Testovi o očekivanju jedne ili više populacija - geometrijski pristup

Sabljak, Mirna

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Mirna Sabljak

TESTOVI O OČEKIVANJU JEDNE ILI VIŠE POPULACIJA - GEOMETRIJSKI PRISTUP

Diplomski rad

Voditelj rada:
doc.dr.sc. Snježana Lubura Strunjak

Zagreb, veljača 2021.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Veliko hvala mojoj obitelji koja mi je cijelo vrijeme studiranja pružala financijsku i psihološku pomoć,
kolegama i prijateljima koji su mi pomagali u rješavanju matematičkih i životnih problema,
mentorici koja je uvijek bila dostupna te svojim znanjem i iskustvom strpljivo pomagala nastanku ovog rada te konačno
cijelom Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu koji mi je još jednom u životu potvrdio da se rad i upornost uvijek isplate te me naučio kako se boriti i ne odustati od svojih ciljeva.

Sadržaj

Uvod

Statistika je grana matematike koja se bavi analizom podataka te metodama izvođenja zaključaka o promatranom fenomenu na osnovi napravljene analize. U svojim istraživanjima koriste ju stručnjaci iz mnogih drugih znanstvenih područja pa nažalost često zbog nedostatka razumijevanja dolazi do toga da se statističke metode koriste kao recepti iz kuharice. Stoga, iako je to pozadina statističkih metoda, kada govorimo o njima teško da ćemo na prvu pomisliti na vektore, ravnine ili algebarski račun koji ih povezuje. U ovom radu predstavit ćemo geometrijski pristup rješavanju jednog statističkog problema, prvenstveno kako bismo bolje shvatili kako je ta kuharica napisana, tj. zašto tradicionalne statističke metode koristimo na način na koji koristimo. Promatrat ćemo problem uspredbe očekivanja populacija koji bismo tradicionalnom metodom riješili pomoću T-testa, tj. računajući T-statistiku. O geometrijskim pristupima rješavanja nekih drugih statističkih problema možemo čitati u [7].

Pokazat ćemo i kako pojedine matematičke grane, iako se to ponekad ne čini tako, mogu jedna drugoj biti koristan alat u rješavanju problema. Na nekoliko primjera ćemo vidjeti kako hipoteze o očekivanju možemo interpretirati, slikovito prikazati te dobiti odgovore na pitanja koja nas interesiraju upravo koristeći geometriju.

U prvom poglavlju prisjetit ćemo se glavnih rezultata iz analitičke geometrije i statistike koji će nam biti potrebni za geometrijski pristup. Ideja je podsjetiti na osnovne pojmove u navedenim grana koje ćemo koristiti u nastavku, pretpostavljajući da se čitatelj nekad prije susreo s njima. Ukoliko nije, detaljnije o tome može pronaći čitajući [2], [3], [4], [6], [5]. Nakon toga u drugom poglavlju objasnit ćemo geometrijski pristup. Treće poglavlje bavi se testovima o očekivanju jedne populacije, dok se četvrto i peto odnose na dvije, tj. više populacija.

Poglavlje 1

Ponavljanje geometrije i statistike

1.1 Geometrija

Vektori

Za početak ćemo definirati osnovne pojmove vezane uz vektore, njihove međusobne odnose te kako s njima računamo. Uvodimo ih jer podatke koje želimo statistički obraditi vrlo jednostavno možemo zapisati pomoću vektora. Tako dalje, poznavajući u nastavku navedene pojmove i pravila, lako možemo baratati željenim podacima. U raznim literaturama naići ćemo na različite definicije vektora, no navest ćemo onu najbazičniju u skladu s temom koju proučavamo.

Definicija 1.1.1. *Vektor v u N-dimenzionalnom prostoru je niz oblika*

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}.$$

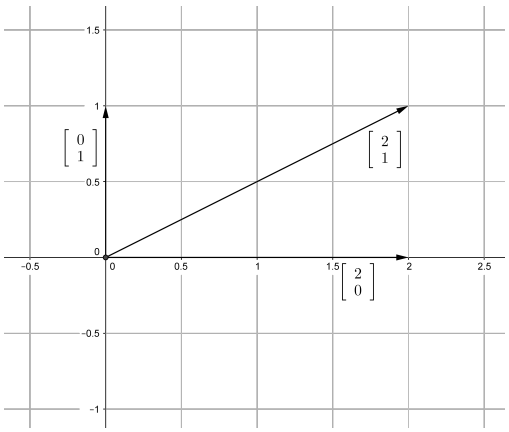
U nastavku teksta vektore ćemo često pisati u transponiranom obliku odnosno

$$\begin{bmatrix} v_1 & v_2 & \dots & v_N \end{bmatrix}^T.$$

Svaki vektor jednoznačno je određen svojom duljinom, smjerom i orijentacijom. Vektori koji su paralelni ili leže na istom pravcu imaju jednak smjer, dok orijentaciju definiramo tek za vektore istog smjera. Vektor suprotne orijentacije od vektora v je $-v$.

Definicija 1.1.2. *Duljina vektora v je* $\|v\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_N^2}$.

Gornja definicija se u specijalnom slučaju, kada je $N = 2$, svodi na Pitagorin poučak pomoću kojeg možemo računati duljinu hipotenuze pravokutnog trokuta. Upravo vektor $\begin{bmatrix} v_1 & v_2 \end{bmatrix}^T$ u dvodimenzionalnom koordinatnom sustavu prikazujemo kao hipotenuzu pravokutnog trokuta s katetama $\begin{bmatrix} v_1 & 0 \end{bmatrix}^T$ i $\begin{bmatrix} 0 & v_2 \end{bmatrix}^T$. Na slici 1.1 vidimo takav prikaz vektora $\begin{bmatrix} 2 & 1 \end{bmatrix}^T$.



Slika 1.1: Prikaz vektora u koordinatnom sustavu

Dakle, ako nas zanima duljina vektora

$$\begin{bmatrix} 2 & 1 \end{bmatrix}^T$$

izračunat ćemo je kao

$$\sqrt{2^2 + 1^2} = \sqrt{5}.$$

Definicija 1.1.3. *Jedinični vektor U je vektor duljine* 1.

Od svakog vektora možemo napraviti jedinični, a proces nazivamo normiranje. Kako bismo to napravili, promatrani vektor v moramo podijeliti njegovom duljinom, tj.

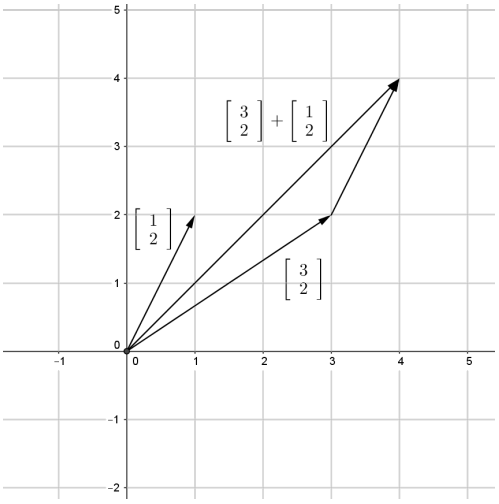
$$U = \frac{v}{\|v\|}.$$

Tako dobiveni vektor ima jednak početni smjer, no promijenjenu duljinu u odnosu na vektor v. Vektori $\begin{bmatrix} 1 & 0 \end{bmatrix}^T$, $\begin{bmatrix} 0 & 1 \end{bmatrix}^T$ su jedinični u smjeru x, odnosno y osi. Tako u N-

dimenzionalnom prostoru skup od N vektora

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

svaki dimenzije $N \times 1$ predstavlja koordinatni sustav.
Zbrajanje vektora odvija se po koordinatama pa je stoga nužno da su vektori koje želimo zbrojiti jednake dimenzije. Na slici 1.2 vidimo kako crtanjem zbrajamo vektore, na kraj prvog vektora stavljamo početak drugog. Oduzimanje shvaćamo kao zbrajanje vektora suprotne orijentacije. Još jedan način mijenjanja duljine vektora je množenje skalarom, a odvija se također po komponentama. Važno je napomenuti da vektor produljujemo množenjem pozitivnim skalarom većim od jedan te skraćujemo množenjem pozitivnim skalarom manjim od jedan. Množenjem negativnim skalarom vektoru uz duljinu mijenjamo i orijentaciju.



Slika 1.2: Zbrajanje vektora

Koordinate vektora također možemo i množiti, a suma tih umnožaka je skalarni produkt.

Definicija 1.1.4. *Skalarni produkt vektora* $\begin{bmatrix} v_1 & v_2 & \dots & v_N \end{bmatrix}^T$ *i* $\begin{bmatrix} w_1 & w_2 & \dots & w_N \end{bmatrix}^T$ *je*

$$v_1w_1 + v_2w_2 + \cdots + v_Nw_N,$$

a označavamo ga $v \cdot w$.

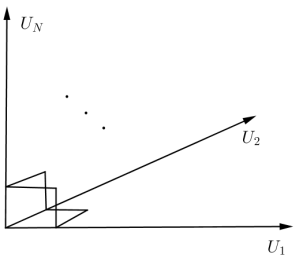
Definicija 1.1.5. *Kut θ između dva vektora v i w definiramo tako da vrijedi*

$$\cos\theta = \frac{v \cdot w}{\|v\| \, \|w\|}.$$

Definicija 1.1.6. *Kažemo da su vektori v i w ortogonalni ako je kut između njih $\theta = 90°$.*

Kako bismo odredili jesu li vektori ortogonalni, dovoljno je provjeriti je li skalarni produkt tih vektora jednak nuli. U nastavku ćemo koristiti koordinatni sustav u kojem su svi vektori međusobno ortogonalni.

Definicija 1.1.7. *Ortogonalni koordinatni sustav u N-dimenzionalnom prostoru je skup od N ortogonalnih, jediničnih vektora $U_1, U_2, ..., U_N$.*



Slika 1.3: Ortogonalni koordinatni sustav

Prostor vektora

Potaknuti uvođenjem operacija zbrajanja vektora i množenja sklarom definiramo potprostor nekog vektorskog prostora koji je za početak i sam vektorski prostor.

Definicija 1.1.8. *Skup vektora S je potprostor ako je za svaka dva vektora v i w iz S te za svaki skalar c iz odgovarajućeg polja njihova suma v + w i umnožak cv također iz S.*

Tako su svaka ravnina i pravac potprostori prostora s tri dimenzije.

Definicija 1.1.9. *Neka je V vektorski prostor i neka je $S = \{v_1, ..., v_k\}$ njegov podskup. Kažemo da je S linearno nezavisan skup vektora ako se nulvektor 0_V može na jedinstven način prikazati pomoću vektora iz S to jest ako iz $c_1a_1 + c_2a_2 + ... + c_ka_k = 0_V$ slijedi da je $c_1 = c_2 = ... = c_k = 0$.*

Za svaki vektorski prostor pa time i potprostor možemo odrediti bazu, najmanji skup linearno nezavisnih vektora pomoću kojih možemo prikazati bilo koji vektor u tom prostoru.

Definicija 1.1.10. *Skup svih linearnih kombinacija oblika* $c_1v_1 + c_2v_2 + ... + c_kv_k$ *vektora* $v_1, v_2, ..., v_k$ *iz* S *naziva se linearna ljuska skupa S.*

Ako je skup $S = \{v_1, ..., v_k\}$, onda njegovu linearnu ljusku zapisujemo kao

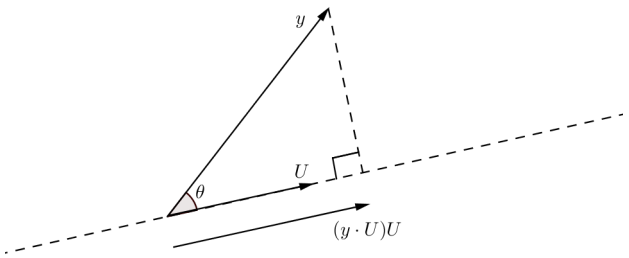
$$[S] = [\{v_1, ..., v_k\}] = \{c_1v_1 + ... + c_kv_k; \quad c_1, ..., c_k \text{ iz nekog polja F}\}.$$

Definicija 1.1.11. *Dimenzija vektorskog prostora* S *je broj vektora u bilo kojoj bazi prostora* S.

Projekcije

Vektor y u nastavku će biti vektor sačinjen od prikupljenih podataka, a projekcije će biti povezane uz prilagodbu statističkog modela.

Definicija 1.1.12. *Projekcija vektora y na jedinični vektor U je vektor* $(y \cdot U)U$.



Slika 1.4: Projekcija y na U

Koristeći se trigonometrijom pravokutnog trokuta možemo izračunati duljinu projekcije $(y \cdot U)U$.

$$||(y \cdot U)U|| = ||y|| \cos\theta = ||y|| \left(\frac{y \cdot U}{||y||\,||U||} \right) = \frac{y \cdot U}{||U||} = y \cdot U.$$

Sada kada znamo što je projekcija vektora na jedinični vektor, želimo to proširiti na projekciju na potprostor, koji ćemo kasnije nazivati prostor modela. Ideja je projicirati vektor na svaku od osi ortogonalnog koordinatnog sustava te zbrojiti dobivene projekcije. Na taj način smo dobili vektor u potprostoru najbliži našem vektoru y.

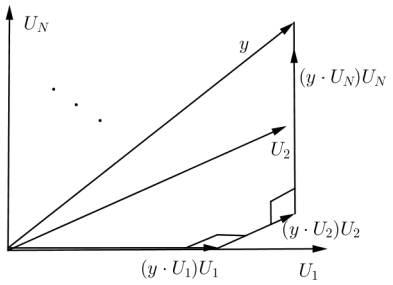
Definicija 1.1.13. *Projekcija od y na potprostor M razapet jediničnim vektorima* $U_1, U_2, ..., U_k$ *je*

$$(y \cdot U_1)U_1 + (y \cdot U_2)U_2 + ... + (y \cdot U_k)U_k.$$

Ako je potprostor upravo cijeli prostor dimenzije N, vidimo da svaki vektor možemo prikazati kao sumu njegovih projekcija na svaku od koordinatnih osi $U_1, U_2, ..., U_N$.

Definicija 1.1.14. *Ortogonalna dekompozicija proizvoljnog vektora y na ortogonalni koordinatni sustav u N-dimenzionalnom prostoru je*

$$y = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + ... + (y \cdot U_N)U_N.$$



Slika 1.5: Ortogonalna dekompozicija vektora y

Već smo spomenuli da pomoću Pitagorinog poučka računamo duljinu vektora u dvije dimenzije, proširimo sada to na N dimenzija.

Definicija 1.1.15. *'Pitagorin poučak' u N-dimenzionalnom prostoru glasi*

$$||y||^2 = (y \cdot U_1)^2 + (y \cdot U_2)^2 + ... + (y \cdot U_N)^2.$$

1.2 Statistika

U svakodnevnom životu riječ populacija najčešće se koristi kada govorimo o živim bićima, no u statistici ta riječ označava skup svih mjerenja na skupu proizvoljnih objekata. Kako je populacija često prevelika i teško je uključiti svaku jedinku u testiranje, iz nje uzimamo uzorak koji promatramo. Taj uzorak možemo shvatiti kao potprostor cijelog prostora, tj. cijele populacije. Želimo da je uzorak reprezentativan, a najbolji način da to postignemo je izabrati ga nasumično. Većina definicija koje navodimo u ovom poglavlju preuzeta je iz [7].

Definicija 1.2.1. *Slučajni uzorak je uzorak određene veličine iz neke populacije takav da smo svaki drugi uzorak te veličine iz iste populacije mogli izabrati s jednakom vjerojatnošću.*

Nas će zanimati neka veličina koju iz uzorka možemo dobiti, npr. aritmetička sredina uzorka visina deset dječaka u nekom gradu u dobi od osam do deset godina.

Definicija 1.2.2. *Slučajna varijabla je veličina čije vrijednosti ovise o ishodu slučajnog pokusa.*

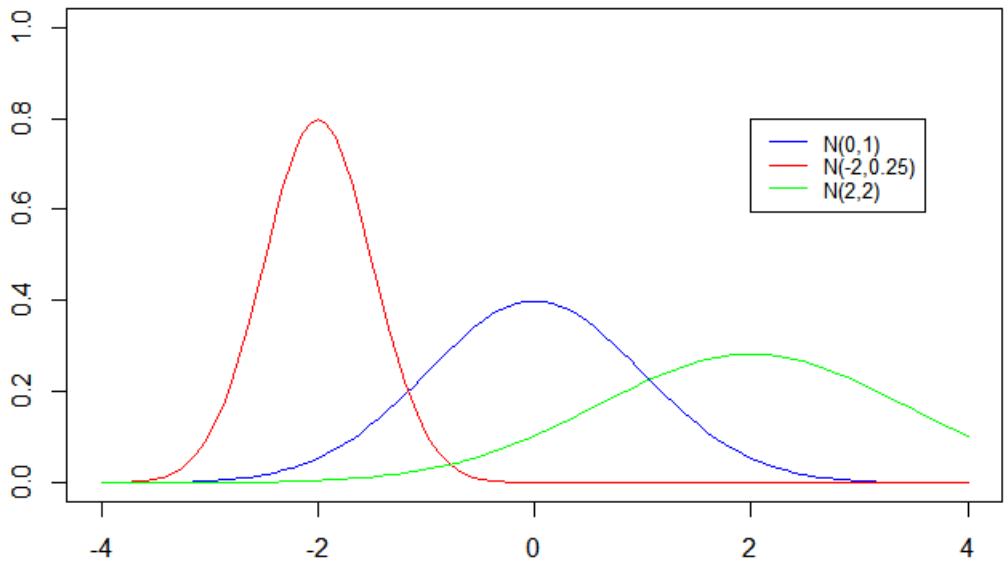
Svako ponavljanje istog pokusa, na primjer računanje aritmetičke sredine visina različitih slučajnih uzoraka, daje različite vrijednosti koje nazivamo realizirane vrijednosti.
Slučajne varijable u nastavku ćemo označavati velikim slovom Y, a realizirane vrijednosti slučajnih varijabli s y.

Svaka diskretna slučajna varijabla poprima vrijednosti u nekom konačnom ili prebrojivom skupu podataka i to svaku od vrijednosti s određenom vjerojatnosti. Raspodjela tih vjerojatnosti naziva se distribucija slučajne varijable. Neprekidne slučajne varijable poprimaju vrijednosti u neprebrojivim skupovima, a distribuciju određuje funkcija gustoće.

Većina realiziranih vrijednosti mjerenja u prirodi prati neku već poznatu distribuciju.
Najčešće se radi o normalno distribuiranim slučajnim varijablama koje ćemo i mi promatrati u nastavku. Graf funkcije gustoće zvonolikog je oblika, a pokazuje nam učestalost pojavljivanja određenih vrijednosti, tj. da se srednja vrijednost nekog mjerenja pojavljuje češće nego male i velike vrijednosti. Funkcija gustoće dana je s

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

gdje je μ bilo koji realan broj, a $\sigma > 0$. Slučajnu varijablu Y normalne razdiobe s očekivanjem μ te varijancom σ^2 pišemo kao $Y \sim N(\mu, \sigma^2)$.



Slika 1.6: Funkcija gustoće normale razdiobe za različite parametre

Na slici 1.6 vidimo grafove različitih funkcija gustoća u ovisnosti o parametrima μ i σ. Parametar μ predstavlja srednju vrijednost distribucije koju još nazivamo i očekivanje, a σ^2 varijabilnost, tj. varijancu. Definirajmo te veličine za slučajnu varijablu Y s funkcijom gustoće f.

Definicija 1.2.3. *Očekivanje slučajne varijable Y definiramo kao*

$$\mu = \frac{zbroj\ svih\ mjerenja}{broj\ mjerenja} = \sum mjerenje \times relativna\ frekvencija = \int y\, f(y)\, dy = E(Y).$$

Definicija 1.2.4. *Varijancu slučajne varijable Y definiramo kao*

$$\begin{aligned}\sigma^2 &= \frac{zbroj\ svih\ (mjerenje\text{-}očekivanje)^{\,2}}{broj\ mjerenja} \\ &= \sum (mjerenje\text{-}očekivanje)^2 \times relativna\ frekvencija\ mjerenja \\ &= \int (y-\mu)^2\, f(y)\, dy = Var(Y).\end{aligned}$$

Još jedan način za izračunati varijancu slučajne varijable je $Var(Y) = E[(Y - \mu)^2]$, tj. kao prosječno kvadratno odstupanje realizirane vrijednosti varijable Y od njene očekivane vrijednosti.
Kada proučavamo dvije slučajne varijable, često nas zanima njihov odnos, tj. utječu li jedna na zavisne ako se promjenom jedne mijenja i druga, npr. visina i težina. U suprotnom imamo nezavisne slučajne varijable, npr. visina i boja kose.

Definicija 1.2.5. *Kovarijancu slučajnih varijabli X i Y definiramo kao*

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))].$$

Definicija 1.2.6. *Normalno distribuirane slučajne varijable X i Y su nezavisne ako im je kovarijanca jednaka nuli.*

Osim odnosa dviju slučajnih varijabli zanimat će nas i varijanca i očekivanje linearne kombinacije slučajnih varijabli. Neka je $W = a_1Y_1 + ... + a_NY_N$ linearna kombinacija nezavisnih slučajnih varijabli, tada vrijedi

1. ako su $Y_1, ..., Y_N$ normalno distribuirane onda je i W normalno distribuirana slučajna varijabla
2. $E(W) = a_1E(Y_1) + ... + a_NE(Y_N)$
3. $Var(W) = a_1^2Y_1 + ... + a_N^2Y_N$

Koristeći gore navedena pravila, lako se pokaže da za normalno distribuirane slučajne varijable $Y_1, ..., Y_N$ iz $N(\mu, \sigma^2)$ aritmetička sredina $\overline{Y} = \frac{Y_1+...+Y_N}{N}$ ima $N(\mu, \sigma^2/N)$ razdiobu.

U našem geometrijskom pristupu zanimat će nas distribucija vektora projekcije, tj. $y \cdot U = a_1y_1 + ... + a_Ny_N$, gdje je $U = \begin{bmatrix} a_1 & \dots & a_N \end{bmatrix}^T$ jedinični vektor. Neka je $Y_1, Y_2, ..., Y_N$ niz nezavisnih, normalno distribuiranih slučajnih varijabli s varijancom σ^2. Primjenjujući ranije navedena pravila, vidimo da je $Y \cdot U = a_1Y_1 + ... + a_NY_N$ uz $a_1^2 + ... + a_N^2 = 1$ normalno distribuirana slučajna varijabla s očekivanjem

$$E(Y \cdot U) = a_1E(Y_1) + ... + a_NE(Y_N)$$

i varijancom

$$Var(Y \cdot U) = \sigma^2.$$

Primijetimo da varijanca ne ovisi o vektoru U, odnosno za svaki U_i, $i = 1, ..., N$ uvijek je jednaka $Var(Y \cdot U_i) = \sigma^2$.

Teorem 1.2.7. *Slučajne varijable $Y \cdot U_c$ i $Y \cdot U_d$ su nezavisne ako i samo ako su vektori $U_c = \begin{bmatrix} c_1 & \dots & c_N \end{bmatrix}^T$ i $U_d = \begin{bmatrix} d_1 & \dots & d_N \end{bmatrix}^T$ ortogonalni.*

Dokaz. Od prije znamo da su slučajne varijable nezavisne ako im je kovarijanca jednaka nuli, tj.

$$\begin{aligned} Cov(Y \cdot U_c, Y \cdot U_d) &= Cov(c_1Y_1 + ... + c_NY_N, d_1Y_1 + ... + d_NY_N) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} c_i d_j Cov(Y_i, Y_j) \\ &= \sum_{i=1}^{N} c_i d_i Var(Y_i) \\ &= \left(\sum_{i=1}^{N} c_i d_i\right)\sigma^2 \\ &= (U_c \cdot U_d)\sigma^2 = 0 \end{aligned}$$

gdje posljednja jednakost vrijedi ako i samo ako su vektori U_c i U_d ortogonalni. Treća jednakost posljedica je nezavisnosti slučajnih varijabli Y_i i Y_j za svaki i, j. □

Procjena parametara

Kako ćemo u nastavku raditi sa slučajnim varijablama normalne razdiobe, no nepoznatih parametara, morat ćemo ih na neki način procijeniti iz podataka.

Definicija 1.2.8. *Nepristrani procjenitelj nekog parametra je slučajna varijabla čije je očekivanje jednako tom parametru.*

Cilj nam je odrediti nepristrane procjenitelje za parametre μ i σ^2. Ako uzmemo uzorak nezavisnih jednako distribuiranih slučajnih varijabli s nekom $N(\mu, \sigma^2)$ razdiobom, parametar μ ćemo procijeniti aritmetičkom sredinom uzorka koju označavamo s $\overline{y}$. Pripadni procjenitelj označavamo s $\overline{Y}$, a lako se pokaže na primjeru uzorka s dva mjerenja da je također i nepristran, tj.

$$E(\overline{Y}) = E(a_1Y_1 + a_2Y_2) = a_1E(Y_1) + a_2E(Y_2) = (a_1 + a_2)\mu = \mu$$

jer smo uzeli $a_1 = a_2 = \frac{1}{2}$. Postoji mnogo nepristranih procjenitelja parametra μ, no cilj nam je pronaći onog s najmanjom varijancom. Pokažimo da je to upravo $\overline{Y}$.

$$Var(\overline{Y}) = Var(a_1Y_1 + a_2Y_2) = a_1^2Var(Y_1) + a_2^2Var(Y_2) = (a_1^2 + a_2^2)\sigma^2 = \sigma^2$$

opet za $a_1 = a_2 = \frac{1}{2}$. Kako je upravo ta kombinacija parametara ona koja u sumi daje jedan te ima najmanju sumu kvadrata, ovu metodu pronalaska procjenitelja nazivamo metodom najmanjih kvadrata.

Sada kada znamo koji procjenitelj koristiti za procjenu parametra μ, ostaje nam još razmotriti što s parametrom σ^2. Slučajne varijable $Y \cdot U_1, ..., Y \cdot U_N$ koje ćemo u nastavku promatrati dijele se u dvije kategorije. U prvu kategoriju pripadaju varijable $Y \cdot U_1, ..., Y \cdot U_p$ povezane s $U_1, ..., U_p$ koji leže u prostoru kojeg ćemo nazivati prostor modela, dok su preostale $Y \cdot U_{p+1}, ..., Y \cdot U_N$ povezane s $U_{p+1}, ..., U_N$ koji leže u takozvanom prostoru greške. Prva kategorija varijabli imat će očekivanje različito od nule, dok će druga imati očekivanje jednako nuli te će nam služiti za procjenu parametra σ^2. Vidimo da je tada, primjenjujući definiciju varijance, za svaki U_i iz prostora greške

$$\begin{aligned} Var(Y \cdot U_i) &= E\{[Y \cdot U_i - E(Y \cdot U_i)]^2\} \\ &= E[(Y \cdot U_i)^2] \end{aligned}$$

gdje posljednja jednakost slijedi jer je $E(Y \cdot U_i) = 0$. Kako uvijek vrijedi $Var(Y \cdot U_i) = \sigma^2$, slijedi da je $E[(Y \cdot U_i)^2] = \sigma^2$ za svaki jedinični vektor iz prostora greške. Zaključujemo sada da je svaka slučajna varijabla $(Y \cdot U_i)^2$ za svaki U_i iz navedenog prostora nepristran procjenitelj za parametar σ^2.
Dakle, kako svaka od $(Y \cdot U_{p+1})^2, ..., (Y \cdot U_N)^2$ slučajnih varijabli predstavlja procjenitelj za σ^2, opet ćemo kao i u slučaju očekivanja najbolji nepristrani procjenitelj dobiti uzimajući njihov prosjek, tj.

$$S^2 = \frac{(Y \cdot U_{p+1})^2 + ... + (Y \cdot U_N)^2}{N - p}.$$

U nastavku će nas zanimati koliko je dobra procjena parametra μ pa ćemo računati pouzdani interval.

Definicija 1.2.9. *Neka je $X_1, ..., X_n$ slučajni uzorak iz parametarskog modela $\mathcal{P} = \{f(x, \theta) : \theta \in \Theta\}, \Theta \subseteq \mathbb{R}$ s jednodimenzionalnim parametrom θ. Kažemo da je slučajan interval $[\hat{\theta}_L, \hat{\theta}_U]$ $(1 - \alpha) \cdot 100\%$ pouzdani interval za parametar θ ako vrijedi*

$$\mathbb{P}_\theta(\hat{\theta}_L \leqslant \theta \leqslant \hat{\theta}_U) \geqslant 1 - \alpha, \quad \forall \theta \in \Theta.$$

F i T-distribucije

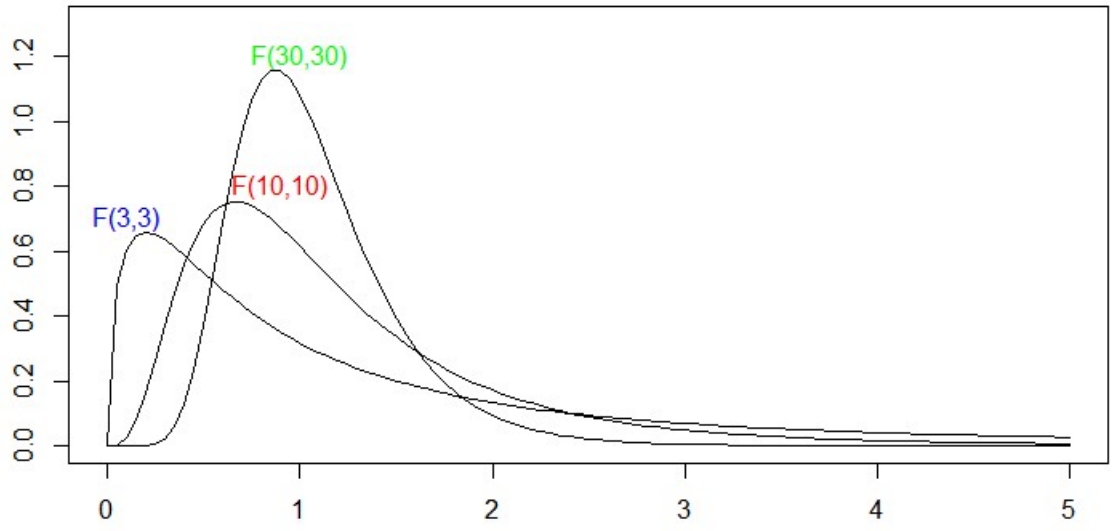
U nastavku ćemo za testiranje hipoteza računati omjere prosjeka kvadratnih duljina projekcija tj. $(Y \cdot U_i)^2$. Ti omjeri uvijek će pratiti F-distribuciju.

Definicija 1.2.10. *Slučajna varijabla*

$$F(p,q) = \frac{\frac{W_1^2+\ldots+W_p^2}{p}}{\frac{W_{p+1}^2+\ldots+W_{p+q}^2}{q}}$$

gdje su $W_1, ..., W_{p+q}$ *nezavisne, jednako distribuirane slučajne varijable razdiobe* $N(0, \sigma^2)$ *naziva se* F*-statistika s* p *i* q *stupnjeva slobode.*

Dakle, F-statistika s p i q stupnjeva slobode prati $F(p,q)$-distribuciju koja se razlikuje za svaki par (p,q). Primijetimo da su i brojnik i nazivnik nepristrani procjenitelji za parametar σ^2, stoga su za male vrijednosti p i q oni prilično varijabilni, što vodi i velikoj varijabilnosti same F-statistike. Kako vrijednosti od p i q rastu, tako se $F(p,q)$-statistika sve više grupira oko 1, a to i vidimo na sljedećoj slici.



Slika 1.7: Funkcije gustoće F-distribucije s različitim stupnjevima slobode

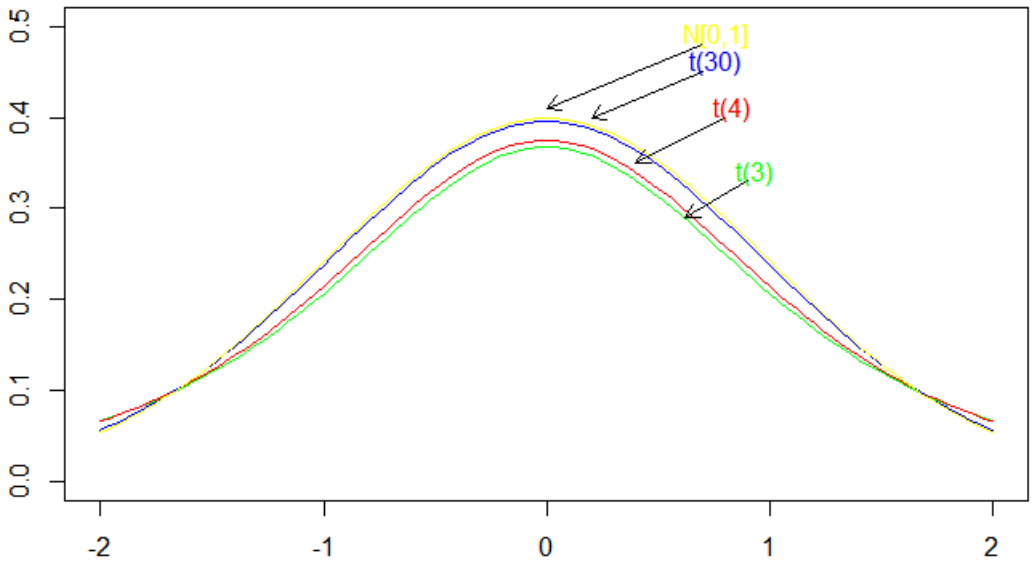
Često se koristi poseban slučaj F-statistike kada je $p = 1$ pa definiramo još i T-statistiku. Mi ćemo je koristiti za računanje intervala pouzdanosti.

Definicija 1.2.11. *Slučajna varijabla*

$$T(q) = \frac{W_1}{\sqrt{\frac{W_2^2+\ldots+W_{q+1}^2}{q}}}$$

gdje su $W_1, ..., W_{q+1}$ *nezavisne, jednako distribuirane slučajne varijable razdiobe* $N(0, \sigma^2)$ *naziva se* T*-statistika s* q *stupnjeva slobode.*

Još kažemo i da slučajna varijabla $T(q)$ ima Studentovu ili T-distribuciju, a pišemo $T(q) \sim t(q)$. Za razliku od $F(p, q)$, $T(q)$-statistika može poprimiti i negativne vrijednosti, a kako je brojnik iz $N(0, \sigma^2)$ simetrična je oko 0. Kako q raste $T(q)$ sve više aproksimira standardnu, normalnu slučajnu varijablu što vidimo i na slici 1.8. Tako već za npr. $q \geqslant 30$ možemo uzeti da je $t(q) \approx N(0, 1)$. Primjetimo da vrijedi $T(q)^2 = F(1, q)$



Slika 1.8: Funkcije gustoće T-distribucije za različite parametre q

T-statistike

Kako smo u uvodnom dijelu i rekli, problem koji rješavamo u ovom radu tradicionalno se rješava pomoću T-testa. Studentov test ili T-test uveo je kemičar William Sealy Gosset 1908. godine u časopisu Biometrika. Radio je u Guinessovoj pivovari te mu je bilo zabranjeno objavljivanje rezultata istraživanja, no ipak ih je objavio pod pseudonimom Student. Nekoliko je inačica T-statistika u upotrebi, ovisno je li riječ o velikim ili malim uzorcima

jedne ili dvije populacije, jesu li uzorci iz različitih populacija jednake veličine, imaju li jednake varijance i konačno jesu li nezavisni ili zavisni. Kroz cijeli rad promatrat ćemo male, nezavisne uzorke, jednake veličine i varijance. U nastavku ćemo računati i tradicionalno korištene statistike, kako bismo se uvjerili da geometrijskim pristupom dobivamo jednake rezultate.
U trećem poglavlju bit će riječ o jednoj populaciji, tj. uzorku

$$y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$$

pa ćemo T-statistiku računati kao

$$\frac{\hat{\mu} - \mu_0}{s} \sqrt{n}$$

gdje nam je n veličina uzorka, $\hat{\mu} = \overline{y}$ procjenitelj očekivanja populacije, a računamo ga kao aritmetičku sredinu uzorka, $\mu_0 = 0$ referentna vrijednost i s procjenitelj standardne devijacije uzorka koji ćemo računati kao

$$\sqrt{\frac{(\overline{y} - y_1)^2 + (\overline{y} - y_2)^2 + ... + (\overline{y} - y_n)^2}{n-1}}.$$

U četvrtom poglavlju imat ćemo dvije populacije, tj. uzorak

$$\begin{bmatrix} y_{11} & \dots & y_{1n} & y_{21} & \dots & y_{2n} \end{bmatrix}^T$$

pa ćemo koristiti statistiku

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{s_{y_{1\cdot}y_{2\cdot}}} \sqrt{n}$$

gdje je opet n veličina uzorka, $\hat{\mu}_1$ i $\hat{\mu}_2$ procjenitelji očekivanja prve i druge populacije, a $s_{y_{1\cdot}y_{2\cdot}} = \sqrt{s_{1\cdot}^2 + s_{2\cdot}^2}$ gdje su $s_{1\cdot}$ i $s_{2\cdot}$ procjene standardnih devijacija.

Poglavlje 2

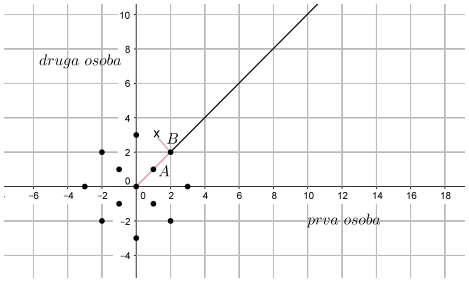
Geometrijski pristup u statistici

2.1 Motivacija

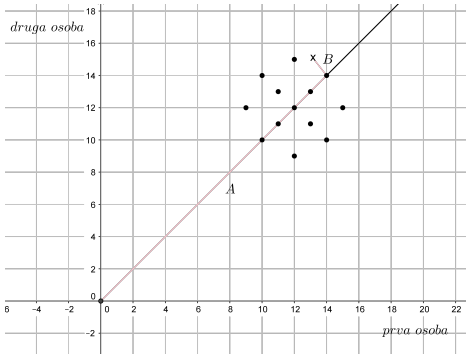
Na vrlo jednostavnom primjeru pokazat ćemo kako doći do statistike koju koristimo kod pitanja vezanih uz očekivanje.

Pitanje: Povećava li se puls nakon konzumiranja šalice kave?

Kako bismo došli do odgovora, izaberemo uzorak od 20 ljudi iz neke populacije te im izmjerimo puls prije i nakon konzumiranja kave. Izračunamo razlike tih vrijednosti te ih označimo s $y_1, y_2, ..., y_{20}$. U ovom trenutku obično bismo izračunali T-statistiku te provjerili upada li vrijednost u kritično područje na temelju čega bismo donijeli zaključak, tj. dobili odgovor na postavljeno pitanje.
Pogledajmo kako drugačije doći do tog odgovora. Pretpostavimo da testiranje provodimo više puta i to tako da uzimamo dvoje ljudi, umjesto da jednom provedemo test na skupini od 20 ljudi. Razmotrimo dva slučaja, prvi da nitko nije popio kavu odnosno da su varali na testu i da je iz tog razloga očekivana vrijednost razlike nakon mnogo testiranja jednaka nuli. Ta situacija prikazana je na slici 2.1 Suprotno tome, pretpostavimo da su svi popili kavu i da je $\mu = 12$ otkucaja u minuti nakon mnogo testiranja, tada na slici 2.2 vidimo da se podaci pomiču po pravcu $y = x$ u desno, tj. prema gore. Zamislimo da na raspolaganju imamo samo jednu vrijednost testiranja npr. točku x te označimo udaljenost od ishodišta do točke $(\mu+2, \mu+2)$ s A i od točke $(\mu+2, \mu+2)$ do x s B. Sada pomoću Pitagorinog poučka lako izračunamo A i B te zaključujemo da je omjer $\frac{A}{B}$ veći u slučaju $\mu = 12$, nego kada je $\mu = 0$. Ovo nam sugerira da za testnu statistiku uzmemo upravo omjer $\frac{A}{B}$ te da promatramo vrijednost tog omjera i sukladno tome donesemo zaključak. Naše početno pitanje možemo preoblikovati u geometrijskom smislu: pripada li vrijednost dobivena testom krugu grupiranom oko $(0, 0)$ ili ne? Zapravo se pitamo je li $\mu = 0$ ili je $\mu \neq 0$. Ako je $\mu = 0$, testna statistika $\frac{A}{B}$ mora biti mala, suprotno slučaju kada je $\mu \neq 0$ i mora biti velika.



Slika 2.1: $\mu = 0$



Slika 2.2: $\mu = 12$

2.2 Postupak

Do sada smo se podsjetili osnovnih definicija i tvrdnji koje ćemo koristiti pa smo spremni objasniti geometrijski pristup rješavanja problema u statistici. U poglavljima koji slijede to ćemo i primijeniti na konkretnim primjerima te slikovito prikazati dobivene rezultate. Pretpostavljat ćemo da su podaci normalno distribuirani, nepoznatog očekivanja i varijance. Nadalje, krajnji cilj nam je procijeniti parametre μ i σ^2, testirati je li očekivanje jednako nuli te izračunati 95% pouzdani intervala za parametar μ.

Uvedimo tri objekta koja će nam biti potrebna.

1. ***Vektor podataka y***

   Prva i osnova stvar je prikupiti podatke istraživanjem ili koristiti neke druge podatke od interesa. Nakon toga prikazujemo ih pomoću vektora koji ćemo nazivati vektor podataka $y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$. Uzorak $y_1, y_2, ..., y_n$ je realizacija slučajnih varijabli $Y_1, Y_2, ..., Y_n$ $N(\mu, \sigma^2)$ distribucije. Ovdje označavamo broj realizacija s n jer će kasnije N biti broj svih realizacija u slučaju više populacija. Naravno, u slučaju jedne populacije vrijedi $N = n$.

   Npr. ako nas zanima koliko se razlikuju prosječne visine muškaraca i žena, možemo uzeti dva uzorka iz svake od navedenih skupina i označiti s y_1 i y_2 visine muškaraca, a s y_3 i y_4 žena. Tada je vektor $y = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \end{bmatrix}^T$ vektor podataka.

2. ***Prostor modela M***

   Sljedeći korak je odrediti potprostor N-dimenzionalnog prostora sačinjen od vektora modela, tj. vektora čije su komponente moguće očekivane vrijednosti naših

opažanja. Prostor koji smo na taj način odredili nazivamo prostor modela i označavamo s M. U ovom koraku određujemo još i jedinične vektore koji razapinju prostor M.

Očekivane vrijednosti mjerenja iz našeg primjera označimo s $\mu_{\check{z}}$ za žene i μ_{m} za muškarce. Sada je svaki vektor oblika $\begin{bmatrix} \mu_{\mathrm{m}} & \mu_{\mathrm{m}} & \mu_{\check{z}} & \mu_{\check{z}} \end{bmatrix}^T$ vektor modela. Takav vektor uvijek se može zapisati u obliku

$$\begin{bmatrix} \mu_{\mathrm{m}} \\ \mu_{\mathrm{m}} \\ \mu_{\check{z}} \\ \mu_{\check{z}} \end{bmatrix} = \mu_{\mathrm{m}} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \mu_{\check{z}} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Sada je prostor modela upravo jednak

$$M = \left[\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right\} \right].$$

To je dvodimenzionalni potprostor četverodimenzionalnog prostora razapet jediničnim vektorima

$$U_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^T$$

i

$$U_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^T$$

koje smo dobili ranije definiranim procesom normiranja.

3. ***Vektor smjera U***

Posljednja stvar koju određujemo je vektor smjera povezan s nul-hipotezom u statističkom problemu. U nastavku ćemo se baviti testovima o očekivanju slučajne varijable pa će svaka nul-hipoteza biti $H_0 : \mu = 0$. Dakle tražimo jedinstveni jedinični vektor U koji leži u prostoru modela M te za kojega vrijedi da je očekivana vrijednost projekcije vektora podataka y na njega jednaka pravoj vrijednosti tog očekivanja do na neki skalar.

U našem primjeru zanima nas razlikuju li se visine muškaraca i žena pa je nul-hipoteza $H_0 = \mu_{\mathrm{m}} - \mu_{\check{z}} = 0$. Vektor U dobivamo tako da oduzmemo vektore U_1 i U_2 za koje smo ustanovili da su jedinični vektori koji razapinju prostor M

$$(U_1 - U_2) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix}.^T$$

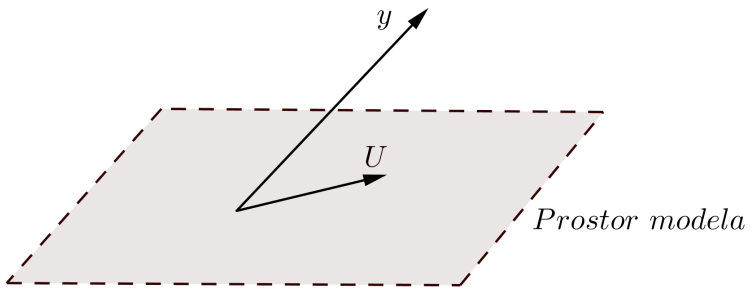
Tako dobiveni vektor normiramo, tj. podijelimo s njegovom duljinom koja iznosi $\sqrt{2}$

$$\frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix}^T = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix}^T = U.$$

Na taj način dobivamo jedinični vektor U koji leži u prostoru M, a kada po definiciji 1.1.12 izračunamo projekciju vektora y na vektor U dobivamo

$$y \cdot U = \frac{1}{2}(y_1 + y_2 - y_3 - y_4) = \overline{y}_{\text{m}} - \overline{y}_{\text{ž}}$$

gdje su $\overline{y}_{\text{m}}$ i $\overline{y}_{\text{ž}}$ prosječne visine muškaraca odnosno žena iz naših podataka. Vidimo sada da je zadovoljen i uvjet da je $E(Y \cdot U) = k(\mu_{\text{m}} - \mu_{\text{ž}})$, gdje je k neki skalar. U našem slučaju je $k = 1$.



Slika 2.3: Vektor podataka, prostor modela i vektor smjera

Nakon što smo pobliže objasnili koje objekte ćemo trebati te na koji način doći do njih, objasnimo dva posljednja koraka geometrijskog pristupa.

1. ***Prilagodba modela***

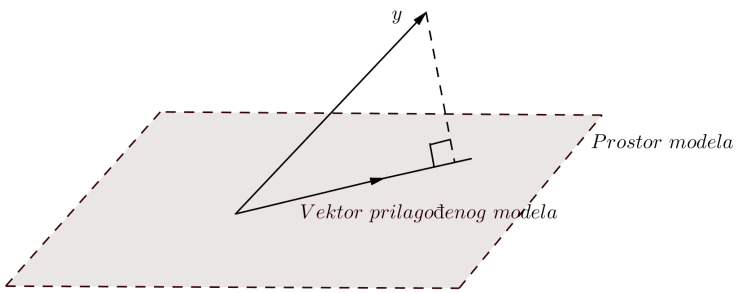
   Postupkom prilagodbe modela, tj. projekcijom vektora y na prostor modela M dobivamo $(y \cdot U_1)U_1 + ... + (y \cdot U_p)U_p$, vektor prilagođenog modela koji je procjena za ranije spomenuti vektor modela. p je dimenzija prostora modela M.

Postupak procjene vektora modela $\begin{bmatrix} \mu_{\text{m}} & \mu_{\text{m}} & \mu_{\check{\text{z}}} & \mu_{\check{\text{z}}} \end{bmatrix}^T$ u prostoru M svodi se na pronalazak vektora što sličnijeg vektoru y u tom prostoru. Iz tog razloga razumno je uzeti projekciju vektora y na M.

U našem primjeru s početka ovog razmatranja dobivamo

$$(y \cdot U_1)U_1 + (y \cdot U_2)U_2 = \begin{bmatrix} \overline{y}_{\text{m}} \\ \overline{y}_{\text{m}} \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \overline{y}_{\check{\text{z}}} \\ \overline{y}_{\check{\text{z}}} \end{bmatrix} = \begin{bmatrix} \overline{y}_{\text{m}} \\ \overline{y}_{\text{m}} \\ \overline{y}_{\check{\text{z}}} \\ \overline{y}_{\check{\text{z}}} \end{bmatrix}$$

gdje su $\overline{y}_{\text{m}}$ i $\overline{y}_{\check{\text{z}}}$ procjenitelji parametara μ_{m} i $\mu_{\check{\text{z}}}$ redom.



Slika 2.4: Prilagodba modela

2. ***Testiranje hipoteza***

   Zadnji korak je izračunati testnu statistiku i donijeti zaključak. Za početak trebamo odrediti prikladan koordinatni sustav sačinjen od jediničnih vektora.

   U našem primjeru biramo četiri vektora

$$U_1 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad U_2 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \quad U_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad U_4 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}.$$

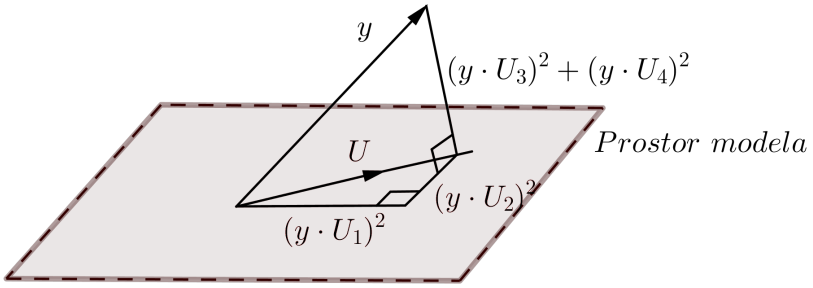
Vektori U_3 i U_4 su očigledno takvi da razapinju ravninu ortogonalnu na ravninu koju razapinju vektori U_1 i U_2 iz točke 2, tj. prostor M. Općenito prostor koji razapinju ti ostali vektori nazivamo prostor greške. U ovoj točki smo izabrali U_2 tako da bude jednak vektoru smjera U, a U_1 da zajedno s njim razapinje prostor modela M. Izabrali smo nove vektor U_1 i U_2 jer ćemo za računanje testne statistike koristiti projekciju vektora y na vektor smjera, a ne na bilo koji vektor koji razapinje prostor modela.

Kada smo odredili koordinatni sustav, izračunamo projekcije vektora y na svaku od osi te primijenom definicije 1.1.15 dobivamo

$$\begin{aligned}\|y\|^2 &= (y \cdot U_1)^2 + (y \cdot U_2)^2 + (y \cdot U_3)^2 + (y \cdot U_4)^2 \\ &= 4\overline{y}^2 + (\overline{y}_m - \overline{y}_{\check{z}})^2 + (y_1 - y_2)^2/2 + (y_3 - y_4)^2/2\end{aligned}$$

Kako bismo testirali našu hipotezu, tj je li $\mu_{\mathrm{m}} \neq \mu_{\check{\mathrm{z}}}$, moramo podijeliti kvadrat duljine projekcije vektora y na vektor smjera U_2 s prosjekom kvadrata duljina projekcija vektora y na koordinatne osi prostora greške. Vidimo da dobivamo upravo realizaciju F-statistike

$$f = \frac{(y \cdot U_2)^2}{[(y \cdot U_3)^2 + (y \cdot U_4)^2]/2} = \frac{(\overline{y}_{\mathrm{m}} - \overline{y}_{\check{\mathrm{z}}})^2}{[(y_1 - y_2)^2 + (y_3 - y_4)^2]/4}$$



Slika 2.5: Primjena Pitagorina poučka

Za kraj je ostalo interpretirati rezultate provedenog testa. Ako je hipoteza da je $\mu_{\text{m}} = \mu_{\text{ž}}$ istinita, brojnik i nazivnik biti će približno jednake vrijednosti, tj. F-statistika će biti mala, dok će u slučaju $\mu_{\text{m}} \neq \mu_{\text{ž}}$ brojnik biti veći pa time i statistika (prisjetimo se primjera s početka ovog poglavlja i statistike $\frac{A}{B}$).

Zaključujemo da će u slučaju da je $\mu_{\text{m}} = \mu_{\text{ž}}$ istina F-statistika biti realizirana vrijednost $F(1, 2)$-distribucije odnosno Studentove s dva stupnja slobode. Odbacujemo hipotezu $H_0 = \mu_{\text{m}} - \mu_{\text{ž}} = 0$ ukoliko je vrijednost od F upala u odgovarajuće kritično područje.

Poglavlje 3

Jedna populacija

Sada ćemo prijeći na konkretne primjere na kojima ćemo korak po korak provesti metodu opisanu u prethodnom poglavlju. Započet ćemo s primjerima vezanim uz jednu populaciju, a kao što smo već i rekli, testirat ćemo je li očekivana vrijednost neke mjere jednaka ili različita od nula.

3.1 Primjer - veličina zrna pšenice

U proljeće 1986. godine znanstvenike s Novog Zelanda zanimalo je utječe li izlaganje sjemena pšenice fungicidima na veličinu zrna i ako da, koliki je prosječni utjecaj. Kako bi to istražili, promatrali su tri farme i na svakoj od njih odvojili dijelove zemlje s posađenim tretiranim i netretiranim sjemenom pšenice. Mjerili su težinu zrna pšenice ubrane na svakom od područja (na svakoj od farmi) te računali razliku tih težina. Dakle radimo s jednom populacijom, populacijom razlika. Podaci mjerenja nalaze se u tablici na slici 3.1, a mjerna jedinica kojom su izraženi je tona/hektar. Više o primjeru možemo pročitati u [7].

	1.farma	2.farma	3.farma
netretirano	5.0	4.3	5.9
tretirano	6.1	5.7	7.0
razlika	1.1	1.4	1.1

Slika 3.1: Tablica podataka

Geometrijski pristup

Koristeći se podacima dobivenim u istraživanju, geometrijskim pristupom doći ćemo do odgovora na pitanje utječe li tretiranje sjemena fungicidima na veličinu zrna pšenice ili ne. Odredimo za početak vektor podataka, prostor modela i vektor smjera.

1. ***Vektor podataka y***

   Vektor podataka sačinjen je od tri razlike između navedena tri para mjerenja.

$$y = \begin{bmatrix} 1.1 & 1.4 & 1.1 \end{bmatrix}^T$$

2. ***Prostor modela M***

   Pretpostavljamo da su naši podaci normalno distribuirani s nepoznatim očekivanjem μ i varijancom σ^2. Kako je očekivanje svakog mjerenja, odnosno razlike μ vektori modela su dani s $\begin{bmatrix} \mu & \mu & \mu \end{bmatrix}^T$. Dakle, prostor modela M je

$$M = \left[\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\} \right]$$

   jednodimenzionalni potprostor trodimenzionalnog prostora razapet jediničnim vektorom

$$U_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

   Vektor podataka y možemo zapisati kao

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

   gdje je $\begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix}^T$ vektor greške.

3. ***Vektor smjera U***

   Želimo testirati

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0$$

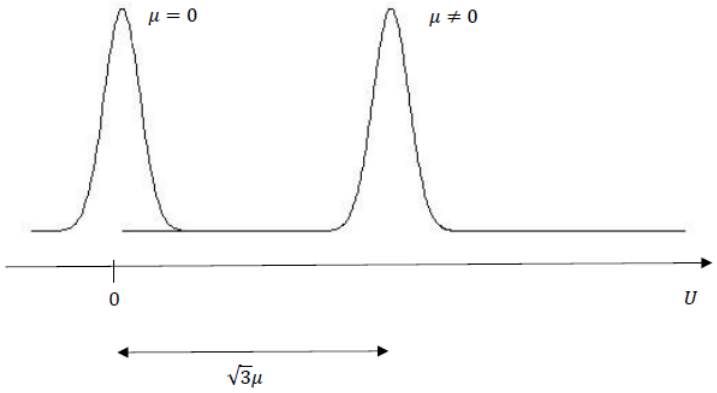
pa moramo pronaći vektor U povezan s hipotezom H_0. Prisjetimo se, moraju biti zadovoljena dva kriterija. Prvi je da vektor U leži u prostoru M, a drugi da vrijedi $E(Y \cdot U) = k\mu$. S obzirom na to da je M jednodimenzionalni prostor, samo je jedan smjer moguć. Uzmimo da je

$$U = U_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$$

vektor smjera. Provjerimo i drugi kriterij

$$y \cdot U = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \cdot \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{3}}(y_1 + y_2 + y_3) = \frac{3\overline{y}}{\sqrt{3}} = \sqrt{3}\overline{y}$$

gdje je $\overline{y}$ aritmetička sredina podataka. Vidimo da je $E(Y \cdot U) = \sqrt{3}\mu$ iz čega slijedi da ako vrijedi $\mu = 0$ duljina projekcije $y \cdot U$ bit će relativno mala i težiti ka nuli nakon mnogo ponavljanja pokusa. S druge strane, ako je $\mu \neq 0$, $y \cdot U$ će težiti ka $\sqrt{3}\mu$, tj. biti veća.



Slika 3.2: Distribucija od $Y \cdot U$

U nastavku pogledajmo kako prilagoditi model te u konačnici i testirati postavljenu hipotezu.

1. ***Prilagodba modela***

Za vektor podataka vrijedi $y_i = \mu + e_i$, gdje su e_1, e_2, e_3 nezavisne, normalno distribuirane slučajne varijable s očekivanjem nula i varijancom σ^2. Kako smo ranije objasnili, u ovom koraku računamo projekciju vektora y na prostor modela M.

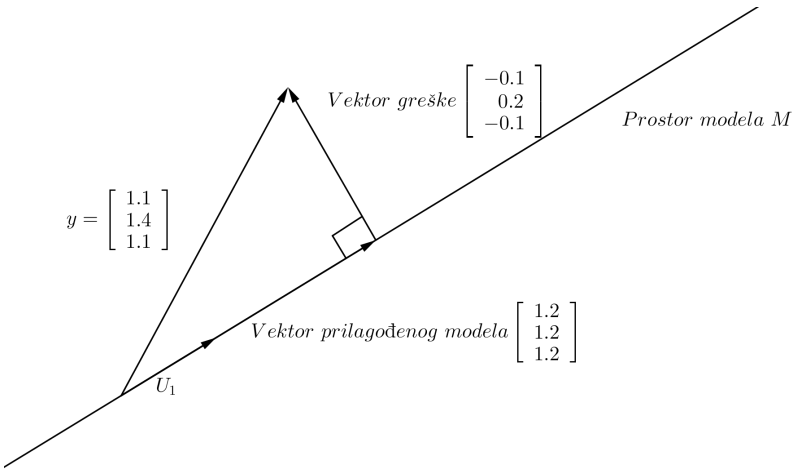
$$(y \cdot U_1)U_1 = \left(\begin{bmatrix} 1.1 \\ 1.4 \\ 1.1 \end{bmatrix} \cdot \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 1.2 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Iz vektora prilagođenog modela dobivamo procjenitelj za očekivanje $\hat{\mu} = \overline{y} = 1.2$. Ovim postupkom smo prilagodili model i došli do rastava vektora podataka na vektor prilagođenog modela i vektor greške

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \overline{y} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ y_3 - \overline{y} \end{bmatrix},$$

tj. u našem slučaju

$$\begin{bmatrix} 1.1 \\ 1.4 \\ 1.1 \end{bmatrix} = 1.2 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -0.1 \\ 0.2 \\ -0.1 \end{bmatrix}.$$



Slika 3.3: Prilagodba modela

2. ***Testiranje hipoteza***

Moramo izabrati dva jedinična vektora koja razapinju prostor greške, koja bi uz vektor U_1 tvorila ortogonalni koordinatni sustav u tri dimenzije. Biramo U_2 i U_3 pa dobivamo sustav

$$U_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad U_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad U_3 = \frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}.$$

Primjenom poopćenja Pitagorina poučka dobivamo

$$\begin{aligned} \|y\|^2 &= (y \cdot U_1)^2 + (y \cdot U_2)^2 + (y \cdot U_3)^2 \\ &= 4.32 + 0.045 + 0.015. \end{aligned}$$

Sada uspoređujemo kvadrat duljine projekcije vektora y na vektor smjera U_1 s prosjekom kvadrata duljina projekcija vektora y na koordinatne osi prostora greške, tj. računamo

$$f = \frac{(y \cdot U_1)^2}{[(y \cdot U_2)^2 + (y \cdot U_3)^2]/2} = \frac{4.32}{(0.045 + 0.015)/2} = 144.$$

Izračunajmo još i T-statistiku kao što smo opisali u 1.2. Iz podataka dobivamo da je

$$\hat{\mu} = 1.2, \ s = 0.1732, \ n = 3$$

pa je

$$\frac{\hat{\mu} - \mu_0}{s}\sqrt{n} = \frac{1.2}{0.1732}\sqrt{3} = 12,$$

tj.

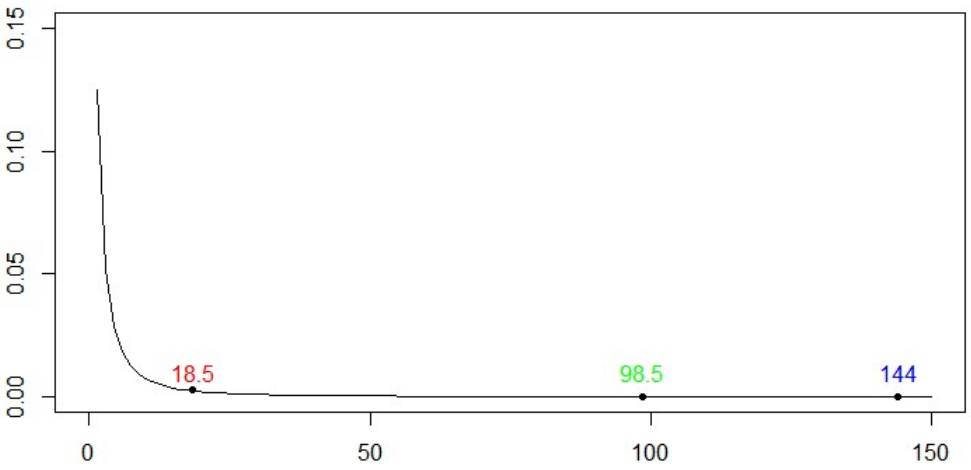
$$t(2)^2 = 12^2 = 144 = f(1, 2).$$

Ako je $\mu = 0$ gornji izraz prati $F(1, 2)$-distribuciju, a ako je $\mu \neq 0$ izraz je velik broj. Vidimo da je izraz velik broj, ali provedimo i test na način da provjerimo upada li vrijednost statistike u kritično područje, tj.

a) Ako upada u kritično područje za $F(1, 2)$-statistiku, odbijamo nul-hipotezu $H_0 : \mu = 0$

b) Ako ne upada u kritično područje za $F(1, 2)$-statistiku, ne odbijamo nul-hipotezu $H_0 : \mu = 0$

Pogledajmo kako izgleda funkcija gustoće $F(1, 2)$-distribucije s označenim 5% i 1% kvantilima.



Slika 3.4: $F(1,2)$-distribucija

U R-u smo izračunali željene kvantile i dobili da je kritično područje za $\alpha = 5\%$

$$[18.5\,, +\infty\rangle,$$

a za $\alpha = 1\%$

$$[98.5\,, +\infty\rangle$$

što znači da za obje razine značajnosti naša testna statistika $f = 144$ upada u kritično područje. Dakle, odbacujemo nul-hipotezu, tj. zaključujemo da je tretiranje sjemena fungicidom u prosjeku povećalo zrno pšenice.

Interval pouzdanosti

Jedan od predmeta interesa bio je odrediti prosječan utjecaj izlaganja pšenice fungicidima. Kako smo ranije izračunali, procjenitelj za očekivanje je $\overline{y} = 1.2$ tona/ha, no zanima nas također i točnost ove procjene. To možemo dobiti računanjem intervala pouzdanosti za parametar μ. U tu svrhu koristit ćemo T-statistiku opisanu u prvom poglavlju.
Kako je općenito $Y \cdot U_1 = \sqrt{n}\,\overline{Y}$ iz $N(\sqrt{n}\mu, \sigma^2)$, slučajna varijabla $\sqrt{n}(\overline{Y} - \mu)$ je iz $N(0, \sigma^2)$. Slučajne varijable $(Y \cdot U_2)^2, ..., (Y \cdot U_n)^2$ su također iz $N(0, \sigma^2)$ pa imamo da je

$$T(n-1) = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sqrt{\frac{(Y \cdot U_2)^2 + ... + (Y \cdot U_n)^2}{n-1}}}$$

T-statistika s $(n-1)$ stupnjeva slobode. Realizirana vrijednost te statistike jednaka je

$$t = \frac{\sqrt{n}(\overline{y}-\mu)}{\sqrt{\frac{(y\cdot U_2)^2+\ldots+(y\cdot U_n)^2}{n-1}}} = \frac{\sqrt{n}(\overline{y}-\mu)}{\sqrt{s^2}} = \frac{\overline{y}-\mu}{\sqrt{s^2/n}}$$

gdje je s^2 procjenitelj varijance. $t(n-1)$ je simetrična pa tražimo $t_{\frac{\alpha}{2}}(n-1)$ takav da vrijedi

$$\mathbb{P}\left(-t_{\frac{\alpha}{2}}(n-1) \leqslant T(n-1) \leqslant t_{\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$

$$\mathbb{P}\left(-t_{\frac{\alpha}{2}}(n-1) \leqslant \frac{\overline{Y}-\mu}{\sqrt{S^2/n}} \leqslant t_{\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$

Preoblikujući gornji izraz dolazimo do

$$\mathbb{P}\left(\overline{Y} - t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{S^2}{n}} \leqslant \mu \leqslant \overline{Y} + t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{S^2}{n}}\right) = 1-\alpha$$

tj. pouzdani interval za μ je

$$\left[\overline{Y} - t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{S^2}{n}}, \overline{Y} + t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{S^2}{n}}\right].$$

Prije nego izračunamo pouzdani interval, moramo procijeniti parametar σ^2. Prisjetimo se da smo u prvom poglavlju rekli da nam za tu procjenu trebaju slučajne varijable $Y\cdot U_2 \sim N(0,\sigma^2)$ i $Y\cdot U_3 \sim N(0,\sigma^2)$. Tada je

$$s^2 = \frac{(y\cdot U_2)^2 + (y\cdot U_3)^2}{2} = \frac{0.045+0.015}{2} = \frac{0.06}{2} = 0.03.$$

Dakle, u našem primjeru imamo da je 95% pouzdani interval upravo jednak

$$\left[1.2 - 4.303\sqrt{\frac{0.03}{3}}, 1.2 + 4.303\sqrt{\frac{0.03}{3}}\right] = \left[0.77, 1.63\right]$$

gdje je $t_{\frac{0.05}{2}}(2) = 4.303$ 5%/2 kvantil T-razdiobe s 2 stupnja slobode. Zaključujemo da je s pouzdanosti od 95% prosječna razlika u veličini zrna tretirane i netretirane pšenice između 0.77 i 1.63.

Primijetimo da će u svakom primjeru jedne populacije vektori modela biti oblika

$$\begin{bmatrix} \mu & \mu & \ldots & \mu \end{bmatrix}^T.$$

Zbog toga će i prostor modela biti

$$M = \left[\left\{ \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T \right\} \right]$$

te jedinični vektor koji ga razapinje $U_1 = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$ gdje je n broj podataka. Vektor U_1 će također biti i vektor smjera povezan s hipotezom H_0.
Jedan od prikladnih izbora ortogonalnog koordinatnog sustava u primjerima s jednom populacijom je

$$U_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, U_3 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \dots, U_n = \frac{1}{\sqrt{n(n-1)}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ -(n-1) \end{bmatrix}$$

Lako se pokaže da su svi vektori međusobno ortogonalni te također normirani.

3.2 Primjer - simulacija

U ovom primjeru ćemo na simuliranim podacima iz normalne razdiobe vidjeti kako funkcionira geometrijska metoda za jednu populaciju. Za simulaciju podataka i daljnje izračune koristit ćemo program R.

1. ***Vektor podataka y***

   Simuliramo dva niza podataka duljine 4 iz dvije različite populacije i to iz $N(0, 0.5)$ i $N(1, 0.5)$ te ćemo odgovarajuće vektore podataka označiti s y_1 i y_2 redom. Funkcija koju koristimo u R-u je *rnorm*.

```
> y1=rnorm(4,0,sqrt(0.5))
> y1
[1]  0.3687886 -0.5250293  0.9679499  1.2516216
> y2=rnorm(4,1,sqrt(0.5))
> y2
[1] 0.9245854 1.2253304 0.4297602 1.5438077
```

Slika 3.5: Vektori podataka

Za lakši izračun i pregledniji prikaz zaokružit ćemo vrijednosti na dvije decimale i tada dobiti vektore podataka

```
> y1=c(0.37,-0.53,0.97,1.25)
> y2=c(0.92,1.23,0.43,1.54)
```

Slika 3.6: Novi vektori podataka

$$y_1 = \begin{bmatrix} 0.37 \\ -0.53 \\ 0.97 \\ 1.25 \end{bmatrix}, \quad y_2 = \begin{bmatrix} 0.92 \\ 1.23 \\ 0.43 \\ 1.54 \end{bmatrix}.$$

2. ***Prostor modela M***

   Kako ćemo u nastavku testirati jesu li očekivanja jednaka nuli, ovdje pretpostavljamo da su ona nepoznata i označavamo ih s μ_1, μ_2 redom. Vektori modela za y_1 i y_2 su redom $\begin{bmatrix} \mu_1 & \mu_1 & \mu_1 & \mu_1 \end{bmatrix}^T$ i $\begin{bmatrix} \mu_2 & \mu_2 & \mu_2 & \mu_2 \end{bmatrix}^T$. Prostor modela je u oba slučaja

$$M = \left[\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\} \right]$$

   jednodimenzionalni potprostor četverodimenzionalnog prostora razapet jediničnim vektorom

$$U = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

3. ***Vektor smjera U***

   Želimo se uvjeriti da podaci iz prve populacije zaista imaju očekivanje jednako nuli, tj. da oni iz druge nemaju. Iz tog razloga promatramo hipoteze

$$H_0 : \mu_1 = 0 \qquad H_0 : \mu_2 = 0$$
$$H_1 : \mu_1 \neq 0 \qquad H_1 : \mu_2 \neq 0$$

Sada je jasno da je vektor smjera u oba slučaja

$$U_1 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T.$$

Za dodatna objašnjenja zašto to vrijedi vratite se na primjer 3.1 u ovom poglavlju.

4. ***Prilagodba modela***

Računamo projekcije vektora podataka na prostor modela

```
> u=c(1/sqrt(4),1/sqrt(4),1/sqrt(4),1/sqrt(4))
> y1.u=crossprod(y1,u)
> y1.u%*%u
      [,1]  [,2]  [,3]  [,4]
[1,] 0.515 0.515 0.515 0.515
>
>
> y2.u=crossprod(y2,u)
> y2.u%*%u
     [,1] [,2] [,3] [,4]
[1,] 1.03 1.03 1.03 1.03
```

Slika 3.7: Projekcije vektora y_1 i y_2 na M

$$(y_1 \cdot U)U = 0.52 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad (y_2 \cdot U)U = 1.03 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

i dobivamo da su procjene za μ_1 i μ_2 redom $\hat{\mu_1} = 0.52$ i $\hat{\mu_2} = 1.03$ Vidimo da procjena za μ_1 nije baš precizna, dok je za μ_2 ipak bolja.

5. ***Testiranje hipoteza***

Ortogonalni koordinatni sustav uzimamo kao što je navedeno na kraju primjera 3.1

$$U_1 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad U_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad U_3 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \end{bmatrix}, \quad U_4 = \frac{1}{\sqrt{12}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ -3 \end{bmatrix}.$$

Računamo sada (također u R-u) testne statistike pa za prvi slučaj dobivamo

$$f^* = \frac{(y_1 \cdot U_1)^2}{[(y_1 \cdot U_2)^2 + (y_1 \cdot U_3)^2 + (y_1 \cdot U_4)^2]/3} = \frac{1.06}{(0.41 + 0.74 + 0.72)/3} = 1.7,$$

a za drugi

$$f^{**} = \frac{(y_2 \cdot U_1)^2}{[(y_2 \cdot U_2)^2 + (y_2 \cdot U_3)^2 + (y \cdot U_4)^2]/3} = \frac{4.24}{(0.05 + 0.28 + 0.35)/3} = 18.71$$

Ako je $\mu = 0$ F-statistika prati $F(1, 3)$ distribuciju, a ako je $\mu \neq 0$ izraz je velik broj. Provjerimo upadaju li testne statistike u kritična područja koja dobivamo računajući odgovarajuće kvantile u R-u. Za razinu značajnosti od 5%

$$[10.13\,, +\infty\rangle$$

i za 1%

$$[34.12\,, +\infty\rangle$$

Vrijednost $f^* = 1.7$ ne upada u kritično područje niti za jednu od promatranih razina značajnosti pa zaključujemo da ne odbacujemo nul-hipotezu $H_0 : \mu_1 = 0$. Prisjetimo se da su podaci iz $N(0, 0.5)$ razdiobe pa su rezultati testa u skladu s našim očekivanjima.

S druge strane $f^{**} = 18.71$ upada u kritično područje za razinu značajnosti od 5%, ali za 1% ne. Dakle za razinu značajnosti od 5% odbacujemo nul-hipotezu $H_0 : \mu_2 = 0$ što smo i očekivali jer su podaci iz $N(1, 0.5)$, dok za 1% to ne možemo zaključiti.

Poglavlje 4

Dvije populacije

U ovom poglavlju proučavat ćemo dvije normalne populacije s očekivanjima μ_1 i μ_2 i zajedničkom varijancom σ^2. Glavno pitanje bit će jesu li očekivanja μ_1 i μ_2 jednaka pa ćemo računati kao i do sada procjenu parametara populacija, interval pouzdanosti za razliku $\mu_1 - \mu_2$ te ćemo testirati hipotezu $H_0 : \mu_1 = \mu_2$. Dakle, imat ćemo dva uzorka $y_{11}, ..., y_{1n}$ i $y_{21}, ..., y_{2n}$ od n mjerenja.
Razlika u odnosu na prethodno poglavlje je u tome što više ne promatramo samo jednu populaciju pa će nam vektor podataka biti popunjen mjerenim veličinama iz obje populacije, tj.

$$\begin{bmatrix} y_{11} & \dots & y_{1n} & y_{21} & \dots & y_{2n} \end{bmatrix}^T,$$

a ne njihovim razlikama. Ako se prisjetimo primjera s visinama muškaraca i žena opisanom u 2. poglavlju vidimo da je to upravo bio primjer s dvije populacije.

4.1 Primjer - kvaliteta vune

Ovaj primjer govori nam o utjecaju duljine tretiranja vune na njezinu kvalitetu, a preuzet je iz 6. poglavlja [7]. Organizacija WRONZ s Novog Zelanda razvila je metodu kojom se mjeri zapremnina prerađene i oprane vune pomoću posebnog stroja. Metoda uključuje niz koraka pripreme vune prije samog mjerenja. Iz velike količine vune uzimaju se mali uzorci koji se potom peru i suše. Nakon toga se uzorci omekšavaju u standardnim atmosferskim uvjetima koji su temperatura od $20 \pm 2°C$ i relativna vlažnost zraka od $65 \pm 2\%$. Uzimajući u obzir dosadašnju praksu, odredili su da faza omekšavanja treba trajati od 12 do 30 sati, a zanimalo ih je ima li razlike u vuni koja je u toj fazi bila dva, tj. tri dana. Dakle, od interesa su nam dvije populacije: ona koja je u navedenoj fazi bila dva i ona koja je bila tri dana. Više o samom postupku uzimanja uzorka i obradi vune možemo pronaći u [7]. Uzeli su deset uzoraka numeriranih brojevima od 1 do 10 te ih nasumično rasporedili u dvije

navedene grupe. Uzorci 2, 4, 5, 6 i 7 su se omekšavali dva dana, a 1, 3, 8, 9 i 10 tri. Podvrgnuti su istim uvjetima ranije obrade te su također izloženi tretmanu omekšavanja u istoj prostoriji s jedinom razlikom u trajanju postupka. Nakon postupka svaki uzorak podijeljen je na tri manja dijela koji su potom mjereni posebnim strojem. Rezultati mjerenja dani su u tablici na slici 4.1. Za uzorak broj 2 jedan od dijelova je izgubljen pa je aritmetička sredina rezultat mjerenja prva dva dijela. Podaci su izraženi u cm^3/g.

broj uzorka	mjerenje			aritmetička sredina
2	29.4	30.2	-	29.80
4	28.2	28.9	28.6	28.57
5	30.0	29.9	30.0	29.97
6	30.3	29.0	29.3	29.53
7	30.4	28.8	30.5	29.90
1	31.2	29.8	30.0	30.33
3	32.0	30.1	31.7	31.27
8	30.5	30.2	30.4	30.37
9	29.8	28.4	29.1	29.10
10	31.7	30.6	30.1	30.80

Slika 4.1: Rezultati mjerenja

1. ***Vektor podataka y***

   Za vektor podataka uzet ćemo prve tri aritmetičke sredine iz obje skupine. Prve tri veličine odgovaraju vuni omekšavanoj dva dana, a preostale onoj omekšavanoj tri dana.

$$y = \begin{bmatrix} 29.80 & 28.57 & 29.97 & 30.33 & 31.27 & 30.37 \end{bmatrix}^T$$

2. ***Prostor modela M***

   Kako smo već i rekli, pretpostavljamo da su podaci iz dvije normalne razdiobe s istom varijancom σ^2, ali različitih očekivanja. Za populaciju iz prvog tretmana koji je trajao dva dana označit ćemo da je očekivanje μ_1, a za drugu μ_2. Sada vektore modela možemo zapisati u obliku

$$\begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \end{bmatrix} = \mu_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Iz toga dobivamo da je prostor modela

$$M = \left[\left\{ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}^T, \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^T \right\} \right]$$

dvodimenzionalni potprostor šesterodimenzionalnog prostora razapet jediničnim vektorima

$$U_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}^T, \quad U_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^T.$$

3. ***Vektor smjera U***

S obzirom na to da nas zanima ima li razlike između dva tretmana, testiramo nulhipotezu nasuprot alternativne

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0.$$

Računamo vektor smjera kao

$$U_1 - U_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}^T$$

te ga još normiramo, tj. podijelimo s njegovom duljinom koja je $\sqrt{2}$ i dobivamo

$$\frac{1}{\sqrt{3}} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}^T = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}^T = U.$$

Vektor U očigledno leži u prostoru M, a kako je

$$y \cdot U = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} \cdot \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix} = \frac{y_{11} + y_{12} + y_{13} - y_{21} - y_{22} - y_{23}}{\sqrt{6}} = \frac{\sqrt{3}(\overline{y}_{1\cdot} - \overline{y}_{2\cdot})}{\sqrt{2}}$$

ispunjeno je i

$$E(Y \cdot U) = \frac{\sqrt{3}(\mu_1 - \mu_2)}{\sqrt{2}}.$$

Koristili smo oznaku $\overline{y}_{i\cdot}$ za $i = 1, 2$ koja predstavlja aritmetičku sredinu i-te populacije.

4. ***Prilagodba modela***

Računamo projekciju vektora y na prostor M na način da ga projiciramo na svaki od vektora kojima je prostor razapet te potom zbrojimo te vrijednosti

$$(y \cdot U_1)U_1 + (y \cdot U_2)U_2 = 29.447 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + 30.657 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 29.447 \\ 29.447 \\ 29.447 \\ 30.657 \\ 30.657 \\ 30.657 \end{bmatrix} = \begin{bmatrix} \overline{y}_{1\cdot} \\ \overline{y}_{1\cdot} \\ \overline{y}_{1\cdot} \\ \overline{y}_{2\cdot} \\ \overline{y}_{2\cdot} \\ \overline{y}_{2\cdot} \end{bmatrix}$$

te dobivamo vektor prilagođenog modela. Naš prilagođeni model sada je jednak

$$\begin{bmatrix} 29.80 \\ 28.57 \\ 29.97 \\ 30.33 \\ 31.27 \\ 30.37 \end{bmatrix} = \begin{bmatrix} 29.447 \\ 29.447 \\ 29.447 \\ 30.657 \\ 30.657 \\ 30.657 \end{bmatrix} + \begin{bmatrix} 0.353 \\ -0.877 \\ 0.523 \\ -0.327 \\ 0.613 \\ -0.287 \end{bmatrix}$$

tj.

$$y = \overline{y}_{i\cdot} + (y - \overline{y}_{i\cdot}).$$

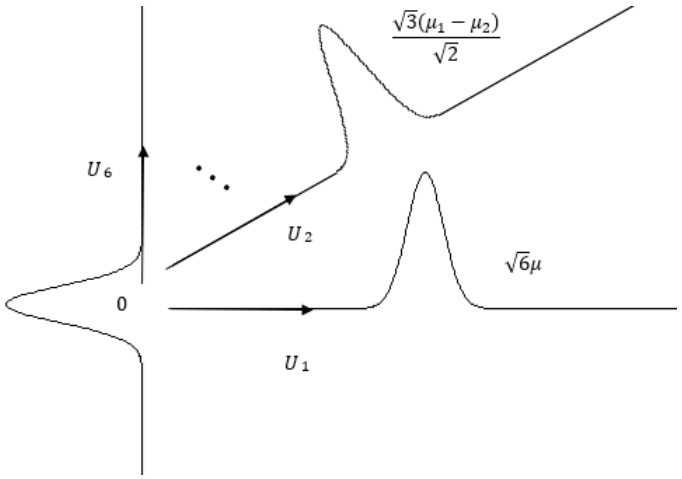
Dakle, procjenitelji za μ_1 i μ_2 su redom 29.447 i 30.657. Primijetimo da su koeficijenti projekcije zapravo procjene za parametre očekivanja populacija pa bi dekompozicija y s obzirom na osi U_1 i U_2 omogućila testiranje nul-hipoteza $\mu_1 = 0$ i $\mu_2 = 0$. Kako to nisu hipoteze od našeg interesa, uzet ćemo drugačiji koordinatni sustav. Takav sustav treba uključivati ranije izračunati vektor smjera povezan s hipotezom $H_0 : \mu_1 = \mu_2$. Jedan takav sustav od šest vektora je $U_1, U_2, ..., U_6$ koji su redom

$$\frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}, \frac{1}{\sqrt{6}} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ -2 \end{bmatrix}.$$

Vidimo da je $U_2 = U$ te da zajedno s U_1 razapinje prostor M. Osi U_3, U_4, U_5 i U_6 razapinju prostor greške. U_1 reflektira zajedničko očekivanje svih mjerenja jer je $y \cdot U_1 = \sqrt{6}\overline{y}_{\cdot\cdot}$ (oznaka $\overline{y}_{\cdot\cdot}$ predstavlja upravo to očekivanje). U_3 i U_4 reflektiraju promjenu u prvom tretmanu, tj.

$$y \cdot U_3 = \frac{y_{11} - y_{12}}{\sqrt{2}}, \quad y \cdot U_4 = \frac{y_{11} + y_{12} - 2y_{13}}{\sqrt{6}},$$

a U_5 i U_6 slično tako u drugom tretmanu.



Slika 4.2: Distribucija koeficijenata projekcije $Y \cdot U_1$, $Y \cdot U_2$,..., $Y \cdot U_6$

Sada ćemo prilagoditi model koristeći novoizabrane vektore U_1 i U_2.

$$(y\cdot U_1)U_1+(y\cdot U_2)U_2 = \begin{bmatrix} 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \end{bmatrix} + \begin{bmatrix} -0.605 \\ -0.605 \\ -0.605 \\ 0.605 \\ 0.605 \\ 0.605 \end{bmatrix} = \begin{bmatrix} \overline{y}_{..} \\ \overline{y}_{..} \\ \overline{y}_{..} \\ \overline{y}_{..} \\ \overline{y}_{..} \\ \overline{y}_{..} \end{bmatrix} + \begin{bmatrix} \overline{y}_{1.} - \overline{y}_{..} \\ \overline{y}_{1.} - \overline{y}_{..} \\ \overline{y}_{1.} - \overline{y}_{..} \\ \overline{y}_{2.} - \overline{y}_{..} \\ \overline{y}_{2.} - \overline{y}_{..} \\ \overline{y}_{2.} - \overline{y}_{..} \end{bmatrix} = \begin{bmatrix} 29.447 \\ 29.447 \\ 29.447 \\ 30.657 \\ 30.657 \\ 30.657 \end{bmatrix}$$

Konačno dobivamo novi prilagođeni model

$$\begin{bmatrix} 29.80 \\ 28.57 \\ 29.97 \\ 30.33 \\ 31.27 \\ 30.37 \end{bmatrix} = \begin{bmatrix} 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \end{bmatrix} + \begin{bmatrix} -0.605 \\ -0.605 \\ -0.605 \\ 0.605 \\ 0.605 \\ 0.605 \end{bmatrix} + \begin{bmatrix} 0.353 \\ -0.877 \\ 0.523 \\ -0.327 \\ 0.613 \\ -0.287 \end{bmatrix}$$

tj.

$$y \quad = \quad \overline{y}_{..} \quad + \quad (\overline{y}_{i.} - \overline{y}_{..}) \quad + \quad (y - \overline{y}_{i.}).$$

5. ***Testiranje hipoteza***

Sada možemo testirati nul-hipotezu nasuprot alternativne

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Kao i prije, računamo F-statistiku

$$f = \frac{(y \cdot U_2)^2}{[(y \cdot U_3)^2 + (y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2]/4} = \frac{2.1962}{0.4331} = 5.07.$$

Izračunajmo opet i T-statistiku. Iz podataka dobivamo da je

$$\hat{\mu}_1 = 29.4467,\ \hat{\mu}_2 = 30.6567,\ s_{y_{1\cdot}} = 0.764,\ s_{y_{2\cdot}} = 0.5316,\ n = 3$$

$$s_{y_{1\cdot}y_{2\cdot}} = \sqrt{s^2_{y_{1\cdot}} + s^2_{y_{2\cdot}}} = 0.9307$$

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{s_{y_{1\cdot}y_{2\cdot}}}\sqrt{n} = \frac{29.4467 - 30.6567}{0.9307}\sqrt{3} = -2.2518$$

Vidimo da zaista vrijedi $t(4)^2 = (-2.2518)^2 = 5.07 = f(1,4)$

Kritično područje za $\alpha = 5\%$ je

$$[7.71\,, +\infty\rangle,$$

a za $\alpha = 1\%$

$$[21.2\,, +\infty\rangle$$

pa niti za jednu razinu značajnosti naša testna statistika $f = 5.07$ ne upada u kritično područje. Dakle, ne odbacujemo nul-hipotezu, tj. nema statistički značajne razlike između vune omekšavane dva ili tri dana.

Interval pouzdanosti

Transformirali smo skup nezavisnih slučajnih varijabli $Y_{11}, Y_{12}, Y_{13} \sim N(\mu_1, \sigma^2)$ i $Y_{21}, Y_{22}, Y_{23} \sim N(\mu_2, \sigma^2)$ u novi skup nezavisnih varijabli

$$Y \cdot U_1 \sim N(\sqrt{6}\mu, \sigma^2), \quad Y \cdot U_2 \sim N(\sqrt{3}(\mu_1 - \mu_2)/\sqrt{2}, \sigma^2)$$

i

$$Y \cdot U_3, Y \cdot U_4, Y \cdot U_5, Y \cdot U_6 \sim N(0, \sigma^2).$$

Prva i druga slučajna varijabla korištene su za procjenu parametara μ i $\mu_1 - \mu_2$, a posljednje četiri služe nam za procjenu parametra σ^2

$$s^2 = \frac{(y \cdot U_3)^2 + (y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2}{4} = \frac{1.7324}{4} = 0.4331.$$

Najbolja procjena za razliku $\mu_1 - \mu_2$ dana je s $\bar{y}_{1\cdot} - \bar{y}_{2\cdot} = 29.447 - 30.657 = -1.21$. Iz toga bismo, a da ne provodimo test, mogli zaključiti da ipak postoji značajna razlika

između vune omekšavane dva i tri dana, no rezultati statističkog testa su suprotni tom zaključku. Stoga nas zanima koliko je precizna naša procjena, a u tu svrhu računamo 95% pouzdani interval za razliku $\mu_1 - \mu_2$. Slučajna varijabla povezana s tom razlikom je $Y \cdot U_2 = \sqrt{3}(\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot})/\sqrt{2}$. Ona ima $N(\sqrt{3}(\mu_1 - \mu_2)/\sqrt{2}, \sigma^2)$ razdiobu pa zato oduzimanjem očekivanja dobivamo slučajnu varijablu $W_1 = \sqrt{3}[(\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}) - (\mu_1 - \mu_2)]/\sqrt{2}$ koja ima $N(0, \sigma^2)$ razdiobu. Kako znamo i da slučajne varijable $Y \cdot U_3, Y \cdot U_4, Y \cdot U_5, Y \cdot U_6$ imaju istu tu distribuciju, imamo da je

$$\frac{\sqrt{3}[(\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}) - (\mu_1 - \mu_2)]/\sqrt{2}}{\sqrt{[(Y \cdot U_3)^2 + (Y \cdot U_4)^2 + (Y \cdot U_5)^2 + (Y \cdot U_6)^2]/4}}$$

$T(4)$-statistika. Realizirana vrijednost te statistike je

$$t = \frac{\sqrt{3}[(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) - (\mu_1 - \mu_2)]/\sqrt{2}}{\sqrt{[(y \cdot U_3)^2 + (y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2]/4}} = \frac{(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{2s^2/3}}.$$

Računamo interval pouzdanosti takav da vrijedi

$$\mathbb{P}\left(-t_{\frac{\alpha}{2}}(n-1) \leqslant T \leqslant t_{\frac{\alpha}{2}}(n-1)\right) = 1 - \alpha$$

$$\mathbb{P}\left(-t_{\frac{\alpha}{2}}(n-1) \leqslant \frac{(\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{2S^2/3}} \leqslant t_{\frac{\alpha}{2}}(n-1)\right) = 1 - \alpha$$

Jednostavnim računskim operacijama dobivamo

$$\mathbb{P}\left((\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}) - t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{2S^2}{3}} \leqslant \mu_1 - \mu_2 \leqslant (\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}) + t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{2S^2}{3}}\right) = 1 - \alpha,$$

tj. pouzdani interval za $\mu_1 - \mu_2$ je

$$\left[(\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}) - t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{2S^2}{3}}, (\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}) + t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{2S^2}{3}}\right].$$

Uvrštavanjem odgovarajućih vrijednosti dobivamo da je 95% pouzdani interval za razliku $\mu_1 - \mu_2$

$$[-2.7, \quad 0.28],$$

tj. sa sigurnošću od 95% možemo reći da je prosječna razlika između zapremnine vune omekšavane dva i tri dana između -2.7 i 0.28.

4.2 Primjer - simulacija

Prisjetimo se primjera 3.2 iz prethodnog poglavlja s jednom populacijom. Simulirali smo podatke iz dvije razdiobe, $N(0, 0.5)$ i $N(1, 0.5)$ te potom primijenili geometrijsku metodu. U ovom primjeru simulirat ćemo podatke iz $N(0, 0.5)$ i $N(3, 0.5)$.

```
> y1=rnorm(3,0,sqrt(0.5))
> y1
[1] 0.2040354 1.0042378 0.2851351
> y2=rnorm(3,3,sqrt(0.5))
> y2
[1] 2.214971 3.555472 3.117927
```

Slika 4.3: Podaci

U trećem poglavlju promatrali smo navedene podatke odvojeno kao dva manja primjera s jednom populacijom, a ovdje ćemo ih promatrati na drugačiji način, u svjetlu ovog poglavlja. Zaokružimo ih opet na dvije decimale radi lakših izračuna.

1. ***Vektor podataka y***

   Sada imamo da je vektor podataka

   $$y = \begin{bmatrix} 0.2 & 1 & 0.29 & 2.21 & 3.56 & 3.12 \end{bmatrix}.$$

2. ***Prostor modela M***

   Kako smo i u prethodnom poglavlju rekli, očekivanje prve populacije označimo s μ_1, a druge s μ_2. Sada vektore modela možemo zapisati u obliku

   $$\begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \end{bmatrix} = \mu_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

   Prostor modela M sada je

   $$M = \left[\left\{ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}^T, \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^T \right\} \right]$$

   razapet jediničnim vektorima

   $$U_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}^T, \quad U_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^T$$

   kao i u prethodnom primjeru.

3. ***Vektor smjera U***

   Želimo utvrditi razlikuju li se naši uzorci pa testiramo

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0.$$

   Vektor smjera je kao i u primjeru 4.1

$$U = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}^T.$$

   Za objašnjenje zašto to vrijedi vratite se na taj primjer.

4. ***Prilagodba modela***

   Projekciju vektora podataka na prostor M računamo kao sumu projekcija na vektore U_1 i U_2 koji su jednaki kao i u prethodnom primjeru.

```
> u1=c(1/sqrt(6),1/sqrt(6),1/sqrt(6),1/sqrt(6),1/sqrt(6),1/sqrt(6))
> u2=c(1/sqrt(6),1/sqrt(6),1/sqrt(6),-1/sqrt(6),-1/sqrt(6),-1/sqrt(6))
> y=c(0.2,1,0.29,2.21,3.56,3.12)
> y.u1=crossprod(y,u1)
> y.u2=crossprod(y,u2)
> y.u1%*%u1+y.u2%*%u2
          [,1]      [,2]      [,3]     [,4]     [,5]     [,6]
[1,] 0.4966667 0.4966667 0.4966667 2.963333 2.963333 2.963333
```

Slika 4.4: Projekcija vektora y na M

   Pa imamo da je

$$(y \cdot U_1)U_1 + (y \cdot U_2)U_2 = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 2.96 \\ 2.96 \\ 2.96 \end{bmatrix}.$$

   Procjene za μ_1 i μ_2 su redom 0.5 i 2.96.

5. ***Testiranje hipoteza***

   Posudit ćemo ortogonalni koordinatni sustav koji se sastoji od vektora $U_1, ..., U_6$ iz prethodnog primjera, gdje imamo i obrazloženje zašto ga koristimo. Vektori su redom

$$\frac{1}{\sqrt{6}}\begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}, \frac{1}{\sqrt{6}}\begin{bmatrix}1\\1\\1\\-1\\-1\\-1\end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\\0\\0\\0\\0\end{bmatrix}, \frac{1}{\sqrt{6}}\begin{bmatrix}1\\1\\-2\\0\\0\\0\end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix}0\\0\\0\\1\\-1\\0\end{bmatrix}, \frac{1}{\sqrt{6}}\begin{bmatrix}0\\0\\0\\1\\1\\-2\end{bmatrix}.$$

Računamo F-statistiku

$$\begin{aligned} f &= \frac{(y \cdot U_2)^2}{[(y \cdot U_3)^2 + (y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2]/4} \\ &= \frac{9.13}{(0.32 + 0.06 + 0.91 + 0.04)/4} \\ &= 27.46. \end{aligned}$$

Kritično područje za $\alpha = 5\%$ je

$$[7.71\,, +\infty\rangle,$$

a za $\alpha = 1\%$

$$[21.2\,, +\infty\rangle$$

pa za obje razine značajnosti naša testna statistika $f = 27.46$ upada u kritično područje. Odbacujemo nul-hipotezu, tj. zaista postoji statistički značajna razlika između dva promatrana uzorka, što smo i očekivali s obzirom na to iz kojih populacija su simulirani.

Poglavlje 5

Više populacija

U prethodnom poglavlju vidjeli smo primjere s dvije populacije te nas je zanimalo razlikuju li se njihova očekivanja odnosno računali smo razliku $\mu_1 - \mu_2$, a sada ćemo slično raditi i na primjerima s više populacija. Naših k populacija bit će normalno distribuirano s očekivanjima $\mu_1, \mu_2, ..., \mu_k$ i zajedničkom varijancom σ^2. Iz svake populacije uzimamo n uzoraka pa su vektor podataka i vektori modela u ovom poglavlju oblika

$$\begin{bmatrix} y_{11} & \dots & y_{1n} & y_{21} & \dots & y_{2n} & \dots & y_{k1} & \dots & y_{kn} \end{bmatrix}^T$$

$$\begin{bmatrix} \mu_1 & \dots & \mu_1 & \mu_2 & \dots & \mu_2 & \dots & \mu_k & \dots & \mu_k \end{bmatrix}^T.$$

Nul-hipoteze koje ćemo testirati su oblika $H_0 : c = 0$, gdje je $c = c_1\mu_1 + ... + c_k\mu_k$ veličina koju definiramo ovisno o primjeru koji proučavamo uz napomenu da vrijedi $c_1 + ... + c_k = 0$. Vektor smjera povezan s nul-hipotezom od interesa je

$$U = \frac{1}{\sqrt{n \sum_{i=1}^{k} c_i^2}} \begin{bmatrix} c_1 \\ \cdot \\ c_1 \\ c_2 \\ \cdot \\ c_2 \\ \vdots \\ c_k \\ \cdot \\ c_k \end{bmatrix}.$$

Točka u vektoru znači ponavljanje vrijednosti c_i za svaki i od 1 do k n puta. Kao i inače, F-statistiku računamo tako da uspoređujemo kvadrat duljine projekcije vektora podataka na vektor smjera povezan s nul-hipotezom s prosjekom kvadrata duljina projekcija istog vektora na prostor greške.

Tipovi veličine c

U praksi razlikujemo četiri različita tipa ranije definirane veličine c te njihove kombinacije. Glavni tipovi su usporedbe razreda, faktorske usporedbe, polinomijalne veličine i usporedbe u parovima. Kratko ćemo ih objasniti kroz primjere o kojima više možemo pronaći u [7].

1. Usporedbe razreda

   Znanstvenici su istraživali zimski usjev četiri sorte cikle i četiri sorte stočne repe, a zanimalo ih je je li cikla imala više ili manje suhe tvari u lukovicama od stočne repe. Posijali su sjeme osam različitih biljaka na 32 parcele nasumičnim redom te ih nakon žetve na određeni način tretirali i mjerili suhu tvar. Veličina c koju koristimo za testiranje hipoteze je

   $$c = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} - \frac{\mu_5 + \mu_6 + \mu_7 + \mu_8}{4}$$

   gdje su μ_i, $i = 1, ..., 4$ očekivanja populacija cikle, a μ_i za $i = 5, ..., 8$ očekivanja populacija stočne repe. Ovdje smo uspoređivali dva razreda, ciklu i repu.

2. Faktorske usporedbe

   U ovom slučaju znanstvenike su zanimali dugoročni zahtjevi za vapnenim dušikom i superfosfatom u uzgoju ječma. Posijali su 20 parcela ječma i slučajnim odabirom ih podijelili u četiri kategorije koje su tretirali godinama na različite načine. Prvi način bio je bez ikakvih gnojiva, drugi sa superfosfatom, treći je uključivao samo vapneni dušik, a četvrti kombinaciju ta dva gnojiva. Zanimalo ih je jesu li vapneni dušik i superfosfat imali ikakav utjecaj na usjev i je li odgovor na superfosfat bio jednak u prisutstvu, tj. odsutstvu vapnenog dušika. U skladu s tim definiramo

   $$c_1 = \frac{\mu_3 + \mu_4}{2} - \frac{\mu_1 + \mu_2}{2}$$

   $$c_2 = \frac{\mu_2 + \mu_4}{2} - \frac{\mu_1 + \mu_3}{2}$$

   $$c_3 = (\mu_4 - \mu_3) - (\mu_2 - \mu_1).$$

   Faktori su u ovom slučaju navedene vrste gnojiva, a $\mu_1, ..., \mu_4$ očekivanja populacija nastalih pojedinim tretmanima.

3. Polinomijalne veličine

   U ovom istraživanju je znanstvenike zanimalo na koji način je na veličinu zrna ječma sijanog na proljeće utjecao stupanj sjetve, tj. količina sjemena posijana po hektaru

zemlje. Konkretnije, je li se veličina zrna povećala s povećanjem stupnja sjetve i je li stopa rasta prinosa zrna opala s daljnjim porastom stope sjetve? Posijali su pet različitih količina sjemena ječma na pet parcela i iz svake od njih na kraju žetve izvagali šest zrna. Odgovarajuće veličine povezane s nul-hipotezom su

$$c_1 = -2\mu_1 - \mu_2 + \mu_4 + 2\mu_5$$

$$c_2 = 2\mu_1 - \mu_2 - 2\mu_3 - \mu_4 + 2\mu_5$$

gdje su $\mu_1, ..., \mu_5$ očekivanja populacija različitih stupnjeva sjetve. To odgovara linearnim i kvadratnim komponentama polinomijalne krivulje, odakle i naziv polinomijalne veličine.

4. Usporedbe u parovima

   Znanstvenici su ispitivali djelovanje četiri nova kemijska sredstva za suzbijanje prugaste hrđe pšenice. Proveden je pokus s četiri tretmana svakom od kemikalija te su uzeli pet mjerenja iz svakog tretmana. Mjerili su težinu tretiranog zrna u kilogramima i računali aritmetičke sredine tih podataka za svaki tretman, a koje su procjene za $\mu_1, ..., \mu_4$. Zanimala ih je usporedba svaka dva tretmana, tj. koji je bolji između njih. Sve usporedbe u parovima dane su kroz veličine povezane s tim pitanjima

$$c_1 = \mu_1 - \mu_2$$

$$c_2 = \mu_1 - \mu_3$$

$$c_3 = \mu_1 - \mu_4$$

$$c_4 = \mu_2 - \mu_3$$

$$c_5 = \mu_2 - \mu_4$$

$$c_6 = \mu_3 - \mu_4.$$

5.1 Primjer - zagađenje zraka

Na primjeru iz [7] vidjet ćemo kako primjenjujemo geometrijsku metodu za testiranje hipoteza u slučaju više populacija. Cilj istraživanja, koje se provodilo 1988. godine, bio je odrediti razlikuje li se onečišćenje zraka zimi u gradovima u odnosu na predgrađa u Christchurchu na Novom Zelandu. Drugo pitanje od interesa bilo je ima li razlike u onečišćenju zraka između brdovitih i ravnih dijelova predgrađa. Znanstveni odbor Novog Zelanda svaki sat tijekom tri mjeseca mjerio je količinu zagađenja dimom u mikrogramima po kubičnom metru. Nasumično su izabrane dvije lokacije brdovitog predgrađa, dvije ravnog predgrađa i dva područja u gradu u kojima su provedena mjerenja. Dakle u ovom primjeru imamo

tri populacije. Aritmetičke sredine mjerenja tijekom tri mjeseca na svakoj od lokacija i dodatno aritmetičke sredine za svaku od skupina područja navedene su u tablici na slici 5.1.

populacija	1.	2.	ukupno
brdovito predgrađe	124	110	117
ravno predgrađe	107	115	111
grad	126	138	132

Slika 5.1: Tablica mjerenja

1. ***Vektor podataka y***

$$y = \begin{bmatrix} 124 & 110 & 107 & 115 & 126 & 138 \end{bmatrix}^T$$

2. ***Prostor modela M***

Pretpostavljamo da su podaci iz normalne razdiobe s različitim očekivanjima μ_1, μ_2 i μ_3 te jednakom varijancom σ^2. Vektore modela zapisujemo kao

$$\begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_3 \\ \mu_3 \end{bmatrix} = \mu_1 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \mu_3 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Stoga je prostor modela M

$$M = \left[\left\{ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}^T, \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}^T \right\} \right]$$

trodimenzionalni potprostor šesterodimenzionalnog prostora razapet jediničnim vektorima

$$U_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad U_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad U_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

3. ***Vektor smjera U***

U ovom primjeru imamo dvije hipoteze od interesa, kako smo i rekli u uvodnom dijelu. U prvom slučaju uspoređujemo predgrađe i grad pa s obzirom na to da imamo dva podatka, za predgrađe uzimamo njihovu aritmetičku sredinu. Drugi se odnosi na brdovito i ravno područje predgrađa, tj. zanima nas jesu li ta očekivanja različita.

$$\begin{aligned} 1.\quad & H_0 : (\mu_1 + \mu_2)/2 = \mu_3 \\ & H_1 : (\mu_1 + \mu_2)/2 \neq \mu_3 \end{aligned}$$

$$\begin{aligned} 2.\quad & H_0 : \mu_1 = \mu_2 \\ & H_1 : \mu_1 \neq \mu_2 \end{aligned}$$

Naravno, možemo ih zapisati i u sljedećem obliku

$$\begin{aligned} 1.\quad & H_0 : (\mu_1 + \mu_2) - 2\mu_3 = 0 \\ & H_1 : (\mu_1 + \mu_2) - 2\mu_3 \neq 0 \end{aligned}$$

$$\begin{aligned} 2.\quad & H_0 : \mu_1 - \mu_2 = 0 \\ & H_1 : \mu_1 - \mu_2 \neq 0 \end{aligned}$$

Vektori smjera povezani s prvim i drugim slučajem su redom

$$U^* = \frac{1}{\sqrt{12}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -2 \\ -2 \end{bmatrix} \quad i \quad U^{**} = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

gdje smo prvi vektor dobili računajući $U_1 + U_2 - 2U_3$, a drugi $U_1 - U_2$ te smo potom dobivene vektore normirali. Očigledno se nalaze u prostor M, a vidimo i da je

$$\begin{aligned} y \cdot U^* &= 2(\overline{y}_{1\cdot} + \overline{y}_{2\cdot} - 2\overline{y}_{3\cdot})/\sqrt{12} \\ y \cdot U^{**} &= 2(\overline{y}_{1\cdot} - \overline{y}_{2\cdot})/\sqrt{4} = \overline{y}_{1\cdot} - \overline{y}_{2\cdot} \end{aligned}$$

točnije, ispunjeno je $E(Y \cdot U^*) = 2(\mu_1 + \mu_2 - 2\mu_3)/\sqrt{12}$ i $E(Y \cdot U^{**}) = \mu_1 - \mu_2$.

Uočimo kako smo drugačijim zapisom hipoteza dobili dvije veličine c za svaki slučaj, npr. u prvom slučaju imamo $c_1 = (\mu_1 + \mu_2)/2 - \mu_3$ i $c_2 = (\mu_1 + \mu_2) - 2\mu_3$. Ako izračunamo vektore smjera za c_1 i c_2 dobivamo isti vektor, tj.

$$U_{c_1} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \\ -1 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{12}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -2 \\ -2 \end{bmatrix} = U_{c_2}.$$

Kako smo dobili jednake vektore smjera zaključujemo da su hipoteze zavisne, štoviše postoji beskonačno mnogo takvih veličina c pa je svejedno koju ćemo koristiti u izračunima, rezultati će biti isti.

4. ***Prilagodba modela***

Računamo projekciju vektora podataka na prostor M

$$(y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 = 117 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + 111 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + 132 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 117 \\ 117 \\ 111 \\ 111 \\ 132 \\ 132 \end{bmatrix} = \begin{bmatrix} \overline{y}_{1\cdot} \\ \overline{y}_{1\cdot} \\ \overline{y}_{2\cdot} \\ \overline{y}_{2\cdot} \\ \overline{y}_{3\cdot} \\ \overline{y}_{3\cdot} \end{bmatrix}$$

Na taj način smo dobili vektor prilagođenog modela, tj. procjenitelje za parametre $\mu_1,\ \mu_2,\ \mu_3$ koji su redom $117,\ 111,\ 132$.

5. ***Testiranje hipoteza***

Biramo odgovarajući ortogonalni koordinatni sustav tako da sadrži vektore U^* i U^{**} koje sada preimenujemo u U_2 i U_3, a koji zajedno s vektorom U_1 razapinju prostor M te vektore U_4, U_5 i U_6 koji razapinju prostor greške. Sada su vektori $U_1, U_2, ..., U_6$ redom dani s

$$\frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{12}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -2 \\ -2 \end{bmatrix}, \frac{1}{\sqrt{4}} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}$$

Sada računamo statistike za ranije navedene hipoteze pa tako imamo da je statistika za prvi slučaj gdje testiramo je li $(\mu_1 + \mu_2)/2 \neq \mu_3$

$$\frac{(y \cdot U_2)^2}{[(y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2]/3} = 6.42.$$

U slučaju gdje nas zanima je li $\mu_1 \neq \mu_2$ testna statistika je

$$\frac{(y \cdot U_3)^2}{[(y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2]/3} = 0.53.$$

Kritično područje za $\alpha = 5\%$ je

$$[10.13\,, +\infty\rangle,$$

a za $\alpha = 1\%$

$$[34.12\,, +\infty\rangle$$

pa za obje razine značajnosti naše testne statistike ne upadaju u kritično područje. Dakle, ne odbacujemo nul-hipoteze u oba slučaja. Ne možemo zaključiti da je zagađenje zraka od dima zimi na testiranim područjima statistički značajno različito u predgrađu i gradu te isto tako za brdovito i ravno područje predgrađa.

Bibliografija

[1] *Primjer parametrijskog testa: Studentov test ili t-test (predavanja)*, https://vub.hr/images/uploads/1471/oirus-statistika-predavanje-sat-12.pdf, posjećena 5.12.2020.

[2] M. Bombardelli i Ž. Milin Šipuš, *Analitička geometrija (skripta)*, (2016), https://web.math.pmf.unizg.hr/nastava/ag/dodatni/AG-predavanja-2016.pdf, posjećena 12.10.2020.

[3] Z. Franušić i J. Šiftar, *Linearna algebra 1 (skripta)*, https://web.math.pmf.unizg.hr/~fran/predavanja-LA1.pdf, posjećena 13.10.2020.

[4] ———, *Linearna algebra 2 (skripta)*, https://web.math.pmf.unizg.hr/~fran/predavanja-LA2.pdf, posjećena 13.10.2020.

[5] M. Huzak, *Vjerojatnost i matematička statistika (predavanja)*, https://www.yumpu.com/xx/document/read/28599991/vjerojatnost-i-matematicka-statistika-poslijediplomski-specijalisticki-, posjećena 3.11.2020.

[6] N. Koceić Bilan, *Primijenjena statistika (skripta)*, https://www.pmfst.unist.hr/odjel-za-matematiku/wp-content/uploads/sites/24/2018/05/n_koceic_b_primijenjena-statistika.pdf, posjećena 30.10.2020.

[7] D. J. Saville i G. R. Wood, *Statistical Methods: The Geometric Approach*, Springer, 1991.

Sažetak

Kada provodimo znanstveno istraživanje, nije dovoljno samo pogledati dobivene rezultate i donijeti zaključak o problemu. Nužno je provesti statistički test i tek tada zaključiti što smo dobili tim istraživanjem. U testiranju postavljenih hipoteza koriste se različiti testovi i statistike, ovisno o problemu koji promatramo. U ovom diplomskom radu promatrali smo F-statistiku, a pitanje koje nas je zanimalo bilo je razlikuju li se promatrana očekivanja populacija. Glavna ideja rada bila je pomoću vektora objasniti zašto koristimo F-statistiku kod rješavanja takvih problema pa smo u prvom poglavlju ponovili osnovna znanja iz područja geometrije i statistike. Niz koraka u kojima smo to predočili nazvali smo geometrijski pristup. Potom smo promatrali jednu, dvije i tri populacije te kroz nekoliko primjera proveli sve korake geometrijskog pristupa te na taj način dobili F-statistiku i donijeli zaključke istraživanja. Tri glavna primjera preuzeta su iz [7], a dva su izmišljena i simulirana u programu R-u. Za bolje razumijevanje rad smo potkrijepili slikama koje su rađene u programima GeoGebra-i i R-u, a ideja za većinu njih preuzeta je iz glavne literature na kojoj se temelji rad [7].

Summary

When we conduct a scientific research it's not enough to just look at the raw data to make conclusions. It is necessary to perform a statistical test and only then we can make a conclusion about the topic. Different tests and statistics are used in testing the hypothesis, depending on the problems we are interested in. In this thesis, we observed F-statistic and research subject was whether the observed population means differ. The main idea of the paper was to explain why we use F-statistics for such problems, while using vectors, so in the first chapter we reminded ourselves of the basic knowledges in the fields of geometry and statistics. We called the series of steps in which we presented this the geometric approach. We then observed one, two, and three populations and through several examples carried out all the steps of the geometric approach thus obtaining F-statistics and drawing research conclusions. Three main examples were taken over from [7] and two were made up and simulated in the R program. For a better understanding, the thesis substantiated with images made in Geogebra and R. The idea for most of them are also taken from the main literature [7] on which the paper is based.

Životopis

Rođena sam 27. lipnja 1996. godine u Našicama. U gradu Orahovici živjela sam od rođenja do upisa na fakultet i preseljenja u Zagreb, gdje i danas živim. Osnovu školu, a potom i opću gimnaziju s odličnim uspjehom završila sam u Orahovici 2015. godine. Nakon toga upisala sam preddiplomski sveučilišni studij Matematike, nastavnički smjer na Prirodoslovno-matematičkom fakultetu u Zagrebu gdje sam 2018. godine stekla titulu sveučilišne prvostupnice edukacije matematike. Sljedeća stepenica u mom obrazovanju bila je upis diplomskog sveučilišnog studija Financijske i poslovne matematike na istom fakultetu.Tijekom školovanja, uz matematiku voljela sam i kemiju i biologiju, a izvan nastave svirala sam saksofon u DVD-u, volontirala u Gradskom društvu Crvenog križa te udruzi mladih PAUK u Orahovici. Slobodno vrijeme provodim aktivno u prirodi, u društvu obitelji i prijatelja ili gledajući filmove i serije.