraživanje na uzorku iz populacije gde građani u proseku zarađuju 1000 evra, uz standardnu devijaciju od 600 evra. Na većini uzoraka iz ove populacije, prosečna zarada će se kretati između 800 i 1200 evra. Ovaj interval nam daje jasan matematički okvir za interpretaciju rezultata i pomaže nam da razumemo prirodnu varijabilnost koju srećemo u praksi.

Vratimo se sada na konkretne podatke koje smo analizirali u prvom poglavlju. Evo šta smo dobili:

podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/492eaf9250b68a35557f224b20e8b310/raw/94329ddd24964e6a9fe0d392a6482d556aa2b54b/primanja.csv")  
AS <- mean(podaci$primanja)  
cat("Aritmetička sredina uzorka je: ", AS)

Aritmetička sredina uzorka je: 799.8

Aritmetička sredina našeg uzorka iznosi 799.8 evra. Pretpostavimo da je ovaj uzorak izvučen iz populacije čija je aritmetička sredina 1000 evra. Ova pretpostavka nam omogućava da izračunamo odstupanje i precizno kvantifikujemo grešku merenja.

Odstupanje je: -200.2

Aritmetička sredina uzorka je 200.2 evra manja od aritmetičke sredine populacije. Da bismo procenili značajnost ovog odstupanja, uporedićemo ga sa standardnom greškom koju smo već izračunali. Ovakvo odstupanje, izraženo u jedinicama standardne greške, predstavlja Z-skor aritmetičke sredine. Ovaj pristup nam omogućava da precizno kvantifikujemo odstupanje našeg uzorka od očekivane vrednosti.

Odnos odstupanja i standardne greške je: -2.584571

Naša aritmetička sredina uzorka odstupa 2.58 standardnih grešaka od aritmetičke sredine populacije. Setimo se pravila 3 sigme koje nam kaže da će 95% svih aritmetičkih sredina uzoraka iz date populacije biti unutar 2 standardne devijacije od aritmetičke sredine populacije. Naš rezultat od -2.58 standardnih devijacija je, dakle, veoma redak događaj koji zahteva pažljivu analizu.

Da bismo bolje razumeli ovo odstupanje, prikažimo ga grafički. Na krivu distribucije aritmetičkih sredina dodaćemo vertikalnu liniju koja označava aritmetičku sredinu našeg uzorka. Ovakav prikaz nam omogućava da jasno vidimo koliko naš uzorak zaista odstupa od očekivane vrednosti i pruža nam jednostavan način da intuitivno shvatimo njegovu poziciju u odnosu na teorijsku distribuciju.

par(family = "Jost")  
  
curve(dnorm(x,mi,s\_AS),  
 from=mi-5\*s\_AS,  
 to=mi+5\*s\_AS,  
 lwd=3,  
 col = "#5C88DAFF",  
 ylab='Gustina verovatnoće',  
 xlab='Aritmetička sredina uzorka',  
 main='Distribucija aritmetičkih sredina uzoraka')  
abline(v=AS,col="#CC0C00FF", lwd = 3)

Line 3

Ovo je teorijska normalna distribucija sa aritmetičkom sredinom 1000 i standardnom devijacijom 40. Ona predstavlja očekivanu raspodelu aritmetičkih sredina uzoraka.

Line 11

Funkcija abline dodaje vertikalnu liniju na grafikon. Parametar v=AS postavlja liniju vertikalno na poziciju koja odgovara aritmetičkoj sredini našeg uzorka. Argument col="red" boji liniju u crveno, što je čini jasno vidljivom.

|  |
| --- |
| Slika 5.9: Distribucija aritmetičkih sredina uzoraka |

Grafikon pokazuje da se aritmetička sredina našeg uzorka nalazi na samom kraju leve strane distribucije, u području gde gustina verovatnoće teži nuli. Ovo otvara zanimljivo pitanje - kako smo dobili uzorak koji je, prema teoriji, toliko malo verovatan da bi trebalo da bude praktično nemoguć? Ovaj neočekivani rezultat ukazuje na ozbiljan problem koji zahteva detaljnu analizu. U narednom poglavlju ćemo istražiti kako je moguće da naš uzorak toliko dramatično odstupa od teorijskih očekivanja.

## 5.5 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1**  Izračunajte uslovnu verovatnoću da je ispitanik viši od 180cm ako je muškarac i ako je žena. Šta nam razlike u tim verovatnoćama govore o odnosu visina muškaraca i žena? |

|  |  |
| --- | --- |
|  | **Zadatak 2**  Preuzmite podatke o primanjima iz datoteke primanja.csv. Iskoristite ove podatke da konstruišete teorijsku distribuciju aritmetičkih sredina uzoraka. Neka distribucija ima identičnu aritmetičku sredinu kao originalni podaci, a standardna greška neka bude jednaka standardnoj grešci uzorka iz podataka. |

|  |  |
| --- | --- |
|  | **Zadatak 3 \***  Nacrtajte grafikon koji prikazuje krive gustine visine za muškarce i žene. Za prvi grafikon koristite funkciju plot(density(...), col="#CC0C00FF"), a za drugi dodajte krivu koristeći lines(density(...), col="#5C88DAFF"). Dodajte legendu koja objašnjava koju liniju predstavlja koja grupa. |

|  |  |
| --- | --- |
|  | **Zadatak 4 \*\***  Ispitajte kako zakon velikih brojeva deluje u praksi. Napravite simulaciju sa 50 uzoraka veličine 100 iz normalne distribucije sa aritmetičkom sredinom 1000 i standardnom devijacijom 400. Za svaki uzorak izračunajte aritmetičku sredinu i prikažite njihov raspored histogramom. Ponovite postupak za 100, 500 i 1000 uzoraka.  Objasnite kako se menja aritmetička sredina svih uzoraka sa povećanjem broja uzoraka. Stavite rezultate u kontekst zakona velikih brojeva. |

# 6. Statistička značajnost, intervali poverenja i Z-test

## 6.1 Retki događaji i pravilo tri sigme

U prethodnom poglavlju susreli smo se sa grafikonom normalne distribucije i vrednošću koja se nalazi na njenom rubu, što ukazuje na to da je ta vrednost izuzetno retka.

Suština statističkog zaključivanja leži upravo u analizi retkih događaja. Kada naiđemo na rezultat koji značajno odstupa od očekivanih vrednosti, to nam pruža mogućnost da preispitamo naše početne pretpostavke o populaciji koju proučavamo.

Vratimo se na naš primer i napravimo pregled ključnih saznanja do kojih smo došli.

Sumiraćemo ključne tačke našeg primera:

1. Početna pretpostavka bila je da prosečna zarada u Srbiji iznosi 1000 EUR.
2. Na osnovu te pretpostavke konstruisali smo normalnu distribuciju aritmetičkih sredina uzoraka ([Slika 5.9](#fig-distribucija-AS)).
3. Analiza stvarnog uzorka pokazala je prosečnu zaradu od 799.8 EUR.
4. Koristeći standardnu grešku, izračunali smo Z-skor koji iznosi -2.58.
5. Ustanovili smo da se ova vrednost nalazi na ekstremnom levom kraju normalne distribucije.

Na osnovu ovih informacija, imamo dve opcije za razmatranje:

1. Imamo problem sa uzorkom koji je zbog svoje specifične strukture doveo je do ekstremnog rezultata.
2. Uzorak je adekvatan, ali je naša početna pretpostavka o populaciji bila netačna.

Drugim rečima, ako pretpostavimo da prosečna zarada iznosi 1000 EUR, a podaci iz našeg uzorka pokazuju drugačije (značajno odstupaju od te pretpostavke), onda neko mora biti „odgovoran“: ili uzorak ili pretpostavka.

U statističkom zaključivanju uvek branimo uzorak. Zbog čega? Prvenstveno zato što se oslanjamo na pretpostavku da je uzorak slučajan i da je proces prikupljanja podataka bio ispravan. Uz to, veliki uzorak donosi manju standardnu grešku, što nam daje čvrstu osnovu da verujemo da uzorak neće previše odstupati od stvarne vrednosti parametra populacije.

Kada se nađemo u situaciji gde rezultat uzorka značajno odstupa od distribucije koja proizlazi iz naše pretpostavke, logičan korak je preispitivanje te pretpostavke.

Kada od pretpostavke dođemo do retkog ili ekstremnog rezultata, logično je zaključiti da je pretpostavka verovatno pogrešna.

Retkost ili ekstremnost nekog rezultata (npr. aritmetičke sredine uzorka) izražavamo kroz standardizovane skorove, kao što je Z-skor. Z-skor predstavlja udaljenost vrednosti od centra distribucije, izraženu u broju standardnih grešaka. Ali kako znamo kada smo ušli u region ekstremnih rezultata? Odgovor leži u **pravilu tri sigme** ([Slika 5.6](#fig-pravilo-3-sigme)). Podsetimo se da se oko 95% svih rezultata nalazi unutar 2 standardne devijacije od centra distribucije, dok se gotovo svi rezultati (99.72%) nalaze unutar 3 standardne devijacije. Ovi procenti definišu granice regiona ekstremnosti. Na primer, ako rezultat ima Z-skor +2.1, on se nalazi među 2.5% najekstremnijih rezultata iznad centra distribucije. Zašto 2.5%? Zato što je 5% najekstremnijih rezultata podeljeno na dva dela, levo i desno od centra distribucije. Prema konvenciji, ovo uzimamo kao jedan od pragova ekstremnosti rezultata.

Zašto baš ovakav prag? Ne postoji čvrst teorijski razlog - reč je o praktičnoj konvenciji koja se zasniva na činjenici da dve standardne devijacije obuhvataju približno 95% svih rezultata u normalnoj distribuciji. Ova konvencija se pokazala korisnom i efikasnom u statističkom zaključivanju tokom proteklog veka.

Postoji i pragmatičniji pristup oceni ekstremnosti rezultata kroz **p-vrednost**. P-vrednost predstavlja verovatnoću dobijanja rezultata koji je jednak ili ekstremniji od onog koji smo dobili u našem uzorku. Kada je p-vrednost manja od 5% (ili 0.05), smatramo da je rezultat prešao prag ekstremnosti o kojem smo malopre govorili.

Da bismo ovo ilustrovali, razmotrimo jednostavan primer. Pretpostavimo da imamo rezultat aritmetičke sredine uzorka koji je udaljen 2.2 standardne greške od centra distribucije, ne ulazeći u detalje o izvoru podataka.

par(family = "Jost")  
  
z <- seq(-3, 3, length=1000)  
  
density <- dnorm(z)  
  
plot(z, density, type="l", lwd=3, col="#5C88DAFF",  
 xlab="Z-skor", ylab="Gustina", xaxt="n")  
  
axis(1, at=c( -3, -2, -1, 0, 1, 2, 2.2, 3), labels=c( -3, -2, -1, 0, 1, 2, 2.2, 3))  
  
polygon(c(-1, seq(-1, 1, length=100), 1), c(0, dnorm(seq(-1, 1, length=100)), 0), col=rgb(0, 0, 1, 0.2))  
polygon(c(-2, seq(-2, 2, length=100), 2), c(0, dnorm(seq(-2, 2, length=100)), 0), col=rgb(0, 0, 1, 0.2))  
polygon(c(-3, seq(-3, 3, length=100), 3), c(0, dnorm(seq(-3, 3, length=100)), 0), col=rgb(0, 0, 1, 0.2))  
  
abline(v=c(-2, 2), col="black", lty=2, lwd=2)  
  
text(0, 0.15, "95%", col="black", font=2)  
  
legend("topright", legend=c("2 SD"),  
 col="black", lty=2, lwd=3, bty="n", cex=0.9, text.font=2)  
  
abline(v=2.2, col="#CC0C00FF", lwd=3)  
  
polygon(c(2.2, seq(2.2, 4, length=100), 4), c(0, dnorm(seq(2.2, 4, length=100)), 0), col=rgb(1, 0, 0, 0.2))  
  
p\_value <- 1 - pnorm(2.2)  
  
text(2.8, 0.2, paste("p-vrednost = \n", round(p\_value, 3)), col="#CC0C00FF", font=2, cex=0.8)

Line 3

Definišemo opseg Z-skorova od -3 do 3.

Line 5

Izračunavamo gustinu distribucije za svaku vrednost Z-skor.

Line 8

Crtamo grafikon gustine distribucije.

Line 10

Postavljamo oznake na x-osi.

Line 14

Prikazujemo regione statističke značajnosti kao osenčene oblasti.

Line 27

p-vrednost dobijamo kada od 1 (ukupna verovatnoća) oduzmemo verovatnoću da se u distribucij nalazi rezultat koji je manji ili jednak 2.2

|  |
| --- |
| Slika 6.1: Primer p-vrednosti |

Rezultat, tj. z-skor 2.2, prikazan je crvenom linijom. Površina ispod krive distribucije koja je ekstremnija od ovog rezultata označena je crvenom bojom (desni rep). Ova osenčena površina predstavlja p-vrednost. Što je p-vrednost manja, to je rezultat ekstremniji jer se nalazi u oblasti retkih vrednosti, tj. u repovima distribucije.

Suština statističkog zaključivanja je donošenje zaključaka o populaciji na osnovu informacija iz uzorka. P-vrednost nam govori koliko je verovatno da dobijemo ovakav ili ekstremniji rezultat ako je naša pretpostavka o populaciji tačna. Kada je naš rezultat u oblasti ekstremnih vrednosti, to nam sugeriše da početna pretpostavka verovatno nije tačna. Nasuprot tome, ako se rezultat našeg uzorka nalazi u središnjem delu distribucije, odnosno u oblasti očekivanih vrednosti, nemamo dovoljno dokaza da sumnjamo u početnu pretpostavku.

Statističko zaključivanje je proces donošenja odluka zasnovan na pretpostavkama i podacima iz uzorka. Dva osnovna oblika statističkog zaključivanja su **testiranje nulte hipoteze** i **konstrukcija intervala poverenja**.

Vratimo se na naš primer sa zaradom - kroz njega ćemo detaljnije objasniti proces testiranja hipoteza i koncept **statističke značajnosti**.

## 6.2 Testiranje hipoteza

Početna pretpostavka u našem prethodnom primeru bila je da prosečna zarada građana Srbije iznosi 1000 EUR. Ova pretpostavka naziva se **nulta hipoteza** i označava se sa . Nulta hipoteza predstavlja polaznu pretpostavku o vrednosti parametra populacije koju proveravamo pomoću **statističkog testa**. Statistički testovi kombinuju informacije iz nulte hipoteze sa informacijama iz uzorka na način koji nam omogućava donošenje odluke na kraju testa.

Princip testiranja je sličan kao u drugim oblastima istraživanja. Uzmimo za primer brzi COVID test u medicini. Početno ili nulto stanje testa je negativno. Kada test dođe u kontakt sa uzorkom krvi pacijenta, dobijamo informaciju koja nas zanima: ili je početno stanje potvrđeno - pacijent je negativan na COVID, ili je test opovrgnuo početno stanje i pokazao da je pacijent pozitivan. Da bi se opovrgnulo početno stanje, mora postojati dovoljno jak signal u uzorku krvi. Drugim rečima, koncentracija virusa mora preći određeni prag značajnosti (ili ekstremnosti) da bismo zaključili da je pacijent inficiran. U statističkom testiranju, taj signal nazivamo **statistička značajnost**.

Vratimo se na naš primer. Nultu hipotezu možemo zapisati ovako:

Šta je alternativa ovoj pretpostavci? Ako prosečna zarada građana Srbije nije 1000 EUR, onda je logična alternativa da je prosečna zarada različita od 1000 EUR. Ovu pretpostavku nazivamo **alternativnom hipotezom** i označavamo je sa .

Međutim, ovakav oblik alternativne hipoteze nije naročito informativan. Zamislite situaciju gde na konferenciji ili poslovnom sastanku predstavljate rezultat koji kaže samo da prosečna zarada u Srbiji nije 1000 EUR. Nastavićemo sa ovakvom alternativnom hipotezom, ali ćemo kasnije razmotriti i drugi, informativniji pristup.

Osnovna ideja testiranja nulte hipoteze je donošenje odluke - možemo li je odbaciti ili ne. Odbacivanje nulte hipoteze znači da smo u uzorku pronašli dovoljno jak signal koji ukazuje na njenu netačnost. Takav rezultat nazivamo **statistički značajnim**.Jednostavnije rečeno: statistički značajan rezultat je onaj koji nam omogućava odbacivanje nulte hipoteze. Ako signal nije dovoljno jak, nultu hipotezu ne možemo odbaciti.

Šta je zapravo taj signal? To je informacija iz uzorka - konkretno, aritmetička sredina uzorka pretvorena u Z-skor. Z-skor meri udaljenost aritmetičke sredine uzorka od centra distribucije, izraženu kroz broj standardnih grešaka. Veliki Z-skor ukazuje da je aritmetička sredina uzorka ekstremna u odnosu na predviđanja nulte hipoteze.

U našem primeru, prosečna zarada u uzorku iznosi 799.8 EUR, sa Z-skorom od -2.58. Prikažimo ovaj rezultat na standardizovanom normalnom rasporedu.

par(family = "Jost")  
x <- seq(-4, 4, length=1000)  
  
density <- dnorm(x)  
  
plot(x, density, type="l", lwd=3, col="#5C88DAFF",  
 xlab="Z-skor", ylab="Gustina", xaxt="n", xlim=c(-4, 4))  
  
axis(1, at=c(-4, -3, -2, -1, 0, 1, 2, 3, 4), labels=c(-4, -3, -2, -1, 0, 1, 2, 3, 4))  
  
abline(v=-2.58, col="#CC0C00FF", lwd=3)  
text(-3.2, 0.2, "Z = -2.58", col="#CC0C00FF", cex=1, font=2)  
  
abline(v=c(-2, 2), col="black", lty=2, lwd=2)

Line 11

Crtamo crvenu liniju koja označava Z-skor našeg rezultata.

Line 12

Dodajemo oznaku za Z-skor.

Line 14

Dodajemo isprekidane linije koje označavaju granice regiona statističke značajnosti.

|  |
| --- |
| Slika 6.2: Standardizovani normalni raspored sa rezultatom iz uzorka |

Kada posmatramo standardizovani normalni raspored, njegova sredina je **uvek 0** i predstavlja informaciju iz **nulte hipoteze**. Ako je nulta hipoteza tačna, informacija iz uzorka (naš Z-skor) trebalo bi da se nađe blizu sredine distribucije, na udaljenosti do dve standardne greške od centra (isprekidane linije na grafikonu). Na grafikonu vidimo da je naš rezultat ekstreman i nalazi se u oblasti retkih vrednosti.

Ovo možemo numerički potvrditi izračunavanjem p-vrednosti za naš rezultat. P-vrednost predstavlja površinu ispod krive distribucije koja odgovara rezultatima ekstremnijim od našeg.

Šta znači „ekstremniji“ rezultat? Odgovor zavisi od oblika nulte hipoteze postavljene na početku testiranja. U slučaju naše nulte hipoteze (), rezultat se smatra ekstremnim ako je značajno manji ili značajno veći od nule. Zbog toga se **regioni statističke značajnosti** nalaze sa obe strane isprekidanih linija na grafikonu. Oni zajedno čine 5% najekstremnijih rezultata u distribuciji.

Vizualizujmo ovo na grafikonu.

x <- seq(-4, 4, length=1000)  
  
density <- dnorm(x)  
  
plot(x, density, type="l", lwd=3, col="#5C88DAFF",  
 xlab="Z-skor", ylab="Gustina", xaxt="n")  
  
axis(1, at=c(-4, -3, -2, -1, 0, 1, 2, 3, 4), labels=c(-4, -3, -2, -1, 0, 1, 2, 3, 4))  
  
polygon(c(-2, seq(-2, 2, length=100), 2), c(0, dnorm(seq(-2, 2, length=100)), 0), col=rgb(0, 0, 1, 0.2))  
  
polygon(c(-4, seq(-4, -2, length=100), -2), c(0, dnorm(seq(-4, -2, length=100)), 0), col=rgb(1, 0, 0, 0.2))  
  
polygon(c(2, seq(2, 4, length=100), 6), c(0, dnorm(seq(2, 4, length=100)), 0), col=rgb(1, 0, 0, 0.2))

|  |
| --- |
| Slika 6.3: Oblasti ispod krive standardizovanog normalnog rasporeda |

Plava oblast prikazuje rezultate koji su u skladu sa nultom hipotezom - to su očekivane vrednosti. Crvena oblast obuhvata ekstremne rezultate koji značajno odstupaju od nulte hipoteze. Da bismo odredili gde se naš rezultat nalazi u odnosu na ove oblasti, možemo izračunati p-vrednost bez potrebe za grafičkim prikazom.

P-vrednost predstavlja verovatnoću koja se dobija izračunavanjem površine ispod krive distribucije. Ovaj proces zahteva kompleksno numeričko integraljenje koje prepuštamo računaru. U R-u to radimo na sledeći način:

p\_vrednost <- pnorm(-2.58)\*2  
round(p\_vrednost,3)

Line 1

P-vrednost računamo kao verovatnoću da dobijemo rezultat manji ili jednak -2.58, pomnoženu sa 2 jer posmatramo regione značajnosti i sa leve i sa desne strane centra distribucije.

Line 2

Zaokružujemo p-vrednost na 3 decimale.

[1] 0.01

P-vrednost iznosi 0.01, odnosno 1%. Šta nam to govori? Ako je nulta hipoteza tačna, verovatnoća da na slučajnom uzorku ove veličine dobijemo ovakav rezultat (prosečnu zaradu od 800 EUR) je samo 1%. Ovo je izuzetno mala verovatnoća, što ukazuje da je naš rezultat ekstreman u odnosu na predviđanja nulte hipoteze. Po konvenciji, ekstremnim rezultatima smatramo sve one čija je p-vrednost manja od 0.05, odnosno 5% (pravilo dve sigme).

Ovakav rezultat nas navodi da **odbacimo nultu hipotezu** i prihvatimo alternativnu: **prosečna primanja građana Srbije (prosek na nivou populacije) nisu jednaka 1000 EUR**. Odstupanje našeg rezultata od nulte hipoteze je jednostavno preveliko da bismo je zadržali.

|  |  |  |
| --- | --- | --- |
|  | **Kako smo došli do ove formule za p-vrednost?**  Da bismo razumeli izraz pnorm(-2.58)\*2, rastavićemo ga na osnovne komponente.  Najlakše ćemo shvatiti šta se događa u ovoj formuli ako analiziramo još jedan grafički prikaz normalne distribucije.   |  | | --- | | Slika 6.4: P-vrednost kao površina ispod krive |   Crvena oblast pokazuje površinu ispod krive distribucije za vrednosti manje ili jednake -2.58. To je verovatnoća dobijanja rezultata manjeg ili jednakog -2.58. Plava oblast predstavlja 1 minus ta verovatnoća, odnosno površinu ispod krive za vrednosti veće od -2.58. Mi želimo da izračunamo verovatnoću dobijanja ekstremnog rezultata - bilo manjeg od -2.58 ili većeg od 2.58. Zato množimo površinu sa 2, da bismo obuhvatili ekstremne rezultate na obe strane centra distribucije.  Funkcija pnorm() u R-u izračunava površinu ispod krive distribucije od minus beskonačno do zadate tačke. |

|  |  |
| --- | --- |
|  | **Koraci testiranja hipoteza**  U ovom primeru smo uradili Z-test za proveru hipoteze o aritmetičkoj sredini. Iako se ovaj test retko koristi u praksi jer zahteva poznavanje standardne devijacije populacije (što je uglavnom nepoznato), on predstavlja odličnu osnovu za razumevanje testiranja hipoteza i koncepta statističke značajnosti.  Proces testiranja hipoteza možemo svesti na pet ključnih koraka:   1. Postavljanje nulte hipoteze i alternativne hipoteze 2. Izračunavanje aritmetičke sredine i standardne greške uzorka 3. Standardizacija aritmetičke sredine u Z-skor (statistika testa) 4. Računanje p-vrednosti 5. Donošenje zaključka: ako je p-vrednost < 0.05, odbacujemo nultu hipotezu; u suprotnom, nemamo dovoljno dokaza za njeno odbacivanje   Jednostavno ali moćno - ovih pet koraka čine srž procesa koji je revolucionirao način na koji donosimo zaključke na osnovu podataka. |

|  |  |
| --- | --- |
|  | **Nivoi značajnosti**  „Standardni“ nivo značajnosti u statističkom zaključivanju iznosi 0.05. Ova vrednost predstavlja granicu za ekstremne rezultate, koja se nalazi na udaljenosti od 2 standardne greške od centra distribucije.  U praksi često koristimo i druge nivoe značajnosti, kao što su 0.01 ili 0.10. Izbor nivoa značajnosti zavisi od konteksta istraživanja i stepena sigurnosti koji želimo da postignemo u zaključivanju. Veći nivo značajnosti (recimo 0.10) povećava verovatnoću odbacivanja nulte hipoteze. Istovremeno, to povećava i rizik da odbacimo nultu hipotezu kada je ona zapravo tačna. O ovoj vrsti rizika detaljnije ćemo govoriti u odeljku o greškama tipa I i tipa II. |

### 6.2.1 Jednostrane i dvostrane hipoteze

Primer koji smo obradili predstavlja dvosmernu nultu i alternativnu hipotezu. Zašto dvosmernu? Zato što se nulta hipoteza može odbaciti sa obe strane, a region statističke značajnosti je podeljen na dve oblasti u levom i desnom repu distribucije. Kao što smo ranije naveli, ovakav oblik hipoteze nije posebno informativan kada je u pitanju aritmetička sredina, jer je zaključak „prosek nije jednak 1000 EUR“ previše uopšten.

Razmotrimo sada drugačiji, precizniji oblik nulte i alternativne hipoteze:

$$
H\_0: \mu \geq 1000 \\
H\_1: \mu < 1000
$$

Ovaj oblik hipoteza je **jednostran** jer nultu hipotezu možemo odbaciti samo sa jedne strane. Region statističke značajnosti nalazi se samo levo od centra distribucije. Pre nego što prikažemo grafički prikaz, razmotrimo logiku ovakvog pristupa.

Nulta hipoteza mora sadržati relaciju jednakosti kako bismo mogli centrirati normalnu distribuciju oko određene vrednosti - u ovom slučaju 1000 EUR. Dakle, nulta hipoteza pretpostavlja da je prosek populacije 1000 ili više. Alternativna hipoteza, suprotno tome, pretpostavlja da je prosek manji od 1000 EUR.

Ovakve hipoteze nazivamo jednostranim jer nultu hipotezu možemo odbaciti samo sa jedne strane distribucije. Hajde da to vizuelno predstavimo.

x <- seq(-4, 4, length=1000)  
  
density <- dnorm(x)  
  
plot(x, density, type="l", lwd=3, col="blue",  
 xlab="Z-skor", ylab="Gustina", xaxt="n")  
  
axis(1, at=c(-4, -3, -2, -1, 0, 1, 2, 3, 4), labels=c(-4, -3, -2, -1, 0, 1, 2, 3, 4))  
  
polygon(c(-4, seq(-4, -2, length=100), -2), c(0, dnorm(seq(-4, -2, length=100)), 0), col=rgb(1, 0, 0, 0.2))  
  
polygon(c(-2, seq(-2, 10, length=100), 10), c(0, dnorm(seq(-2, 10, length=100)), 0), col=rgb(0, 0, 1, 0.2))

|  |
| --- |
| Slika 6.5: Oblasti ispod krive standardizovanog normalnog rasporeda za jednostrane hipoteze |

Plava oblast označava prihvatanje nulte hipoteze, dok crvena oblast predstavlja njeno odbacivanje. Desni rep sada pripada oblasti prihvatanja nulte hipoteze jer visoka aritmetička sredina uzorka (značajno veća od 1000 EUR) potvrđuje nultu hipotezu koja tvrdi da je prosek 1000 EUR ili više.

Naš rezultat iz uzorka ostaje -2.58, ali p-vrednost se menja jer računamo površinu ispod krive distribucije samo za vrednosti manje od -2.58. Razlog je jednostavan - region značajnosti se nalazi samo sa leve strane centra distribucije, pa ne množimo vrednost sa 2.

p\_vrednost\_jednostrana <- pnorm(-2.58)  
round(p\_vrednost\_jednostrana,3)

Line 1

P-vrednost za jednostrane hipoteze računamo kao verovatnoću da dobijemo rezultat manji ili jednak -2.58.

Line 2

Zaokružujemo p-vrednost na 3 decimale.

[1] 0.005

P-vrednost je, kao što smo očekivali, dva puta manja. Odbacivanje nulte hipoteze je sada jednostavnije jer posmatramo samo jedan deo distribucije. P-vrednost iznosi 0.005, odnosno 0.5%. Ovaj rezultat je i dalje ekstreman, što nas vodi do odbacivanja nulte hipoteze.

Ono što je posebno važno jeste da je sada zaključak mnogo informativniji: „prosečna zarada u Srbiji je ispod 1000 EUR“. Ovakav zaključak je toliko jasan i direktan da ga lako možete zamisliti kao naslov na nekom internet portalu.

|  |  |
| --- | --- |
|  | **Lična karta metoda: Z-test za aritmetičku sredinu**  **Šta radi?** Ispituje da li se aritmetička sredina uzorka značajno razlikuje od pretpostavljene aritmetičke sredine populacije (nulte hipoteze).  **Kada se koristi?** U situacijama kada poznajemo standardnu devijaciju populacije i želimo da proverimo da li se aritmetička sredina uzorka statistički značajno razlikuje od početne pretpostavke.  **Koliko varijabli analiziramo?** Jednu kvantitativnu varijablu (). Ona može biti merena intervalnom ili racio skalom. U određenim slučajevima, prihvatljiva je i ordinalna varijabla.  **Kako formulišemo hipoteze?** Ako označimo pretpostavljenu vrednost aritmetičke sredine populacije sa , hipoteze možemo formulisati kao dvosmerne ili jednosmerne.  Za dvosmerni test:  $$ H\_0: \mu = \mu\_0 \\ H\_1: \mu \neq \mu\_0 $$  Za jednosmerni test:  $$ H\_0: \mu \geq \mu\_0 \\ H\_1: \mu < \mu\_0 $$  (ili obrnuto, zavisno od konteksta istraživanja)  **Kako izgleda statistika testa?** Statistika testa je Z-skor koji se računa kao:  gde je aritmetička sredina uzorka, pretpostavljena vrednost aritmetičke sredine populacije, standardna devijacija populacije i veličina uzorka.  **Kako računamo p-vrednost?** Za dvostrani test, p-vrednost je jednaka dvostrukoj površini ispod krive distribucije za vrednosti manje ili veće od Z-skora, u zavisnosti od toga da li je on negativan ili pozitivan.    Za jednostrani test, p-vrednost predstavlja površinu ispod krive distribucije za vrednosti manje od Z-skora (kada je Z negativan) ili veće od Z-skora (kada je Z pozitivan).    **Kako donosimo zaključak?** Kada je p-vrednost manja od nivoa značajnosti (najčešće 0.05), odbacujemo nultu hipotezu. U suprotnom slučaju, nemamo dovoljno dokaza da odbacimo nultu hipotezu. |

## 6.3 Greške pri testiranju hipoteza

Kao i kod svakog procesa donošenja odluka, suočavamo se sa rizikom pogrešnog izbora. Isto važi i za statističko testiranje. Odbacivanje nulte hipoteze nosi rizik greške, baš kao što i njeno prihvatanje može biti pogrešno. Jedinstvena prednost statističkog zaključivanja leži u tome što možemo precizno kvantifikovati verovatnoće ovih grešaka, čineći naše zaključke jasnijim i pouzdanijim.

Za ilustraciju ovih grešaka, počnimo sa jednostavnim primerom van statistike — testovima za trudnoću.

Ovi testovi detektuju hormon hCG u urinu. U normalnim okolnostima, ovaj hormon nije prisutan kod muškaraca i žena koje nisu trudne. Ipak, u retkim slučajevima, kod obe grupe mogu se javiti minimalne koncentracije ovog hormona. To znači da je očekivana vrednost (prosek) 0, ali prirodne varijacije mogu proizvesti veoma niske, ne-nulte vrednosti — ovo je izvor varijabiliteta. Test daje pozitivan rezultat samo kada koncentracija hormona pređe određeni prag. Koncentracije ispod tog praga rezultiraju negativnim testom. Ovaj mehanizam direktno odgovara konceptu statističke značajnosti.

Razmotrimo test na dva subjekta:

1. trudnu ženu
2. muškarca kao kontrolni subjekt

Za oba subjekta moguća su dva ishoda testa: pozitivan i negativan. Pogledajmo rezultate u sledećoj tabeli.

| Osoba | Test | Zaključak | Interpretacija |
| --- | --- | --- | --- |
| Trudnica | Pozitivan | Tačno | Ispravan rezultat |
| Muškarac | Pozitivan | **Greška** | Lažno pozitivan test |
| Trudnica | Negativan | **Greška** | Lažno negativan test |
| Muškarac | Negativan | Tačno | Ispravan rezultat |

U dve situacije doneli smo pogrešan zaključak. U prvom slučaju, test je pokazao pozitivan rezultat kod muškarca, iako trudnoća nije moguća. Ovo je **lažno pozitivan test**. U drugom slučaju, test je pokazao negativan rezultat kod trudnice, iako je trudna. Ovo je **lažno negativan test**.

Kod lažno pozitivnog testa, prag za pozitivan rezultat je često postavljen prenisko. To znači da test detektuje čak i minimalne koncentracije hormona, koje se prirodno javljaju kod muškaraca ili žena koje nisu trudne. Podizanjem praga za pozitivan rezultat smanjuje se verovatnoća lažno pozitivnih rezultata i povećava specifičnost testa.

Specifičnost testa predstavlja njegovu sposobnost da pravilno identifikuje osobe koje nemaju stanje koje se testira - u ovom slučaju, da razlikuje ne-trudnoću od trudnoće. Što je test specifičniji, to bolje izbegava lažno pozitivne rezultate.

U drugom slučaju, kod lažno negativnog testa, granica za pozitivan rezultat je postavljena previsoko. To znači da test možda neće detektovati ni prisustvo hormona kod trudnice ako je njegova koncentracija ispod te visoke granice. Snižavanjem praga za pozitivan rezultat smanjuje se verovatnoća dobijanja lažno negativnih rezultata i povećava se osetljivost testa.

Osetljivost testa je njegova sposobnost da pravilno identifikuje osobe koje imaju stanje koje se testira - u ovom slučaju, da potvrdi trudnoću kod trudnica. Što je test osetljiviji, to je bolji u otkrivanju stvarnih pozitivnih slučajeva i smanjivanju lažno negativnih rezultata.

Međutim, **nema besplatnog ručka** - povećanje osetljivosti često dolazi na račun specifičnosti. To znači da postoji neizbežan kompromis između osetljivosti i specifičnosti testa. Granicu za pozitivan rezultat treba pažljivo podesiti kako bi se postigao optimalan balans koji će svesti na minimum i lažno pozitivne i lažno negativne rezultate. Ne postoji savršen test koji će biti 100% osetljiv i 100% specifičan.

Vratimo se sada na statističko zaključivanje i razmotrimo greške koje možemo napraviti pri testiranju hipoteza.

Nulta hipoteza predstavlja negativan rezultat testa, slično situaciji kada osoba nije trudna i kada je koncentracija hormona jednaka nuli. Alternativna hipoteza odgovara pozitivnom rezultatu testa.

Lažno pozitivan test („muškarac je trudan“) u statističkom zaključivanju naziva se **greška tipa I**. To je situacija kada odbacimo nultu hipotezu iako je ona tačna. Jednostavnije rečeno, greška tipa I je pogrešno odbacivanje nulte hipoteze - vidimo signal tamo gde ga zapravo nema. Verovatnoću pravljenja greške tipa I označavamo sa i nazivamo je **nivo značajnosti** testa. Ovu vrednost smo već spominjali - ona je standardno postavljena na 0.05 i predstavlja prag za donošenje odluke o odbacivanju nulte hipoteze.

Lažno negativan test („trudnica nije trudna“) u statističkom zaključivanju nazivamo **greškom tipa II**. To je situacija kada prihvatimo nultu hipotezu iako je ona netačna. Jednostavnije rečeno, greška tipa II nastaje kada ne vidimo signal koji zaista postoji. Verovatnoću pravljenja greške tipa II označavamo sa . Za razliku od , beta nije fiksna vrednost - ona zavisi od nekoliko faktora, uključujući veličinu uzorka, jačinu efekta koji pokušavamo da detektujemo i nivo značajnosti testa. O vrednosti i njenom značaju detaljnije ćemo govoriti u narednom poglavlju, kada budemo analizirali **snagu testa**.

## 6.4 Intevali poverenja

Pređimo sada na drugi pristup statističkom zaključivanju. Testiranje hipoteza možemo posmatrati kao **binarni** proces: nultu hipotezu ili odbacujemo ili ne odbacujemo. Analiziramo da li se informacija iz uzorka nalazi u regionu značajnosti, procenjujući verovatnoću retkih događaja unutar distribucije.

Rezultat testiranja može biti informativan („Prosečna zarada u Srbiji je ispod 1000 EUR“), ali ponekad je i previše uopšten („Prosečna zarada u Srbiji nije jednaka 1000 EUR“). Zamislite da takav zaključak prezentujete na sastanku. Koje bi bilo prvo pitanje koje biste čuli? Verovatno: „U redu, ako je manja od 1000, koliko tačno iznosi?“.

Tačan i sasvim precizan odgovor na ovo pitanje ne možemo dati jer radimo sa podacima iz uzorka. Međutim, možemo napraviti procenu proseka na nivou populacije na osnovu podataka iz uzorka. Ova procena se naziva **interval poverenja**. Ilustrujmo ovo primerom: „Sa 95% pouzdanosti možemo reći da se prosečna zarada u Srbiji kreće između 750 i 850 EUR“. Ovo je primer rezultata statističkog zaključivanja koji nije zasnovan na testiranju hipoteza, već pripada oblasti **statističkog ocenjivanja**. Hajde da vidimo kako možemo konstruisati ovakav interval.

### 6.4.1 Konstrukcija intervala poverenja

Vratimo se na naš primer. Na uzorku od 60 ispitanika istraživali smo zarade građana Srbije. Nakon obrade podataka dobili smo prosečnu zaradu od 799.8 EUR, sa standardnom devijacijom 600 EUR. Pošto je uzorak dovoljno veliki, možemo pretpostaviti da je standardna devijacija uzorka približno jednaka standardnoj devijaciji populacije.

Šta nam ovi podaci omogućavaju da izračunamo? Prvo ćemo izračunati standardnu grešku aritmetičke sredine koristeći formulu:

U R-u to izgleda ovako:

s <- 600  
n <- 60  
  
s\_x <- s/sqrt(n)  
round(s\_x, 2)

Lines 1-2

Unosimo vrednosti standardne devijacije i veličine uzorka.

Line 4

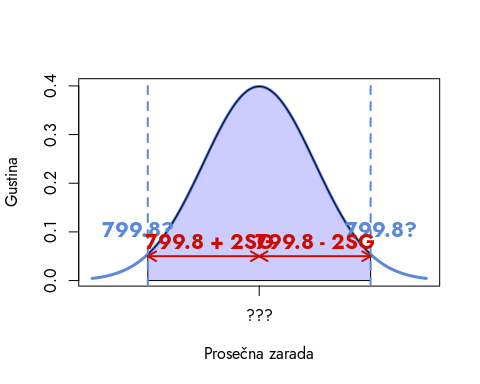
Računamo standardnu grešku aritmetičke sredine prema formuli.

[1] 77.46

Šta možemo da uradimo sa standardnom greškom? Vratimo se na analizu normalne distribucije aritmetičkih sredina. Kod testiranja hipoteza, u centar te distribucije postavili smo vrednost iz nulte hipoteze (1000 EUR) i posmatrali verovatnoću da se naša vrednost iz uzorka nalazi u regionu značajnosti.

Međutim, sada ne znamo koja je aritmetička sredina populacije. Zato ćemo analizirati granične slučajeve.

Ako pretpostavimo da se vrednost iz našeg uzorka nalazi unutar 95% vrednosti koje okružuju centar distribucije, koji su granični slučajevi? To su vrednosti udaljene približno 2 standardne greške od centra distribucije. Ove vrednosti nazivamo **kritičnim vrednostima**. Hajde da to vizuelno prikažemo.



Kritične vrednosti za Z distribuciju

Plave isprekidane linije označavaju granične slučajeve za aritmetičku sredinu. Pretpostavljamo da ona može odstupati najviše dve standardne greške od izmerene vrednosti. Sa ovom pretpostavkom možemo izračunati interval poverenja. Crvene linije nam pokazuju moguće pozicije aritmetičke sredine populacije.

Konstruisanjem intervala oko ovih vrednosti možemo „uhvatiti“ pravu aritmetičku sredinu populacije. Iako ne znamo tačnu vrednost, ona će biti obuhvaćena ovim intervalom ako je naša početna pretpostavka ispravna.

Interval poverenja se računa prema formuli:

ili

gde su:

* - aritmetička sredina uzorka
* - standardna greška aritmetičke sredine
* - kritična vrednost za normalnu distribuciju (u našem slučaju 2)
* - aritmetička sredina populacije

U R-u to izgleda ovako:

Z <- 2  
  
donja\_granica <- 799.8 - Z \* s\_x  
  
gornja\_granica <- 799.8 + Z \* s\_x  
  
interval\_poverenja <- c(donja\_granica, gornja\_granica)  
  
cat("Intervala poverenja je od ", interval\_poverenja[1], " do ", interval\_poverenja[2], " EUR.")

Line 1

Postavljamo kritičnu vrednost na 2.

Lines 3,5

Računamo donju i gornju granicu intervala poverenja.

Line 7

Spajamo ove dve granice u vektor koji predstavlja interval poverenja.

Line 9

Ispisujemo interval poverenja.

Intervala poverenja je od 644.8807 do 954.7193 EUR.

Kritična vrednost od 2 označava da smo konstruisali interval poverenja od približno 95%. Na osnovu toga možemo zaključiti da se, sa 95% pouzdanosti, prosečna zarada u Srbiji nalazi između 644.88 i 954.72 EUR.

Ovaj interval nam pruža realnu sliku o mogućem opsegu prosečne zarade u populaciji. Primetićete da interval nije naročito precizan, što je direktna posledica relativno visoke standardne greške koja proizilazi iz male veličine uzorka. Veći uzorak bi doveo do manje standardne greške, što bi rezultiralo užim i preciznijim intervalom poverenja.

|  |  |
| --- | --- |
|  | **Pogrešna interpretacija intervala poverenja**  Kada govorimo o intervalima poverenja, u udžbenicima statistike često se sreću određene zablude. Jedna od njih je poređenje sa testiranjem hipoteza, gde se tvrdi da intervali poverenja, za razliku od testiranja hipoteza, ne polaze od početne pretpostavke.  Ovo nije tačno. Konstrukcija intervala poverenja zapravo se zasniva na ključnoj pretpostavci - da se naš uzorak nalazi unutar centralnih 95% vrednosti distribucije (ili 99% ili 90%, zavisno od izabranog nivoa značajnosti). Ova pretpostavka je implicitna i ne moramo je formalno navoditi, ali je fundamentalna za ceo proces. Glavna razlika u odnosu na testiranje hipoteza je u tome što se ova pretpostavka ne proverava kroz sam statistički postupak.  Druga česta greška tiče se samog koncepta pouzdanosti. Često ćete čuti tvrdnje poput: „Verovatnoća da se aritmetička sredina populacije nalazi u ovom intervalu je 95%, a postoji verovatnoća od 5% da se ona nalazi izvan njega.“ Ovo nije tačno.  Vrednost koju smo dobili iz uzorka i na osnovu koje smo konstruisali interval poverenja predstavlja samo jednu od mogućih vrednosti iz populacije. Mi smo pretpostavili da se naš uzorak nalazi unutar centralnih 95% distribucije. Međutim, ne možemo obrnuti ovu logiku i tvrditi da se centar distribucije nalazi u intervalu oko naše vrednosti sa određenom verovatnoćom.  Evo ispravne interpretacije pouzdanosti: „Kada bismo ponovili uzorkovanje veliki broj puta i svaki put izračunali interval poverenja na isti način, u približno 95% slučajeva prava vrednost populacione sredine našla bi se unutar tog intervala.“  Ova interpretacija je složenija za razumevanje i često je ne naglašavamo jer je podrazumevana u samoj konstrukciji intervala poverenja. Ipak, važno je razumeti da je ovo precizna interpretacija pouzdanosti intervala poverenja.  U praksi je dovoljno reći da, na osnovu intervala poverenja od 95% pouzdanosti, prosečna zarada u Srbiji iznosi između 644.88 i 954.72 EUR. Interpretacija postaje kompleksnija kada shvatimo da je naš interval, nakon prikupljanja podataka i izračunavanja na osnovu jednog uzorka, zapravo fiksan - on je ili uspeo da obuhvati pravu vrednost populacije ili nije. Ne možemo proceniti verovatnoću tog uspeha na osnovu samo jednog uzorka. |

Testiranje nulte hipoteze i konstrukcija intervala poverenja predstavljaju snažne alate za donošenje zaključaka o nepoznatim parametrima populacije. Izbor metoda zavisi od ciljeva istraživanja, a često se oba pristupa kombinuju kako bi zaključci bili precizniji i informativniji.

U svakodnevnom životu ne donosimo odluke na ovaj način, zbog čega nam statističko zaključivanje može delovati apstraktno i neprirodno. Ipak, ovakav pristup nam omogućava da jasno sagledamo granice naših zaključaka - granice koje proizilaze iz nepotpunih informacija - kao i rizik od pogrešnih odluka. To je fundamentalna razlika između statističkog i intuitivnog zaključivanja, i glavni razlog zašto je statistika postala nezaobilazan alat u savremenoj nauci i šire.

Naučna istraživanja često imaju za cilj identifikaciju specifičnih efekata ili uticaja: efikasnost neke intervencije, delotvornost leka, razlike između društvenih grupa ili povezanost između simptoma. U narednim poglavljima pokazaćemo kako se ovi ciljevi mogu ostvariti primenom statističkih testova i metoda ocenjivanja.

## 6.5 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1**  Novosadski vodovod saopštio je da je u uzorku od 150 merenja pijaće vode pronađena povišena koncentracija nitrata iz otpadnih voda. Koncentracija u uzorku iznosi 55 mg/L, dok je standardna devijacija 10 mg/L.  Maksimalna dozvoljena koncentracija nitrata u pijaćoj vodi je 50 mg/L. Vodovod tvrdi da nema razloga za zabrinutost i da je ovaj rezultat samo izolovani slučaj, te da je koncentracija nitrata u pijaćoj vodi u Novom Sadu i dalje na granici ili ispod nje.  Uzmimo ovu tvrdnju vodovoda kao nultu hipotezu i testiraćemo je na nivou značajnosti od 0.05.   * Izračunajte Z-skor * Izračunajte p-vrednost * Testirajte nultu hipotezu * Pod pretpostavkom da je vodovod u pravu, kolika je verovatnoća da se ovakav uzorak pojavi slučajno? |

|  |  |
| --- | --- |
|  | **Zadatak 2**  Na uzorku od 750 domaćinstava u Srbiji utvrđeno je da je prosečan broj dece u domaćinstvu 1.2. Standardna devijacija populacije iznosi 0.5.  Proverite hipotezu da prosečno domaćinstvo u Srbiji ima najmanje jedno dete. Koristićemo nivo značajnosti 0.05.  Da biste odgovorili na ovo pitanje, potrebno je da:   * Izračunate Z-skor * Izračunate p-vrednost * Vizuelno prikažete rezultat na grafikonu normalne distribucije koristeći kod iz poglavlja |

|  |  |
| --- | --- |
|  | **Zadatak 3 \***  Vratimo se na primer sa prosečnim zaradama u Srbiji i interval poverenja od 95% pouzdanosti. Dobili smo interval od 644.88 do 954.72 EUR.  Na vašoj prezentaciji rezultata, šefica/mentorka nije zadovoljna širinom ovog intervala. Potrebna joj je preciznija ocena prosečne zarade, sa maksimalnim odstupanjem od ±50 EUR u oba smera.  Odredite minimalnu veličinu uzorka potrebnu za postizanje ovog cilja. Koji interval poverenja očekujete da dobijete sa novim uzorkom?  *Napomena*: Budući da nije određen nivo pouzdanosti intervala, možete prilagoditi i taj parametar tokom analize.  *Dodatni izazov*: Umesto korišćenja postojeće aritmetičke sredine, simulirajte novu vrednost iz normalne distribucije aritmetičkih sredina pomoću funkcije rnorm(). |

|  |  |
| --- | --- |
|  | **Zadatak 4 \***  Administrator jedne Instagram stranice želi da bolje razume svoju ciljnu grupu i ispita starost pratilaca. Na anketu o godinama odgovorilo je 132 pratioca. Prosečna starost je 28.2 godine, sa standardnom devijacijom (populacije) od 15.3 godine.   1. Konstruišite interval poverenja od 95% za prosečnu starost pratilaca stranice. 2. Konstruišite interval poverenja od 90% za prosečnu starost pratilaca stranice (Z vrednost za 90% je 1.64). 3. Administrator stranice planira da plati reklamu na Instagramu usmerenu ka određenoj starosnoj grupi. Dostupne opcije za ciljanje su: 18-24, 25-34, 35-44 godina. Na osnovu dobijenih intervala poverenja, koju starosnu grupu biste preporučili? 4. Analizirajte granice oba intervala. Kako biste opisali odnos između preciznosti i pouzdanosti intervala poverenja? |

|  |  |
| --- | --- |
|  | **Zadatak 5 \*\***  U našim ranijim razmatranjima naveli smo da oblast od +/- 2 standardne greške oko aritmetičke sredine obuhvata približno 95% vrednosti u distribuciji aritmetičkih sredina. Ova aproksimacija, iako korisna, nije potpuno precizna.  Hajde da pronađemo tačne vrednosti:   1. Izračunajte Z vrednost koja tačno definiše 95% vrednosti u distribuciji aritmetičkih sredina. Za ovo možete koristiti R funkciju qnorm(). 2. Odredite Z vrednost koja definiše oblast koja obuhvata 99% vrednosti u distribuciji aritmetičkih sredina. 3. Konstruišite 99% intervale poverenja za aritmetičke sredine iz zadataka 2 i 4. |

# 7. t-test

## 7.1 Veličina uzorka

Analiza prethodnog poglavlja otkriva da smo malo pažnje posvetili veličini uzorka. Ona se pojavila samo u formuli za standardnu grešku, gde smo videli da se greška smanjuje s povećanjem uzorka.

Z-test iz prethodnog poglavlja propustio je da uvaži ključnu činjenicu: neizvesnost u donošenju zaključaka se smanjuje s rastom uzorka. Statističko zaključivanje je u suštini borba s neizvesnošću koja proizilazi iz ograničenih informacija o populaciji dobijenih iz uzorka. Jednostavno rečeno: veći uzorak znači manju neizvesnost zaključaka, dok mali uzorak donosi veliku neizvesnost.

Kada govorimo o neizvesnosti, mislimo na verovatnoću grešaka tipa I ili tipa II. Kod malog uzorka, verovatnoća greške raste. Kod velikog uzorka, ona opada. Ovo je fundamentalna istina statističkog zaključivanja: veličina uzorka direktno određuje pouzdanost naših zaključaka.

|  |  |  |
| --- | --- | --- |
|  | **Veliki i mali uzorci**  U statistici postoji važan kriterijum za razlikovanje malih i velikih uzoraka. Uzorci s više od 30 jedinica smatraju se velikim. Jednostavnije rečeno - u društvenim istraživanjima, dovoljno je 30 ispitanika za „veliki“ uzorak. Sve manje od toga smatra se malim uzorkom.  Broj 30 može delovati mali. Često čujemo o istraživanjima sa 300, 600 ili čak 1000 ispitanika. Zašto onda baš 30? Razlog je matematički elegantan.  Ova granica nije proizvoljna, već direktno proizlazi iz ponašanja standardne greške aritmetičke sredine. Hajde da istražimo zašto je broj 30 poseban.  Zamislite praktičan primer - merimo prosečna primanja ljudi. Pretpostavimo da standardna devijacija primanja u populaciji iznosi 100 evra. Koristeći formulu za standardnu grešku ([Formula 5.1](#eq-sx)), možemo ispitati šta se dešava kada menjamo veličinu uzorka od 2 do 100 ispitanika.  n <- seq(2, 100, by = 1) se <- 100 / sqrt(n)  round(head(se),2)  Line 1  Pravimo sekvencu brojeva od 2 do 100.  Line 2  Računamo standardnu grešku za svaku veličinu uzorka.  Line 4  Prikazujemo prvih nekoliko vrednosti standardne greške. Vidite da standardna greška opada kako uzorak raste. Za 2 ispitanika, ona je oko 70 EUR, ali brzo pada na 40 EUR za 5 ispitanika. Dramatično smanjenje, ali šta se dešava dalje? Istražimo.  [1] 70.71 57.74 50.00 44.72 40.82 37.80  par(family = "Jost")  plot(n, se, type = "l", col = "#5C88DAFF", lwd = 3, xlab = "Veličina uzorka", ylab = "Standardna greška") abline(v = 30, col = "#CC0C00FF", lty = 2, lwd=3)  Line 3  Crtamo grafikon koji pokazuje kako se standardna greška menja s veličinom uzorka.  Line 4  Dodajemo crvenu liniju na 30 ispitanika.   |  | | --- | | Slika 7.1: Standardna greška za različite veličine uzorka |   Grafik otkriva jasan obrazac. U opsegu od 2 do 30 ispitanika, kriva standardne greške pokazuje oštar pad. Nakon 30 ispitanika, kriva postaje gotovo horizontalna. Ovo je matematički razlog zašto broj 30 predstavlja granicu između malih i velikih uzoraka - nakon ove tačke, standardna greška opada toliko postepeno da dodatni ispitanici donose minimalno poboljšanje preciznosti. Zbog toga uzorke veće od 30 jedinica klasifikujemo kao velike.  Da bismo preciznije razumeli ovu dinamiku, uporedimo standardnu grešku za 10, 30 i 90 ispitanika:  cat("Standardna greška za 10 ispitanika:", se[9], "\n")  cat("Standardna greška za 30 ispitanika:", se[29], "\n")  cat("Standardna greška za 90 ispitanika:", se[89], "\n")  Standardna greška za 10 ispitanika: 31.62278  Standardna greška za 30 ispitanika: 18.25742  Standardna greška za 90 ispitanika: 10.54093  Pogledajmo kako se greška menja s povećanjem uzorka. Kada povećamo uzorak s 10 na 30 ispitanika (samo 20 više), greška se smanjuje za oko 13 EUR. Značajan napredak. Ali šta se dešava kad nastavimo? Kada dodamo još tri puta više ispitanika (s 30 na 90), greška opada za samo 8 EUR. Iz ovoga izvlačimo važan zaključak: nakon određene tačke, dodavanje novih ispitanika donosi sve manje poboljšanja. To je glavni razlog zašto uzorke preko 30 jedinica smatramo „velikim“ - potrebno je mnogo dodatnih ispitanika za relativno malo smanjenje greške. Kod malih uzoraka situacija je drugačija - svaki novi ispitanik značajno doprinosi smanjenju greške. |

Z-test koji smo ranije koristili nije obuhvatio ovu dinamiku. On primenjuje standardizovani normalni raspored sa fiksnim parametrima - aritmetičkom sredinom 0 i standardnom devijacijom 1. Ne postoji direktna veza između veličine uzorka i neizvesnosti zaključaka, osim kroz Z-statistiku koja zavisi od veličine uzorka. Drugim rečima, rezultat testa se menja sa veličinom uzorka, ali distribucija za zaključivanje ostaje nepromenjena. Ovo je fundamentalna razlika: Z-test ne prilagođava parametre veličini uzorka, što ga čini nepreciznim za male uzorke.

Kao odgovor na ovaj problem, razvijena je modifikacija normalne distribucije koja uzima u obzir veličinu uzorka. Reč je o Studentovoj t-distribuciji koja predstavlja osnovu za t-test. Ova distribucija nam omogućava da preciznije modelujemo neizvesnost kada radimo s malim uzorcima.

## 7.2 Studentova t-distribucija

T-distribucija je adaptivna verzija normalne distribucije koja se prilagođava veličini uzorka. S porastom veličine uzorka, t-distribucija konvergira ka normalnoj distribuciji. Kada radimo s velikim uzorcima, ona postaje gotovo identična normalnoj - simetrična je i ima tanke repove. Međutim, kod malih uzoraka t-distribucija pokazuje svoje prave karakteristike: postaje šira i ima teže repove, što precizno odražava veću neizvesnost koja proizilazi iz ograničenog broja podataka.

|  |  |
| --- | --- |
|  | **Student i Ginis pivo**  Ime „Studentova t-distribucija“ ima zanimljivu pozadinu. Sve je počelo sa Vilijamom Sejmourom Gosetom, koji je 1908. godine objavio značajan rad pod pseudonimom „Student“ (Student, 1908).  Goset je radio kao hemičar u pivari „Guinness“, gde se bavio unapređenjem kvaliteta piva. Kompanija je u to vreme primenjivala naučne metode za optimizaciju proizvodnje. Gosetov zadatak je bio analiza hemijskih svojstava ječma, ključnog sastojka u proizvodnji. Zbog ekonomskih ograničenja, radio je sa izuzetno malim uzorcima - testirao je samo 3-4 džaka ječma koji su predstavljali celokupnu proizvodnju.  Kroz rad je uočio fundamentalni problem - Z-test nije davao pouzdane rezultate na ovako malim uzorcima. Test jednostavno nije mogao da obuhvati velike varijacije koje su prirodne kada radite sa malim brojem opservacija. Motivisan ovim izazovom, Goset je razvio novu distribuciju koja je eksplicitno uzimala u obzir veličinu uzorka. Rezultat njegovog rada je t-distribucija, koja je postala osnova za t-test.  Zašto je izabrao pseudonim „Student“? Najverovatnije zato što Guinness nije dozvoljavao svojim zaposlenima da objavljuju naučne radove pod pravim imenom. Možda su se plašili da će konkurencija iskoristiti njihove inovacije. Ili su želeli da prikriju činjenicu da rade sa malim uzorcima. Bez obzira na pravi razlog, pseudonim je postao neraskidivo povezan sa jednom od najvažnijih distribucija u statistici. |

Da bismo bolje razumeli ovo ponašanje, hajde da ga vizualizujemo.

par(family = "Jost")  
  
x <- seq(-4, 4, length = 1000)  
  
y1 <- dnorm(x, mean = 0, sd = 1)  
  
y2 <- dt(x, df = 4)  
  
y3 <- dt(x, df = 11)  
  
y4 <- dt(x, df = 31)  
  
plot(x, y1, type = "l", col = "black", lwd = 3, xlab = "t", ylab = "Gustina")  
lines(x, y2, col = "#5C88DAFF", lwd = 3)  
lines(x, y3, col = "#CC0C00FF", lwd = 3)  
lines(x, y4, col = "#00AF66FF", lwd = 3)  
legend("topright", legend = c("Normalna", "t (n=5)", "t (n=10)", "t (n=30)"),  
 col = c("black", "#5C88DAFF", "#CC0C00FF", "#00AF66FF"), lwd = 2, cex = 1.1)

Line 3

Definišemo opseg x vrednosti.

Line 5

Definišemo gustinu normalne distribucije funkcijom dnorm().

Lines 7,9,11

Definišemo gustinu t-distribucije funkcijom dt(). Argument df je broj stepeni slobode.

Line 13

Crtamo gustinu normalne distribucije.

Lines 14-16

Crtamo gustinu t-distribucije za tri različita broja stepeni slobode.

Line 18

Dodajemo legendu.

|  |
| --- |
| Slika 7.2: Studentova t-distribucija za različite veličine uzorka |

Na grafikonu vidimo ponašanje t-distribucije u poređenju s normalnom distribucijom. Kada imamo mali broj stepeni slobode, t-distribucija pokazuje šire „repove“ i veću varijabilnost - to je ilustrovano plavom linijom. Ona je niža od ostalih, što ukazuje na manju koncentraciju verovatnoće oko centra, uz „deblje“ repove koji odražavaju veću neizvesnost u oceni i povećanu verovatnoću ekstremnih vrednosti. T-distribucija zadržava simetriju oko nule, baš kao i standardizovana normalna distribucija.

Kako raste broj stepeni slobode, t-distribucija konvergira ka normalnoj. Zelena i crna linija se praktično preklapaju, pokazujući da već pri t-distribucija postaje gotovo identična standardizovanoj normalnoj distribuciji. Ipak, nastavljamo da je koristimo jer ona precizno modeluje neizvesnost koja proizilazi iz veličine uzorka - ključna prednost kada radimo s uzorcima različitih veličina.

Pre nego što nastavimo dalje, neophodno je uvesti novi koncept - **stepene slobode**. Iako ih je intuitivno teško razumeti, oni predstavljaju fundamentalan pojam za t-distribuciju i mnoge druge statističke koncepte. Pogledajmo detaljnije šta oni predstavljaju.

### 7.2.1 Stepeni slobode

Pre formalne definicije, razmotrite jednostavan primer koji će vam pomoći da intuitivno shvatite koncept. Zamislimo: cimerka se vraća iz kupovine s dve kese. Kaže da je potrošila 1000 dinara i traži da pogodite koliko je platila sadržaj svake kese. Šta možete proceniti? U suštini, samo vrednost jedne kese, jer je vrednost druge automatski određena prvom procenom. Ako procenite da prva kesa vredi 700 dinara, druga mora vredeti 300. Nemate „slobodu“ da nezavisno procenite vrednost druge - ona je već definisana. U ovom slučaju imate **jedan stepen slobode**.

Sada zamislimo da je došla s tri kese, a ukupna kupovina je i dalje 1000 dinara. Šta se menja? Sada možete proceniti vrednost prve (recimo 500 dinara) i druge kese (300 dinara), a vrednost treće je automatski određena (200 dinara). U ovom slučaju imate **dva stepena slobode**.

Generalizujući ovaj princip, stepeni slobode su jednaki broju kesa umanjenom za 1. S jednim stepenom slobode, situacija je deterministička - kada znate vrednost jedne kese, znate i vrednost druge. S više stepeni slobode, otvara se prostor za varijacije u procenama vrednosti pojedinačnih kesa.

U formalnom smislu, stepeni slobode predstavljaju broj nezavisnih vrednosti koje mogu slobodno varirati u izračunavanju neke statistike. U primeru s kesama, ako imate kesa i znate ukupnu cenu, imate stepeni slobode. Zašto? Zato što kada odredite cene kesa, cena poslednje je fiksirana kako bi se održala ukupna suma. U statistici, stepeni slobode se označavaju kao (degrees of freedom) i izražavaju formulom:

gde je broj opservacija (vrednosti u uzorku), a broj ograničenja ili procenjenih parametara. U primeru s tri kese čija je ukupna vrednost ograničena na 1000 dinara, imamo jedno ograničenje (), što nam daje dva stepena slobode ().

A kako stepeni slobode funkcionišu u t-distribuciji? Ovo je ključno pitanje za razumevanje. Hajde da pogledamo konkretan primer koji će nam pokazati kako t-test koristi stepene slobode za testiranje hipoteza o aritmetičkoj sredini. Na ovaj način ćemo videti direktnu primenu teorije koju smo obradili.

## 7.3 t-test za aritmetičku sredinu

Započnimo s praktičnim primerom t-testa, koji strukturalno nalikuje z-testu.

Posmatrajmo uzorak od 10 studenata kojima smo merili prosečan broj sati utrošenih na pripremu ispita. Pretpostavka je da student može spremiti ispit za minimalno 30 sati. Formulišemo nultu hipotezu:

Alternativna hipoteza je:

U jednom ispitnom roku, profesor je od 10 studenata saznao da je njihov prosek učenja bio 25 sati uz standardnu devijaciju od 5 sati. Pretvorimo to u R kod.

n <- 10  
as <- 25  
sd <- 5

Line 1

Broj studenata.

Line 2

Prosečan broj sati učenja u uzorku.

Line 3

Standardna devijacija uzorka.

Za razliku od Z-testa gde smo koristili pretpostavljenu vrednost standardne devijacije populacije, kod t-testa koristimo standardnu devijaciju uzorka. Razlog je jednostavan - kada ne koristimo stvarnu vrednost populacijske standardne devijacije (), uvodimo dodatnu neizvesnost jer se greška zasniva na proceni standardne devijacije iz uzorka. Međutim, ovu dodatnu neizvesnost obuhvatamo kroz t-distribuciju koja se prilagođava veličini uzorka.

Nakon toga, izračunavamo standardnu grešku i t-statistiku. T-statistika je standardizovana mera koja pokazuje koliko je aritmetička sredina uzorka udaljena od vrednosti iz nulte hipoteze, izražena u jedinicama standardne greške.

sg <- sd/sqrt(n)  
  
cat("Standardna greška:", round(sg, 2), "\n")  
  
t <- (as - 30)/sg  
  
cat("t-statistika:", round(t, 2), "\n")

Line 1

Računamo standardnu grešku.

Line 5

Standardizujemo razliku između uzorka i nulte hipoteze, dobijamo t-statistiku.

Standardna greška: 1.58   
t-statistika: -3.16

Na ovako malom uzorku, greška procene prosečnog vremena učenja je 1.58. Apsolutna razlika između uzorka i nulte hipoteze iznosi 5 sati (studenti prosečno uče 25 sati, dok je pretpostavka 30). Standardizovanjem te razlike dobijamo t-statistiku -3.16.

Hajde da vizualizujemo ovu t-statistiku na t-distribuciji. Za to nam je prvo potrebno da izračunamo **stepene slobode t-distribucije**. Stepeni slobode predstavljaju razliku između broja opservacija i broja procenjenih parametara. U našem slučaju imamo 10 opservacija (10 studenata) i jedan procenjeni parametar - aritmetičku sredinu uzorka (jedan ukupan zbir), što nam daje 9 stepeni slobode. Zašto je aritmetička sredina ekvivalentna zbiru? Jednostavno - kada kažete da je 10 studenata učilo prosečno 25 sati, to je identično tvrdnji da je ukupan broj sati učenja za 10 studenata 250. Prema tome, radimo sa 9 stepeni slobode.

par(family = "Jost")  
df <- n - 1  
  
x <- seq(-4, 4, 0.01)  
y <- dt(x, df)  
  
plot(x, y, type = "l", col = "#5C88DAFF", lwd = 3, xlab = "t-skor", ylab = "Gustina", xaxt = "n")  
  
axis(1, at=seq(-4, 4, 1), labels=seq(-4, 4, 1))  
abline(v = t, col = "#CC0C00FF", lwd = 3, lty = 1)  
  
text(-2.3, 0.15, "t = -3.16 ", col = "red", cex = 1.2, font = 2)

Line 2

Računamo stepene slobode t-distribucije.

Line 4

Definišemo opseg x vrednosti.

Line 5

Računamo gustinu t-distribucije.

Line 7

Crtamo gustinu t-distribucije.

Lines 9-10

Dodajemo oznake na x osu i crvenu liniju za t-skor.

|  |
| --- |
| Slika 7.3: t-distribucija za t-statistiku |

T-statistika od -3.16 nalazi se na levoj strani distribucije. Da bismo odredili p-vrednost, računamo površinu ispod krive t-distribucije. P-vrednost predstavlja površinu ispod krive levo od t-skora. U R-u ovo računamo koristeći funkciju pt() koja daje kumulativnu verovatnoću za zadati t-skor i stepene slobode.

Ovo je suštinska razlika u odnosu na Z-distribuciju - ovde eksplicitno uvažavamo veličinu uzorka pri računanju p-vrednosti. Za male uzorke, repovi distribucije postaju širi, a p-vrednost se povećava ([Slika 9.5](#fig-t-distribucija)). Ova osobina nam omogućava preciznije modelovanje neizvesnosti zaključaka kada radimo sa malim uzorcima.

Da vidimo konkretnu p-vrednost za naš uzorak.

p <- pt(t, df)  
cat("p-vrednost:", round(p, 4), "\n")

Line 1

Računamo p-vrednost za našu t-statistiku. Parametri su t-skor i stepeni slobode.

p-vrednost: 0.0058

Naš test daje p-vrednost od 0.006, što je značajno manje od standardnih pragova 0.05 i 0.01. Ovako niska p-vrednost pruža snažan dokaz za odbacivanje nulte hipoteze na nivou značajnosti 0.05. Šta nam to konkretno govori? Ako pretpostavimo da studenti u populaciji zaista uče 30 sati, verovatnoća da dobijemo ovakav uzorak je izuzetno mala. Logično je, dakle, da odbacimo tu pretpostavku. Zaključak je jasan: **studenti u proseku provode manje od 30 sati pripremajući ovaj ispit**.

Da bismo preciznije kvantifikovali ovaj zaključak, konstruišemo interval poverenja za aritmetičku sredinu populacije. Slično Z-testu, interval poverenja ima oblik:

gde je granična t-distribucije sa stepeni slobode i nivoom značajnosti . Izračunajmo 95% interval poverenja za naš uzorak.

alpha <- 0.05  
  
t <- qt(1 - alpha/2, df)  
  
donja\_granica <- as - t \* sg  
  
gornja\_granica <- as + t \* sg  
  
cat("Interval poverenja:", "\n")  
cat("(", round(donja\_granica, 2), ",", round(gornja\_granica, 2), ")")

Line 1

Nivo značajnosti 0.05. Naš prag za odbacivanje nulte hipoteze.

Line 3

Računamo graničnu vrednost t-distribucije za 95% interval poverenja. Zašto delimo alfa na dva? T-distribucija je simetrična, pa „odsecamo“ 2.5% sa svake strane.

Line 5

Donja granica intervala poverenja. Naša „pesimistična“ procena.

Line 7

Gornja granica. „Optimistična“ procena. Šta nam govori 95% interval poverenja (21.42, 28.58)? Kada bismo ponovili istraživanje mnogo puta, u 95% slučajeva bi stvarni prosek vremena učenja bio između 21.42 i 28.58 sati.

Interval poverenja:   
( 21.42 , 28.58 )

Ključno: ovaj zaključak jasno odražava ograničenja malog uzorka - samo 10 studenata. To smo precizno obuhvatili kroz graničnu vrednost t-distribucije, koja se prilagođava veličini uzorka. Razmislite o sledećem: šta bi se desilo da smo anketirali 100 ili 1000 studenata? Uz istu standardnu devijaciju, interval poverenja bi se značajno suzio, dajući nam mnogo precizniju procenu stvarne vrednosti u populaciji.

Primetićete da su rezultati t-testa i intervala poverenja uporedivi sa onima iz Z-testa. Međutim, t-test otvara vrata ka složenijim istraživačkim problemima - njegova puna snaga dolazi do izražaja pri poređenju dve aritmetičke sredine.

## 7.4 t-test razlike dve aritmetičke sredine

Istraživanja u društvenim naukama najčešće imaju komparativnu prirodu. Osnovni primer je analiza **razlika** između **dve** populacije. Populacije uglavnom predstavljaju različite društvene grupe, a cilj je otkriti razlike između njih po određenom obeležju.

Ovakve populacije zovemo **nezavisnim** jer ne postoji preklapanje između grupa. Fokusiraćemo se upravo na nezavisne populacije i uzorke, budući da oni predstavljaju okosnicu većine socioloških istraživanja.

U psihologiji i medicini češće srećemo **zavisne** populacije i uzorke, što je karakteristično za eksperimentalna istraživanja. Uzmimo primer medicinskog istraživanja efikasnosti tablete za snižavanje krvnog pritiska: dva uzorka čine **isti ispitanici**, a merimo njihov pritisak **pre** i **posle** terapije. Pošto su jedinice uzorka identične osobe, nazivamo ih **zavisnim**. U širem smislu govorimo o zavisnim populacijama, jer to obuhvata sve potencijalne pacijente pre i posle terapije.

Kod nezavisnih uzoraka, definišemo dve odvojene grupe već pri konstrukciji uzorka. Ispitanike raspoređujemo u jedan ili drugi uzorak prema jasno definisanim kriterijumima grupa koje istražujemo. Primer koji ćemo analizirati koristi klasičnu podelu u sociologiji: uzorke prema polu, odnosno uzorak muškaraca i žena.

Vratimo se na konkretan primer studije o prosečnim zaradama u Srbiji. Naši podaci su organizovani tako da je uzorak podeljen na muškarce i žene, a cilj nam je da utvrdimo postoji li **značajna razlika u prosečnim zaradama između ove dve grupe**. Pogledajmo kako izgledaju podaci:

podaci <- "https://gist.githubusercontent.com/atomashevic/8c30bc0a11ba93ec934a12d0e08a6d17/raw/4924fc12814787602001e5058ec774f3487fad64/primanja-pol.csv"  
podaci <- read.csv(podaci)  
  
head(podaci)

Line 2

Učitavamo podatke s URL-a.

Line 4

Prikazujemo prvih nekoliko redova tabele.

primanja pol  
1 902 Z  
2 986 M  
3 890 Z  
4 456 Z  
5 831 Z  
6 1091 M

Analizirajmo podatke sistematično. Naš problem sadrži **dve varijable**. Varijabla pol je **kategorijalna** i ima dve vrednosti: M i Z. Ovo je jasan primer **nezavisnih** uzoraka, jer muškarci i žene predstavljaju dve različite grupe. Jednostavnije rečeno - jedan ispitanik ne može istovremeno biti i muškarac i žena u varijabli pol.

Naša početna pretpostavka glasi: **prosečne zarade muškaraca i žena u Srbiji su jednake**. Drugim rečima, pol nema efekta na prosečnu zaradu. Ovo je **nulta hipoteza** koju ćemo testirati. Ona pretpostavlja da je razlika između proseka u populaciji nula, odnosno da ne postoji efekat. Upravo zbog toga je i zovemo nultom hipotezom. U matematičkoj notaciji to zapisujemo:

ili

gde je prosečna zarada muškaraca, a prosečna zarada žena u populaciji, odnosno među svim radno sposobnim građanima Srbije.

**Alternativna hipoteza** tvrdi suprotno - da postoji razlika između prosečnih zarada muškaraca i žena u Srbiji. Drugim rečima, prosečne zarade nisu jednake.

Hipoteza podrazumeva postojanje **efekta** - pol utiče na prosečnu zaradu. To znači da postoji veza između pola osobe i njene zarade. U uzorku će se ovo manifestovati kroz različite prosečne zarade muškaraca i žena. Naravno, ovo ne znači da svaka žena zarađuje manje ili više od svakog muškarca, već da se **prosečne vrednosti** razlikuju. To ukazuje na činjenicu da u populaciji, odnosno u našem društvu, **postoji mehanizam** koji dovodi do toga da jedna grupa prosečno zarađuje više od druge. Koji je to mehanizam? Možemo izneti neke pretpostavke - možda je reč o diskriminaciji, društvenim normama, razlikama u obrazovanju ili nekim drugim faktorima. Međutim, na ovom nivou statističke analize ne možemo dati konačan odgovor; za to je potrebno dublje istraživanje i teorijski okvir.

Kako testirati ove hipoteze? Prvo moramo detaljno pogledati podatke. Potrebno je izračunati broj muškaraca i žena u uzorku, a zatim i prosečne zarade za svaku grupu. Krenimo s tim:

muskarci <- podaci[podaci$pol == "M", ]  
zene <- podaci[podaci$pol == "Z", ]  
  
n\_m <- nrow(muskarci)  
n\_z <- nrow(zene)  
  
cat("Broj muškaraca u uzorku:", n\_m, "\n")  
cat("Broj žena u uzorku:", n\_z, "\n")

Line 1

Izdvajamo podatke o muškarcima. Pravimo novu varijablu muskarci koja sadrži sve redove gde je vrednost varijable pol jednaka "M". Uzimamo sve kolone.

Line 2

Izdvajamo podatke o ženama. Pravimo novu varijablu zene koja sadrži sve redove gde je vrednost varijable pol jednaka "Z". Uzimamo sve kolone.

Lines 4-5

Računamo broj muškaraca i broj žena u uzorku koristeći funkciju nrow koja broji redove u tabeli podataka.

Broj muškaraca u uzorku: 80   
Broj žena u uzorku: 40

Vidimo da uzorak nije balansiran - broj muškaraca je dvostruko veći od broja žena. Ovo može biti rezultat načina uzorkovanja ili jednostavno slučajnost. Za trenutnu analizu taj disbalans nije kritičan, pa nastavljamo dalje. Fokusirajmo se na ono što je bitno - računanje prosečnih zarada za muškarce i žene:

X\_m <- mean(muskarci$primanja)  
X\_z <- mean(zene$primanja)  
  
sd\_m <- sd(muskarci$primanja)  
sd\_z <- sd(zene$primanja)  
  
cat("Prosečna zarada muškaraca:", round(X\_m, 2), "\n")  
cat("Prosečna zarada žena:", round(X\_z, 2), "\n")  
  
cat("Standardna devijacija zarada muškaraca:", round(sd\_m, 2), "\n")  
cat("Standardna devijacija zarada žena:", round(sd\_z, 2), "\n")

Line 1

**Računamo prosečnu zaradu muškaraca.** Funkcija mean će izračunati aritmetičku sredinu zarada za sve muškarce.

Line 2

**Računamo prosečnu zaradu žena.** Funkcija mean će izračunati aritmetičku sredinu zarada za sve žene.

Line 4

**Računamo standardnu devijaciju zarada muškaraca.** Funkcija sd će izračunati standardnu devijaciju zarada za sve muškarce.

Line 5

**Računamo standardnu devijaciju zarada žena.** Funkcija sd će izračunati standardnu devijaciju zarada za sve žene.

Prosečna zarada muškaraca: 844.78   
Prosečna zarada žena: 781.58   
Standardna devijacija zarada muškaraca: 230.93   
Standardna devijacija zarada žena: 168.86

Rezultati pokazuju da žene u našem uzorku imaju prosečno za nešto više od 60 EUR manja primanja u odnosu na muškarce. Ovo je prvi signal da postoji razlika između prosečnih primanja muškaraca i žena. Međutim, ključno je pitanje: da li je ta razlika dovoljno izražena da bismo mogli zaključiti da postoji stvarni efekat pola na prosečna primanja? Da li je razlika statistički značajna? Za odgovore na ova pitanja moramo primeniti statističke testove.

Pre nego što to uradimo, vizualizujmo naše podatke pomoću dva histograma.

par(family = "Jost", mar = c(5, 4, 4, 2) + 0.1)  
  
hist(podaci$primanja[podaci$pol == "M"],  
 breaks = 30,  
 col = rgb(0, 0, 1, 0.3),  
 main = "Distribucija zarada po polu",  
 xlab = "Zarada (EUR)",  
 ylab = "Frekvencija",  
 xlim = range(podaci$primanja),  
 ylim = c(0, 20),  
 border = "blue")  
  
hist(podaci$primanja[podaci$pol == "Z"],  
 breaks = 15,  
 col = rgb(1, 0, 0, 0.3),  
 add = TRUE,  
 border = "red")  
  
abline(v = X\_m, col = "blue", lwd = 2, lty = 2)  
abline(v = X\_z, col = "red", lwd = 2, lty = 2)  
  
arrows(X\_z, 16, X\_m, 16, code = 3, length = 0.1, lwd = 2)  
text((X\_m + X\_z)/2 - 100, 16,  
 paste0(round(X\_m - X\_z, 2), " EUR"),  
 pos = 3, font = 2)  
  
legend("topright",  
 legend = c("Muškarci", "Žene"),  
 fill = c(rgb(0, 0, 1, 0.3), rgb(1, 0, 0, 0.3)),  
 border = c("blue", "red"))

Line 3

Funkcija hist prikazuje distribuciju zarada za sve ispitanike sa vrednošću pol jednakom "M". Histogram je plave boje.

Line 13

Funkcija hist prikazuje distribuciju zarada za sve ispitanike sa vrednošću pol jednakom "Z". Histogram je crvene boje.

Lines 19-20

Linija plave boje označava prosečne zarade muškaraca, dok crvena linija označava prosečne zarade žena.

Line 22

Strelica pokazuje razliku od 60 EUR između muškaraca i žena.

Line 25

Tekst prikazuje konkretnu vrednost razlike zarada.

Line 30

Legenda pomaže u razlikovanju histograma muškaraca i žena.

|  |
| --- |
| Slika 7.4: Distribucija zarada po polu |

Na grafikonu vidimo da se aritmetičke sredine (prikazane isprekidanim linijama) nalaze sa različitih strana podeoka od 800 EUR. Iako zarade žena izgledaju pomerene ulevo u odnosu na muškarce, distribucija otkriva složeniju sliku. Crveni histogram je uži, što ukazuje da su zarade žena koncentrisanije u nižem delu distribucije, dok plavi histogram pokazuje širi raspon zarada muškaraca.

Centralno pitanje je sledeće: da li je ova uočena razlika između prosečnih zarada dovoljno izražena da bismo mogli zaključiti da postoji stvaran efekat pola na primanja? Za odgovor na ovo pitanje koristićemo t-test za dva nezavisna uzorka.

Za t-test nam je potrebna statistika testa. U prethodnim primerima, statistika testa merila je odstupanje uzorka od vrednosti nulte hipoteze. Sada ćemo se usmeriti na razliku između dve aritmetičke sredine.

Trenutno, razlika između prosečnih zarada iznosi -63.2 EUR. Šta predviđa nulta hipoteza? Kao što sam naziv sugeriše - nulu. Dakle, statistiku testa računamo prema formuli:

Ovde uzimamo razliku između dve aritmetičke sredine i delimo je sa standardnom greškom razlike (). Standardnu grešku razlike između dve aritmetičke sredine računamo na sledeći način:

**Zašto računamo standardnu grešku na ovaj način?** Ova formula proizilazi iz činjenice da imamo **dva izvora** neizvesnosti jer proučavamo dve različite grupe. Svaka grupa donosi svoju aritmetičku sredinu i standardnu devijaciju iz zasebnih uzoraka. Kada analiziramo razliku između ovih sredina, moramo obuhvatiti sve izvore neizvesnosti, što znači da se standardne greške moraju kombinovati.

Matematički elegantno rešenje je sabiranje kvadrata standardnih grešaka i izračunavanje kvadratnog korena, čime dobijamo jedinstvenu, kombinovanu standardnu grešku koja precizno odražava ukupnu neizvesnost naše procene.

|  |  |  |
| --- | --- | --- |
|  | **Geometrijska interpretacija**  Da li vam je izraz poznat? Ako se sećate Pitagorine teoreme, prepoznaćete da on predstavlja dužinu hipotenuze pravouglog trougla čije su katete i . Ovo je elegantna geometrijska interpretacija standardne greške razlike između dve aritmetičke sredine.   |  | | --- | | Slika 7.5: Geometrijska interpretacija standardne greške razlike |   Obe standardne greške aritmetičkih sredina iznose približno 26 EUR. Kada ih kombinujemo, dobijamo standardnu grešku razlike između dve aritmetičke sredine koja prelazi 36 EUR. Ovo jasno pokazuje zašto moramo uzeti u obzir sve izvore neizvesnosti kada analiziramo razlike između grupa. Matematički gledano, hipotenuza pravouglog trougla je uvek duža od kateta - analogno tome, standardna greška razlike između dve aritmetičke sredine je nužno veća od pojedinačnih standardnih grešaka. |

Standardna greška razlike je uvek veća od pojedinačnih standardnih grešaka aritmetičkih sredina. Kad god poredimo dve grupe, moramo uzeti u obzir sve izvore neizvesnosti koji se kombinuju i povećavaju ukupnu grešku merenja. Iz ovoga sledi važna pouka: iako je standardna greška aritmetičke sredine jednog uzorka (oko 800 EUR) manja od standardne greške razlike dve sredine (oko 60 EUR), procena pojedinačne sredine je preciznija. Praktična posledica je sledeća - teže je statistički dokazati postojanje razlike između dve grupe nego proceniti pojedinačnu sredinu, jer je standardna greška razlike veća od standardnih grešaka pojedinačnih sredina.

|  |  |
| --- | --- |
|  | **Jednake i nejednake varijanse**  Način na koji dolazimo do informacije o standardnoj greški razlike zavisi od pretpostavke o varijansama u populacijama. U našem slučaju, pretpostavljamo da je varijansa, odnosno varijabilitet zarada, različit u obe populacije. Razlike u varijabilitetu unutar populacije žena i muškaraca mogu biti posledica različitih faktora, kao što su obrazovna i starosna struktura, kao i razlike u zaposlenju. Ovakav pristup se naziva **pretpostavka o nejednakim varijansama**. Ovo je složeniji slučaj i zbog toga imamo „inflaciju“ standardne greške razlike, što je direktna posledica dodatne neizvesnosti.  Postoji i jednostavniji pristup - možemo pretpostaviti da su varijanse u obe populacije jednake. Ovo je **pretpostavka o jednakim varijansama**. U tom slučaju, standardna greška razlike između dve aritmetičke sredine se računa na elegantniji način.  Prvi korak je računanje kombinovane varijanse:  Ova formula izgleda kompleksnije, ali zapravo predstavlja **ponderisanu** varijansu. Ona kombinuje veličinu uzorka i standardne greške pojedinačnih grupa, dajući veći značaj većem uzorku.  Zašto je to važno? Pod pretpostavkom jednakih varijansi, standardne devijacije iz uzoraka služe kao procenitelji jedne populacijske varijanse. Logično je da veći uzorak daje precizniju procenu te varijanse, pa mu dodeljujemo veću težinu. Ova ponderacija je ključna jer minimizuje **pristrasnost** u proceni varijanse.  Na kraju, standardnu grešku razlike izračunavamo koristeći ovu varijansu:  Uzimamo u obzir veličinu i jednog i drugog uzorka. Razlika u veličini uzoraka (duplo više muškaraca nego žena) nije problem jer oba poduzorka doprinose našem razumevanju varijabiliteta u populaciji.  Kad je pretpostavka o jednakoj varijansi opravdana? Najčešće je primenjujemo kada koristimo standardizovane merne instrumente ili kada su grupe koje poredimo vrlo slične po svim karakteristikama osim one koju proučavamo. Zamislite, na primer, merenje IQ-a na studentskim uzorcima iz dve države. Nema teorijskog razloga zašto bi varijabilitet IQ-a studenata iz Švajcarske bio drugačiji od varijabiliteta studenata iz Indije. U takvom slučaju, pretpostavka o jednakim varijansama je logična.  U našem slučaju znamo vrlo malo o tome kako zarade variraju između muškaraca i žena. Stoga je metodološki ispravnije krenuti od pretpostavke o nejednakim varijansama. |

Hajde da vidimo kako izgleda ova razlika u praksi na našem primeru.

SG\_M <- sd\_m / sqrt(n\_m)  
SG\_Z <- sd\_z / sqrt(n\_z)  
SG\_MZ <- sqrt(SG\_M^2 + SG\_Z^2)  
  
cat("Standardna greška razlike između dve aritmetičke sredine: ", round(SG\_MZ, 2), "\n")

Standardna greška razlike između dve aritmetičke sredine: 37.14

t <- (X\_m - X\_z) / SG\_MZ  
  
cat("t statistika: ", round(t, 4), "\n")

t statistika: 1.7016

Statistika našeg testa iznosi 1.7. Šta nam ova vrednost govori o nultoj hipotezi? Odgovor pronalazimo kroz konstrukciju t-distribucije. Za to nam je prvo potreban broj stepeni slobode. Naš uzorak ima ukupno opservacija. Međutim, broj stepeni slobode je manji jer procenjujemo **dve aritmetičke sredine**. Konkretno, imamo stepeni slobode.

Sa ovom informacijom možemo vizuelno predstaviti t-distribuciju i locirati našu test statistiku na njoj.

par(family = "Jost")  
x <- seq(-4, 4, 0.01)  
  
y <- dt(x, df = 118)  
  
plot(x, y, type = "l", col = "#5C88DAFF", lwd = 3, xlab = "t-statistika", ylab = "Gustina")  
abline(v = t, col = "#CC0C00FF", lwd = 3, lty = 1)  
  
text(t + 0.7, 0.2, paste("t =",round(t, 2)), col = "#CC0C00FF", pos = 3, cex = 1.2, font = 2)

Line 4

Funkcija dt računa gustinu t-distribucije za datu vrednost t i broj stepeni slobode. U našem slučaju, broj stepeni slobode je 118.

|  |
| --- |
| Slika 7.6: t-distribucija i t-statistika |

Naša statistika se nalazi u desnom delu distribucije, što ukazuje da je dobijena razlika veća od one koju bismo očekivali pod nultom hipotezom. Za donošenje konačnog zaključka moramo izračunati **p-vrednost**. Ova vrednost će nam precizno reći da li je signal koji smo dobili iz uzorka dovoljno jak (statistički značajan) da bismo mogli odbaciti nultu hipotezu.

Vizualizujmo p-vrednost na t-distribuciji.

par(family = "Jost", mar = c(5, 4, 2, 2) + 0.1)  
x <- seq(-4, 4, 0.01)  
y <- dt(x, df = 118)  
  
plot(x, y, type = "l", col = "#5C88DAFF", lwd = 3, xlab = "t-statistika", ylab = "Gustina", ylim = c(0, 0.4), yaxs = "i")  
  
segments(t, 0, t, dt(t, df = 118), col = "#CC0C00FF", lwd = 2)  
  
polygon(c(x[x > t], rev(x[x > t])), c(y[x > t], rep(0, sum(x > t))), col = rgb(1, 0, 0, 0.3), border = NA)  
  
segments(-t, 0, -t, dt(-t, df = 118), col = "#CC0C00FF", lwd = 2)  
  
polygon(c(x[x < -t], rev(x[x < -t])), c(y[x < -t], rep(0, sum(x < -t))), col = rgb(1, 0, 0, 0.3), border = NA)  
  
p\_vrednost <- 2 \* pt(-t, df = 118)  
  
text(0, 0.05, paste("p =", round(p\_vrednost, 4)), col = "red", pos = 3, cex = 1.2, font = 2)

Lines 2-3

Definišemo opseg x-ose.

Line 5

Crtamo t-distribuciju i označavamo t-statistiku.

Line 15

Računamo p-vrednost za t-test. Vrednost funkcije pt množimo sa dva jer računamo površinu u oba repa distribucije.

|  |
| --- |
| Slika 7.7: t-distribucija i p-vrednost |

Budući da koristimo dvostrani test, razmatramo ukupnu površinu u oba repa distribucije: levo i desno od naše t-statistike (-1.7 i 1.7). P-vrednost za t-test računamo u R-u pomoću funkcije pt. Ova funkcija izračunava površinu ispod t-distribucije do date t-statistike, uzimajući u obzir stepene slobode. Dobijeni rezultat množimo sa dva jer, kao što je prikazano na grafikonu, posmatramo oba ekstremna dela t-distribucije.

p\_vrednost <- 2 \* pt(-t, df = 118)  
  
cat("P-vrednost:", round(p\_vrednost, 4), "\n")

Line 1

Računamo p-vrednost za t-test. Vrednost funkcije pt množimo sa dva. Koristimo -t zato što ova funkcija računa površinu ispod krive sa leva na desno.

Line 3

Zaokružujemo p-vrednost na četiri decimale i prikazujemo je.

P-vrednost: 0.0915

P-vrednost je približno 0.09 i veća je od standardnih nivoa značajnosti 0.05 i 0.01. To nam ukazuje da naš rezultat nije dovoljno ekstreman u odnosu na ono što predviđa nulta hipoteza. Preciznije, razlika od 63 EUR koju smo izmerili u uzorku nije dovoljno snažan signal da bismo zaključili da postoji stvarna razlika između prosečnih zarada muškaraca i žena u populaciji. Ovo je suštinska informacija koju dobijamo iz našeg statističkog testa.

Prema tome, nemamo dovoljno dokaza da odbacimo nultu hipotezu i zaključujemo da **ne postoje razlike u prosečnim zaradama muškaraca i žena u Srbiji**.

|  |  |
| --- | --- |
|  | **Lična karta metoda: t-test za dva nezavisna uzorka**  **Šta radi?** Proverava da li je razlika između prosečnih vrednosti dva nezavisna uzorka statistički značajno različita od nule.  **Kada se koristi?** Kada želimo da uporedimo dve grupe ispitanika i utvrdimo postojanje statistički značajne razlike između njih.  **Koliko varijabli imamo?** Dve: jednu binarnu kategorijalnu (npr. pol) merenu nominalnom ili ordinalnom skalom i jednu kvantitativnu (npr. zarada) merenu intervalnom ili racio skalom, u posebnim slučajevima i ordinalnom skalom. Grupe ispitanika formiraju se na osnovu binarne kategorijalne varijable.  **Kako glase hipoteze?** Za dvosmerni test:  Za jednosmerni test:  (ili obrnuto, zavisno od konteksta istraživanja)  **Kako izgleda statistika testa?** Statistika t-testa računa se kao količnik između razlike aritmetičkih sredina uzoraka i standardne greške razlike.  gde je standardna greška razlike između dve aritmetičke sredine.  **Kako računamo p-vrednost?** Za dvosmerni test, p-vrednost se računa kao dvostruka površina ispod t-distribucije do vrednosti t-statistike sa leve strane distribucije. Za jednosmerni test, p-vrednost se računa kao površina ispod t-distribucije do vrednosti t-statistike. |

## 7.5 Interval poverenja za razliku aritmetičkih sredina

Sada možemo primeniti drugi oblik statističkog zaključivanja - interval poverenja za razliku dve aritmetičke sredine. Ovaj pristup nam daje drugačiju perspektivu o razlici između prosečnih zarada muškaraca i žena.

Konstrukcija intervala poverenja počinje od osnovne tačke - razlike aritmetičkih sredina uzoraka: . Nakon toga dodajemo i oduzimamo kritičnu vrednost , koja je određena željenim nivoom poverenja i brojem stepeni slobode.

Kritične vrednosti definišu granice u kojima očekujemo da se nalazi naša statistika testa, uz pretpostavku da obuhvata 95% ili 99% svih mogućih vrednosti.

Pri konstrukciji intervala poverenja od 95%, granične slučajeve postavljamo tako da „odsecamo“ po 2.5% vrednosti sa obe strane distribucije. Za dobijanje kritične vrednosti potrebni su nam granična vrednost t-statistike i standardna greška razlike aritmetičkih sredina.

Kritičnu vrednost izračunavamo pomoću R funkcije qt().

alfa <- 0.05  
  
df = n\_m + n\_z - 2  
  
t <- qt(1 - alfa/2, df = df)  
  
cat("Kritična vrednost t-statistike: ", round(t, 2), "\n")

Line 1

Nivo poverenja od 95% odgovara alfa = 0.05.

Line 3

Broj stepeni slobode je 118.

Line 5

Računamo kritičnu vrednost t-statistike za nivo poverenja od 95%.

Kritična vrednost t-statistike: 1.98

Konkretno, proširujemo interval od za , gde je standardna greška razlike između dve aritmetičke sredine. S obzirom da je t-statistika približno 2, interval efektivno proširujemo za .

razlika <- X\_m - X\_z  
  
donja\_granica <- razlika - t \* SG\_MZ  
  
gornja\_granica <- razlika + t \* SG\_MZ  
  
cat("Interval poverenja: (", round(donja\_granica, 2), ", ", round(gornja\_granica, 2), ")\n")

Line 1

Računamo razliku aritmetičkih sredina.

Lines 3,5

Računamo donju i gornju granicu intervala poverenja.

Interval poverenja: ( -10.35 , 136.75 )

Ovaj interval obuhvata vrednosti od -10 do +136 EUR, što ukazuje na značajnu neizvesnost u proceni razlike. Interval pokriva tri moguća scenarija:

1. Negativna razlika do -10 EUR - žene imaju veća prosečna primanja od muškaraca.
2. Nulta razlika - ne postoji razlika u prosečnim primanjima između muškaraca i žena.
3. Pozitivna razlika do 136 EUR - muškarci imaju veća prosečna primanja od žena.

Činjenica da interval obuhvata sva tri scenarija jasno pokazuje ograničenja našeg uzorka. Na osnovu ovih podataka ne možemo pouzdano utvrditi ni postojanje ni nepostojanje razlike u primanjima između polova. Drugim rečima, neizvesnost u našim procenama je toliko velika da ne možemo isključiti mogućnost da razlike između ove dve populacije zapravo ne postoje.

Na osnovu intervala poverenja, naš zaključak je jasan: **Sa nivoom pouzdanosti od 95%, razlika između prosečnih zarada muškaraca i žena kreće se u intervalu od -10 do 136 EUR**. U praksi to znači: ako bismo ponavljali ovo istraživanje i izvlačili nove uzorke, u 95% slučajeva razlika između prosečnih zarada bi se našla unutar ovog intervala.

Šta nam ovaj širok interval govori? U matematičkom smislu, on obuhvata sve moguće scenarije. Posebno je zanimljiva asimetrija intervala - proteže se dalje u pozitivnom smeru, što ukazuje da ćemo u ponovljenim uzorcima verovatnije dobiti pozitivnu razliku između prosečnih zarada muškaraca i žena. Ipak, jasno je da naši podaci nisu dovoljno precizni da bismo mogli tvrditi kako je ova razlika sistematska na nivou cele populacije.

Ova situacija je kao kada u nekom kriminalističkom slučaju, ubica biva uhvaćen, svima je očigledno da je on kriv, ali prosto nema materijalnih dokaza koji bi to potvrdili na sudu. Isto, tako mi imamo teorijske razloge da verujemo da su prosečna primanja muškaraca veća, imamo rezultate prethodnih istraživanja koja to potvrđuju, imamo rezultat iz uzorka koji to sugeriše, ali nedovoljno dobro za standarde statističkog testa.

## 7.6 Veličina efekta

Postoji još jedan način da izmerimo jačinu dokaza koje imamo u korist postojanja efekta, odnosno uticaja jedne varijable na drugu. Reč je o indikatoru koji nazivamo **jačina** ili **veličina** efekta, a dobijamo ga pomoću **Koenove d statistike**.

Logika iza ove statistike je gotovo identična statistici t-testa. Jedina razlika je u tome što u formuli ne koristimo standardnu grešku, već združenu ili kombinovanu standardnu devijaciju. Kao rezultat dobijamo vrednost koja se direktno i jednostavno može interpretirati. Formula glasi:

gde je združena standardna devijacija, koja se računa na sledeći način:

Formula je složena, ali njena složenost nam trenutno nije bitna. Pošto se ova kombinovana standardna devijacija često koristi u različitim analizama, napisaćemo R funkciju koja će nam olakšati izračunavanje kad god nam zatreba.

sd\_komb <- function(grupa1, grupa2){  
 n1 <- length(grupa1)  
 n2 <- length(grupa2)  
  
 sd1 <- sd(grupa1)  
 sd2 <- sd(grupa2)  
  
 return(sqrt(((n1 - 1) \* sd1^2 + (n2 - 1) \* sd2^2) / (n1 + n2 - 2)))  
}

Lines 2-3

Računamo veličinu prve i druge grupe.

Lines 5-6

Računamo standardnu devijaciju za obe grupe.

Line 8

Primenjujemo formulu iznad i vraćamo izračunatu vrednost putem funkcije return()

Sada jednostavno možemo da izračunamo veličinu efekta za naš primer.

d <- (X\_m - X\_z) / sd\_komb(muskarci$primanja, zene$primanja)  
  
cat("Koenova d statistika: ", round(d, 2), "\n")

Line 1

Primenjujemo formulu za Koenovu d statistiku.

Koenova d statistika: 0.3

Rekli smo da je vrednost Koenove d statistike lako interpretirati. Ove vrednosti nam služe kao referentne tačke za interpretaciju:

* 0.2: mali efekat
* 0.5: srednji efekat
* 0.8: veliki efekat

Odakle dolaze ove referentne vrednosti? One su postavljene po konvenciji, na osnovu empirijskih rezultata primene Koenove d statistike u slučajevima gde imamo jasnu potvrdu postojanja ili nepostojanja efekta.

Evo konkretnijeg načina interpretacije ovih vrednosti. Znamo da je aritmetička sredina muškaraca u uzorku veća od aritmetičke sredine žena. Postavimo ključno pitanje: koji procenat žena ima zaradu manju od prosečne zarade muškaraca?

Teorijski maksimalni efekat bi postojao kada bi sve žene imale zaradu manju od prosečne zarade muškaraca. Takve ekstremne razlike su retke u realnim podacima. U praksi su razlike suptilnije, ali koliko suptilnije? Koenova d statistika nam daje precizan odgovor na ovo pitanje. U tabeli ispod možete videti procente koji odgovaraju

Odnos veličine efekta i procenta opservacija i jedne grupe koji se nalaze ispod proseka druge grupe.

| Veličina efekta | % |
| --- | --- |
| 0.2 | 58 |
| 0.5 | 69 |
| 0.8 | 79 |

Kada bismo dobili Koenov d od 0.2, to bi značilo da otprilike 60% žena ima zaradu manju od prosečne zarade muškaraca. Sa druge strane, ako bismo dobili Koenov d od 0.8, to bi značilo da 80% žena ima zaradu manju od prosečne zarade muškaraca.Imajte u vidu da, ako su podaci približno normalno raspoređeni ili približno simetrični, 50% muškaraca takođe ima zaradu manju od prosečne zarade muškaraca. Zato se 60% uzima kao prvi korak, odnosno prva granica za veličinu efekta.

### 7.6.1 Više od dve grupe?

Primer koji smo obradili u ovom poglavlju je primer jednostavnog **statističkog modela** koji opisuje uticaj jedne **nezavisne kategorijale** binarne varijable na **zavisnu** kvantitativnu varijablu. Ovaj uticaj smo istraživali tako što smo uporedili prosečne vrednosti zavisne varijable između dve grupe. Imamo samo dve grupe zbog toga što smo kao nezavisnu varijablu izabrali biološki pol ispitanika, koji može da ima samo dve vrednosti.

|  |  |
| --- | --- |
|  | **Nezavisne i zavisne varijable**  Kada definišemo statistički model, vrlo često pravimo razliku između **nezavisnih** i **zavisnih** varijabli. Nezavisne varijable su one koje koristimo da bismo objasnili varijabilnost zavisne varijable. U našem slučaju, nezavisna varijabla je pol ispitanika, a zavisna varijabla su prosečna primanja.  Šta znači opisati varijabilinost? To znači da hoćemo da identifikujemo pravilnosti ili obrasce u razlikama koje postoje u zavisnoj varijabli. Konkretno, pokušali smo da objasnimo razlike u primanjima, tako što bismo podelili uzorak na dve grupe, muškarce i žene, i uporedili prosečne vrednosti primanja između njih. Da smo uspeli u tome, saznali bismo nešto novo o obrascu koji postoji u našim podacima – u proseku žene imaju niža primanja od muškaraca.  Da je to slučaj, to bi imalo važne implikacije. Na primer, značilo bi da kada bismo nasumično izabrali jednog muškarca i jednu ženu iz populacije, postoji veća verovatnoća da će muškarac imati veća primanja. To nam omogućava da **predvidimo** vrednost zavisne varijable (prosečna primanja) na osnovu nezavisne varijable (pol ispitanika). Naravno, ovo predviđanje nije savršeno precizno, ali nam pruža korisnu informaciju o primanjima ispitanika samo na osnovu njihovog pola.  Upravo zbog ove prediktivne moći, nezavisne varijable se često nazivaju **prediktorima**, a zavisne varijable **kriterijumima**. U našem primeru, pol ispitanika je prediktor, a prosečna primanja su kriterijum. |

Šta ako želimo da analiziramo kompleksniju nezavisnu varijablu? Uzmimo logičan primer: očekujemo da zarada zaposlenog raste sa godinama iskustva u određenoj delatnosti. Teško je zamisliti da apsolutni početnik ima istu platu kao neko sa 10, 20 ili 30 godina iskustva.

Kako konstruisati statistički model koji može objasniti ove razlike? Naš trenutni alat zahteva grubu podelu ispitanika na „početnike“ i „iskusne“, nakon čega bismo uporedili prosečne zarade između grupa. Međutim, ova podela je previše pojednostavljena jer ignoriše činjenicu da je iskustvo kontinualna, a ne kategorička varijabla.

Idealno, trebalo bi da izračunamo prosečna primanja za svaku godinu iskustva - od 0 godina, preko 1 godine, 2 godine, i tako dalje. Ali šta sa osobama koje imaju 1.5 ili 2.25 godina iskustva? Kako uporediti prosečna primanja između ovih preciznih vrednosti?

Odgovor leži u jednom od najmoćnijih alata statističke analize - **regresionoj analizi**.

## 7.7 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1 \***  U zanimljivom pedagoškom eksperimentu, studenti su slušali isti jednosemestralni kurs. Kontrolna grupa je brojala 120 studenata koji su pratili nastavu i radili predispitne zadatke na klasičan način. Eksperimentalna grupa od 30 studenata imala je mogućnost korišćenja AI alata pri rešavanju predispitnih zadataka.  Obe grupe studenata polagale su identičan završni ispit. U kontrolnoj grupi, prosečna ocena iznosila je 7.80 sa standardnom devijacijom od 1.5. Studenti iz eksperimentalne grupe ostvarili su prosečnu ocenu 7.35 uz standardnu devijaciju od 1.2.  Istraživačka hipoteza pretpostavlja da će studenti koji su koristili AI alate postići niže prosečne ocene na završnom ispitu u poređenju sa studentima iz kontrolne grupe. Testirajte ovu hipotezu primenom t-testa za dva nezavisna uzorka, uz nivo značajnosti 0.05. |

|  |  |
| --- | --- |
|  | **Zadatak 2 \***  Da bismo bolje razumeli rezultate eksperimenta, potrebno je da proširimo analizu.   1. Izračunajte interval poverenja od 95% za razliku u prosečnim ocenama između kontrolne i eksperimentalne grupe. Šta nam dobijeni interval govori o efektima primene AI alata? 2. Izračunajte i interpretirajte veličinu efekta (Koenov d). Koliko je snažan uticaj AI alata na ocene studenata? 3. Na osnovu dobijenih rezultata, koje konkretne preporuke biste dali za dizajn budućih istraživanja koja bi preciznije merila efekte AI alata na akademski uspeh? |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Zadatak 3 \*\***  U narednoj fazi istraživanja o uticaju AI alata na akademski uspeh, istraživački dizajn je proširen. Kontrolna grupa je zadržana, ali su uvedene dve eksperimentalne grupe. Prva eksperimentalna grupa od 50 studenata koristila je AI alate tokom pripreme ispita, dok je druga eksperimentalna grupa od 30 studenata dobila mogućnost korišćenja AI alata i tokom samog ispita.  Rezultati studije prikazani su u tabeli:   | Grupa | Prosečna ocena | Standardna devijacija | | --- | --- | --- | | Kontrolna grupa | 7.80 | 1.5 | | Eksperimentalna grupa 1 | 7.35 | 1.2 | | Eksperimentalna grupa 2 | 7.95 | 0.3 |   Pretpostavka je da će studenti iz eksperimentalne grupe 2 imati više prosečne ocene na završnom ispitu od studenata iz kontrolne grupe, kao i od studenata iz eksperimentalne grupe 1.   1. Testirajte ove dve hipoteze koristeći t-test za dva nezavisna uzorka na nivou značajnosti od 0.05. Pri analizi obratite posebnu pažnju na pretpostavke o (ne)jednakim varijansama. 2. Interpretirajte oba rezultata i formulišite celovit zaključak o uticaju AI alata na akademski uspeh studenata. 3. Ako biste želeli da uporedite sve grupe studenata međusobno, koliko bi statističkih testova bilo potrebno? Razmislite o nedostacima ovakvog pristupa. |

# 8. Regresija

## 8.1 Osnove regresione analize

U prethodnom poglavlju smo ispitivali uticaj jedne varijable na drugu. Sada prelazimo na analizu odnosa između dve kvantitativne varijable. Istraživačko pitanje ostaje isto: postoji li uticaj nezavisne varijable na zavisnu varijablu? Međutim, metod kojim ispitujemo taj uticaj je drugačiji.

Razmotrimo konkretan primer. Imamo uzorak od 60 država sa dva ključna podatka:

* godišnji budžet za zdravstvo po glavi stanovnika (u dolarima) i
* ocenu kvaliteta zdravstvenog sistema (na skali od 1 do 100).

Najbolju ocenu zdravstvenog sistema dobila je Norveška (100), dok je najlošiju dobila Nigerija (29). Ocene su formirali međunarodni eksperti na osnovu više ključnih parametara. Naš zadatak je jasan - utvrditi vezu između budžeta za zdravstvo i kvaliteta zdravstvenog sistema. Preciznije, želimo odgovoriti na pitanje: imaju li države koje više ulažu u zdravstvo bolje zdravstvene sisteme?

Ovo je klasičan primer komparativne analize. Cilj nam je da ispitamo razlike u kvalitetu zdravstvenih sistema u odnosu na budžete koje države izdvajaju za zdravstvo. Međutim, za razliku od standardne komparativne analize gde poredimo jasno definisane grupe, ovde se suočavamo sa kontinuiranom skalom budžetskih vrednosti.

Krenimo od samih podataka.

podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/1ea72e8ea912083492b60af7ecd07fbf/raw/7d658eb07d321126989acf774a8428dc712bfc9d/zdravstvo.csv")  
  
n <- nrow(podaci)  
cat("Broj opservacija:", n, "\n")  
  
min\_budzet <- min(podaci$budzet)  
max\_budzet <- max(podaci$budzet)  
  
cat("Min i max budzet:", min\_budzet, max\_budzet, "\n")  
  
min\_kvalitet <- min(podaci$kvalitet)  
max\_kvalitet <- max(podaci$kvalitet)  
  
cat("Min i max kvalitet:", min\_kvalitet, max\_kvalitet, "\n")  
  
X\_budzet <- mean(podaci$budzet)  
X\_kvalitet <- mean(podaci$kvalitet)  
  
cat("Prosečan budžet:", X\_budzet, "\n")  
cat("Prosečan kvalitet:", X\_kvalitet, "\n")

Broj opservacija: 60   
Min i max budzet: 292.99 9922.57   
Min i max kvalitet: 29 100   
Prosečan budžet: 4893.517   
Prosečan kvalitet: 64.58333

Imamo podatke za 60 država. Budžeti se kreću od oko 300 do skoro 10000 dolara po glavi stanovnika. Kvalitet zdravstvenog sistema varira od 29 do 100.

Da bismo bolje razumeli distribuciju budžeta, pogledajmo histogram.

par(family = "Jost")  
  
hist(podaci$budzet, breaks = 10,  
 main = "Histogram budžeta za zdravstvo",  
 xlab = "Budžet za zdravstvo (u $)",  
 ylab = "Broj država",  
 col = "#5C88DAFF")

Line 3

Crtamo histogram koji predstavlja varijablu izdeljenu u 10 intervala, tj. 9 kategorija.

|  |
| --- |
| Slika 8.1: Histogram budžeta za zdravstvo |

Distribucija budžeta za zdravstvo nije normalna. Samo jedan interval, od 3000 do 4000 dolara, ima veći broj država, dok su frekvencije u ostalim intervalima ispod 10.

Jedan pristup bi bio da „kompresujemo“ ove intervale i napravimo 2-3 intervala sa većim frekvencijama. Na primer, mogli bismo podeliti države na one koje izdvajaju ispod i iznad 5000 dolara. To bi nam omogućilo da primenimo t-test i uporedimo ove dve grupe.

par(family = "Jost")  
  
hist(podaci$budzet, breaks = 3,  
 main = "Histogram budžeta za zdravstvo",  
 xlab = "Budžet za zdravstvo (u $)",  
 ylab = "Broj država",  
 col = "#5C88DAFF")

Line 3

Delimo podatke u 3 intervala, tj. dve kategorije.

|  |
| --- |
| Slika 8.2: Histogram dve kategorije budžeta za zdravstvo |

Ipak, ovakav pristup dovodi do gubitka važnih informacija. Razmotrimo suštinu: postoji značajna razlika između država koje izdvajaju 100, 1000 ili 3000 dolara po glavi stanovnika za zdravstvo. Imamo precizne podatke o budžetima - zašto ih ne bismo iskoristili u punom obimu? Grupisanje u ovom slučaju ne samo da nije neophodno, već može biti i kontraproduktivno jer zamagljuje fine razlike između država.

## 8.2 2D analiza

Umesto grupisanja država, preciznije je analizirati svaku pojedinačno. Za svaku državu imamo dva kvantitativna podatka, što nam omogućava njihovo predstavljanje u dvodimenzionalnom koordinatnom sistemu. Na horizontalnoj osi prikazaćemo budžet za zdravstvo, a na vertikalnoj kvalitet zdravstvenog sistema. Svaka tačka u ovom prostoru predstavljaće jednu državu. Pogledajmo kako to izgleda:

par(family = "Jost")  
plot(podaci$budzet, podaci$kvalitet,  
 xlab = "Budžet za zdravstvo (u $)",  
 ylab = "Kvalitet zdravstvenog sistema",  
 main = "Budžet za zdravstvo i kvalitet zdravstvenog sistema",  
 col = "#CC0C00FF",  
 pch = 19)

Line 2

Koristimo funkciju sa dva osnovna argumenta: prvi je vrednost za x-osu, a drugi za y-osu.

|  |
| --- |
| Slika 8.3: Dijagram raspršenosti |

Ovaj grafikon se zove **dijagram raspršenosti** (eng. *scatter plot*). On prikazuje prostorni raspored dve varijable i njihov međusobni odnos. Na prvi pogled može delovati složen, ali postaje jasniji kada razumemo njegovu logiku.

Hajde da analiziramo šta nam dijagram govori:

1. U donjem levom uglu nalazi se grupa od pet država. One imaju izrazito nizak budžet za zdravstvo i loše ocene kvaliteta zdravstvenog sistema.
2. Centralni deo grafikona sadrži najveću koncentraciju tačaka, gde se budžeti kreću između 3000 i 8000 dolara, a kvalitet zdravstvenog sistema između 50 i 80.
3. U gornjem desnom uglu vidimo tri države koje izdvajaju približno 10 hiljada dolara po glavi stanovnika za zdravstvo, a kvalitet njihovih zdravstvenih sistema je između 90 i 100. Ove države prednjače po oba parametra.

Ipak, samo vizuelna analiza nije dovoljna za pouzdane zaključke o uticaju budžeta na kvalitet zdravstvenog sistema. Da bismo preciznije sagledali ovaj odnos, iskoristićemo aritmetičke sredine ovih varijabli kao referentne tačke na dijagramu.

par(family = "Jost")  
  
budzet\_sredina <- mean(podaci$budzet)  
kvalitet\_sredina <- mean(podaci$kvalitet)  
  
plot(podaci$budzet, podaci$kvalitet,  
 xlab = "Budžet za zdravstvo (u $)",  
 ylab = "Kvalitet zdravstvenog sistema",  
 main = "Budžet za zdravstvo i kvalitet zdravstvenog sistema",  
 col = "#CC0C00FF",  
 pch = 19)  
  
abline(v = budzet\_sredina,  
 col = "#5C88DAFF",  
 lty = 2,  
 lwd = 4)  
  
abline(h = kvalitet\_sredina,  
 col = "#5C88DAFF",  
 lty = 2,  
 lwd = 4)  
  
text("A", x = budzet\_sredina - 3000, y = 90, col = "#5C88DAFF", cex = 2, font = 2)  
  
text("B", x = budzet\_sredina + 3000, y = 90, col = "#5C88DAFF", cex=2, font = 2)  
  
text("C", x = budzet\_sredina + 3000, y = 40, col = "#5C88DAFF", cex=2, font = 2)  
  
text("D", x = budzet\_sredina - 3000, y = 40, col = "#5C88DAFF", cex=2, font = 2)

Line 11

Koristimo funkciju plot sa dva osnovna argumenta: prvi je vrednost za x-osu, a drugi za y-osu.

Lines 16,21

Dodajemo dve linije koje prolaze kroz aritmetičke sredine budžeta i kvaliteta zdravstvenog sistema.

|  |
| --- |
| Slika 8.4: Dijagram raspršenosti sa četiri kvadranta |

Isprekidane plave linije na grafikonu prikazuju aritmetičke sredine obe varijable. Prosečan budžet iznosi oko 4900 dolara po glavi stanovnika, dok je prosečna ocena kvaliteta približno 65. Presekom ovih linija formiraju se četiri karakteristična kvadranta: A, B, C i D. Svaki kvadrant nosi posebno značenje i pruža nam uvid u različite obrasce odnosa između budžeta i kvaliteta zdravstvenog sistema.

* Kvadrant A: Države koje imaju budžet ispod proseka, ali pokazuju iznadprosečan kvalitet zdravstvenog sistema. Ove zemlje postižu značajne rezultate uz ograničena sredstva, verovatno zahvaljujući efikasnom upravljanju ili visokoj stručnosti zdravstvenog osoblja.
* Kvadrant B: Države sa budžetom i kvalitetom zdravstvenog sistema iznad proseka. Ovo su lideri u našoj analizi - sistemi koji kombinuju visoka ulaganja sa visokim kvalitetom usluge.
* Kvadrant C: Države koje uprkos iznadprosečnom budžetu beleže ispodprosečan kvalitet zdravstvenog sistema. Ovo ukazuje na moguće probleme u efikasnosti sistema i korišćenju raspoloživih sredstava.
* Kvadrant D: Države sa budžetom i kvalitetom zdravstvenog sistema ispod proseka. Ove zemlje se suočavaju sa dvostrukim izazovom - ograničenim resursima i nezadovoljavajućim kvalitetom zdravstvene zaštite.

Fokusirajmo se na kvadrante B i D.

|  |
| --- |
| Slika 8.5: Dijagram raspršenosti sa kvadrantima B i D |

Pogledajmo šta nam kvadranti govore.

1. Države u donjem kvadrantu (D) pokazuju jasnu vezu - nizak budžet za zdravstvo prati nizak kvalitet zdravstvenog sistema.
2. Gornji kvadrant (B) pokazuje komplementarnu sliku - države sa većim budžetom za zdravstvo održavaju viši kvalitet zdravstvenog sistema.

Ovakav raspored podataka je logičan i očekivan. Kada pratimo tačke od leve ka desnoj strani, uočavamo da se one kreću od donjeg levog ka gornjem desnom uglu, što je predstavljeno plavom strelicom na grafikonu.

Ako posmatramo horizontalno kretanje od leve strane, prvo uočavamo ispodprosečne budžete. Kako se pomeramo udesno, primećujemo i vertikalno kretanje ka većim ocenama kvaliteta. Da imamo samo ove vrednosti, naš zaključak bi bio nedvosmislen: budžet i kvalitet zdravstvenog sistema su pozitivno povezani. Ipak, moramo biti oprezni - ovo je samo vizuelna procena zasnovana na delimičnom skupu podataka.

Razmotrimo sada preostala dva kvadranta zasebno.

|  |
| --- |
| Slika 8.6: Dijagram raspršenosti sa kvadrantima A i C |

Pogledajmo sada suprotnu situaciju:

1. Određene države sa manjim budžetima uspevaju da održe visok kvalitet zdravstvenog sistema (kvadrant A).
2. Pojedine države, uprkos većim izdvajanjima za zdravstvo, beleže niži kvalitet zdravstvenog sistema (kvadrant C).

Ovaj obrazac nam otkriva kako drugi faktori poput efikasnosti upravljanja sredstvima, stručnosti zdravstvenog osoblja i specifične zdravstvene politike značajno utiču na vezu između budžeta i kvaliteta zdravstva.

Kada posmatramo ove tačke, uočavamo nešto neočekivano - u nekim slučajevima povećanje budžeta prati pad kvaliteta zdravstvenog sistema. Plava strelica koja se pruža od gornjeg levog do donjeg desnog ugla jasno ilustruje ovaj kontraintuitivni odnos. Kako se krećemo desno na grafikonu, pratimo i vertikalni pad ka nižim ocenama kvaliteta.

Plave linije na prethodnim grafikonima prikazuju trend zajedničke promene dve varijable. No, vidimo da ovi trendovi ukazuju na dva različita obrasca u odnosu između budžeta i kvaliteta zdravstva. Jedan segment podataka sugeriše pozitivnu vezu, dok drugi ukazuje na negativnu povezanost. Ova dualnost nas dovodi do ključnih pitanja:

1. Da li zaista postoji veza između budžeta i kvaliteta zdravstvenog sistema?
2. Ako postoji, da li je ta veza pozitivna ili negativna?
3. Ili možda veza uopšte ne postoji?

Iako na našim grafikonima možemo pronaći dokaze za sve ove tvrdnje, oni nam ne daju jasan i konačan odgovor.

Da bismo bolje razumeli situaciju, vratimo se dijagramu raspršenosti. Da bi linija predstavljala trend zajedničke promene dve varijable, ona mora prolaziti kroz tačku koja predstavlja aritmetičke sredine obe varijable. Ova tačka, presek sredina budžeta i kvaliteta zdravstvenog sistema, opisuje centralnu tendenciju podataka. Linija treba da nam pokaže šta se dešava kada imamo podatke koji se kreću ispod i iznad tih proseka. Koje sve linije trenda možemo povući na ovom grafikonu? Pogledajmo.

|  |
| --- |
| Slika 8.7: Dijagram raspršenosti sa različtim potencijalnim linijama trenda |

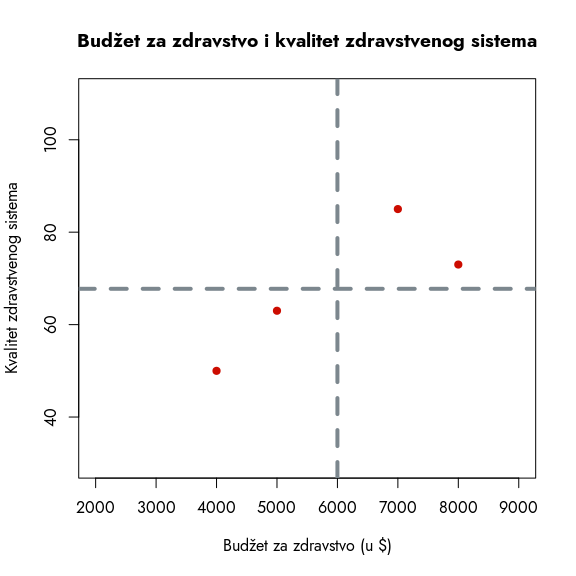
Koja od ovih linija najbolje opisuje naše podatke? Trebalo bi da to bude linija koja dobro predstavlja opšti trend podataka. U praksi, tražimo liniju koja prolazi kroz središnji deo grafikona i nalazi se na optimalnoj udaljenosti od svih tačaka. Sa 60 tačaka u našem skupu podataka, ručno pronalaženje takve linije bilo bi neprecizno i neefikasno.

Za pronalaženje optimalne linije koristimo metod **linearne regresije**. Ovaj pristup nam omogućava da matematički precizno odredimo liniju koja najbolje opisuje odnos između dve varijable. Regresiona linija nam pruža bogat skup informacija i pomaže nam da kvantifikujemo prirodu veze između varijabli koje proučavamo.

U nastavku ćemo detaljno objasniti konstrukciju regresione linije. Nakon toga, pokazaćemo kako nam ona omogućava da testiramo hipoteze o odnosu između varijabli koristeći rigorozne statističke metode.

## 8.3 Konstrukcija linije

Da bismo razumeli kako se pronalazi optimalna linija, započnimo sa jednostavnijim primerom. Razmotrimo situaciju sa samo četiri države i njihovim vrednostima budžeta i kvaliteta zdravstvenog sistema.



Dijagram raspršenosti za pojednostavljeni primer

U ovom primeru sve tačke su raspoređene tako da ukazuju na potrebu za rastućom linijom. Kao što smo ranije objasnili, ta linija mora proći kroz presek sredina budžeta i kvaliteta.

Iz geometrije znamo da su za definisanje prave potrebne dve tačke. U našem slučaju, jedna tačka je već određena - to je presek proseka budžeta i kvaliteta. Izbor druge tačke određuje **nagib** linije. Pošto još nemamo sistematičan način za određivanje optimalne pozicije druge tačke, počećemo sa dva proizvoljna nagiba. Rezultujuće linije su prikazane na grafikonu.

|  |
| --- |
| Slika 8.8: Dve linije trenda |

Ove dve linije su rastuće i prolaze kroz tačku preseka aritmetičkih sredina. Njihovi nagibi se, međutim, značajno razlikuju:

1. Puna plava linija je strmija, sa izraženijim nagibom.
2. Isprekidana linija ima blaži nagib i približava se horizontalnom položaju.

Nijedna od ovih linija ne opisuje savršeno raspored tačaka u prostoru - neke tačke su bliže jednoj, druge drugoj liniji. Koja je onda bolja opcija?

Odgovor leži u preciznom merenju udaljenosti tačaka od svake linije. Počnimo sa analizom odstupanja od isprekidane linije:

|  |
| --- |
| Slika 8.9: Udaljenost tačaka od isprekidane linije |

Plave vertikalne linije koje povezuju crvene tačke (opservacije) sa linijom trenda predstavljaju grešku, odnosno odstupanje stvarne vrednosti od predviđene vrednosti. Dužina ovih linija direktno ukazuje na preciznost modela - duže linije znače veću grešku u predviđanju.

Proverimo sada da li isprekidana linija bolje opisuje naše podatke:

|  |
| --- |
| Slika 8.10: Udaljenost tačaka od isprekidane linije |

Ovo nije komplikovano ako razumemo osnovne principe geometrije. Svaka prava linija može se predstaviti jednostavnom jednačinom: . Ova jednačina ima dva ključna parametra:

* je odsečak na y-osi (presečna tačka)
* je koeficijent nagiba (strmina linije)

Kada znamo ove parametre, možemo za svaku vrednost izračunati odgovarajuću vrednost , tako da tačka leži na liniji. U našem konkretnom primeru, plavu isprekidanu liniju definišu sledeći parametri:

Na primer, ako je , tada je vrednost y-koordinate tačke na liniji .

Izračunajmo udaljenost između prve tačke (donji levi ugao) i isprekidane linije. Empirijske vrednosti za prvu tačku su:

budzet <- c(4000, 5000, 7000, 8000)  
  
kvalitet <- c(50, 63, 85, 73)  
  
x1 <- budzet[1]  
y1 <- kvalitet[1]  
cat("Vrednost x-koordinate prve tačke:", x1, "\n")  
  
cat("Vrednost y-koordinate prve tačke:", y1, "\n")

Lines 5-6

Izdvajamo prvu vrednost iz vektora budžeta i kvaliteta.

Vrednost x-koordinate prve tačke: 4000   
Vrednost y-koordinate prve tačke: 50

Budžet za prvu tačku je 4000. Kada tu vrednost ubacimo u jednačinu linije, dobićemo procenjeni kvalitet za tu tačku na liniji:

y\_linija <- 37.75 + 0.005 \* x1  
  
cat("Procenjena vrednost y-koordinate na liniji:", y\_linija, "\n")

Line 1

Računamo y-koordinatu tačke na liniji prema regresionoj jednačini.

Procenjena vrednost y-koordinate na liniji: 57.75

Dobijamo 57.75, što je više od stvarne vrednosti (kao što vidimo na grafikonu).

Razlika, odnosno greška linije, izračunava se kao razlika između stvarne vrednosti i one koju daje linija:

razlika <- y1 - y\_linija  
cat("Razlika između opservacije i linije je:", razlika, "\n")

Razlika između opservacije i linije je: -7.75

Da bismo izračunali ukupnu grešku linije, primenićemo ovaj postupak na sve tačke. Pošto ćemo ovaj proces ponavljati više puta, napišimo funkciju koja računa ukupnu grešku za svaku liniju:

ukupna\_greska <- function(x, y, a, b) {  
 y\_linija <- a + b \* x  
 razlika <- y - y\_linija  
 return(sum(razlika^2))  
}

Line 1

Ulazni parametri funkcije su vektori x i y (vrednosti opservacija) i dva parametra linije a i b.

Line 2

Definišemo liniju i računamo y-koordinate tačaka na liniji.

Line 3

Računamo razliku između stvarne vrednosti i vrednosti na liniji.

Line 4

Funkcija vraća sumu kvadrata razlika.

|  |  |
| --- | --- |
|  | **Zašto kvadriramo grešku?**  Kada računamo grešku, koristimo kvadrat razlike između stvarne vrednosti i one koju daje linija. Zašto?   1. Kvadriranje čini sve vrednosti pozitivnim. Tako možemo sabirati odstupanja koja su i ispod i iznad linije bez poništavanja. 2. Kvadriranje daje veću težinu većim greškama. Na primer: i . Ovo znači da će veće greške imati veći uticaj na ukupnu grešku. |

Sada ćemo primeniti ovu funkciju na naše podatke. Pre nego što je vidimo na delu, napisaćemo jednačinu druge linije.

greska\_linija1 <- ukupna\_greska(budzet, kvalitet, 37.75, 0.005)  
  
greska\_linija2 <- ukupna\_greska(budzet, kvalitet, -52.5, 0.02)  
  
print("Greška isprekidane linije:")  
  
greska\_linija1  
  
print("Greška pune linije:")  
  
greska\_linija2

[1] "Greška isprekidane linije:"  
[1] 232.75  
[1] "Greška pune linije:"  
[1] 1943

R je izračunao da je greška pune linije značajno veća nego isprekidane ([Slika 8.8](#fig-dve-linije)). Gde je puna linija najviše promašila? Imamo dva izrazita primera odstupanja.

1. Kada je budžet oko 4000 dolara, stvarni kvalitet zdravstvenog sistema je značajno viši od onoga što predviđa ova linija.
2. Kada je budžet oko 8000 dolara, stvarni kvalitet zdravstvenog sistema je značajno niži od onoga što predviđa ova linija.

Ova odstupanja zajedno pokazuju da ova linija trenda ne opisuje adekvatno odnos između dve varijable.

Obe linije smo nacrtali bez jasne strategije za pronalaženje optimalne linije – one kod koje je greška najmanja. Metod koji koristimo za pronalaženje te optimalne linije naziva se **metod najmanjih kvadrata** i predstavlja temelj proste linearne regresije.

## 8.4 Prosta linearna regresija

Regresiona analiza je metod koji opisuje kako se zavisna varijabla menja pod uticajem drugih varijabli. Kroz regresionu analizu izvodimo jednačinu koja povezuje **zavisnu varijablu** sa jednom ili više **nezavisnih varijabli**. Jednostavno rečeno, pokušavamo da razumemo kako jedna varijabla reaguje na promene u drugoj varijabli.

U ovom poglavlju bavimo se **prostom linearnom regresijom** - osnovnim modelom regresione analize. Ovaj model koristi jednu nezavisnu varijablu da objasni varijacije u zavisnoj varijabli. Kada koristimo više od jedne nezavisne varijable, ulazimo u područje multivarijacione analize.

Zašto linearna regresija? Zato što pretpostavljamo da se odnos između varijabli može opisati pravom linijom. Linije trenda koje smo videli ranije predstavljaju primere regresionih linija.

Rezultat linearne regresije je regresiona jednačina. Ova jednačina ima dve primene: opisuje jačinu uticaja nezavisne varijable (opisna funkcija) i omogućava nam da predvidimo vrednosti zavisne varijable, što ćemo detaljnije obraditi u narednom poglavlju.

Matematički zapisano, ako zavisnu varijablu označimo sa , a nezavisnu sa , jednačina **regresione linije** ima sledeći oblik:

Ova jednačina je slična jednačini prave linije koju smo videli ranije, ali regresija koristi drugačiju notaciju. Ovde, označava slobodni koeficijent, dok je koeficijent nagiba. Pre nego što detaljno objasnimo kako se izračunavaju ovi koeficijenti, razmotrimo njihovu geometrijsku interpretaciju.

|  |
| --- |
| Slika 8.11: Interpretacija koeficijenata regresione linije |

Na grafikonu vidimo interpretaciju oba koeficijenta. Prvi, , poznat kao slobodni koeficijent ili **odsečak na y-osi**, pokazuje vrednost zavisne varijable kada je nezavisna varijabla jednaka 0.

Konkretno, ako je budžet za zdravstvo 0 dolara, očekivana ocena kvaliteta zdravstvenog sistema bi bila , što je u ovom slučaju približno 27. Iako su ove vrednosti hipotetičke i retko imaju praktičnu primenu (nijedna država nema budžet za zdravstvo od 0 dolara), one nam daju polaznu tačku za razumevanje odnosa između varijabli.

Drugi koeficijent, , predstavlja nagib linije. Za razliku od koji označava tačku, opisuje ugao pod kojim se linija pruža. Pogledajte sliku ispod koja detaljno prikazuje geometrijsku interpretaciju ovog koeficijenta.

|  |
| --- |
| Slika 8.12: Interpretacija koeficijenata regresione linije |

Ovaj koeficijent nam otkriva jednostavnu istinu: kada se vrednost nezavisne varijable poveća za 1, vrednost zavisne varijable će se povećati za . Možemo skalirati ovaj odnos множењем sa proizvoljnim brojem. U našem primeru, ako se budžet poveća za 1000 dolara, kvalitet zdravstvenog sistema će se, u proseku, povećati za poena. Kroz ovaj koeficijent dobijamo jasan i precizan opis odnosa između dve varijable. Postavlja se pitanje: šta možemo očekivati da će se desiti sa zavisnom varijablom kada promenimo nezavisnu?

Fokusirajmo se na reč „očekivanje“. U kontekstu našeg primera, ovo znači da će se kvalitet u proseku povećati za 6.8 poena kada se budžet poveća za 1000 dolara. Naglasak je na „u proseku“ - neke države će odstupati od ovog obrasca, ali ovo je generalni trend koji možemo očekivati.

Ovime se vraćamo na početnu komparativnu dimenziju problema. Koeficijent nagiba nam omogućava da odgovorimo na konkretno pitanje: kolika je prosečna razlika u kvalitetu zdravstvenog sistema između država koje imaju budžet od 5000$ i 6000$? Odgovor je 6.8 poena. Ova informacija predstavlja srž regresione analize i jasno pokazuje njen komparativni karakter.

## 8.5 Metod najmanjih kvadrata

Sada dolazimo do suštinskog pitanja: kako odrediti ova dva koeficijenta? Za prostu linearnu regresiju, rešenje nam nudi elegantni **metod najmanjih kvadrata**.

Ovaj metod nam omogućava da precizno odredimo koeficijent . Već znamo da regresiona linija mora proći kroz tačku sredine, ali kako odabrati optimalan nagib? Metod najmanjih kvadrata ima jasan kriterijum - biramo onaj koeficijent koji minimizuje sumu kvadrata odstupanja između stvarnih vrednosti i vrednosti predviđenih linijom. Matematički rečeno, tražimo koeficijent koji daje najmanju moguću vrednost **sume kvadrata greške** (eng. *Sum of Squared Errors*, SSE).

|  |  |
| --- | --- |
|  | **Optimizacija i funkcija cilja**  Većina statističkih metoda u svojoj osnovi funkcioniše po principu optimizacije određene funkcije. U našem slučaju, optimizujemo funkciju greške regresije. Ova funkcija zavisi od koeficijenata regresione linije i naš cilj je da je minimizujemo. Takvu funkciju nazivamo **funkcijom cilja** (eng. *objective function*).  Statistika je nastala kao grana primenjene matematike, i mnogi statistički metodi su zapravo matematički postupci optimizacije. To podrazumeva korišćenje numeričkih metoda, integrala i diferencijalnih jednačina za pronalaženje rešenja.  U naprednijim oblastima statistike i računarstva, poput mašinskog učenja i veštačke inteligencije, ovaj pristup je zamenjen **algoritmima** koji se zasnivaju na iterativnom pristupu. Ovi algoritmi su računarski efikasniji i brži, ali su istovremeno manje transparentni i teži za razumevanje. Ipak, osnovni princip ostaje isti: pronaći vrednosti parametara koji minimizuju funkciju cilja kako bi se ostvario željeni cilj modela. |

Ključno je razumeti da postupak minimizacije ima jedinstveno analitičko rešenje. Šta to znači u praksi? To znači da kroz određene matematičke operacije (parcijalni izvodi greške u odnosu na regresione koeficijente) dolazimo do jedinstvenog rešenja. Drugim rečima, postoji precizna formula koja nam omogućava izračunavanje koeficijenata i tako da greška regresije bude minimalna.

|  |
| --- |
| Slika 8.13: Minimizacija greške regresije |

Nećemo ulaziti u detalje, ali osnovna logika pronalaženja optimalne vrednosti je jasno prikazana na slici iznad. Grafikon prikazuje sumu kvadrata greške (SSE) kao **funkciju** različitih koeficijenata nagiba. Pošto radimo sa prostom regresijom i pravom linijom, funkcija uvek ima ovakav oblik. Cilj metoda najmanjih kvadrata je da pronađe minimum ove krive - da se spusti do najniže tačke i odredi vrednost koja daje najmanju moguću grešku. Do tog minimuma dolazimo formulom:

Donji deo razlomka predstavlja varijansu nezavisne varijable, dok gornji deo predstavlja **kovarijansu**, označenu kao . Kovarijansa meri zajedničku varijabilnost dve varijable - drugim rečima, govori nam koliko dosledno promena jedne varijable prati promenu druge.

Razmotrimo situaciju koju ilustruje [Slika 8.4](#fig-kvadranti). Imamo države sa budžetom za zdravstvo oko 2000 dolara. Te države se nalaze u kvadrantu D, sa budžetima i ocenama kvaliteta ispod proseka. Kada se pomerimo desno na X-osi, vidimo države sa budžetom oko 8000 dolara. Od četiri takve države, dve su u kvadrantu C, jedna u kvadrantu B, a jedna na granici B i C. Iako smo budžet povećali četiri puta, dve države su i dalje ispod proseka, jedna je na nivou proseka, a jedna iznad proseka u pogledu kvaliteta zdravstvenog sistema. Ovo jasno pokazuje da povećanje budžeta **nije dosledno** praćeno povećanjem kvaliteta zdravstvenog sistema.

Kovarijansa kvantifikuje doslednost zajedničkih promena dve varijable. Savršena doslednost postoji kod **determinističkog** odnosa. Uzmimo primer fiksnog kursa između dve valute - 100 evra uvek vredi 11800 dinara. Udvostručenje iznosa u evrima neizbežno znači udvostručenje iznosa u dinarima jer je odnos deterministički. Međutim, kod **stohastičkog** odnosa, udvostručenje u jednoj varijabli ne garantuje proporcionalno povećanje u drugoj.

Kovarijansa upravo meri konzistentnost stohastičkog odnosa između varijabli. Pogledajmo formulu:

Na prvi pogled, kovarijansa deluje složeno, ali je suština jednostavna - ona meri udaljenost tačaka u različitim kvadrantima ([Slika 8.4](#fig-kvadranti)). U formuli za kovarijansu, brojilac predstavlja proizvod odstupanja opservacija obe varijable od njihovih aritmetičkih sredina. Na našem grafikonu sa kvadrantima, to je proizvod udaljenosti svake tačke od centra grafikona (preseka aritmetičkih sredina). Šta nam ovaj proizvod govori?

Opservacije u kvadrantima B i D daju pozitivan doprinos kovarijansi. U kvadrantu B, obe udaljenosti su pozitivne, što rezultira pozitivnim proizvodom. U kvadrantu D, obe udaljenosti su negativne, ali njihov proizvod je opet pozitivan. Nasuprot tome, opservacije u kvadrantima A i C umanjuju kovarijansu. Konačna vrednost kovarijanse predstavlja ishod „nadmetanja“ između ova četiri kvadranta. Mogući su sledeći ishodi:

1. Dominiraju pozitivne vrednosti (B i D kvadranti). Više opservacija se nalazi u ovim kvadrantima, a njihove udaljenosti su ukupno veće nego u drugim kvadrantima. Rezultat je relativno visoka pozitivna kovarijansa.
2. Prevladavaju negativne vrednosti (A i C kvadranti). Ishod je relativno niska pozitivna kovarijansa.
3. Izjednačen rezultat. Pozitivne i negativne udaljenosti se približno poništavaju, što vodi ka kovarijansi bliskoj nuli.

Na narednom grafikonu vidimo kako se ovi odnosi manifestuju u našem primeru budžeta i kvaliteta zdravstvenog sistema.

|  |
| --- |
| Slika 8.14: Dijagram raspršenosti sa prikazom dorpinosa opservacija kovarijansi. Veličina tačke ukazuje na doprinos opservacije kovarijansi. |

Brojevi su veliki. Pozitivna suma dominira, što potvrđuje da kvadranti B i D imaju veći uticaj na model. Koji kvadranti nose najveću težinu? Iz veličine tačaka na grafikonu jasno je da su to države sa niskim budžetom i niskim kvalitetom zdravstva, kao i tri države u gornjem desnom uglu koje kombinuju velike budžete sa visokom ocenom kvaliteta.

Kovarijansu možemo jednostavno izračunati u R-u koristeći funkciju cov.

kovarijansa <- cov( podaci$budzet, podaci$kvalitet)  
  
print("Kovarijansa:")  
  
print(kovarijansa)

[1] "Kovarijansa:"  
[1] 21669.57

Vidimo da je kovarijansa veliki, pozitivan broj. Sam podatak o pozitivnom predznaku je koristan, ali numerička vrednost nam ne govori mnogo. Slično kao kod varijanse, kovarijansa ima jedinicu mere koja nema praktičnu interpretaciju.

Hajde da se vratimo na formulu za koeficijent nagiba. Pozitivna kovarijansa i dominacija kvadranata B i D vode ka pozitivnom koeficijentu nagiba. To znači da će regresiona linija prolaziti kroz ove dominantne kvadrante.

Sada možemo jednostavno izračunati vrednost koeficijenta nagiba koji minimizuje sumu kvadrata greške.

varijansa\_X <- var(podaci$budzet)  
  
b1 <- kovarijansa / varijansa\_X  
  
print("Koeficijent nagiba (b1):")  
print(round(b1,4))

[1] "Koeficijent nagiba (b1):"  
[1] 0.0032

Ovo je vrednost koju smo uočili na grafikonu minimizacije sume kvadrata greške: . Kroz ovaj proces smo otkrili da je koeficijent nagiba, glavni regresioni koeficijent, zapravo odnos kovarijanse i varijanse nezavisne varijable. On kvantifikuje koliki je zajednički varijabilitet u odnosu na varijabilitet nezavisne varijable. Što je apsolutna vrednost koeficijenta veća, to je veza između varijabli jača i konzistentnija.

Kada imamo koeficijent nagiba, jednostavno možemo izračunati i drugi regresioni koeficijent. Znamo da regresiona linija mora proći kroz tačku koja predstavlja aritmetičke sredine obe varijable (). Primenom ovih vrednosti u jednačini regresione linije, dobijamo:

Ovu jednostavnu formulu za drugi koeficijent možemo odmah da primenimo u R-u.

b0 <- X\_kvalitet - b1 \* X\_budzet  
  
print('Slobodni koeficijent:')  
print(round(b0,2))

Line 1

Izračunavamo slobodni koeficijent pomoću formule koja koristi aritmetičke sredine obe varijable.

[1] "Slobodni koeficijent:"  
[1] 49.07

Dakle, naša regresiona jednačina izgleda ovako:

Na dijagramu rasprešnosti možemo najzad ucrtati liniju za koju znamo da sigurno predstavlja liniju koja je prolazi kroz tačku proseka obe varijable i najmanje je udaljena od svih drugih tačaka.

par(family = "Jost")  
  
b0 <- 49.07  
b1 <- 0.0032  
  
plot(podaci$budzet, podaci$kvalitet,  
 xlab = "Budžet za zdravstvo (u $)",  
 ylab = "Kvalitet zdravstvenog sistema",  
 main = "Dijagram raspršenosti sa regresionom linijom",  
 col = "#CC0C00FF",  
 pch = 19)  
  
abline(a = b0, b = b1, col = "#5C88DAFF", lwd = 4)  
  
text(2000, 90, expression(bold(paste("Y = 49.07 + 0.0032X"))), col = "#5C88DAFF", cex = 0.9, font = 2)

Lines 3-4

Postavljamo vrednosti koeficijenata i .

Line 11

Crtamo dijagram raspršenosti.

Line 13

Crtamo regresionu liniju sa odgovarajućim koeficijentima.

|  |
| --- |
| Slika 8.15: Dijagram raspršenosti sa regresionom linijom |

## 8.6 Regresioni model

Regresiona linija proizilazi iz uzorka i opisuje samo informacije koje u njemu nalazimo. Međutim, ako želimo da izvedemo šire zaključke i opišemo odnos između dva fenomena na nivou populacije, potreban nam je **regresioni model**.

Model možemo zapisati na sledeći način:

Kao što vidite, postoje značajne razlike u načinu kako formulišemo regresionu liniju. Prvo, koristimo grčka slova i , što označava da se regresioni model odnosi na nepoznate **parametre** populacije. Slobodni koeficijent i koeficijent nagiba su sada nepoznate vrednosti o kojima ćemo, kao i u prethodnim poglavljima, pokušati da donesemo zaključke uz pomoć statističkog testa. Preostaje nam - nepoznati parametar koji opisuje ukupnu **grešku regresije**.

Ranije smo naglasili da je odnos koji opisuje regresija stohastički. To znači da regresiona linija, odnosno njen nagib , neće savršeno opisati odstupanja svih tačaka od linije. Ta odstupanja su zapravo greške regresije . One predstavljaju sve faktore koji utiču na kvalitet zdravstvenog sistema, a koje nismo obuhvatili budžetom. Ove greške su stohastičke - slučajne su i nezavisne od budžeta, rezultat svih ostalih faktora koji oblikuju kvalitet zdravstvenog sistema.

Na nivou uzorka, lako smo izračunali grešku regresije. Međutim, na nivou populacije to nije moguće jer ne znamo sve opservacije koje postoje u „velikom svetu“. Upravo zato je nepoznati parametar populacije.

Pred nama su dva ključna koraka. Prvi je testiranje regresione hipoteze na osnovu rezultata iz uzorka - želimo utvrditi da li je regresioni model **statistički značajan** na nivou populacije. Drugi je **dijagnostika modela** - proveravamo da li regresioni model adekvatno opisuje odnos između varijabli i da li su pretpostavke o grešci ispunjene.

### 8.6.1 Testiranje statističke značajnosti regresionog modela

Vratimo se našem primeru. Na nivou uzorka dobili smo sledeću regresionu liniju:

Koeficijenti i iz uzorka su naše najbolje procene parametara populacije i . Kritičan korak je testiranje statističke značajnosti koeficijenta nagiba, jer on definiše prirodu veze između varijabli. Vrednost bliska 0 ukazuje na odsustvo linearne veze (efekti iz kvadranata A i C neutrališu efekte iz kvadranata B i D).

U našem primeru, dobili smo koeficijent koji nije tačno 0, ali je blizu te vrednosti. Ovo nas dovodi do poznatog problema - kako utvrditi da li je dobijena vrednost dovoljno različita od nule da bismo mogli doneti pouzdan zaključak? Odgovor leži u primeni statističkog testa ili intervala poverenja. Razmotrićemo oba pristupa.

Prvo definišimo nultu hipotezu. Ona postulira odsustvo efekta, što u kontekstu regresije znači da je koeficijent nagiba jednak nuli:

Alternativna hipoteza tvrdi da efekat postoji, odnosno da koeficijent nagiba nije jednak 0. U ovom dvosmernom obliku ne razmatramo smer uticaja, već samo njegovo postojanje.

Kako testiramo ovu hipotezu? Potrebna nam je test statistika - standardizovana vrednost koeficijenta nagiba. Do nje dolazimo deljenjem koeficijenta nagiba njegovom **standardnom greškom**. Standardna greška koeficijenta nagiba meri preciznost naše procene. Manja greška ukazuje na precizniju procenu.

Formula za standardnu grešku glasi:

Hajde da pokušamo da razumemo ovu formulu. Imenilac predstavlja koren sume kvadrata odstupanja nezavisne varijable - to je suština formule za varijansu. Manja odstupanja povećavaju grešku regresije. Zašto je to tako? Kada nezavisna varijabla nema dovoljno varijabiliteta, ne može dobro objasniti promene zavisne varijable.

Zamislite ekstremnu situaciju: sve države u uzorku imaju identičan budžet. Imali bismo 60 tačaka vertikalno poređanih na istoj X koordinati. Nijedna regresiona linija ne bi mogla smisleno objasniti varijacije u kvalitetu zdravstva - jednostavno nema osnove za to. Greška regresije bi bila ogromna (tehnički, beskonačna zbog deljenja nulom). Zato je ključno da naše nezavisne varijable pokazuju značajan varijabilitet.

Brojilac predstavlja grešku regresije. Izračunavamo je deljenjem sume kvadrata odstupanja od regresione linije (SSE) sa stepenima slobode. Model ima dva nepoznata parametra ( i ), što znači da je broj stepeni slobode .

Dakle, **standardna greška koeficijenta b1 ()** predstavlja odnos greške regresione linije i varijabiliteta nezavisne varijable. Mi zapravo merimo grešku regresije u odnosu na preciznost merenja . Za nisku vrednost standardne greške potrebna su nam dva elementa: regresiona linija koja minimizira odstupanja (što automatski dobijamo metodom najmanjih kvadrata) i nezavisna varijabla sa dovoljnim varijabilitetom.

Izračunajmo standardnu grešku koeficijenta nagiba za naš primer. Potrebni su nam sledeći elementi:

* : suma kvadrata greške regresije
* : standardna greška regresije (rezidualna standardna greška)
* : standardna devijacija nezavisne varijable

Prvo, suma kvadrata greške regresije:

b0 <- 49.07  
b1 <- 0.0032  
  
SSE <- ukupna\_greska(podaci$budzet, podaci$kvalitet, b0, b1)  
  
cat("Suma kvadrata greške regresije:", round(SSE, 0), "\n")

Line 4

Računamo SSE koristeći ranije definisanu funkciju ukupna\_greska.

Suma kvadrata greške regresije: 12654

Sada izračunajmo grešku regresije:

s <- sqrt(SSE / (nrow(podaci) - 2))  
  
cat("Greška regresije:", round(s, 2), "\n")

Line 1

Greška regresije je koren SSE podeljenog brojem stepeni slobode.

Greška regresije: 14.77

Standardnu devijaciju nezavisne varijable dobijamo funkcijom sd. Sada imamo sve za izračun standardne greške koeficijenta nagiba:

standardna\_greska <- s / sqrt(sum((podaci$budzet - X\_budzet)^2))  
cat("Standardna greška koeficijenta nagiba:", round(standardna\_greska, 5), "\n")

Line 1

Primenjujemo formulu za standardnu grešku - imenilac je koren sume kvadrata odstupanja X od njegove srednje vrednosti.

Standardna greška koeficijenta nagiba: 0.00074

Za testiranje koristimo t-test. Ako je nulta hipoteza tačna (nema regresionog efekta), vrednost koeficijenta nagiba je 0. U tom slučaju, vrednosti nagiba u uzorku variraju oko nule, a te standardizovane varijacije prate t-raspodelu sa n-2 stepena slobode.

Jednostavnije rečeno - moguće je da regresija na nivou populacije ne postoji, a da mi u uzorku dobijemo malu (pozitivnu ili negativnu) vrednost zbog slučajnosti. T-testom proveravamo da li je vrednost koeficijenta nagiba iz uzorka statistički značajno različita od nule.

Test statistika je:

t\_statistika <- b1 / standardna\_greska  
  
cat("t-statistika:", round(t\_statistika, 2), "\n")

Line 1

Računamo t-statistiku deljenjem koeficijenta nagiba sa standardnom greškom.

t-statistika: 4.35

Vrednost 4.35 se nalazi daleko u desnom repu t-raspodele. Za donošenje zaključka potrebna nam je p-vrednost. Naš model ima 58 stepeni slobode (60 opservacija - 2 nepoznata parametra). P-vrednost računamo pomoću funkcije pt koja izračunava vrednost t-raspodele. Pošto testiramo dvosmernu hipotezu, zanima nas verovatnoća da je t-statistika veća od 4.31 ili manja od -4.31.

p\_vrednost <- 2 \* pt(-abs(t\_statistika), df = n - 2)  
  
cat("p-vrednost:", round(p\_vrednost, 3), "\n")

Line 1

Računamo p-vrednost koristeći funkciju pt.

p-vrednost: 0

P-vrednost je toliko niska da je R zaokružuje na 0. Pred nama je jak dokaz za odbacivanje nulte hipoteze. Možemo zaključiti da postoji statistički značajan regresioni efekat na nivou populacije. Drugim rečima, linearna zavisnost između budžeta i kvaliteta zdravstvenog sistema nije ograničena samo na naš uzorak, već se može generalizovati na širu populaciju.

### 8.6.2 Interval poverenja za koeficijent nagiba

Sada kada imamo sve potrebne elemente, konstruisaćemo interval poverenja za koeficijent nagiba. Struktura ovog intervala je slična onoj iz prethodnog poglavlja, ali se ovde fokusiramo konkretno na i njegovu standardnu grešku.

Formula za interval poverenja glasi:

Hajde da ovo primenimo u praksi. Izabraćemo nivo poverenja od 95% i izračunati raspon vrednosti za koeficijent nagiba koristeći R.

t\_vrednost <- qt(0.975, df = n - 2)  
  
interval\_poverenja <- c(b1 - t\_vrednost \* standardna\_greska, b1 + t\_vrednost \* standardna\_greska)  
  
cat("Interval poverenja za koeficijent nagiba:", round(interval\_poverenja, 4), "\n")

Line 1

Izračunavamo kritičnu vrednost t-raspodele za 95% nivo poverenja.

Line 3

Konstruišemo interval poverenja za koeficijent nagiba.

Interval poverenja za koeficijent nagiba: 0.0017 0.0047

Zamislite da smo uzeli 100 različitih uzoraka iz populacije i za svaki izračunali regresioni model. U 95 od tih 100 slučajeva, koeficijent nagiba bi se našao između 0.0017 i 0.0046. Ovo nije samo teorijski koncept - možemo ga jasno prikazati na grafiku. Pogledajmo vizuelnu reprezentaciju:

|  |
| --- |
| Slika 8.16: Interval poverenja za koeficijent nagiba. |

Pogledajmo pažljivo plavu osenčenu oblast na ovom grafikonu. Ona predstavlja opseg mogućih regresionih linija, uzimajući u obzir 95% interval poverenja i veličinu našeg uzorka. Ova vizualizacija nam elegantno prikazuje neizvesnost u našoj proceni regresione linije.

|  |  |
| --- | --- |
|  | **Razlika između linije i modela**  U čemu je razlika između regresione linije i regresionog modela? Pogledajte pažljivo gornji grafikon - on elegantno ilustruje tu razliku. Plava neprekidna linija je regresiona linija koju smo dobili iz uzorka. Ali primetite plavu osenčenu oblast oko nje - to je regresioni model u punom smislu. Model obuhvata mnogo više od same linije jer uključuje sve neizvesnosti vezane za regresionu liniju. Te neizvesnosti su prikazane kroz interval poverenja.  Dakle, model nam pruža potpuniju sliku time što uzima u obzir varijabilnost koju sama linija ne može da predstavi. Ovo je ključni uvid: regresioni model nije samo jedna linija, već čitav prostor mogućih linija koje su konzistentne sa našim podacima. Kada shvatimo ovu razliku, jasnije vidimo zašto je model moćniji alat od proste linije - on nam govori ne samo šta očekujemo, već i koliko smo sigurni u ta očekivanja. |

Primetićemo da se mnoge opservacije nalaze van plave oblasti mogućih regresionih linija. Čak i kada u obzir uzmemo neizvesnost koeficijenata, tačno predviđanje svih vrednosti kvaliteta zdravstvenog sistema ostaje izazov. Ovo je značajan uvid. Da bismo razumeli kako ove greške utiču na pouzdanost našeg modela, moramo sprovesti detaljnu dijagnostiku. Ovaj korak je ključan za razumevanje stvarnih mogućnosti i ograničenja regresionog modela pred nama.

## 8.7 Dijagnostika proste linearne regresije

Do sada smo primenili metod najmanjih kvadrata da bismo našli optimalnu regresionu liniju. Testiranjem hipoteza potvrdili smo da je nagib statistički značajan i konstruisali interval poverenja oko linije.

Sada prelazimo na analizu grešaka regresione linije - **reziduale**. Reziduali predstavljaju razlike između stvarnih i predviđenih vrednosti zavisne varijable. Centralna pretpostavka regresionog modela je da su **reziduali slučajni**. Drugim rečima, tretiramo ih kao čist šum - rezultat nasumičnih faktora koji onemogućavaju savršeno predviđanje. U našem primeru, to su svi nekontrolisani faktori koji utiču na kvalitet zdravstva, a nisu obuhvaćeni modelom. Pošto je ova greška slučajna, očekujemo da bude nezavisna od objašnjavajuće varijable i da prati normalnu raspodelu.

Zašto baš normalnu? Ako su reziduali zaista slučajni, prirodno je očekivati da budu simetrično raspoređeni iznad i ispod linije. Većina reziduala trebalo bi da bude umereno udaljena od linije, dok bi ekstremna odstupanja trebalo da budu retka. Ovaj obrazac odgovara upravo normalnoj raspodeli. Kada reziduali prate normalnu raspodelu, to je snažan indikator da naš model dobro opisuje podatke i da greška ne narušava njegovu validnost.

Hajde da ispitamo da li naši reziduali zadovoljavaju ove pretpostavke.

|  |  |
| --- | --- |
|  | **Zašto nenormalnost ugrožava validnost modela?**  Nenormalna distribucija reziduala je jasan signal da naš regresioni model ne hvata nešto važno. To je kao da nam podaci šapuću: „Propustili ste nešto bitno“. Najčešće, ovo ukazuje na postojanje trećeg faktora koji utiče na naš model, a koji nismo uzeli u obzir.  Ovo direktno narušava našu ključnu pretpostavku o slučajnosti greške. Kada greška nije slučajna, već sistemska posledica izostavljenog faktora, naš model gubi pouzdanost. Ne možemo se više osloniti na njegova predviđanja, a posebno ne na koeficijent nagiba koji bi trebalo da nam pokaže jačinu veze između varijabli.  Kako rešiti ovaj problem? Odgovor leži u konstrukciji složenijeg modela koji će obuhvatiti više faktora. Međutim, to zalazi u područje naprednije statistike i prevazilazi okvire ovog udžbenika.  Drugi mogući uzrok nenormalnosti je nelinearan odnos između varijabli. Zamislite da pokušavate da opišete krivudavu liniju pravom - to jednostavno neće dati dobre rezultate. U takvim slučajevima, linearna regresija nije pravi alat za naš problem. Kako to možemo proveriti? Dijagram raspršenosti je naš prvi indikator - jasna zakrivljenost u rasporedu tačaka sugeriše da treba razmotriti nelinearne modele regresije. |

Hajde da pogledamo kako izgleda distribucija reziduala u našem slučaju.

b0 <- 49.07  
b1 <- 0.0032  
  
par(family = "Jost")  
reziduali <- podaci$kvalitet - (b0 + b1 \* podaci$budzet)  
  
hist(reziduali,  
 breaks = 15,  
 col = adjustcolor("#5C88DAFF", alpha.f = 0.7),  
 border = "white",  
 main = "Histogram reziduala",  
 xlab = "Reziduali",  
 xlim = c(-50, 50),  
 ylim = c(0,0.05),  
 ylab = "Frekvencija",  
 prob = TRUE)  
  
lines(density(reziduali), col = "#CC0C00FF", lwd = 3)  
  
curve(dnorm(x, mean = mean(reziduali), sd = sd(reziduali)),  
 add = TRUE, col = "black", lty = 2, lwd = 3)  
  
legend("topright",  
 legend = c("Histogram", "Kriva gustine", "Normalna raspodela"),  
 fill = c(adjustcolor("#5C88DAFF", alpha.f = 0.7), NA, NA),  
 col = c(NA, "#CC0C00FF", "black"),  
 lty = c(NA, 1, 2),  
 lwd = c(NA, 2, 2),  
 border = c(NA, NA, NA),  
 cex = 0.8)

Lines 1-2

Postavljamo vrednosti koeficijenata i .

Line 5

Računamo rezidualne vrednosti

Line 16

Pravimo histogram reziduala

Line 18

Dodajemo krivu gustine reziduala

Line 21

Dodajemo krivu normalne raspodele sa istom srednjom vrednošću i standardnom devijacijom kao i reziduali

Line 30

Legenda u gornjem desnom uglu

|  |
| --- |
| Slika 8.17: Histogram i kriva gustine reziduala |

Šta nam otkriva ovaj grafikon? Crvena kriva distribucije reziduala odstupa od isprekidane normalne distribucije. Kako se razlikuju? Kriva reziduala je znatno spljoštenija. Ovo ukazuje da je greška više ravnomerno raspoređena u intervalu, bez ekstremnih vrednosti u repovima. Međutim, to takođe znači da nema koncentracije reziduala oko nule. Vizuelno, kriva gustine reziduala je relativno simetrična, slično normalnoj raspodeli.

Visoka spljoštenost ne narušava značajno validnost zaključka, ali upozorava da greška nije koncentrisana oko nule, već je ravnomerno raspoređena. Posledično, greška predviđanja može biti veća nego što bismo očekivali kod normalne raspodele.

Drugi ključni dijagnostički pristup je analiza odnosa između predviđenih vrednosti i reziduala. U idealnom slučaju, ne bi trebalo da postoji veza između ovih veličina. Hajde da ispitamo taj odnos. Na x-osi ćemo prikazati predviđene vrednosti zavisne varijable iz regresione jednačine, a na y-osi reziduale. Ako su reziduali zaista slučajni, ne bi trebalo da uočimo nikakav obrazac. Prisustvo obrasca bi ukazivalo da je greška sistemski povezana sa nekim aspektom zavisne varijable koji nismo obuhvatili modelom.

b0 <- 49.07   
b1 <- 0.0032   
  
regresija\_kvalitet <- b0 + b1 \* podaci$budzet  
  
par(family = "Jost")  
  
plot(regresija\_kvalitet, reziduali,  
 xlab = "Predviđene vrednosti",  
 ylab = "Reziduali",  
 main = "Dijagram predviđenih vrednosti i reziduala",  
 col = "#5C88DAFF",  
 pch = 19)  
  
abline(h = 0, col = "#CC0C00FF", lwd = 2)  
  
fit <- lm(reziduali ~ regresija\_kvalitet)  
  
abline(fit, col = "black", lwd = 2)  
  
legend("topright",  
 legend = c("Reziduali", "Linija y = 0", "Linearna regresija"),  
 col = c(adjustcolor("#5C88DAFF", alpha.f = 0.7), "#CC0C00FF", "black"),  
 pch = c(19, NA, NA),  
 lty = c(NA, 1, 1),  
 lwd = c(NA, 2, 2),  
 cex = 0.8)

Line 4

Računamo predviđene vrednosti zavisne varijable

Line 13

Prikazujemo dijagram predviđenih vrednosti i reziduala

Line 15

Dodajemo horizontalnu liniju na nuli

Line 17

Pravimo regresiju između predviđenih vrednosti i reziduala

Line 19

Dodajemo regresionu liniju

|  |
| --- |
| Slika 8.18: Dijagram predviđenih vrednosti i reziduala |

Na grafikonu uočavamo da su tačke, odnosno parovi (predviđena vrednost, rezidual), raspoređeni gotovo nasumično i ne formiraju nikakav očigledan obrazac. Recimo, ne vidimo da se najveća negativna greška javlja kada je predviđena vrednost najmanja. Kada bismo podelili grafikon na kvadrante, videli bismo da su tačke ravnomerno raspoređene u sva četiri kvadranta.

Kako da potvrdimo ovu tvrdnju matematički? Na dijagramu, pored crvene horizontalne linije koja označava nulu, vidimo i crnu liniju. Ta crna linija predstavlja linearnu regresiju između predviđenih vrednosti i reziduala. Njen oblik nam govori sve - kada bi postojala značajna veza između predviđenih vrednosti i reziduala, linija bi imala primetan nagib. No, ona je gotovo savršeno horizontalna, što matematički potvrđuje odsustvo povezanosti između ovih veličina. To je jak dokaz da naš model adekvatno opisuje podatke.

|  |  |
| --- | --- |
|  | **Regresija za dijagnostiku regresije?**  Možda se pitate kako je moguće koristiti novu regresiju (predviđene vrednosti i reziduali) za dijagnostiku postojeće regresije? Ovo je prirodno pitanje. Odgovor je jednostavan - ne kreiramo novi model, već samo vizuelno prikazujemo trend između predviđenih vrednosti i reziduala pomoću regresione linije.  U sledećem poglavlju ćemo istražiti naprednije metode za analizu veze između predviđenih vrednosti i reziduala, ali trenutno je ovaj pristup sasvim dovoljan. Primetićete da smo postupak dobijanja regresione linije sveli na samo dve linije koda. Detaljno objašnjenje funkcije lm sledi na kraju ovog poglavlja, a njenu punu primenu ćete upoznati kroz ostatak udžbenika. |

Ove metode predstavljaju osnovne oblike dijagnostike regresionog modela, s naglaskom na reziduale. Važno je razumeti da dijagnostika nije metod za donošenje zaključaka poput testiranja hipoteza. Umesto toga, ona nam otkriva dublju strukturu modela i njegove granice. Kako budemo napredovali kroz gradivo, upoznaćemo dodatne tehnike za merenje efikasnosti regresionog modela.

## 8.8 Regresioni model u R-u

Do sada smo korak po korak gradili regresioni model i vršili dijagnostiku koristeći osnovne R funkcije. R nam ipak nudi efikasnije alate za rad sa regresijom. Najvažnija među njima je funkcija lm (eng. *linear model*), koja nam omogućava brzu konstrukciju regresionog modela. Pogledajmo kako ova funkcija radi na našem primeru:

model <- lm(kvalitet ~ budzet, data = podaci)  
  
model

Line 1

Konstruišemo regresioni model koristeći funkciju lm. Prvi argument je formula regresije, a drugi argument je skup podataka.

Line 3

Ispisujemo osnovne rezultate regresionog modela.

Call:  
lm(formula = kvalitet ~ budzet, data = podaci)  
  
Coefficients:  
(Intercept) budzet   
 49.073523 0.003169

Funkcija lm koristi formulu za definisanje regresionog modela. U našem primeru, kvalitet ~ budzet označava da predviđamo kvalitet zdravstvenog sistema pomoću budžeta. R primenjuje ovakvu notaciju za sve modele: levo od znaka ~ nalazi se zavisna varijabla, a desno je formula modela, koja je u ovom slučaju jednostavno budzet. Ova R notacija direktno odgovara regresionoj jednačini koju smo ranije definisali.

Nakon što prikažemo model, dobijamo ključne komponente regresione jednačine - njene koeficijente. Intercept predstavlja (slobodni član), dok ispod njega, uz oznaku budzet, nalazimo .

Za detaljniju analizu modela koristimo funkciju summary.

summary(model)

Call:  
lm(formula = kvalitet ~ budzet, data = podaci)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-29.523 -11.245 -0.986 9.639 36.324   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 4.907e+01 4.073e+00 12.05 < 2e-16 \*\*\*  
budzet 3.170e-03 7.354e-04 4.31 6.4e-05 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 14.77 on 58 degrees of freedom  
Multiple R-squared: 0.2426, Adjusted R-squared: 0.2295   
F-statistic: 18.58 on 1 and 58 DF, p-value: 6.404e-05

Hajde da proučimo rezultate našeg modela. Svaka linija ispisa nosi važne informacije:

1. Reziduali: Brojevi koji pokazuju koliko naš model odstupa od stvarnosti. Medijana od -0.986 ukazuje da je model prilično precizan. Najveće odstupanje je 36, a najmanje -29.523. Ovi brojevi nam daju jasan uvid u opseg grešaka modela.
2. Koeficijenti: Ovo je centralni deo našeg modela. Fokusirajmo se na red označen sa budzet. Tu nalazimo vrednost koeficijenta, standardnu grešku, t-statistiku i p-vrednost. Ovi brojevi potvrđuju rezultate koje smo ranije dobili manuelnim računanjem.
3. Rezidualna standardna greška (s): Vrednost koja kvantifikuje prosečno odstupanje predviđenih od stvarnih vrednosti. Manja greška ukazuje na precizniji model.
4. Koeficijent determinacije: Mera koja kvantifikuje koliko naš model objašnjava varijabilitet podataka. O ovome detaljnije u nastavku.
5. F-statistika: Test koji meri značajnost celokupnog modela. Ovaj koncept ćemo detaljnije obraditi u narednim poglavljima.

Zašto je ovaj pristup koristan? R nam je dao objekat model koji sadrži sve ključne informacije o našoj regresiji. Ovaj objekat je moćan alat za dalju analizu i testiranje.

Na primer, možemo lako dobiti reziduale koristeći funkciju residuals(model). Ovo nam omogućava brzu dijagnostiku i vizualizaciju rezultata.

reziduali\_model <- residuals(model)  
  
cat("Prvih 5 reziduala: \n")  
print(reziduali\_model[1:5])

Prvih 5 reziduala:   
 1 2 3 4 5   
 -2.111039 -11.042392 -1.563762 7.672095 -2.899660

Vredi napomenuti da funkcija residuals ne zahteva direktno podatke iz matrice, već koristi kompletan regresioni model. Ovaj model već sadrži sve potrebne informacije za izračunavanje reziduala.

Takav pristup nas prirodno vodi ka još jednom bitnom koraku u dijagnostici reziduala - analizi **Kukove distance**. Kukova distanca je precizan alat koji identifikuje opservacije sa ekstremnim rezidualima. Ovakve opservacije mogu narušiti normalnost reziduala i iskriviti odnos između reziduala i predviđenih vrednosti. Formula za Kukovu distancu je jednostavna:

U ovoj formuli, su reziduali, je greška regresije, a elementi dijagonale matrice , poznate kao **matrica uticaja**. Matrica meri koliko pojedinačne opservacije utiču na regresioni model. Kako? Proces je jasan: izračunamo regresioni model bez svake pojedinačne opservacije i uporedimo ga sa originalnim modelom. Ako postoji velika razlika, ta opservacija ima značajan uticaj na model. Kukova distanca precizno kvantifikuje ovaj uticaj.

Opservacije sa visokom Kukovom distancom zaslužuju detaljniju analizu. One mogu značajno narušiti validnost modela i biti izvor sistemskih problema. Izračunajmo Kukove distance za naš model i analizirajmo rezultate.

kukove\_distance <- cooks.distance(model)  
  
cat("Prvih 6 Kukovih distanci: \n")  
print(kukove\_distance[1:6])

Line 1

Izračunavamo Kukove distance koristeći funkciju cooks.distance.

Prvih 6 Kukovih distanci:   
 1 2 3 4 5 6   
0.0002087541 0.0063536274 0.0001704526 0.0025945777 0.0009470525 0.0182036078

Dobijene vrednosti su relativno male. Da bismo ih ispravno protumačili, potrebno je izračunati kritičnu vrednost za Kukove distance. Ova vrednost se računa kao , gde je broj opservacija, a broj nezavisnih varijabli. Opservacije čija je Kukova distanca iznad ove kritične vrednosti smatraju se značajno uticajnim za model.

U našem slučaju, s obzirom da koristimo prostu regresiju, imamo i . Izračunajmo kritičnu vrednost:

Sada možemo identifikovati sve opservacije čija je Kukova distanca veća od 0.069. Te opservacije zahtevaju detaljniju analizu jer mogu bitno uticati na parametre našeg modela.

granica <- 4 / (nrow(podaci) - 2)  
  
problemi <- podaci[kukove\_distance > granica, ]  
  
cat("Opservacije sa velikom Kukovom distancom: \n")  
  
print(problemi)

Line 1

Izračunali smo gornju granicu za Kukove distance

Line 3

Identifikovali smo opservacije koje prelaze tu granicu

Opservacije sa velikom Kukovom distancom:   
 budzet kvalitet  
34 9922.57 51  
37 8637.97 100  
46 8467.98 100  
51 559.90 29  
56 8344.78 46

Pronašli smo pet država koje značajno odstupaju od modela: one pod brojevima 34, 37, 46, 51 i 56. Ovde vidimo zanimljiv obrazac - neke od njih imaju visoke budžete i visok kvalitet zdravstvenog sistema, dok druge beleže izrazito niske vrednosti za obe varijable.

Za bolju vizuelnu analizu, konstruisaćemo dijagram raspršenosti sa dodatnom dimenzijom. Na njemu će veličina svake tačke odgovarati njenoj Kukovoj distanci - što je tačka veća, to je njen uticaj na model značajniji. Problematične opservacije biće

|  |
| --- |
| Slika 8.19: Dijagram raspršenosti sa regresionom linijom i Kukovim distancama |

Tačke koje smo identifikovali nalaze se na obodima dijagrama i značajno odstupaju od regresione linije. Ova situacija nije jedinstvena za regresiju - svi metodi koji koriste kovarijansu su osetljivi na ekstremne vrednosti.

Kako pristupiti ovim odstupajućim opservacijama? Prvi korak je provera tačnosti podataka. Greške pri unosu su česte, posebno u anketnim istraživanjima gde ispitanik može napraviti jednostavnu grešku - na primer, uneti „3“ umesto „33“ za godine starosti, što dramatično menja regresioni model.

Često istraživači eliminišu opservacije sa visokom Kukovom distancom. Ovo je metodološki problematično iz dva razloga: veštački poboljšava regresioni model i narušava slučajnost uzorka. Ispravan pristup je drugačiji - ako otkrijemo grešku u podacima, ispravljamo je i ponovno računamo model.

U našem primeru, identifikacija odstupajućih opservacija (tri sa najvećom Kukovom distancom su označene na dijagramu) poboljšava naše razumevanje zavisnosti. Država 51 pokazuje izrazito niske vrednosti obe varijable, dok opservacije 34 i 56 demonstriraju da visok budžet nije dovoljan uslov za visok kvalitet zdravstvenog sistema. Ovi uvidi su ključni za razumevanje složenosti odnosa koji analiziramo.

## 8.9 Objašnjenje i predviđanje

Interpretacija regresionog modela ima dva ključna aspekta: objašnjenje i predviđanje.

1. Objašnjenje: Regresioni model nam otkriva kako se jedna varijabla menja u zavisnosti od druge. U našem primeru jasno vidimo da kvalitet zdravstvenog sistema raste sa povećanjem budžeta. Kroz ovaj uvid saznajemo: - Da li postoji zajednička promena dve varijable? - U kom smeru se odvija ta promena? - Koliko je ta promena intenzivna?

Model nam takođe omogućava da identifikujemo koje opservacije značajno oblikuju ovaj odnos i da procenimo da li naši podaci pružaju dovoljno informacija za razumevanje veze između varijabli. Ovo nas vodi ka dubljoj teorijskoj analizi.

1. Predviđanje: Zamislite situaciju gde predstavljamo analizu donosiocima odluka. Njihovo ključno pitanje će verovatno biti: „Ako povećamo budžet za zdravstvo na 8500 dolara po glavi stanovnika godišnje, kakav kvalitet zdravstvenog sistema možemo očekivati?“. Regresiona analiza na ovo pitanje daje precizan, ali ograničen odgovor.

Regresija pokazuje izuzetnu preciznost u predviđanju prosečnih vrednosti zavisne varijable. To znači da pouzdano možemo proceniti prosečan kvalitet zdravstvenog sistema za zemlje sa određenim budžetom. Na primer, države koje izdvajaju 8500 dolara po glavi stanovnika u proseku dostižu ocenu kvaliteta oko 73. Kritično je razumeti - ovo ne znači da će svaka država koja poveća budžet na 8500 dolara nužno dostići ocenu 73.

Za procenu preciznosti regresije u predviđanju koristimo intervale poverenja. Počnimo sa intervalom poverenja za prosečnu vrednost regresije. Formula je složena, ali vredi je razumeti:

Hajde da razumemo komponente ove formule:

1. - predviđena vrednost regresije
2. - kritična vrednost t-raspodele za 95% interval poverenja
3. - standardna greška regresije
4. - broj opservacija
5. - vrednost nezavisne varijable za koju pravimo predviđanje (u našem primeru 8500)
6. - srednja vrednost nezavisne varijable

Izraz pod korenom je ključan za razumevanje preciznosti regresije. On nam pokazuje da je regresija najpreciznija kada je blizu . To znači da regresija najpouzdanije predviđa vrednosti koje su bliske proseku nezavisne varijable. Što se više udaljavamo od proseka, preciznost predviđanja opada.

Ovo je fundamentalan uvid u prirodu regresije - ona nije pouzdan alat za predviđanje ekstremnih ili neuobičajenih vrednosti. Regresija je najsnažnija u domenu prosečnih vrednosti, gde daje najpreciznije rezultate.

U narednom koraku ćemo vizuelno prikazati ovaj interval poverenja da bismo bolje sagledali njegove karakteristike.

|  |
| --- |
| Slika 8.20: Intervali poverenja za prosečnu vrednost i indivdiualne predikcije |

Crvena linija predstavlja regresionu liniju. Crvene isprekidane linije označavaju intervale poverenja za **srednju predviđenu vrednost**. Primetite kako se ove linije savijaju sa leve i desne strane grafikona - nisu prave. To ilustruje važan koncept: interval je najuži u centru grafikona (u oblasti prosečnog budžeta) i postepeno se širi kako se udaljavamo od sredine. Ovo empirijski potvrđuje ranije objašnjenje - regresija daje najpreciznije rezultate u oblasti proseka. Osobina direktno proizlazi iz člana u formuli intervala poverenja.

Na grafikonu uočavamo i plave isprekidane linije. One su značajno udaljenije od centra i predstavljaju **interval predviđanja za pojedinačna opažanja**. Ovaj interval je primetno širi od intervala poverenja za srednju vrednost, što jasno ukazuje na manju preciznost pri predviđanju pojedinačnih opservacija.

Kada želimo predvideti specifičnu vrednost kvaliteta zdravstvenog sistema za određenu državu, neophodno je konstruisati interval poverenja oko proseka. Konstrukcija se zasniva na pretpostavci da opservacije oko predviđene vrednosti prate t-raspodelu. Praktično, to znači da će neke države imati rezultate iznad proseka, druge ispod, ali ne očekujemo ekstremna odstupanja. Na osnovu ove pretpostavke formiramo **interval poverenja za predviđenu pojedinačnu vrednost**. Ovaj interval je širi od intervala poverenja za srednju vrednost, odražavajući inherentnu neizvesnost pri predviđanju pojedinačnih slučajeva.

Razlika između ova dva intervala poverenja leži u dodatku „+1“ ispod korena. Ovaj mali dodatak značajno proširuje interval, ukazujući na znatno veću grešku individualnog predviđanja. Zašto? Kada predviđamo pojedinačnu vrednost, moramo uzeti u obzir ne samo grešku u proceni srednje vrednosti, već i slučajnu grešku - uticaj svih faktora koje nismo obuhvatili modelom. Širi interval pokazuje da na kvalitet zdravstva utiču mnogi faktori osim budžeta. Posledično, interval postaje toliko širok da gubi praktičnu upotrebljivost za individualna predviđanja.

Uzmimo konkretan primer sa grafikona. Za državu koja izdvaja 6000 dolara po glavi stanovnika za zdravstvo, naš model predviđa kvalitet zdravstvenog sistema između 35 i 95. Raspon je toliko širok da ne pruža nikakvu korisnu informaciju - zaključak koji bismo mogli izvesti i bez regresione analize.

Ovo nije mana regresionog modela, već prirodno ograničenje proste regresije. U stvarnom svetu društvenih nauka, retko koji fenomen možemo objasniti samo jednim prediktorom. Za preciznija predviđanja potrebne su naprednije metode multivarijacione analize koje mogu obuhvatiti više relevantnih faktora istovremeno.

## 8.10 Zaključak

Ovo poglavlje je zahtevalo strpljenje i fokus jer smo prošli kroz niz fundamentalnih koncepata. Razmotrimo glavni cilj regresione analize: opisivanje linearne zavisnosti jedne varijable od druge. U našem primeru, istraživali smo kako kvalitet zdravstvenog sistema zavisi od budžeta za zdravstvo.

Prevedeno na jezik geometrije, naš zadatak je bio pronaći pravu liniju koja optimalno opisuje odnos između dve varijable. Ova regresiona linija minimizuje odstupanja od svih opservacija. Za njeno pronalaženje primenili smo metod najmanjih kvadrata koji sistemski smanjuje grešku regresije.

Regresiona linija proizlazi iz podataka uzorka. Međutim, nas zanima zavisnost na nivou cele populacije, pa moramo proveriti da li linija iz uzorka verno predstavlja taj širi odnos. To postižemo testiranjem hipoteza i konstruisanjem intervala poverenja.

Nakon potvrde značajnosti modela na nivou populacije, prelazimo na dijagnostiku. Ona se temelji na analizi reziduala - odstupanja stvarnih od predviđenih vrednosti. Ključna pretpostavka je slučajnost reziduala. Ako su reziduali zaista slučajni, model je pouzdan. To proveravamo kroz tri ključne analize: normalnost reziduala, njihov odnos sa predviđenim vrednostima i Kukove distance.

U završnom delu razmotrili smo prediktivnu moć modela. Prosta linearna regresija efikasno predviđa prosečne vrednosti zavisne varijable, ali pokazuje ograničenja u predviđanju pojedinačnih vrednosti.

Svi ovi elementi zajedno otkrivaju složenost odnosa između dve kvantitativne varijable. Postoji još jedan važan aspekt ovog odnosa koji opisuje regresioni model - korelacija, koju ćemo detaljno istražiti u narednom poglavlju.

|  |  |
| --- | --- |
|  | **Lična karta metoda: prosta linearna regresija**  **Šta radi?** Testira uticaj jedne kvantitativne varijable na drugu.  **Kada se koristi?** Kada želimo da predvidimo ili objasnimo vrednost jedne kvantitativne varijable pomoću druge.  **Koliko varijabli imamo?** Dve kvantitativne varijable merene na intervalnom ili racio mernom nivou. U posebnim slučajevima, varijable mogu biti ordinalne.  **Kako glase hipoteze?** Nulta hipoteza tvrdi da je koeficijent nagiba jednak nuli, odnosno da ne postoji uticaj nezavisne varijable na zavisnu. Alternativna hipoteza tvrdi suprotno - da postoji uticaj.  **Test statistika:** T-test koristimo da proverimo da li je koeficijent nagiba različit od nule. Test statistiku izračunavamo deljenjem procenjenog koeficijenta nagiba njegovom standardnom greškom.  gde je procenjeni koeficijent nagiba, a njegova standardna greška.  **Kako računamo p-vrednost?** P-vrednost predstavlja dvostuku površinu ispod t-distribucije od vrednosti t-statistike do beskonačnosti. Drugim rečima, merimo kolika je verovatnoća da dobijemo vrednost t-statistike koja je jednaka ili ekstremnija od izračunate, pod pretpostavkom da je nulta hipoteza tačna. |

## 8.11 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1**  U ovom zadatku analiziramo odnos između vremena provedenog na društvenim mrežama i akademskog uspeha studenata. Naši podaci sadrže informacije o prosečnom broju sati koje studenti dnevno provode na društvenim mrežama i njihovim prosečnim ocenama. Cilj nam je da pomoću proste linearne regresije ispitamo kako vreme na društvenim mrežama utiče na akademski uspeh.  Postupak ćemo izvesti kroz tri koraka:   1. Najpre ćemo kreirati dijagram raspršenosti koji vizuelno prikazuje odnos između ove dve varijable. 2. Zatim ćemo, korak po korak, izračunati koeficijente regresione linije koristeći metod najmanjih kvadrata. 3. Na kraju ćemo ucrtati dobijenu regresionu liniju na dijagram raspršenosti i interpretirati rezultate.   Za početak, učitajmo podatke:  podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/17d9f2b7432ba72c926859924e5a9753/raw/b36866440b70d6098dc2081da332a63a312e8682/dm-prosek.csv") |

|  |  |
| --- | --- |
|  | **Zadatak 2**  Testirajtee hipotezu o postojanju linearne zavisnosti između vremena provedenog na društvenim mrežama i prosečne ocene studenata. Nivo značajnosti je 0.05. Bitno je da razumete i ispravno protumačite rezultate testa. |

|  |  |
| --- | --- |
|  | **Zadatak 3 \***  Testirajte hipotezu da vreme provedeno na društvenim mrežama ima **negativan uticaj** na prosečnu ocenu studenata. Postavite nivo značajnosti na 0.05. Protumačite rezultate testa i analizirajte kako se razlikuju od rezultata iz Zadatka 2. |

|  |  |
| --- | --- |
|  | **Zadatak 4 \***  Uradite punu dijagnostiku dobijenog regresionog modela.   1. Normalnost reziduala 2. Odnos između reziduala i predviđenih vrednosti 3. Kukove distance   Na osnovu sva tri metoda dijagnostike, napišite jedan paragraf teksta koji objašnjava aspekte odnos korišćenja društvenih mreža i akademskog uspeha u odnosu na ono što nam govori dijagnostika reziduala. |

|  |  |
| --- | --- |
|  | **Zadatak 5 \*\***  Zamislite da se od vas traži (iako nema previše smisla) obrnuta analiza - da predviđate vreme provedeno na društvenim mrežama na osnovu prosečne ocene studenata.  Koristeći lm funkciju, konstruišite regresioni model i interpretirajte rezultate. Razmislite kako se koeficijenti razlikuju u odnosu na prethodnu analizu. Kako interpretirate rezultate ovog modela? |

# 9. Korelacija

## 9.1 Međusobne regresije

U prethodnom poglavlju detaljno smo se upoznali sa regresionom analizom i opisali uticaj nezavisne varijable na zavisnu. Međutim, u mnogim istraživačkim situacijama nije uvek jasno koja varijabla treba da bude nezavisna, a koja zavisna. Za takve slučajeve postoji efikasan alat za istraživanje međusobnog odnosa dve kvantitativne varijable.

Krenimo od poznatog - proste linearne regresije. Razmotrimo jednostavan primer sa dve kvantitativne varijable. Istraživanje se fokusiralo na svakodnevne navike adolescenata koristeći podatke iz mobilnih uređaja. Prikupljanje ovakvih podataka zahteva metodološki precizan pristup. Da bi se izbegla narušavanja privatnosti, istraživači ispitanicima daju nove mobilne telefone sa instaliranim samo onim aplikacijama koje su predmet istraživanja. Ispitanici ne koriste svoje lične naloge. Uređaje koriste 10-30 dana, nakon čega ih vraćaju istraživačima.

Analiziraćemo podatke o korišćenju telefona tokom jednog dana od strane 80 ispitanika. Varijable su:

1. netflix: broj sati dnevno koje ispitanici provedu gledajući Netflix
2. koraci: broj koraka koje ispitanici pređu dnevno

Logično je pretpostaviti da su ove dve aktivnosti negativno povezane - teško je istovremeno gledati seriju i šetati. No, postavlja se suštinsko pitanje: koja aktivnost utiče na koju? Da li više vremena na Netflix-u rezultira manjom fizičkom aktivnošću, ili manja fizička aktivnost vodi ka dužem gledanju serija? Ovo je klasičan primer problema uzročnosti u statistici.

Pitanje uzroka i posledice ovde nije trivijalno. Možemo razmišljati u oba smera: vreme provedeno na Netflix-u moglo bi uticati na broj koraka - više gledanja znači manje vremena za kretanje. Istovremeno, niska fizička aktivnost može voditi ka većoj konzumaciji streaming sadržaja - kada smo neaktivni, verovatnije je da ćemo slobodno vreme ispuniti gledanjem serija.

Kako pristupiti ovom problemu? Kao i u prethodnom poglavlju, počećemo od vizuelne analize podataka.

podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/a08682e874ac33291538aa57091b5dac/raw/a62d76f267481158a49964c6736668a5feec5aa4/netflix-koraci.csv")  
  
par(family = "Jost")  
  
plot(podaci$netflix, podaci$koraci,  
 xlab = "Netfliks (u satima)",  
 ylab = "Broj koraka",  
 main = "Netfliks i broj koraka",  
 col = "#CC0C00FF",  
 pch = 19)

Line 1

Učitavanje podataka iz CSV fajla

Line 10

Kreiranje dijagrama raspršenosti

|  |
| --- |
| Slika 9.1: Dijagram raspršenosti. Netflix je na x-osi, a broj koraka na y-osi. |

Dijagram raspršenosti nam pruža početni uvid u odnos između ove dve varijable. Jasno je vidljiva tendencija opadajuće linije trenda, što ukazuje na negativnu povezanost između vremena provedenog uz Netflix i broja pređenih koraka.

Za precizniju analizu ove veze, primenićemo regresionu analizu. Ispitaćemo kako vreme provedeno uz Netflix utiče na broj pređenih koraka. U R-u, ovu regresiju zapisujemo formulom: koraci ~ netflix.

par(family = "Jost")  
  
regresija1 <- lm(koraci ~ netflix, data = podaci)  
  
plot(podaci$netflix, podaci$koraci,  
 xlab = "Netfliks (u satima)",  
 ylab = "Broj koraka",  
 main = "Netfliks i broj koraka",  
 col = "#CC0C00FF",   
 pch = 19)  
  
abline(regresija1, col = "#5C88DAFF", lwd = 4)  
  
text(4, 15000,  
 labels = paste("y =", round(regresija1$coefficients[1], 2),  
 "-", abs(round(regresija1$coefficients[2], 2)),"x"),  
 font = 2, cex=1.1)

Line 3

Konstruišemo regresioni model putem funkcije lm

Line 10

Kreiramo dijagram raspršenosti

Line 12

Dodajemo regresionu liniju sa koeficijentima iz objekta regresija1

|  |
| --- |
| Slika 9.2: Regresiona linija. Netfliks je na x-osi, a broj koraka na y-osi. |

Kako razumeti ovaj rezultat? Koeficijent iznosi približno 13753 koraka. To predstavlja vrednost kada je jednako 0. U našem primeru, to znači da kada ispitanici uopšte ne gledaju Netflix, u proseku prelaze oko 13700 koraka.

Koeficijent iznosi -344. Ovaj koeficijent prikazuje promenu u za jediničnu promenu u . Bitno je uočiti da je vrednost negativna. Konkretno, to znači da svaki dodatni sat proveden u gledanju Netflix-a smanjuje prosečan broj koraka za 344.

Umesto dalje analize ovog regresionog modela, fokusiraćemo se na inverzni odnos između ove dve varijable. Ispitaćemo regresionu liniju koja pokazuje kako broj koraka utiče na vreme provedeno u gledanju Netflix-a: netflix ~ koraci.

par(family = "Jost")  
  
regresija2 <- lm(netflix ~ koraci, data = podaci)  
  
plot(podaci$koraci, podaci$netflix,  
 xlab = "Broj koraka",  
 ylab = "Netfliks (u satima)",  
 main = "Broj koraka i Netfliks",  
 col = "#CC0C00FF",   
 pch = 19)  
  
abline(regresija2, col = "#5C88DAFF", lwd = 4)   
  
text(5000, 17,  
 labels = paste("y =", round(regresija2$coefficients[1], 2),  
 "-", abs(round(regresija2$coefficients[2], 5)), "x"),  
 font = 2, cex=1.1)

|  |
| --- |
| Slika 9.3: Regresiona linija. Broj koraka je na x-osi, a Netfliks na y-osi. |

Pogledajmo sada drugačiju interpretaciju regresionih koeficijenata. Koeficijent iznosi 15.2 - ovo predstavlja vrednost kada je jednako 0. U našem kontekstu, ako ispitanik ne napravi nijedan korak, očekujemo da će u proseku gledati 15.2 sati Netflix-a.

Koeficijent je izuzetno mali, približno -0.00042. Praktično, to znači da svaki dodatni korak smanjuje vreme provedeno uz Netflix za 0.00042 sata. Prevedeno u sekunde, svaki dodatni korak smanjuje vreme gledanja za 1.5 sekunde.

Radi jasnijeg razumevanja, razmotrimo efekat na nivou od 1000 koraka. Za svakih 1000 dodatnih koraka, vreme provedeno uz Netflix se u proseku smanjuje za 0.42 sata.

Kada analiziramo međusobni odnos ove dve varijable, uočavamo sledeće:

* Nagib regresione linije koja opisuje zavisnost broja sati Netflix-a od broja koraka iznosi -344.46
* Nagib regresione linije koja opisuje zavisnost broja koraka od broja sati Netflix-a iznosi -0.00042

Ovde se prirodno nameće pitanje: kako bismo mogli da kombinujemo ova dva pokazatelja u **jedinstveni indikator**? Potrebna nam je mera koja će dati jasan odgovor na dva ključna pitanja:

1. Postoji li **međusobna povezanost** između ove dve varijable?
2. Ako postoji, kakav je njen smer i intenzitet?

Problem bismo mogli rešiti uzimanjem aritmetičke sredine dva regresiona koeficijenta. Ali tu nailazimo na prepreku - vrednosti koeficijenata se drastično razlikuju. Prvi je oko 100.000 puta veći od drugog. Zašto? Razlog je u različitim jedinicama mere. Dnevni broj koraka varira od 0 do 20.000, dok se sati gledanja Netflix-a kreću od 0 do 24. Zbog toga dobijamo značajno različite vrednosti regresionih koeficijenata, zavisno od izbora zavisne varijable.

Kako prevazići ovaj problem različitih vrednosti? Rešenje je u geometrijskoj sredini. Označimo prvi regresioni koeficijent kao (uticaj sati Netflix-a na broj koraka), a drugi kao (uticaj broja koraka na sate Netflix-a). Objedinjeni indikator njihovog međusobnog odnosa označićemo sa . Primenom geometrijske sredine dobijamo:

Implementacijom ove formule u R-u dobijamo sledeći rezultat:

bNK <- regresija1$coefficients[2]  
bKN <- regresija2$coefficients[2]  
r = sqrt(bNK\* bKN)  
  
cat("Rezultat je: ",round(r,4), "\n")

Line 1

Izdvajanje regresionog koeficijenta za netflix iz prvog modela

Line 2

Izdvajanje regresionog koeficijenta za koraci iz drugog modela

Line 3

Izračunavanje koeficijenta korelacije kao geometrijske sredine

Rezultat je: 0.3792

Šta nam ovaj rezultat govori? Kako da interpretiramo vrednost 0.3792? Odgovor na ova pitanja postaje jasan kada detaljnije razmotrimo pokazatelj koji smo izračunali. Ovaj pokazatelj je poznat kao **Pirsonov koeficijent korelacije**.

## 9.2 Pirsonov koeficijent korelacije

Pirsonov koeficijent korelacije je mera koja pokazuje **intenzitet i smer** povezanosti između dve kvantitativne varijable. Za razliku od regresije, koja analizira uticaj jedne varijable na drugu, korelacija se usmerava na **međusobnu povezanost**. Ne zanimaju nas objašnjenje i predviđanje, već merenje jačine veze između dva fenomena koje smo opisali kvantitativnim varijablama.

Suštinska odlika korelacije je njen intenzitet. Šta to konkretno znači? Pirsonov koeficijent korelacije uzima vrednosti od -1 do 1. Ako zanemarimo predznak, apsolutna vrednost koeficijenta (između 0 i 1) određuje intenzitet povezanosti. Vrednosti bliske 0 ukazuju na odsustvo povezanosti, dok vrednosti koje teže ka 1 ili -1 pokazuju snažnu vezu. Ovo vam verovatno zvuči poznato iz prethodnog poglavlja.

Pirsonov koeficijent predstavlja elegantno rešenje - standardizovanu verziju kovarijanse. Kod kovarijanse imamo samo jednu referentnu tačku: vrednosti oko 0 koje ukazuju na odsustvo doslednih zajedničkih promena (balansirana distribucija između 4 kvadranta). Problem nastaje jer za ostale vrednosti kovarijanse nemamo jasne smernice za tumačenje. Upravo zato nam je potreban koeficijent korelacije.

Pomoću kovarijanse i standardnih devijacija, kovarijansu možemo pretvoriti u standardizovanu meru koja ima jasnu interpretaciju. Korelacija zadržava predznak kovarijanse, što nam pokazuje smer veze - pozitivan ili negativan. Intenzitet veze određujemo prema apsolutnoj vrednosti koeficijenta korelacije:

* : slaba povezanost
* : umerena povezanost
* : jaka povezanost

U našem primeru sa Netflix-om i brojem koraka, dobili smo r = -0.38. Kako ovo tumačimo?

* Negativan predznak ukazuje na obrnutu vezu: više vremena na Netflix-u znači manje koraka
* Vrednost 0.38 govori o umerenoj povezanosti

Ovaj rezultat je intuitivan: ljudi koji više vremena provode gledajući serije, manje se kreću. Umerena korelacija nam govori da postoji srednji nivo doslednosti između vremena provedenog na Netflix-u i fizičke aktivnosti. Ali šta zapravo znači ta „doslednost“ u ovom kontekstu?

Korelacija i regresija su komparativni metodi. Konzistentnost merimo poređenjem podataka unutar uzorka. Zamislite da nasumično izaberemo dva ispitanika:

* Prvi gleda Netflix manje od proseka
* Drugi gleda Netflix više od proseka

Kolika je verovatnoća da će prvi imati više koraka od proseka, a drugi manje? Što je ta verovatnoća veća, veza je konzistentnija i korelacija jača.

Slabe i umerene korelacije ukazuju na postojanje drugih faktora koji utiču na vezu. U našem primeru, to mogu biti pol ili starost ispitanika. Recimo, dvadesetogodišnjaci možda uspevaju da kombinuju natprosečno gledanje Netflix-a sa natprosečnom fizičkom aktivnošću, za razliku od starijih ispitanika. Ovakvi „skriveni“ faktori smanjuju korelaciju jer umanjuju direktnu vezu između gledanja Netflix-a i broja koraka.

|  |  |
| --- | --- |
|  | **Poreklo povezanosti između varijabli**  Pri analizi povezanosti dve varijable, tri osnovna pristupa nam pomažu da razumemo njihov odnos:   1. **Zajednički uzrok**: Varijable mogu biti povezane zbog treće varijable koja utiče na obe. Na primer, starost utiče i na vreme provedeno na Netflix-u i na fizičku aktivnost. Ovo je čest scenario u društvenim naukama i glavni razlog zašto moramo biti oprezni pri interpretaciji korelacija. 2. **Refleksivna veza**: Ponekad su varijable povezane jer mere različite aspekte istog fenomena. Na primer, visina i težina su povezane jer obe mere fizičke karakteristike tela. U ovom slučaju, korelacija pokazuje koliko konzistentno merimo isti fenomen kroz različite indikatore. 3. **Kauzalni lanac**: Varijable mogu biti povezane jer su deo istog uzročno-posledičnog lanca. Na primer: vreme učenja → ocene → mogućnosti za zaposlenje. Bitno je napomenuti da sama korelacija ne ukazuje na direktnu uzročnost.   Razumevanje izvora povezanosti je ključno za preciznu interpretaciju korelacije i izbegavanje pogrešnih zaključaka. Svaka statistička analiza mora početi od jasnog razumevanja prirode veze između varijabli. |

Da bismo bolje razumeli različite intenzitete korelacije, korisno je videti kako izgledaju dijagrami raspršenosti za različite vrednosti koeficijenta korelacije.

|  |
| --- |
| Slika 9.4: Primeri različitih nivoa korelacije |

Gornji grafikoni prikazuju različite nivoe korelacije:

1. **Jaka korelacija** (|r| > 0.7): Tačke se grupišu oko jasne linije, formirajući izražen obrazac. Vrednost jedne varijable pouzdano predviđa vrednost druge varijable.
2. **Umerena korelacija** (0.3 ≤ |r| ≤ 0.7): Tačke prate prepoznatljiv trend, ali pokazuju veće odstupanje od linije. Povezanost je primetna, ali manje izražena nego kod jake korelacije.
3. **Slaba korelacija** (|r| < 0.3): Tačke pokazuju značajno rasipanje, bez jasnog obrasca. Veza između varijabli postoji, ali je slabog intenziteta.

Korelacija može biti i negativna - u tom slučaju dijagrami raspršenosti pokazuju opadajući trend. Intenzitet korelacije tumačimo na isti način, pri čemu negativan predznak ukazuje na inverznu vezu između varijabli. Do kraja poglavlja objasnićemo logiku iza graničnih vrednosti 0.7 i 0.3.

Vratimo se Pirsonovom koeficijentu korelacije. Prethodno smo ga definisali kao geometrijsku sredinu dva regresiona koeficijenta.

Podsetimo se formule za regresione koeficijente:

Hajde da vidimo šta će se desiti ako u formuli za zamenimo regresione koeficijente njhovim formulama.

Ako korenujemo ovu formulu, dobijamo:

Ovo je formula za izračunavanje empirijskog Pirsonovog koeficijenta korelacije. Formula je praktična, sa jasnom interpretacijom.

Korelacija predstavlja količnik između:

* **kovarijanse** - mere zajedničkog varijabiliteta dve varijable
* **proizvoda standardnih devijacija** - mere individualnog varijabiliteta dve varijable

Jednostavnije rečeno, korelacija pokazuje koliko se dve varijable konzistentno menjaju zajedno u odnosu na njihove individualne promene. To je odnos zajedničkog varijabiliteta (kovarijansa) i individualnog varijabiliteta (standardne devijacije).

|  |  |
| --- | --- |
|  | **Ko je bio Pirson?**  Karl Pirson (eng. Karl Pearson, 1857-1936) je bio engleski matematičar i biostatističar koji je značajno doprineo razvoju moderne statistike. Kao jedan od osnivača statističke nauke, razvio je mnoge metode koje su i danas temelj statističke analize. Njegovo ime vezujemo za brojne koncepte, uključujući i **Pirsonov koeficijent korelacije** koji smo upravo objasnili.  Interesantna istorijska činjenica je da Pirson nije originalni tvorac ovog koeficijenta - on ga je učinio široko poznatim i primenjivim. Koncept je prvobitno razvio njegov mentor Frensis Galton krajem 19. veka. Galton je eksperimentisao sa terminologijom, krenuvši od pojma „reverzija“, preko „regresije“, da bi konačno usvojio termin „korelacija“. Sama matematička formula je još starija - prvi put se pojavila u radu francuskog fizičara Ogista Bravea.  Posebno je važno istaći da su ovi naučnici matematički dokazali da količnik kovarijanse i proizvoda standardnih devijacija uvek ima vrednost između -1 i 1. Iako Pirson nije originalni tvorac ovog koeficijenta, njegov doprinos leži u njegovoj sistematizaciji i širokoj primeni u praksi. Ovo je karakterističan primer kako u nauci često pripisujemo zasluge onima koji su uspešno implementirali i popularizovali određene metode, a ne nužno njihovim originalnim autorima. |

Ova formula nam omogućava da analiziramo korelaciju između dve varijable nezavisno od regresionih modela. Upravo zato i pravimo razliku između regresije i korelacije. Regresija nam pokazuje uticaj jedne varijable na drugu, dok korelacija opisuje njihovu međusobnu povezanost. Međutim, sa čisto statističkog stanovišta, koeficijenti regresije i korelacije su veoma slični - oba predstavljaju odnose između kovarijanse i standardnih devijacija. Ključna razlika je u tome što regresija koristi odnos kovarijanse i standardne devijacije **nezavisne** varijable, dok korelacija uzima u obzir standardne devijacije obe varijable.

Vrednost koeficijenta nam govori o smeru i intenzitetu veze između varijabli. Kada je , govorimo o negativnoj korelaciji - varijable se menjaju u suprotnim smerovima, baš kao u našem Netflix primeru gde više sati gledanja serija prati manji broj koraka. Kada je , imamo pozitivnu korelaciju - varijable se menjaju u istom smeru. Vrednosti blizu 0 ukazuju na odsustvo linearne veze. Vrednosti 0.3 i 0.7 predstavljaju korisne granice za procenu jačine korelacije.

Bitno je istaći da se ovi zaključci odnose samo na naš uzorak. Za izvođenje zaključaka o vezi između varijabli na nivou cele populacije, neophodno je primeniti metode statističkog zaključivanja kroz testiranje hipoteza i intervale poverenja.

## 9.3 Test za Pirsonov koeficijent korelacije

Kao i u slučaju regresije, nakon što izračunamo koeficijent korelacije na nivou uzorka, potrebno je proveriti njegovu značajnost na nivou populacije. Nulta i alternativna hipoteza su:

Pri čemu je koeficijent korelacije dve varijable na nivou čitave populacije. Nulta hipoteza označava odsustvo korelacije u populaciji, dok alternativna potvrđuje da korelacija postoji na nivou populacije.

Za proveru ovih hipoteza koristimo test, pri čemu t-statistika predstavlja odnos između koeficijenta korelacije izračunatog na nivou uzorka i standardne greške koeficijenta korelacije:

Donji deo razlomka predstavlja standardnu grešku koeficijenta korelacije. Ova formula je slična formuli za test u regresiji, s tim što je ovde standardna greška koeficijenta korelacije definisana kao:

Standardna greška koeficijenta korelacije meri preciznost procene koeficijenta korelacije na nivou populacije. U slučaju korelacije, ovo izračunavanje je jednostavno. Što je koeficijent korelacije bliži maksimalnoj apsolutnoj vrednosti (1 ili -1), preciznost je veća, a greška manja. Kao i kod svake standardne greške, veći uzorak daje precizniju procenu.

Demonstriraćemo kako možemo izračunati standardnu grešku koeficijenta korelacije i testirati značajnost korelacije na nivou populacije koristeći R. Prvo ćemo ponovo izračunati koeficijent korelacije za naš uzorak koristeći formulu iz prethodnog odeljka.

r = cov(podaci$netflix, podaci$koraci) / (sd(podaci$netflix) \* sd(podaci$koraci))  
cat("Korelacija je: ", round(r, 4), "\n")

Line 1

Izračunavanje koeficijenta korelacije prema standardnoj formuli.

Korelacija je: -0.3792

Kao što smo očekivali, dobijamo identičan rezultat. Sada računamo standardnu grešku i statistiku.

sr = sqrt((1 - r^2) / (n - 2))  
  
cat("Standardna greška koeficijenta korelacije je: ", round(sr, 4), "\n")

Line 1

Izračunavanje standardne greške koeficijenta korelacije.

Standardna greška koeficijenta korelacije je: 0.0935

Standardna greška je oko 0.1. Sada računamo statistiku.

t = r / sr  
cat("t-statistika je: ", round(t, 4), "\n")

Line 1

Izračunavanje t-statistike.

t-statistika je: -4.0565

Iz dosadašnjeg iskustva sa t-distribucijom, možemo pretpostaviti kakvom zaključku nas vodi vrednost statistike koja je približno . Ovaj rezultat je ekstremno niži u odnosu na vrednost koju pretpostavlja nulta hipoteza (). Da bismo bili sigurni u naš zaključak, proverićemo p-vrednost.

|  |
| --- |
| Slika 9.5: Vizualizacija t-statistike. Crvena vertikalna linija označava izračunatu t-vrednost. |

1. Izračunavanje stepena slobode za t-distribuciju.
2. Izračunavanje p-vrednosti.

P-vrednost je praktično nula, što znači da je korelacija statistički značajna. Odbacujemo i zaključujemo da korelacija postoji u populaciji.

Važna napomena: korelaciju interpretiramo samo kada odbacimo . Ako je korelacija nije značajna (prihvatimo ), ne treba je interpretirati jer je verovatno proizvod slučajnosti u uzorku.

## 9.4 Interval poverenja za Pirsonov koeficijent korelacije

Interval poverenja za Pirsonov koeficijent korelacije je izuzetno koristan alat jer nam omogućava da preciznije zaključimo o intenzitetu korelacije.

Formula za interval poverenja ima standardni oblik:

U R-u to izgleda ovako.

alpha <- 0.05  
t\_alpha <- qt(1 - alpha/2, df)  
  
donja\_granica <- r - t\_alpha \* sr  
gornja\_granica <- r + t\_alpha \* sr  
  
cat("Interval poverenja za koeficijent korelacije je: (", round(donja\_granica, 4), ", ", round(gornja\_granica, 4), ")\n")

Lines 1-2

Izračunavanje vrednosti za odgovarajući stepen slobode.

Lines 4-5

Izračunavanje donje i gornje granice intervala poverenja.

Interval poverenja za koeficijent korelacije je: ( -0.5653 , -0.1931 )

Vidimo da se interval poverenja kreće od -0.56 do -0.19. Na ovom nivou pouzdanosti možemo zaključiti da je korelacija u populaciji negativna i da varira od slabe do umerene. Ovaj interval nam jasno pokazuje da nije reč o snažnoj korelaciji između dve varijable.

R nam nudi elegantno rešenje za testiranje hipoteze o Pirsonovom koeficijentu korelacije i računanje intervala poverenja kroz funkciju cor.test. Jednom linijom koda dobijamo sve potrebne rezultate!

cor.test(podaci$netflix, podaci$koraci)

Pearson's product-moment correlation  
  
data: podaci$netflix and podaci$koraci  
t = -3.6189, df = 78, p-value = 0.0005234  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 -0.5528280 -0.1739408  
sample estimates:  
 cor   
-0.3791668

Iz ovih rezultata možemo brzo izdvojiti sve bitne informacije:

* t-statistiku, stepene slobode i p vrednost
* interpretaciju alternativne hipoteze
* 95% interval poverenja za koeficijent korelacije
* vrednost koeficijenta korelacije u uzorku

Ova elegantna jednostavnost u primeni i interpretaciji čini koeficijent korelacije izuzetno korisnim alatom u istraživanjima, naročito kada ispitujemo povezanost varijabli bez zalaženja u kompleksnu problematiku uzročnosti.

No, tu nije kraj. Zbog svoje neraskidive veze sa regresijom, korelacija služi i kao moćan dijagnostički alat. Razmotrimo formulu za standardnu grešku korelacije - u njoj se pojavljuje kvadrat koeficijenta korelacije. Šta dobijamo ovim kvadriranjem? S obzirom da se korelacija kreće između -1 i 1, njen kvadrat će uvek biti između 0 i 1, što možemo izraziti i kao procenat od 0% do 100%. Ovaj broj nazivamo **koeficijentom determinacije** i on predstavlja ključni element u dijagnostici regresije.

## 9.5 Koeficijent determinacije

Koeficijent determinacije, koji označavamo sa , jeste kvadrat koeficijenta korelacije. Za njegovo bolje razumevanje, predstavićemo ga vizuelno kroz jednostavan primer. Umesto da nastavimo sa Netflix primerom, posmatraćemo odnos dve standardizovane varijable merene na skali od 0 do 1. Iako je ovo pojednostavljen primer koji olakšava interpretaciju i vizuelizaciju, ista logika se primenjuje i na složenije varijable merene na drugim skalama.

|  |
| --- |
| Slika 9.6: Vizuelizacija koeficijenta determinacije (R²) |

Grafikon prikazuje dve standardizovane varijable (na skali od 0 do 1) čija korelacija iznosi približno 0.7. Koeficijent determinacije (), koji je kvadrat ove vrednosti, iznosi oko 0.5. Plavi kvadrat na grafikonu ima površinu jednaku , što vizuelno predstavlja procenat varijanse jedne varijable koji možemo objasniti drugom varijablom.

Ova vizuelizacija nam pomaže da razumemo zašto su vrednosti 0.3 i 0.7 bitne granice za interpretaciju korelacije:

* Kada je , je manji od 0.1 (10% objašnjene varijanse)
* Kada je , je između 0.1 i 0.5 (10-50% objašnjene varijanse)
* Kada je , je veći od 0.5 (preko 50% objašnjene varijanse)

Granica , odnosno , ima posebno intuitivno značenje. Ona nam govori da određeni regresioni model može da objasni najmanje polovinu varijabiliteta zavisne varijable. Jednostavnije rečeno, model objašnjava više nego što ostavlja neobjašnjeno. Zato korelacije preko 0.7 nazivamo jakim - one nam omogućavaju da sa visokom pouzdanošću predvidimo promene jedne varijable na osnovu promena druge.

Granica predstavlja konvenciju - označava korelacije koje objašnjavaju manje od 10% varijabilnosti zavisne varijable. Ovako slab stepen povezanosti ukazuje da postoji mnogo prostora za druge faktore koji utiču na zavisnu varijablu. Kada otkrijemo slabu korelaciju između dve varijable, to nas usmerava ka istraživanju dodatnih faktora koji bi mogli bolje objasniti promene u zavisnoj varijabli. Jasno je da u takvim slučajevima trenutna nezavisna varijabla nije dovoljno precizan prediktor.

Kod analize bivarijatnih korelacija (korelacija između dve varijable), potreban je poseban oprez pri interpretaciji slabih korelacija. U praksi se često fokusiramo na korelacije koje se približavaju ili prelaze prag od 0.7, jer one pružaju čvršću empirijsku osnovu za zaključivanje o povezanosti varijabli.

Koeficijent determinacije predstavlja prirodan uvod u analizu varijabiliteta koji se može objasniti statističkim modelom. U prethodnom poglavlju smo naveli da je jedan od glavnih ciljeva regresionog (i donekle korelacionog) modela objašnjenje varijabiliteta zavisne varijable. Šta to konkretno znači?

Objašnjavanje varijabiliteta neke varijable znači da razumemo razlike koje postoje među opservacijama tako što ih direktno povezujemo sa razlikama u vrednostima nezavisne varijable. Konzistentnost tih razlika je naše glavno eksplanatorno sredstvo - alat pomoću kojeg gradimo objašnjenje. Kada utvrdimo da razlike u broju koraka dosledno prate razlike u vremenu provedenom uz Netflix, možemo preciznije objasniti zašto neki ljudi provode više ili manje vremena gledajući serije.

U svakom statističkom modelu razlikujemo dva tipa varijabiliteta:

1. **Objašnjeni varijabilitet**: deo varijacija zavisne varijable koji možemo direktno povezati sa promenama u nezavisnoj varijabli. Ovo je onaj deo koji naš model uspešno zahvata i opisuje.
2. **Neobjašnjeni varijabilitet**: preostali deo varijacija zavisne varijable koji nije povezan sa nezavisnom varijablom i ostaje van domašaja našeg modela.

Kako bismo precizno razumeli odnos između objašnjenog i neobjašnjenog varijabiliteta, koristimo moćan alat - **analizu varijanse**. O njoj detaljno govorimo u sledećem poglavlju.

## 9.6 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1**  Imamo podatke o broju sati provedenih na društvenim mrežama nedeljno (X) i prosečnoj oceni na studijama (Y) za uzorak od 50 studenata.   1. Izračunajte kovarijansu i Pirsonov koeficijent korelacije u uzorku. Šta nam oni govore o vezi između vremena provedenog na društvenim mrežama i akademskog uspeha? 2. Testirajte značajnost ove korelacije na nivou α = 0.05. Postavite hipoteze i sprovedite test. 3. Konstruišite 95% interval poverenja za populacionu korelaciju. Šta nam on govori o jačini veze u populaciji?   Podatke možete učitati na sledeći način:  zadatak1 <-  read.csv("https://gist.githubusercontent.com/atomashevic/a2512864088dc27bfe64e6261688de6d/raw/a9be32d28f543c861598ca1d556370b5a1a86918/dm-prosek-2.csv") |

|  |  |
| --- | --- |
|  | **Zadatak 2 \***  U istraživanju o fizičkoj aktivnosti i mentalnom zdravlju, na uzorku od 40 osoba, merene su dve ključne varijable:   * X: broj sati fizičke aktivnosti nedeljno * Y: skor na skali depresivnosti (viši skor označava izraženije simptome)   Vaš zadatak je da:   1. Izračunate Pirsonov koeficijent korelacije između ove dve varijable. 2. Na nivou značajnosti α = 0.01, proverite da li postoji **pozitivna korelacija** između fizičke aktivnosti i mentalnog zdravlja. 3. Protumačite dobijene rezultate u kontekstu istraživačkog problema. Obratite posebnu pažnju na praktične implikacije vaših nalaza.   Podatke možete učitati korišćenjem sledećeg koda:  zadatak2 <- read.csv("https://gist.githubusercontent.com/atomashevic/2553fd69f0526dbc901190d0078371c0/raw/ecf5a8233d3cf56777a37ed844677c487583ef0d/aktivnost-depresija.csv") |

|  |  |
| --- | --- |
|  | **Zadatak 3 \*\***  Istraživači su sproveli studiju o povezanosti između tri varijable na uzorku od 60 zaposlenih: - X: broj sati prekovremenog rada mesečno - Y: nivo stresa (skala 1-10) - Z: zadovoljstvo poslom (skala 1-10)   1. Izračunajte korelacije između svih parova varijabli. 2. Testirajte značajnost svih korelacija na nivou . 3. Šta možete zaključiti o odnosu između prekovremenog rada, stresa i zadovoljstva poslom? Koji elementi ovih odnosa su ključni za razumevanje radnog okruženja? 4. Sa kojim izazovima se suočavamo pri interpretaciji ovakvih korelacionih podataka? Objasnite moguća ograničenja u zaključivanju.   Podatke možete učitati na sledeći način:  zadatak3 <-  read.csv("https://gist.githubusercontent.com/atomashevic/bc0d4201301ddce7d1a63c03374d0ace/raw/a722ad5c63573106440d997e5eb36a5583876071/rad-stres.csv") |

# 10. Analiza varijanse (ANOVA)

## 10.1 Povratak na (diskretne) grupe

Vreme je da se vratimo na jedan od fundamentalnih problema statističke analize – poređenje između grupa koje su određene prema diskretnim obeležjima. Kada kažem diskretnim, mislim na to da su naše jedinice uzorka podeljene prema nekim jasnim kategorijama. Zašto je ovo važno? Zato što je ovo jedan od najčešćih scenarija u praksi.

Klasičan primer u društvenim istraživanjima je podela ispitanika na osnovu biološkog pola, što smo već videli kod t-testa. Drugi sjajan primer je obrazovanje. Gotovo svaki obrazovni sistem na svetu deli obrazovanje na osnovno, srednje i visoko. Ovo je primer ordinalne merne skale - svaka naredna kategorija označava viši stepen obrazovanja.

Hajde da odmah skočimo u konkretan primer koji će nam pomoći da razumemo zašto je ova tema važna. Zamislimo da istražujemo cene kafe u Novom Sadu. Imamo podatke o cenama iz XX kafeterija i znamo u kom delu grada se svaka nalazi. Grad je podeljen na tri zone:

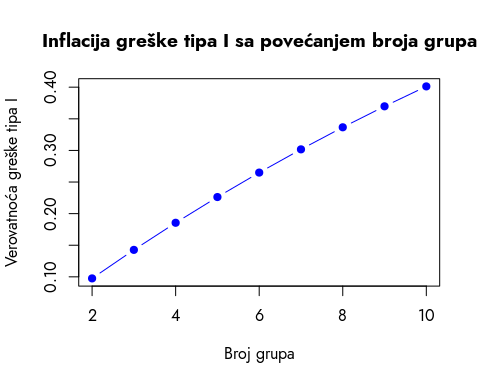
1. Uži centar (područje oko Zmaj Jovine, Dunavske, Pozorišnog trga)
2. Širi centar (Grbavica, Liman, Sajam, Bulevar)
3. Nova naselja (Detelinara, Novo Naselje, Telep, Satelit)

Ako nas zanima da proverimo da li prosečna cena kafe zavisi od lokacije kafterije, možemo da primenimo ono što smo dobro savladali – t-test. Međutim, potrebna su nam tri t-testa da bismo utvrdili da li postoji razlika između svake dve grupe.

1. Uži centar vs. Širi centar
2. Uži centar vs. Nova naselja
3. Širi centar vs. Nova naselja

Dakle, trebaju nam tri zasebna zaključka na osnovu tri različita testa da bismo odgovorili na jedno pitanje. Sa druge strane, pitanje koje nas zanima je jednostavno – da li postoji zavisnost ili uticaj lokacije na cenu kafe? Korišćenje tri testa za odgovor na jedno pitanje čini se kao previše posla. Pri tome, ne zaboravimo da su to statistički testovi - svaki do njih poseduje rizik greške tipa I i tipa II. Ako povećavamo broj testova, povećavamo i rizik greške tipa I.

Razmislite, ako razlike na nivou populacije postoje između sve tri grupe, potrebno je da ispravno odbacimo tri različite nulte hipoteze. Ako je pri riziku greške od 5% verovatnoća da ćemo pri svakom testu ispravno odbaciti nultu hipotezu 95%, onda je verovatnoća da ćemo uspeti u svim tri testa . Dakle, rizik greške tipa I se povećava sa brojem statističkih testova. Što je broj grupa za poređenje veći, to je veća verovatnoća da doneti ispravan zaključak kada objedinimo sve rezultate. Ovaj fenomen se naziva **inflacija greške tipa I**.



Inflacija greške tipa I sa povećanjem broja grupa

Šta je alternativa ovakvom pristupu? Potreban nam je novi metod koji će nam omogućiti da jednim udarcem rešimo problem poređenja više grupa. Međutim, postoji jedan veliki problem koji treba prevazići u ovom pristupu. Kada imamo više grupa, nije najjasnije razdvojiti razlike u podacima koje proističu iz različitih izvora.

## 10.2 Dva izvora varijabiliteta

Zamislite sledeću situaciju: dva studenta raspravljaju o cenama kafe. Prvi tvrdi: „U centru je kafa uvek skuplja nego u novim naseljima“. Drugi odgovara: „Ma nije to tako jednostavno. Poznajem jeftine kafiće u centru, a neki lanci su skupi gde god da se nalaze.“

Ova rasprava savršeno ilustruje dva ključna koncepta koja moramo razumeti:

1. Varijabilitet između grupa (prvi student)
2. Varijabilitet unutar grupa (drugi student)

Hajde da pažljivo analiziramo ovaj drugi argument. Saznajemo dve stvari:

* Među kafićima u centru postoje značajne varijacije u ceni kafe
* Postoji neki faktor koji nismo razmotrili (npr. da li je to lokalni kafe ili lanac kafeterija), a koji utiče na cenu kafe

Ova dva studenta zapravo opisuju dva izvora varijabiliteta u podacima.

1. Prvi student opisuje varijabilitet **između grupa** - razlike u prosečnim cenama kafe između različitih delova grada.
2. Drugi student opisuje varijabilitet **unutar grupa** - razlike u cenama kafe unutar samih grupa. Ovo nisu razlike između aritmetskih sredina, već razlike između pojedinačnih opservacija unutar iste grupe.

Oba tipa varijabiliteta su usko povezana sa podelom na grupe koje smo konstruisali. Razlike između grupa su posledica uticaja obeležja na osnovu kojih smo podelili jedinice uzorka u grupe. Ako između različitih delova grada postoje razlike u ceni kafe, to znači da nešto u tim delovima grada ostvaruje uticaj i pravi razliku u prosečima.

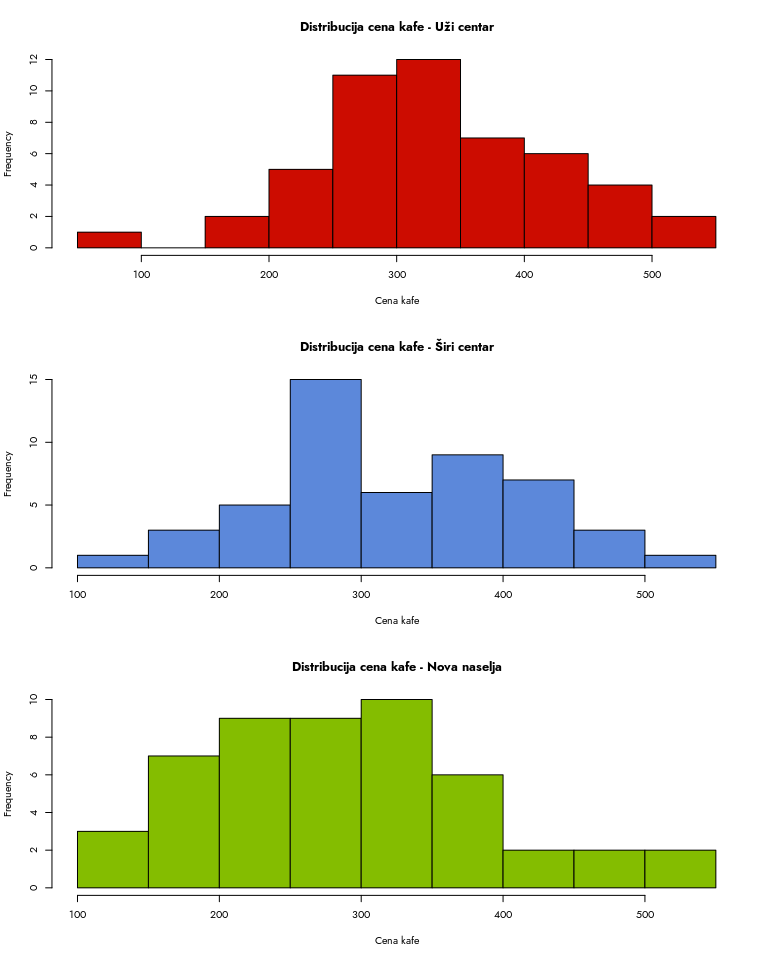
Problem u statističkoj analizi je što u svakom setu podataka mi možemo veoma lako pronaći i jedne i druge razlike. Hajde da vidimo kako to izgleda na našem primeru sa kafama. Hajde da vidimo kako izgleda distribucija cena kafe u kafeterijama u centru i novim naseljima.

Prvo učitavamo podatke.

podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/4b430a26e5a73651d316792a87d90255/raw/2ec415f6cbd15c36ba572b86310f46307c72dfdb/kafa-ns.csv")  
  
head(podaci)

cena lokacija  
1 264 Širi centar  
2 297 Širi centar  
3 476 Širi centar  
4 327 Širi centar  
5 333 Širi centar  
6 492 Širi centar

Zatim ćemo za svaku grupu podataka nacrtati histogram distribucije cena kafe.



Distribucija cena kafe po lokaciji

1. Podešavamo raspored grafikona na 3 reda i 1 kolonu
2. Crtamo histogram za cene kafe u Užem centru
3. Crtamo histogram za cene kafe u Širem centru
4. Crtamo histogram za cene kafe u Novim naseljima

Na osnovu histograma možemo videti da postoji značajna varijabilnost u cenama kafe unutar svake grupe, što ide u prilog tezi koju zagovara student. U centru grada možete popiti kafu i za 100 dinara (ako znate gde da je pronađete) i za 500 dinara.

Međutim, vidimo i da između grupa postoje razlike u cenama kafe. Kafu za manje od 200 dinara je laše naći u novim naseljima nego u centru. Hajde da vidimo da li se to ogleda i u aritmetičkim sredinama.

AS\_siri\_centar <- mean(podaci$cena[podaci$lokacija == "Širi centar"])  
  
AS\_uzi\_centar <- mean(podaci$cena[podaci$lokacija == "Uži centar"])  
  
AS\_nova\_naselja <- mean(podaci$cena[podaci$lokacija == "Nova naselja"])  
  
cat("Aritmetička sredina cena kafe u Širem centru: ", round(AS\_siri\_centar, 2), "\n")  
  
cat("Aritmetička sredina cena kafe u Užem centru: ", round(AS\_uzi\_centar, 2), "\n")  
  
cat("Aritmetička sredina cena kafe u Novim naseljima: ", round(AS\_nova\_naselja, 2), "\n")

Aritmetička sredina cena kafe u Širem centru: 323.38   
Aritmetička sredina cena kafe u Užem centru: 330.64   
Aritmetička sredina cena kafe u Novim naseljima: 286.64

Vidimo da je prosečna cena kafe u Širem centru najviša (oko 330 dinara), dok je najniža u Novim naseljima (oko 286 dinara). Međutim, da li nas proseci varaju? Da li su razlike **između grupa** veće nego što su razlike **unutar grupa**? Koji student je u pravu?

Hajde da pokušamo da vizualizaujemo odnos između ovih razlika.

|  |
| --- |
| Slika 10.1: Varijabilitet unutar i između grupa |

Šta vidimo na ovom grafikonu? Svaka od tačaka predstavlja jednu opservaciju, tj. jednu zabeleženu cenu kafe u različitim delovima grada. Crni kvadrati predstavljaju aritmetičku sredinu svake grupe. Unutar svake grupe vidimo da su opservacije udaljene različito od proseka svoje grupe.

Međutim, vidimo da su i proseci svake grupe udaljeni od **opšteg proseka** (isprekidana horizontalna linija) i te udaljenosti su predstavljene crvenim linijama sa strelicom.

Na prvi pogled deluje da je situacija veoma jednostavna. Sive linije koje označavaju udaljenost opservacija od svojih grupa su značajno duže od malih crvenih linija. Međutim, ako želimo zaista da na osnovu ovih podataka donesemo zaključak, potrebno je da kvantifikujemo ovu razliku. Takođe, potrebno je da ovim izvorima varijabiliteta dodelimo „službena“ imena: faktorski i rezidualni varijabilitet.

## 10.3 Faktorski i rezidualni varijabilitet

Varijabilitet između grupa nazivamo **faktorski varijabilitet**. Ovaj varijabilitet je posledica razlika između grupa koje su definisane na osnovu nekog faktora. U našem slučaju, faktor je lokacija kafeterije. Faktori su veoma česti naziv za nezavisne varijable koje su merene kategorijalno, tj. nominalnom ili ordinalnom mernom skalom.

Faktorski varijabilitet je predmet našeg istraživanja. Na kraju, definisali smo i iskoristili varijablu lokacija kako bismo ispitali **uticaj** lokacije na cenu kafe. Taj uticaj se manifestuje kroz faktorski varijabilitet.

[Slika 10.1](#fig-varijabilitet) daje privid da je rezidualni varijabilitet značajno veći, sudeći po dužini i broju sivih linija. Međutim, taj vizuelni utisak može lako da nas prevari. Zašto? Ova dva tipa varijabiliteta nisu direktno uporedivi jer poseduju različite cd**stepene slobode**.

Stepene slobode upoznali smo kada smo govorili o t-testu ([Odeljak 7.2](#sec-tdistribucija)). Oni predstavljaju odnos između broj opservacija i broja nepoznatih parametara.

U slučaju faktorskog varijabiliteta, broj stepeni slobode je jednak broju grupa minus jedan. Zašto baš ovako? Ovo je situacija koja je analogna problemu sa tri kese koji smo spominjali ranije. Imamo tri aritmetičke sredine grupa i imamo opštu aritmetičku sredinu varijable cena koja predstavlja nepoznati parametar. Samim tim, broj stepeni slobode biće: , gde je broj grupa.

Sa druge strane, kod rezidualnog varijabiliteta broj stepeni slobode je jednak broju opservacija minus broj grupa. U ovom slučaju sve opservacije čine naš uzorak, a imamo tri nepoznata parametra - aritmetičke sredine grupa. Dakle, broj stepeni slobode za rezidualni varijabilitet je: , gde je ukupan broj opservacija, a broj grupa. U našem slučaju imamo 150 opservacija i 3 grupe, pa je . Kao što možemo videti, broj stepeni slobode će uvek biti značajno veći u slučaju rezidualnog varijabiliteta.

U oba slučaja broj nepoznatih parametara označava broj aritmetičkih sredina u odnosu na koje posmatramo razlike, tj. varijabilitet. U prvom slučaju je to jedna aritmetička sredina, a u drugom su to aritmetičke sredine svake grupe.

OK, sada kada znamo šta su faktorski i rezidualni varijabilitet, hajde da vidimo kako možemo da ih kvantifikujemo i uporedimo. Metod koji se koristi za ovu svrhu naziva se **analiza varijanse** ili skraćeno **ANOVA**.

## 10.4 ANOVA postupak

Postupak izračunavanja varijabiliteta smo ranije videli kod primera izračunavanja varijanse, kovarijanse i greške regresije. Sada ćemo primeniti sličan postupak, ali na grupisane podatke.

Počećemo od faktorskog varijabiliteta. On predstavlja udaljenost aritmetičkih sredina grupa od opšte aritmetičke sredine. Faktorski varijabilitet je u svojoj suštini izražen kroz kvadrirano odstupanje svake aritmetičke sredine od opšte aritmetičke sredine svih opservacija. Ovaj tip varijabiliteta kvantifikujemo putem sledeće formule:

gde je faktorski varijabilitet, broj opservacija u -toj grupi, aritmetička sredina -te grupe, a opšta aritmetička sredina svih opservacija. U formuli vidimo da se kvadrat razlike između aritmetičke sredine grupe i opšte aritmetičke sredine množi sa brojem opservacija u grupi. Ovakav pristup ima svoju logiku - veće grupe nam daju više informacija o odnosu između proseka te grupe i opšteg proseka. Koristimo oznaku da naglasimo da računamo sumu kvadriranih odstupanja, što je osnovni princip izračunavanja varijabiliteta kvantitativnih varijabli.

Rezidualni varijabilitet, sa druge strane, predstavlja sumu kvadrata odstupanja svake opservacije od aritmetičke sredine svoje grupe. Izračunavamo ga pomoću formule:

Ovaj izraz izgleda komplikovano zbog dvostrukog sumiranja i . Prvi indeks sumiranja prolazi kroz sve grupe, dok drugi indeks sumira sve opservacije unutar svake grupe. Suština rezidualnog varijabiliteta ogleda se u procesu koji ima dva koraka:

1. prolazimo kroz svaku grupu posebno ,
2. unutar svake grupe prolazimo kroz sve opservacije i računamo kvadrat odstupanja te opservacije od aritmetičke sredine svoje grupe .

Hajde da demonstriramo izračunavanje oba tipa varijabiliteta u R-u.

# Izračunavanje faktorskog varijabiliteta  
  
opsta\_sredina <- mean(podaci$cena)  
  
sredine\_grupa <- c(AS\_nova\_naselja, AS\_siri\_centar, AS\_uzi\_centar)  
  
velicine\_grupa <- table(podaci$lokacija)  
  
SSF <- sum(velicine\_grupa \* (sredine\_grupa - opsta\_sredina)^2)  
  
cat("Faktorski varijabilitet: ", round(SSF, 2), "\n")

Line 3

Izračunavamo opštu aritmetičku sredinu svih opservacija

Line 5

Spajamo aritmetičke sredine grupa u jedan vektor

Line 7

Iz tabulacije varijable lokacija dobijamo veličine grupa u jednom vektoru

Line 9

Izračunavamo faktorski varijabilitet prema formuli koju smo videli ranije

Faktorski varijabilitet: 55642.25

Kao i svaka druga suma kvadrata, faktorski varijabilitet nam ne daje direktno interpretabilnu vrednost.

Sada prelazimo na izračunavanje rezidualnog varijabiliteta. Da bismo to uradili precizno i metodično, prvo ćemo razdvojiti podatke po grupama.

# Izdvajanje podataka po grupama  
  
uzi\_centar <- podaci$cena[podaci$lokacija == "Uži centar"]  
  
siri\_centar <- podaci$cena[podaci$lokacija == "Širi centar"]  
  
nova\_naselja <- podaci$cena[podaci$lokacija == "Nova naselja"]  
  
SSR <- sum((uzi\_centar - AS\_uzi\_centar)^2) + sum((siri\_centar - AS\_siri\_centar)^2) + sum((nova\_naselja - AS\_nova\_naselja)^2)  
  
cat("Rezidualni varijabilitet: ", round(SSR, 2), "\n")

Lines 3,5,7

Izdvajamo cene kafe po grupama

Line 9

Izračunavamo rezidualni varijabilitet prema formuli iznad

Rezidualni varijabilitet: 1302325

Rezidualni varijabilitet u našem primeru je takođe velika cifra, ali kao i faktorski varijabilitet, nema direktnu interpretaciju. Kao što smo već naglasili, ova dva varijabiliteta ne možemo direktno porediti. Da bismo ih mogli uporediti, potrebno je da ih transformišemo u faktorsku i rezidualnu varijansu.

|  |  |
| --- | --- |
|  | **Varijabilitet i varijansa**  U toku ovog udžbenika naizmenično smo koristili pojmove varijabiliteta i varijanse. U pretposlednjem poglavlju važno je da napravimo formalnu razliku između ova dva pojma, jer oni nisu sinonimi.  Varijabilitet ili varijabilnost je osnovna karakteristika skupa podataka koja opisuje koliko su podaci raspršeni oko neke centralne vrednosti. Kada govorimo o „sirovom“ varijabilitetu, mislimo na jednu od dve stvari:   1. Sumu apsolutnih odstupanja od centralne vrednosti (npr. aritmetičke sredine) 2. Sumu kvadrata odstupanja od centralne vrednosti   U oba slučaja varijabilitet predstavlja **sumu** odstupanja od centralne tendencije. Najčešće koristimo sumu kvadriranih odstupanja, pa ove osnovne mere varijabilnosti nazivamo **sumama kvadrata**.  Varijansa je specifična transformacija ovog sirovog varijabiliteta. Ona predstavlja odnos između varijabiliteta (sume kvadrata odstupanja) i odgovarajućih stepeni slobode. Preciznije, varijansa je prosečno kvadratno odstupanje po stepenu slobode i dobijamo je deljenjem sume kvadrata sa brojem stepeni slobode. Na ovaj način dobijamo **međusobno uporedive** mere varijabiliteta za grupe različitih veličina.  Logika ove razlike je slična tome zašto ne koristimo sumu svih vrednosti da opišemo centralnu tendenciju varijable, već koristimo aritmetičku sredinu. Same sume (vrednosti, apsolutnih vrednosti ili kvadriranih vrednosti) daju brojeve koje ne možemo smisleno interpretirati niti uporediti jer direktno zavise od veličine uzorka. |

Da bismo došli do faktorske i rezidualne varijanse, faktorski i rezidualni varijabilitet delimo sa odgovarajućim stepenima slobode.

Hajde da vidimo kako to izgleda u R-u.

sF <- SSF / (3-1)  
sR <- SSR / (150-3)  
  
cat("Faktorska varijansa: ", round(sF, 2), "\n")

Faktorska varijansa: 27821.13

cat("Rezidualna varijansa: ", round(sR, 2), "\n")

Rezidualna varijansa: 8859.35

Sada imamo međusobno uporedive vrednosti. Faktorska varijansa iznosi približno 28000, dok rezidualna varijansa iznosi približno 8900. Na osnovu ovoga možemo zaključiti da je na nivou našeg uzorka faktorska varijansa značajno veća od rezidualne, tj. da su razlike između grupa mnogo izraženije nego razlike unutar grupa. Ako je verovati našem uzorku, prvi student je bio u pravu.

Da bismo upotpunili naše razumevanje varijabiliteta, pogledajmo i kolika je varijansa same varijable koja meri cenu kafe.

Varijansa cene kafe: 9113.87

U kakvom odnosu je ova ukupna varijansa sa faktorskom i rezidualnom varijansom? Ako saberemo faktorski i rezidualni varijabilitet, dobićemo ukupan varijabilitet. Kada taj ukupan varijabilitet podelimo sa brojem stepeni slobode za ocenu varijanse uzorka (), dobićemo upravo varijansu čitave varijable cena. Dakle, vrednost koju smo dobili direktnim izračunavanjem varijanse je zapravo kombinovani zbir faktorske i rezidualne varijanse.

SSU <- SSF + SSR  
  
s2 <- SSU / (150 - 1)  
  
cat("Ukupan varijabilitet: ", round(SSU, 2), "\n")  
  
cat("Varijansa cene kafe: ", round(s2, 2), "\n")

Line 1

Izračunavamo ukupan varijabilitet kao zbir faktorskog i rezidualnog varijabiliteta

Line 3

Izračunavamo varijansu cene kafe kao ukupan varijabilitet podeljen sa brojem stepeni slobode

Ukupan varijabilitet: 1357967   
Varijansa cene kafe: 9113.87

Kao što vidite dobili smo identičan broj. Ovaj proces izračunavanja faktorskog i rezidualnog varijabiliteta naziva se **dekompozicija** ili **razlaganje** varijanse. Sam metod koji ćemo uskoro upoznati zove se **analiza varijanse** upravo zato što razdvaja ukupnu varijansu na onu koju proizvodi faktor i onu koja ostaje unutar grupa.

Koji je cilj ovakve analize? Želimo da utvrdimo koliko je puta faktorska varijansa veća od rezidualne. Taj odnos faktorske i rezidualne varijanse naziva se **F-statistika** i predstavlja osnovu za testiranje statističke značajnosti faktorskog varijabiliteta.

Pre nego što objasnimo prirodu F-statistike, važno je naglasiti razliku u odnosu na t-statistiku. Ako pogledamo t-test za 2 uzorka, nulta hipoteza je fokusirana na vrednost 0 unutar t-distribucije jer nula označava odsustvo razlika između grupa, pa samim tim i odsustvo efekta. Kod F-statistike situacija je drugačija - ne očekujemo vrednost 0. Zašto? Pošto je F-statistika količnik mera varijabiliteta, ona može biti bilo koja pozitivna vrednost. Nula ne predstavlja samo odsustvo efekta, već potpuno odsustvo varijabiliteta iz faktora, što je izuzetno retko u empirijskim istraživanjima.

Ključna vrednost za razumevanje ove statistike je . Ako je F-statistika jednaka 1, to znači da je faktorska varijansa jednaka rezidualnoj varijansi - razlike između grupa su jednake razlikama unutar grupa. Ako je F-statistika veća od 1, faktorska varijansa nadmašuje rezidualnu varijansu, što znači da su razlike između grupa izraženije od razlika unutar grupa. Suprotno tome, F-statistika manja od 1 ukazuje da su razlike unutar grupa dominantnije od razlika između grupa, što odgovara scenariju odsustva efekta.

Dakle, da bismo demonstrirali postojanje statistički značajnog efekta faktora na nivou populacije, F mora prevazići vrednost 1, pokazujući da faktorska varijansa nadmašuje rezidualnu varijansu. Koliko velika F-statistika mora biti? Odgovor na ovo pitanje leži u još jednoj distribuciji verovatnoće - F-distribuciji.

## 10.5 F-distribucija

Pre nego što vidimo F-distribuciju na delu, važno je istaći po čemu se ona suštinski razlikuje od normalne i t-distribucije. Kada pogledamo formulu za F-statistiku, uočićemo jednu fundamentalnu činjenicu - F-statistika je uvek pozitivan broj. Pošto predstavlja količnik dve varijanse, matematički je nemoguće da bude negativna.

Kroz F-distribuciju pokušavamo da opišemo situaciju koja odgovara nultoj hipotezi i odsustvu efekta. U tom slučaju, najveća koncentracija verovatnoće nalazi se između 0 i 1 (analogno vrednosti 0 u t-distribuciji). U odnosu na ovu vrednost, F-distribucija pokazuje jasnu asimetriju u desno, što znači da ekstremne vrednosti nalazimo isključivo na desnoj strani distribucije.

Šta je logika ove asimetrije? Ponovo, odgovor leži u prirodi F-statistike. Kada pretpostavljamo odsustvo efekta, očekujemo vrednosti manje od ili jednake 1, jer to znači da je faktorski varijabilitet jednak ili manji od rezidualnog. Značajno veći faktorski varijabilitet predstavlja direktan dokaz protiv nulte hipoteze, zbog čega distribucija mora biti asimetrična u desno.

Kao i t-distribucija, F-distribucija mora biti adaptivna prema karakteristikama uzorka koji analiziramo. U ovom slučaju imamo dva ključna parametra koja oblikuju F-distribuciju: broj stepeni slobode faktora i broj stepeni slobode reziduala. Promena ovih parametara direktno utiče na oblik F-distribucije.

Hajde da vizualizujemo kako bi izgledala F-distribucija u našem primeru.

par(family = "Jost")  
df1 <- 2  
df2 <- 150 - 3  
  
x <- seq(0, 5, length.out = 200)  
  
y <- df(x, df1, df2)  
  
plot(x, y, type = "l", lwd = 3, col = "#5C88DAFF",  
 xlab = "F vrednost",   
 ylab = "Gustina verovatnoće",  
 main = "F distribucija (N = 150)")

Line 2

Broj stepeni slobode faktora

Line 3

Broj stepeni slobode reziduala

Line 5

Generišemo vrednosti za x-osu; uzimamo samo pozitivne vrednosti

Line 7

Izračunavamo gustinu verovatnoće za svaku vrednost x pomoću ugrađene funkcije df koja računa gustinu verovatnoće F-distribucije

|  |
| --- |
| Slika 10.2: F distribucija za N=150 opservacija |

Vidimo da se najveća koncentracija vrednosti nalazi između 0 i 1. Razmotrimo sada koliki deo površine ispod krive zauzima interval od 0 do 1, a koliki deo se nalazi desno od 1.

|  |
| --- |
| Slika 10.3: F distribucija sa označenim oblastima F < 1 i F > 1 |

Plavi deo obuhvata 63% svih mogućih F-statistika i predstavlja oblast koja je najkonzistentnija sa nultom hipotezom. Vrednosti F-statistike iznad 1 obuhvataju preostalih 37% svih mogućih vrednosti - ovo je oblast od posebnog interesa za našu analizu. Međutim, kao i kod drugih statističkih distribucija, fokusiramo se na površinu ispod krive koja sadrži mali procenat najekstremnijih vrednosti. Na primer, ako bismo želeli da pomerimo granicu udesno tako da plava oblast obuhvata 95% svih mogućih vrednosti, kolika bi trebalo da bude F-statistika?

|  |
| --- |
| Slika 10.4: F distribucija sa označenim kritičnim oblastima (95% i 5%) |

Na grafikonu vidimo da je to vrednost . Dakle, da bismo na ovakvom uzorku i sa ovim brojem grupa mogli zaključiti da postoji značajan efekat faktora, faktorska varijansa mora biti približno tri puta veća od rezidualne. Ovo je kritična tačka koja razdvaja oblast u kojoj nemamo dovoljno dokaza protiv nulte hipoteze od oblasti u kojoj moramo odbaciti nultu hipotezu.

|  |  |  |
| --- | --- | --- |
|  | **Izgled F-distribucije**  Sa samo tri grupe, F-distribucija ima oblik koji podseća na desni rep normalne distribucije. Međutim, promena broja grupa značajno utiče na oblik distribucije. Kako broj grupa raste, F-distribucija dobija svoj prepoznatljiv oblik - postaje sve više asimetrična i verovatnoća se koncentriše oko vrednosti 1.   |  | | --- | | Slika 10.5: F distribucija za različit broj grupa (N = 150) |   Zašto dolazi do ovog pomeranja oblasti najveće koncentracije vrednosti udesno? Sa povećanjem broja grupa, faktički povećavamo i faktorski varijabilitet. Što je veći broj grupa, to više prostora imaju međugrupne razlike da dođu do izražaja, čak i kada je dejstvo faktora zanemarljivo. Zbog samog povećanja broja grupa koje poredimo, sredina distribucije pomera se ka 1, odnosno ka izjednačenom faktorskom i rezidualnom varijabilitetu.  Konačno, kada radimo sa velikim brojem grupa, F distribucija poprima oblik pozitivno asimetrične modifikacije normalne distribucije. |

Sada smo spremni da pokrenemo formalnu ANOVA analizu. Najpre ćemo uvesti F-test i prikazati njegovu primenu u R-u.

## 10.6 F-test

Nulta hipoteza F-testa označava odsustvo dejstva nezavisne varijable (faktora) na zavisnu varijablu. Preciznije, nulta hipoteza tvrdi da ne postoji razlika između aritmetičkih sredina grupa. Pošto tih aritmetičkih sredina može biti 3, 5, 10 ili više, ovaj tip hipoteze nazivamo **omnibus hipoteza**. Sam termin potiče iz latinskog i znači „za sve“. U našem slučaju, nulta hipoteza tvrdi da je efekat faktora nepostojeći za sve grupe.

Alternativna hipoteza ima podjednako zanimljiv oblik. Šta bi bilo suprotno onome što predviđa nulta hipoteza? Dovoljno je da se samo jedna aritmetička sredina razlikuje od ostalih i već imamo osnovu za odbacivanje nulte hipoteze. Alternativna hipoteza F-testa tvrdi da postoji **bar jedna** grupa koja se razlikuje od ostalih. Ove hipoteze formalno zapisujemo na sledeći način:

Kada jedna grupa odstupa od ostalih, to direktno proizvodi značajan porast faktorskog varijabiliteta, što dovodi do veće vrednosti F statistike. Jednostavno rečeno, F-test proverava da li je faktorska varijansa dovoljno puta veća od rezidualne varijanse.

Razmotrimo rezultate koje smo dobili:

* Faktorska varijansa: 2.782113^{4}
* Rezidualna varijansa: 8859.35

Sledeći korak je izračunavanje F-statistike.

F <- sF / sR  
  
cat("F-statistika: ", round(F, 2), "\n")

F-statistika: 3.14

Sada ćemo vizuelno prikazati poziciju naše F-statistike u odnosu na F-distribuciju koja je definisana parametrima našeg primera - faktorskim stepenima slobode (2) i rezidualnim stepenima slobode (147, jer imamo 150 opservacija i 3 grupe).

par(family = "Jost")  
  
df1 <- 2 #<1  
df2 <- 150 - 3  
  
x <- seq(0, 5, length.out = 200)  
  
y <- df(x, df1, df2)  
  
plot(x, y, type = "l", lwd = 3, col = "#5C88DAFF",  
 xlab = "F vrednost",   
 ylab = "Gustina verovatnoće",  
 main = "F distribucija (N = 150)")  
  
abline(v = F, lty = 2, lwd = 3, col = "#CC0C00FF")  
  
text(2.5, 0.2, paste0("F = ", round(F, 2)), font = 2, cex = 1.5, col = "#5C88DAFF")  
  
x\_right <- seq(F, 5, length.out = 100)  
y\_right <- df(x\_right, df1, df2)  
  
polygon(c(x\_right, 5, F), c(y\_right, 0, 0), col = "#CC0C00FF")  
  
p\_value <- 1 - pf(F, df1, df2)  
  
text(4, 0.2, paste0("p = ", round(p\_value, 4)), font = 2, cex = 1.5, col = "#CC0C00FF")

Line 4

Definišemo broj stepeni slobode faktora i reziduala

Line 6

Generišemo vrednosti za x-osu. F-distribucija je definisana za vrednosti veće od 0

Line 8

Izračunavamo gustinu verovatnoće prema F-distribuciji za svaku vrednost x koristeći ugrađenu funkciju df

Line 13

Kreiramo grafikon F-distribucije

Line 15

Dodajemo vertikalnu liniju koja označava vrednost F-statistike

Line 22

Bojimo oblast desno od F-statistike

Line 24

Izračunavamo p-vrednost kao površinu ispod krive desno od F-statistike korišćenjem ugrađene funkcije pf

|  |
| --- |
| Slika 10.6: F distribucija sa označenom vrednošću F-statistike |

Vidimo da se naša F-statistika nalazi u desnom delu distribucije, što znači da je faktorska varijansa značajno veća od rezidualne. Da li je ta razlika dovoljno velika da bismo odbacili nultu hipotezu? To nam otkriva p-vrednost koja iznosi 0.0462. Ova vrednost nam pokazuje kolika je verovatnoća da dobijemo ovakvu ili ekstremniju vrednost F-statistike pod pretpostavkom da je nulta hipoteza tačna. Ta verovatnoća je u našem slučaju manja od 5%, odnosno . To znači da posedujemo dovoljno dokaza da odbacimo nultu hipotezu i zaključimo da postoji bar jedna grupa koja se statistički značajno razlikuje od ostalih.

Praktično, veća vrednost F-statistike ukazuje na snažnije dejstvo faktora na zavisnu varijablu. Kada F-statistika premaši kritičnu vrednost, kao u našem primeru, možemo zaključiti da je efekat faktora statistički značajan.

Međutim, sama ANOVA nam ne otkriva koja grupa (ili grupe) se razlikuje od ostalih. Da bismo dobili precizniju sliku o prirodi ovih razlika, neophodno je sprovesti **post-hoc analizu**.

## 10.7 Post-hoc analiza

Naziv post-hoc nam govori da je reč o analizi koja se sprovodi nakon što dobijemo značajne rezultate ANOVA analize. Post-hoc analiza je niz testova koji se izvršavaju kada ANOVA test pokaže postojanje značajne razlike između grupa. Glavni cilj post-hoc analize je da precizno identifikuje koje grupe se međusobno razlikuju i kolika je veličina tih razlika. Na taj način dobijamo precizniju sliku o prirodi uticaja faktora na zavisnu varijablu.

Postoji nekoliko pristupa post-hoc analizi, a izbor zavisi od broja grupa, veličine uzorka i prirode podataka. U ovom udžbeniku predstavićemo jedan od najčešće korišćenih metoda - **Tjukijev post-hoc test** (Tukey, 1949). Ovaj test je efikasan i pruža sve ključne informacije o međugrupnim razlikama.

Tjukijev test spada u „konzervativne“ testove jer rešava problem višestrukog testiranja. Kada poredimo više grupa međusobno, povećavamo verovatnoću greške tipa I. Tjukijev test elegantno rešava ovaj problem korigujući nivo značajnosti za svako pojedinačno poređenje. Zato nosi i naziv Tjukijev test „poštenih značajnih razlika“ (engl. HSD – Honest Significant Differences).

Osnovni princip testa je sledeći: svaki par razlika između grupa poredi se sa kritičnom vrednošću određenom za taj specifični par. Matematička pozadina određivanja kritičnih vrednosti Tjukijevog kriterijuma je izvan okvira ovog udžbenika, ali u R-u možemo brzo i efikasno sprovesti i ANOVA analizu i Tjukijev post-hoc test.

U R-u, ANOVA analizu i Tjukijev post-hoc test izvršavamo pomoću funkcija aov i TukeyHSD. Pogledajmo kako to funkcioniše u praksi.

anova\_model <- aov(cena ~ lokacija, data = podaci)  
  
summary(anova\_model)

Line 1

Prvo definišemo model pomoću funkcije aov. Model zadajemo u obliku zavisna\_varijabla ~ faktor. U našem slučaju, zavisna varijabla je cena, a faktor je lokacija.

Line 3

Funkcijom summary dobijamo sve relevantne informacije ANOVA analize. Posebno je važna poslednja kolona koja sadrži p-vrednost za F-test.

Df Sum Sq Mean Sq F value Pr(>F)   
lokacija 2 55642 27821 3.14 0.0462 \*  
Residuals 147 1302325 8859   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Rezultate dobijamo u vidu ANOVA tabele. U ovom jednostavnom slučaju analize varijanse imamo jedan faktor i jednu zavisnu varijablu. Za složenije modele sa više faktora, tabela će sadržati dodatne redove.

Prvi red prikazuje mere faktorskog varijabiliteta, dok drugi red sadrži mere rezidualnog varijabiliteta. F-statistika iznosi 3.14, uz odgovarajuću p-vrednost od 0.0462.

Nakon ANOVA analize, sprovodimo Tjukijev post-hoc test jednom linijom koda.

tukey\_test <- TukeyHSD(anova\_model)  
  
tukey\_test

Line 1

Funkcijom TukeyHSD sprovodimo Tjukijev post-hoc test nad prethodno definisanim modelom

Line 3

Ispisujemo rezultate testa

Tukey multiple comparisons of means  
 95% family-wise confidence level  
  
Fit: aov(formula = cena ~ lokacija, data = podaci)  
  
$lokacija  
 diff lwr upr p adj  
Širi centar-Nova naselja 36.74 -7.8314172 81.31142 0.1280701  
Uži centar-Nova naselja 44.00 -0.5714172 88.57142 0.0538469  
Uži centar-Širi centar 7.26 -37.3114172 51.83142 0.9213366

Rezultati nam otkrivaju zanimljivu sliku. U ispisu vidimo tri reda koji prikazuju poređenja između grupa. Svaki red sadrži četiri ključne informacije:

* diff: razlika između aritmetičkih sredina grupa
* lwr: donja granica intervala poverenja od 95% za razliku
* upr: gornja granica intervala poverenja od 95% za razliku
* p adj: prilagođena p-vrednost za razliku dve aritmetičke sredine

Analiza pokazuje da je jedina razlika koja se približava statističkoj značajnosti ona između prosečne cene kafe u lokalima u užem centru grada i lokalima u novim naseljima. P-vrednost za ovu razliku iznosi 0.054, što je malo iznad 0.05, te možemo zaključiti da je razlika značajna samo na nivou . Na nivou uzorka, ta razlika iznosi 44 dinara u prosečnoj ceni kafe.

Faktorski varijabilitet koji smo otkrili ANOVA analizom proizlazi upravo iz ove razlike između lokala u užem centru i novih naselja. Za sva ostala poređenja nemamo dovoljno dokaza da bismo tvrdili da postoje značajne razlike u cenama kafe - drugim rečima, te razlike nisu dovoljno izražene u odnosu na varijabilnost cena unutar svake grupe.

Vraćajući se na početnu raspravu između dva studenta, možemo reći da su obojica delimično bili u pravu. Prvi student je ispravno primetio postojanje razlike u cenama između određenih delova grada, ali drugi student je takođe bio u pravu kada je ukazao na značajan varijabilitet cena unutar svake lokacije. ANOVA i post-hoc testovi nam omogućavaju da precizno kvantifikujemo ove odnose i bolje razumemo prirodu uticaja lokacije na cenu kafe.

### 10.7.1 Veličina efekta

U prethodnom delu videli smo da postoje relativno male razlike u prosečnoj ceni kafe između lokacija. Možemo li preciznije izmeriti ove razlike? Kao i kod t-testa, rešenje leži u konceptu **veličine efekta**.

Za kvantifikovanje veličine efekta u ANOVA analizi koristimo meru koja se naziva **eta kvadrat** (). Eta kvadrat izračunavamo kao količnik faktorskog i ukupnog varijabiliteta. Vrednosti blizu 0 ukazuju na zanemarljiv uticaj faktora na zavisnu varijablu, dok vrednosti koje se približavaju 1 otkrivaju snažan uticaj faktora.

U našem slučaju, eta kvadrat iznosi:

eta2 <- SSF / (SSF + SSR)  
  
cat("Eta kvadrat: ", round(eta2, 2), "\n")

Line 1

Izračunavamo eta kvadrat kao količnik faktorskog i ukupnog varijabiliteta

Eta kvadrat: 0.04

Slično tumačenju korelacije i koeficijenta determinacije, postoje opšteprihvaćene granične vrednosti za interpretaciju eta kvadrata:

* : mali efekat
* : srednji efekat
* : veliki efekat

U našem primeru dobili smo mali efekat, što znači da faktor lokacije ima ograničen uticaj na cenu kafe u Novom Sadu. Ovaj rezultat je u skladu sa post-hoc analizom, koja je otkrila da je jedina značajna razlika u cenama između lokala u užem centru grada i novih naselja. Eta kvadrat nam precizno pokazuje da lokacija objašnjava samo 4% ukupne varijabilnosti u cenama kafe.

## 10.8 ANOVA i regresija

Analiza varijanse je fundamentalni alat za razumevanje statističkih modela. Ona nam omogućava da razdvojimo varijabilitet koji pripisujemo nezavisnoj varijabli od varijabiliteta koji potiče iz drugih izvora: slučajnosti i dejstva neidentifikovanih ili nekontrolisanih faktora.

Ako pažljivo razmislimo, ovaj cilj je gotovo identičan ciljevima metoda koje smo ranije obradili u ovom udžbeniku. F-test smo već susreli u poglavlju o regresiji (odeljak [Odeljak 8.8](#sec-regresioni-model-r)). Vreme je da otkrijemo njegovu pravu prirodu.

Hajde da ponovo konstruišemo regresioni model iz tog poglavlja.

podaci\_reg <- read.csv("https://gist.githubusercontent.com/atomashevic/1ea72e8ea912083492b60af7ecd07fbf/raw/7d658eb07d321126989acf774a8428dc712bfc9d/zdravstvo.csv")  
  
model\_reg <- lm(kvalitet ~ budzet, data = podaci\_reg)  
  
summary(model\_reg)

Call:  
lm(formula = kvalitet ~ budzet, data = podaci\_reg)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-29.523 -11.245 -0.986 9.639 36.324   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 4.907e+01 4.073e+00 12.05 < 2e-16 \*\*\*  
budzet 3.170e-03 7.354e-04 4.31 6.4e-05 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 14.77 on 58 degrees of freedom  
Multiple R-squared: 0.2426, Adjusted R-squared: 0.2295   
F-statistic: 18.58 on 1 and 58 DF, p-value: 6.404e-05

Vidimo da F-statistika iznosi 18.58, uz stepene slobode 1 i 58. P-vrednost ovog testa je veoma blizu nule (6.404e-05, što je zapravo 0.00006404). Ovo nas direktno vraća na ANOVA analizu. Tačnije, regresija je specifičan slučaj ANOVA analize gde faktor predstavlja kontinuiranu kvantitativnu varijablu.

ANOVA nalazi svoju primenu i pri analizi regresionog modela. Kada konstruišemo regresioni model, varijabilitet zavisne varijable možemo razdvojiti na deo koji je objašnjen regresionim modelom i deo koji nije. ANOVA analiza nam omogućava da precizno procenimo koliko faktor (nezavisna varijabla) doprinosi objašnjenju varijabiliteta zavisne varijable. Sam postupak razlaganja varijabiliteta je identičan kao u primeru koji smo demonstrirali u ovom poglavlju, što znači da je i primena u R-u gotovo ista.

Koristićemo ponovo funkciju anova(), pri čemu je jedini ulazni parametar regresioni model.

anova\_reg <- anova(model\_reg)  
  
anova\_reg

Analysis of Variance Table  
  
Response: kvalitet  
 Df Sum Sq Mean Sq F value Pr(>F)   
budzet 1 4052.2 4052.2 18.576 6.404e-05 \*\*\*  
Residuals 58 12652.4 218.1   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Vidimo da je regresiona ili faktorska varijansa 18 puta veća od rezidualne, što ukazuje na to da je ovaj regresioni model uspešno objasnio varijabilitet zavisne varijable. Ovo predstavlja dodatni korak u dijagnostici regresionog modela.

U poređenju sa klasičnom analizom varijanse, u ovom specifičnom regresionom slučaju stepeni slobode odgovaraju broju nezavisnih varijabli u modelu. Pošto imamo jednu nezavisnu varijablu, faktorski stepen slobode je 1, a rezidualni stepen slobode je 58. Drugim rečima, umesto broja grupa, označava broj nezavisnih varijabli u modelu. Takođe, primena post-hoc analize nema smisla u ovom slučaju jer nemamo diskretno definisane grupe koje bismo mogli analizirati.

Ipak, činjenica da se regresija može posmatrati kao poseban slučaj analize varijanse otkriva nam dublju povezanost ovih metoda. Tu vezu ćemo detaljnije istražiti u završnom poglavlju ovog udžbenika, gde ćemo otkriti jednu neočekivanu istinu skrivenu u samom nazivu udžbenika.

### 10.8.1 Lična karta metoda: ANOVA

**Šta radi?** Testira postojanje uticaja faktora na zavisnu varijablu, tj. provera da li postoje statistički značajne razlike između aritmetičkih sredina grupa.

**Kada koristiti?** Kada imamo više od dve grupe i želimo da proverimo da li postoji statistički značajan uticaj faktora na zavisnu varijablu.

**Koliko varijabli imammo?** Jedna kategorijalna varijabla (faktor) i jedna kvantitativna varijabla (zavisna).

**Kako glase hipoteze?** Nulta hipoteza tvrdi da ne postoji razlika između aritmetičkih sredina grupa, dok alternativna hipoteza tvrdi da postoji bar jedna grupa koja se razlikuje od ostalih.

**Kako izgleda statistika testa?** Koristimo F-test da testiramo odnos između faktorske i rezidualne varijanse.

**Kako računamo p-vrednost?** P-vrednost se računa kao površina ispod krive F-distribucije koja se nalazi desno od naše F-statistike.

## 10.9 Zadaci

|  |  |
| --- | --- |
|  | **Zadatak 1**  U ovom zadatku koristimo podatke iz komparativnog istraživanja sa prostora bivše Jugoslavije. Istraživači su merili stepen jugonostalgije kod ispitanika na skali od 0 („Ne osećam nostalgiju prema vremenu nekadašnje Jugoslavije“) do 10 („Osećam veliku nostalgiju prema vremenu nekadašnje Jugoslavije“). Ispitanici su podeljeni u 4 grupe prema državi iz koje dolaze: Srbija, Hrvatska, Crna Gora i Slovenija.  Koristeći varijable drzava i jugonostalgija, sprovedite analizu varijanse i post-hoc testove kako biste utvrdili postoje li statistički značajne razlike u stepenu jugonostalgije između građana različitih država.  Razmislite: Kako u ovom slučaju interpretiramo dejstvo faktora i po čemu se ono razlikuje od primera sa cenama kafe koji smo obradili ranije?  Podatke možete učitati na sledeći način:  podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/51ec08bc58476230f59f51ef7435efaa/raw/27e3ee1ad8f783ec6b040874b000784d2fe96c40/jugonostalgija.csv") |

|  |  |
| --- | --- |
|  | **Zadatak 2**  U ovom zadatku analiziramo podatke o subjektivnom zadovoljstvu životom (subjektivnom blagostanju) građana sa različitim stepenima obrazovanja. U istraživanju su definisane četiri grupe ispitanika prema stepenu obrazovanja:   1. bez završene osnovne škole 2. završena osnovna škola 3. završena srednja škola 4. završen fakultet ili visoka škola   Koristeći varijable obrazovanje i zadovoljstvo, potrebno je sprovesti analizu varijanse i post-hoc testove kako bi se utvrdilo postojanje statistički značajnih razlika u subjektivnom blagostanju između grupa ispitanika sa različitim nivoima obrazovanja.  Podatke možete učitati na sledeći način:  podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/c3189aa3191a7c7c2306bb726f1cb80d/raw/d7472ad5736d165a3084a4f3c5b028818c111447/obrazovnje-zivot.csv") |

|  |  |
| --- | --- |
|  | **Zadatak 3 \*\***  U podacima iz prethodnog zadatka uočavamo problem - nije ispunjena pretpostavka o **homogenosti varijanse** između grupa. Variance zadovoljstva životom značajno se razlikuju među grupama. Ovo može biti posledica nekog nekontrolisanog faktora koji nismo uzeli u obzir ili jednostavno priroda uticaja obrazovanja na zadovoljstvo životom.  Za prevazilaženje ovog problema, potrebno je ponoviti analizu koristeći Velčov ANOVA metod (eng. Welch’s ANOVA). Ovaj metod je elegantan jer ne zahteva pretpostavku o homogenosti varijanse između grupa.  Velčova ANOVA je modifikacija standardne ANOVA analize koja koristi korekciju za različite varijanse grupa. U R-u je implementirana kroz funkciju oneway.test() sa parametrom var.equal = FALSE.  Sprovedite analizu koristeći Velčov metod i nakon toga uradite post-hoc testove. Obratite posebnu pažnju na sledeća pitanja: - Kakve razlike između grupa otkriva ova analiza? - Da li možemo govoriti o linearnom uticaju obrazovanja na zadovoljstvo životom? |

# 11. Razotkrivanje

## 11.1 Sve je povezano

U prethodnim poglavljima upoznali smo različite statističke metode. Krenuli smo od t-testa za jedan uzorak, prešli na t-test za dva uzorka, zatim na regresiju, korelaciju i ANOVA-u, i konačno stigli do regresione analize. Kroz ovaj proces otkrivali smo duboke veze između ovih naizgled različitih metoda.

Posebno je zanimljivo kako su koeficijent nagiba regresije i Pirsonov koeficijent korelacije zapravo dva lica iste pojave - načini da kvantifikujemo kovarijansu između dve varijable. Kada izračunamo koeficijent korelacije, mi zapravo dobijamo geometrijsku sredinu dva koeficijenta nagiba: jedan iz regresije gde je prva varijabla zavisna, a drugi iz regresije gde je ona nezavisna promenljiva.

Videli smo da ANOVA efikasno rešava problem višestrukog testiranja hipoteza. Ona nam omogućava da istovremeno testiramo više grupa, bez povećanja verovatnoće greške tipa I. Post-hoc testovi nam precizno pokazuju koje grupe se međusobno razlikuju, ispravljajući ograničenja standardnog t-testa.

Regresiona analiza se prirodno nadovezuje na ANOVA-u, pri čemu kategoričku varijablu zamenjujemo kvantitativnom. Dekompozicija varijanse, koja je ključna za ANOVA-u, ima svoju paralelu u regresionoj analizi gde ispitujemo odnos između regresionog i rezidualnog dela ukupnog varijabiliteta zavisne varijable.

Ove veze nisu slučajne niti su tu da zbune istraživača pri odabiru metoda. One ukazuju na duboku povezanost između svih analiza koje smo predstavili u ovom udžbeniku. Na kraju, one razotkrivaju i prividnu nejasnoću u samom nazivu udžbenika - sve ove analize su manifestacije jednog jedinstvenog statističkog okvira: **opšteg linearnog modela**.

## 11.2 Opšti linearni model

Ponovo se vraćamo na linije. Sve što smo radili kroz ovaj udžbenik svodi se na jedan princip - pronalaženje linije koja najbolje opisuje odnos između dve varijable. U regresiji i korelaciji smo to direktno prikazali kroz vizualizacije na dijagramu raspršenosti. Iako za neke druge analize, poput ANOVA-e i t-testa, nismo eksplicitno crtali linije, osnovni princip ostaje isti.

Da bismo razumeli kako konstruišemo ove linije, moramo sagledati problem na opštijem nivou. U svim predstavljenim analizama radimo sa dve varijable: jednom zavisnom i jednom nezavisnom. Opšti linearni model je statistički okvir koji može predstaviti odnos između ove dve varijable kroz linearnu funkciju.

Opšti linearni model u svom osnovnom obliku možemo zapisati kao:

gde je zavisna varijabla, nezavisna varijabla, predstavlja koeficijente modela, a njegovu grešku. Ovaj model je osnova za sve analize koje smo predstavili u ovom udžbeniku. Koristimo kao jedinstvenu oznaku koja sadrži koeficijente modela jer, kao što ćemo videti, u nekim situacijama model može imati više od jednog koeficijenta.

Interpretacija koeficijenata modela zavisi od prirode varijabli koje koristimo. Ovo možemo najbolje ilustrovati na primeru koji obuhvata različite tipove varijabli, što će nam omogućiti da demonstriramo primenu svih analiza kroz perspektivu linearnog modela.

Razmotrimo podatke iz anketnog istraživanja o medijskom ponašanju građana. Glavni fokus je na samoproceni ispitanika o vremenu koje provode prateći političke vesti kroz različite medije (TV, radio, štampani mediji, internet). Uz to, prikupljeni su i podaci o izlaznosti na prethodnim parlamentarnim izborima, starosti i obrazovanju ispitanika. Varijable su organizovane na sledeći način:

* pracenje\_vesti - vreme provedeno u praćenju političkih vesti u minutima dnevno
* glasao - učešće ispitanika na prethodnim izborima (vrednosti „Da“ i „Ne“)
* starost - starost ispitanika u godinama
* obrazovanje - stepen obrazovanja ispitanika (vrednosti „Osnovno“, „Srednje“, „Više“)

Pogledajmo kako možemo primeniti opšti linearni model na ovim podacima. Počećemo sa analizom odnosa između praćenja vesti i izlaznosti na izbore.

## 11.3 Alternativni t-test

U ovom primeru konstruišemo model: pracenje\_vesti ~ glasao. Zavisna varijabla je pracenje\_vesti, a nezavisna varijabla je glasao. Nezavisna varijabla je binarna i kategorička.

Ovo je klasičan primer t-testa, ali ovde ga posmatramo kao deo opšteg linearnog modela. Kako primeniti ovaj model u našem slučaju? Model možemo zapisati u sledećem obliku:

Pogledajmo šta se ovde dešava. Model ima dva koeficijenta: i . Grupa ispitanika koji su odgovorili „Da“ na pitanje o glasanju je **referentna grupa**. To znači da je koeficijent modela (koji u regresiji nazivamo slobodnim koeficijentom) zapravo aritmetička sredina vrednosti zavisne varijable za ovu grupu. Koeficijent predstavlja razliku između referentne grupe i grupe ispitanika koji nisu glasali. U regresiji, ovo je ono što nazivamo koeficijentom nagiba.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Indikatorske varijable**  Varijablu nazivamo **indikatorska** ili *dummy* varijabla. Ona nastaje transformacijom početne varijable glasao u numeričku binarnu varijablu gde vrednost 1 označava da ispitanik nije glasao, a 0 da jeste. Ova transformacija nam omogućava da koristimo kategoričke varijable u linearnim modelima.  Za neke kategoričke varijable nije dovoljna samo jedna indikatorska varijabla. U tim slučajevima koristimo više indikatorskih varijabli, pri čemu svaka predstavlja jednu kategoriju. Na primer, za varijablu obrazovanje referentna kategorija je „Osnovno“, pa koristimo dve indikatorske varijable za „Srednje“ i „Visoko“. Kombinacijom ove dve varijable, koje mogu imati vrednost 0 ili 1, možemo da rekonstruišemo sve tri kategorije obrazovanja.   | Srednje obrazovanje | Visoko obrazovanje | Kategorija | | --- | --- | --- | | 0 | 0 | Osnovno | | 1 | 0 | Srednje | | 0 | 1 | Visoko |   Ovim pristupom transformišemo kategoričke podatke u format koji linearni modeli mogu da obrade. Kada kodiramo kategoričke varijable na ovaj način, možemo precizno meriti kako svaka kategorija utiče na zavisnu varijablu. Na primer, možemo direktno izmeriti koliko kategorija „Visoko“ utiče na zavisnu varijablu u poređenju sa referentnom kategorijom „Osnovno“. |

Kroz opšti linearni model u ovom slučaju saznajemo prosečnu razliku u vremenu praćenja političkih vesti između onih koji su glasali i onih koji nisu. Model nam daje jasan uvid u to koliko minuta dnevno više ili manje provode apsitenti u odnosu na birače prateći političke vesti. Ako je koeficijent koji opisuje ovu razliku statistički značajan, možemo zaključiti da uočena razlika nije slučajna već postoji i na nivou cele populacije.

Pogledajmo kako to izgleda u praksi koristeći R.

podaci <- read.csv("https://gist.githubusercontent.com/atomashevic/8cfc7453347471c26ea0892a1a946c4e/raw/0c2682066f2a298d73d52405ca87e7ae45c28726/politika.csv")  
  
model\_1 <- lm(pracenje\_vesti ~ glasao, data = podaci)  
  
summary(model\_1)

Line 3

Primenjujemo linearni model na podatke koristeći funkciju lm. Zavisna varijabla je pracenje\_vesti, a nezavisna varijabla je glasao.

Call:  
lm(formula = pracenje\_vesti ~ glasao, data = podaci)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-52.589 -21.589 -0.837 20.163 94.411   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 52.589 1.235 42.584 <2e-16 \*\*\*  
glasaoNe -3.751 1.897 -1.978 0.0482 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 29.64 on 998 degrees of freedom  
Multiple R-squared: 0.003905, Adjusted R-squared: 0.002907   
F-statistic: 3.912 on 1 and 998 DF, p-value: 0.04821

Pogledajmo tabelu sa koeficijentima. Drugi red nosi oznaku glasaoNe, gde vidimo da je vrednost koeficijenta jednaka -3.751. Ova vrednost nam pokazuje da ispitanici koji nisu glasali u proseku prate vesti o politici 3.751 minuta manje dnevno u poređenju sa onima koji su glasali. Standardna greška ovog koeficijenta iznosi 1.897, što daje t-vrednost od -1.98. P-vrednost je nešto manja od 0.05, što nas dovodi do zaključka da je uočena razlika statistički značajna.

Dakle, pokazali smo da postoji merljiva razlika u praćenju vesti između birača i apstinenata. Preciznije, utvrdili smo da izlazak na izbore ima jasan efekat na vreme provedeno u praćenju političkih vesti - oni koji nisu glasali u proseku provedu nekoliko minuta manje dnevno prateći politička dešavanja.

Kada pogledamo rezultate, vidimo da je Intercept koeficijent jednak 52.589, što predstavlja , odnosno aritmetičku sredinu vrednosti zavisne varijable za referentnu grupu. Rezultati F-testa, koji upoređuje objašnjenu i neobjašnjenu varijansu modela, takođe su zanimljivi. Test je statistički značajan (), a F-statistika od 3.91 nam pokazuje da model uspešno objašnjava značajan deo ukupne varijanse.

Da bismo potvrdili valjanost ove specifikacije modela, uporedićemo rezultate sa klasičnim t-testom koristeći funkciju t.test u R-u.

biraci <- podaci$pracenje\_vesti[podaci$glasao == "Da"]  
apsistenti <- podaci$pracenje\_vesti[podaci$glasao == "Ne"]  
  
t.test(biraci, apsistenti)

Line 1

Izdvajamo vrednosti zavisne varijable za birače.

Line 2

Izdvajamo vrednosti zavisne varijable za apstinente.

Line 4

Primenjujemo t-test na ove dve grupe.

Welch Two Sample t-test  
  
data: biraci and apsistenti  
t = 1.9895, df = 930.23, p-value = 0.04694  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 0.05090431 7.45165072  
sample estimates:  
mean of x mean of y   
 52.58854 48.83726

Dobijamo identične rezultate koji nas vode ka istom zaključku. Svaki linearni model proizvodi najmanje dva koeficijenta koje možemo iskoristiti za konstrukciju linije. Ali kakva je to linija u slučaju t-testa?

|  |
| --- |
| Slika 11.1: Linearni model za praćenje vesti prema glasanju |

Ova linija povezuje dve aritmetičke sredine i predstavlja suštinu t-testa kroz perspektivu linearnog modela. Slobodni koeficijent označava aritmetičku sredinu referentne grupe, dok pokazuje razliku između aritmetičkih sredina dve grupe. Pošto je koeficijent nagiba negativan, linija opada i pokazuje da ispitanici koji nisu glasali u proseku manje vremena posvećuju praćenju vesti.

Upravo ova jednostavna geometrijska interpretacija - linija koja povezuje dve aritmetičke sredine - daje nam jasan uvid u to kako t-test za dva nezavisna uzorka funkcioniše kao poseban slučaj opšteg linearnog modela.

## 11.4 Regresija i korelacija kao liearni model

Sada ćemo ispitati vezu između vremena provedenog u praćenju političkih vesti i starosti ispitanika. Ovaj odnos možemo predstaviti elegantno prostim linearnim modelom:

gde je vreme provedeno u praćenju političkih vesti, starost ispitanika, slobodni koeficijent, koeficijent nagiba, a greška modela.

Slobodni koeficijent ima jasnu interpretaciju - to je aritmetička sredina zavisne varijable kada je vrednost nezavisne varijable nula. U našem slučaju ta vrednost nema praktičnog smisla jer bi predstavljala procenu vremena koje novorođenče provodi prateći političke vesti. Bez obzira na to, koeficijent nam daje korisnu informaciju - on pokazuje prosečnu promenu u vremenu praćenja vesti za svaku dodatnu godinu starosti ispitanika.

Pogledajmo kako to izgleda u R-u.

model\_2 <- lm(pracenje\_vesti ~ starost, data = podaci)  
  
summary(model\_2)

Line 1

Primenjujemo linearni model na podatke koristeći funkciju lm. Zavisna varijabla je pracenje\_vesti, a nezavisna varijabla je starost.

Call:  
lm(formula = pracenje\_vesti ~ starost, data = podaci)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-61.085 -20.929 -0.429 19.563 93.422   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 30.74442 3.03477 10.131 < 2e-16 \*\*\*  
starost 0.44772 0.06395 7.001 4.66e-12 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 28.99 on 998 degrees of freedom  
Multiple R-squared: 0.04681, Adjusted R-squared: 0.04586   
F-statistic: 49.01 on 1 and 998 DF, p-value: 4.662e-12

Vidimo da slobodni koeficijent iznosi približno pola sata, dok je koeficijent nagiba 0.4. To nam pokazuje da se vreme provedeno u praćenju vesti o politici povećava za 0.4 minuta dnevno sa svakom dodatnom godinom starosti ispitanika. T-statistika od 7 i p-vrednost bliska nuli jasno ukazuju na statističku značajnost ove veze. F-statistika od 49 potvrđuje snagu modela - varijansa objašnjena modelom značajno nadmašuje neobjašnjenu varijansu.

Model nam ne daje direktnu informaciju o korelaciji između varijabli, ali poznajemo koeficijent determinacije . Iz njega možemo jednostavno izračunati koeficijent korelacije u R-u.

r = sqrt(summary(model\_2)$r.squared)  
  
cat("Koeficijent korelacije je:", round(r, 2))

Line 1

Izračunavamo koeficijent korelacije iz koeficijenta determinacije dobijenog iz rezultata linearnog modela.

Koeficijent korelacije je: 0.22

Dobili smo pozitivnu korelaciju umerenog intenziteta između vremena provedenog u praćenju političkih vesti i starosti ispitanika. Linearni model nam je tako pružio sve informacije potrebne za interpretaciju i regresionog i korelacionog odnosa između dve kvantitativne varijable. Ova veza je očekivana i intuitivna - sa godinama raste interesovanje za političke teme i vesti.

## 11.5 Alternativna ANOVA

Zamenom starosti sa obrazovanjem u prethodnom primeru dobijamo linearni model koji odgovara modelu opisanom u prethodnom poglavlju. S obzirom na to da obrazovanje ima tri kategorije, model će imati sledeći oblik:

gde su , i koeficijenti modela, a i indikatorske varijable koje označavaju da li ispitanik ima srednji ili visoki stepen obrazovanja. predstavlja aritmetičku sredinu vrednosti zavisne varijable za referentnu grupu, dok i predstavljaju razlike između aritmetičke sredine referentne grupe i grupa koje se upoređuju.

Pogledajmo kako to izgleda u praksi koristeći R.

model\_3 <- lm(pracenje\_vesti ~ obrazovanje, data = podaci)  
  
summary(model\_3)

Line 1

Primenjujemo linearni model na podatke koristeći funkciju lm. Zavisna varijabla je pracenje\_vesti, a nezavisna varijabla je obrazovanje.

Call:  
lm(formula = pracenje\_vesti ~ obrazovanje, data = podaci)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-68.233 -20.388 -0.388 18.843 87.612   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 40.463 2.207 18.334 < 2e-16 \*\*\*  
obrazovanjeSrednje 6.925 2.482 2.790 0.00538 \*\*   
obrazovanjeVisoko 27.771 2.889 9.612 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 28.09 on 997 degrees of freedom  
Multiple R-squared: 0.1061, Adjusted R-squared: 0.1043   
F-statistic: 59.17 on 2 and 997 DF, p-value: < 2.2e-16

Analiza rezultata pokazuje nekoliko ključnih elemenata. Prvo, oba koeficijenta su pozitivna i statistički značajna, što ukazuje na jasnu razliku u prosečnom vremenu praćenja političkih vesti između posmatranih grupa i referentne grupe. F-statistika od 59.17 sa p-vrednošću blizu 0 odbacuje omnibus nultu hipotezu o jednakosti svih koeficijenata modela. Koeficijent determinacije približno iznosi 0.1, što znači da model objašnjava 10% varijanse zavisne varijable.

Rezultate ovog linearnog modela možemo predstaviti grafički kroz linije na dijagramu raspršenosti. Postavlja se pitanje - kako konstruisati liniju kada imamo tri koeficijenta umesto uobičajena dva?

Rešenje leži u konstrukciji dva povezana linearna segmenta. Prvi segment prikazuje razliku između aritmetičke sredine referentne grupe i grupe sa srednjim obrazovanjem, dok drugi segment pokazuje razliku između aritmetičkih sredina grupa sa srednjim i visokim obrazovanjem. Ova dva segmenta spajaju se u tački koja predstavlja aritmetičku sredinu grupe sa srednjim obrazovanjem, formirajući jedinstvenu sliku odnosa između nivoa obrazovanja i vremena praćenja vesti.

|  |
| --- |
| Slika 11.2: Linearni model za praćenje vesti prema obrazovanju |

Na osnovu ovog grafičkog prikaza jasno se vidi porast vremena provedenog u praćenju političkih vesti sa višim nivoom obrazovanja ispitanika. Crveno osenčena oblast oko linije predstavlja interval poverenja - našu procenu neizvesnosti u ovom odnosu. Kada kombinujemo vizualni prikaz sa numeričkim rezultatima linearnog modela, dobijamo preciznu sliku o tome kako stepen obrazovanja utiče na vreme koje ispitanici posvećuju praćenju političkih vesti.

Kroz ova tri primera demonstrirali smo kako se naizgled različite statističke analize mogu svesti na različite specifikacije opšteg linearnog modela. Model nam pruža moćan i jednostavan alat za opisivanje odnosa između dve varijable kroz linearnu funkciju. Ovakav pristup nam otkriva suštinsko jedinstvo svih statističkih analiza koje smo obradili u ovom udžbeniku.

## 11.6 Šta dalje?

Stigli smo do kraja udžbenika gde naše putovanje kroz statističko zaključivanje i konstrukciju statističkih modela završavamo prostim linearnim modelom.

Važno je razumeti da ovaj udžbenik nije zamišljen kao priručnik za primenu statističkih metoda, već kao vodič za razumevanje njihove suštine. Nismo se fokusirali na mehaničko praćenje koraka ili proveru uslova za primenu određenih metoda. Umesto toga, cilj je bio da shvatite logiku koja stoji iza statističkih metoda - ne da biste ih samo primenjivali, već da biste mogli da koristite to razumevanje kada **razmišljate** o istraživačkim problemima.

Mnogi izazovi u primeni statistike u društvenim i drugim naukama nastaju jer se statistička analiza često tretira kao završna faza istraživanja, a ne kao njegov integralni deo. Ovo dovodi do problema poput pogrešne upotrebe p-vrednosti, nerazumevanja koncepta efekta ili nagomilavanja grešaka u zaključivanju. Jedini put ka rešenju je razvijanje dublje intuicije o procesu statističkog zaključivanja i njegovim ključnim elementima.

Kada osmišljavamo istraživanje, biramo varijable i njihove merne skale, potrebna nam je intuicija o tome kako bi mogla izgledati kovarijansa između određenih varijabli, kakva bi mogla biti distribucija razlika između grupa, koja je očekivana veličina efekta, i da li je potrebno testirati jednosmernu ili dvosmernu hipotezu. Ova intuicija je ključna za kvalitetno istraživanje.

## 11.7 Preporuke za nastavak putovanja

Materijal predstavljen u ovom udžbeniku je prvi korak u izgradnji intuicije. Ipak, on je ograničen na relativno jednostavne primere koji služe za ilustraciju osnovnih ideja.

Ako želite da nastavite sa produbljivanjem razumevanja statističkog zaključivanja, pred vama su tri glavna pravca:

1. multivarijantna analiza,
2. strukturno modeliranje,
3. bejzijanska statistika.

Multivarijantna analiza će vam omogućiti da u potpunosti savladate opšti linearni model koji može obuhvatiti različite kombinacije zavisnih i nezavisnih varijabli. Ova analiza se može proširiti i na latentne konstrukte - varijable koje ne možemo direktno meriti, što je posebno bitno u društvenim naukama. Ovaj pravac će vas voditi kroz niz naprednih metoda: eksploratornu faktorsku analizu, analizu glavnih komponenti, kanoničku korelaciju, diskriminantnu analizu, logističku regresiju i hijerarhijske modele.

Odličan prvi korak u ovom pravcu je udžbenik *Regression and Other Stories* (Gelman i ostali, 2020) (dostupan besplatno [ovde](https://users.aalto.fi/~ave/ROS.pdf)). Iako ne pokriva sve navedene metode i modele, predstavlja izvrstan vodič kroz multivarijantnu analizu zasnovanu na opštem linearnom modelu i linearnoj regresiji.

Drugi pravac je strukturno modeliranje koje se bavi modelovanjem složenih odnosa između varijabli, posebno kada su te varijable latentne. Ovaj pristup je od ključnog značaja u psihologiji, sociologiji, ekonomiji i drugim društvenim naukama. On vas vodi ka upoznavanju modela kao što su: SEM (strukturno jednačenje modela), LCA (latentna klasna analiza), LTA (latentna tranziciona analiza) i drugi.

Za temeljno razumevanje strukturalnog modeliranja, preporučujem serijal predavanja Saše Epskampa (Epskamp, 2019a, 2019b) dostupnih kroz kurseve [SEM1](https://www.youtube.com/watch?v=fGdsiugwO0k&list=PLliBbGBc5nn0NdoFO4nlbwZrT9TJ5V9sJ) i [SEM2](https://www.youtube.com/watch?v=QOUHlPz2wtk&list=PLliBbGBc5nn3tQiIBHFd-YvWK6dXb4ryr).

Na kraju, bejzijanska statistika predstavlja sasvim drugačiji pristup statističkom zaključivanju koji uspostavlja neposredniju vezu između osnovnih principa verovatnoće (koje smo upoznali u ovom udžbeniku) i procesa zaključivanja. Najbolji resurs i najprirodniji način za izgradnju intuicije o bilo kom obliku statističkog zaključivanja je knjiga i kurs *Statistical Rethinking* (McElreath, 2020). Iako knjiga nije besplatno dostupna, kompletan sadržaj kursa zasnovanog na knjizi, uključujući video snimke predavanja, možete besplatno pronaći [ovde](https://github.com/rmcelreath/stat_rethinking_2024).

Na kraju, ovaj udžbenik dopunjuje kolekciju izdvanja Filozofskog fakulteta u Novom Sadu koja se bave statističkim metodama u društvenim naukama. Ovi resursi su besplatno dostupni studentima i svim zainteresovanima. Udžbenici pokrivaju oblasti deskriptivne statistike (Sokolovska, 2013), tehnike vizualizacije u bazičnoj statistici (Pajić, 2020) i tehnike vizualizacije u naprednoj statistici i statističkom zaključivanju (Jevremov, 2023).

Koji god put daljeg učenja da odaberete, nadam se da ćete u njemu pronaći inspiraciju i intelektualno zadovoljstvo!

## 11.8 Zadaci

Ovo poglavlje nema nove zadatke. Umesto toga, vratite se na zadatke iz prethodnih poglavlja i rešite ih koristeći funkciju lm u R-u. Time ćete utvrditi razumevanje opšteg linearnog modela i njegove povezanosti sa svim statističkim metodama koje smo obradili.

# Literatura

Allaire, J. J., Teague, C., Xie, Y., & Dervieux, C. (2022). *Quarto*. Zenodo. <https://doi.org/10.5281/ZENODO.5960048>

Epskamp, S. (2019a). *SEM1 (2019) - Stats Recap*.

Epskamp, S. (2019b). *SEM2 Lecture 1 - Expectation and Covariance Algebra*.

Gauss, C.-F. (1823). *Theoria combinationis observationum erroribus minimis obnoxiae*. Henricus Dieterich.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and Other Stories* (1st edition). Cambridge University Press.

Jaynes, E. T. (2003). *Probability theory: The logic of science* (G. L. Bretthorst, Ur.). Cambridge University Press. <https://doi.org/10.1017/cbo9780511790423>

Jevremov, T. (2023). *Primena tehnika vizualizacije u naprednoj statistici*. Filozofski fakultet.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Second edition). CRC Press.

Pajić, D. (2020). *Primena tehnika vizualizacije u bazičnoj statistici* (str. 290). Filozofski fakultet.

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Sokolovska, V. (2013). *Deskriptivna statistika*. Filozofski fakultet.

Stagg, G. W., Lionel, H., & Others. (2023). *webR: The statistical language R compiled to WebAssembly via Emscripten* (Version 0.2.2). <https://docs.r-wasm.org/webr/latest/>

Stuart, A. (1987). *Ideas of Sampling* (3rd izd.). Oxford University Press.

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1. <https://doi.org/10.2307/2331554>

Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, *5*(2), 99. <https://doi.org/10.2307/3001913>