

# Analiza varijance ponovljenih mjerena

---

**Pažin, Ivan**

**Master's thesis / Diplomski rad**

**2014**

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: <https://urn.nsk.hr/um:nbn:hr:217:530462>

Rights / Prava: [In copyright/Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-04-26**



Repository / Repozitorij:

[Repository of the Faculty of Science - University of Zagreb](#)



**SVEUČILIŠTE U ZAGREBU**  
**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**  
**MATEMATIČKI ODSJEK**

Ivan Pažin

**ANALIZA VARIJANCE PONOVLJENIH  
MJERENJA**

Diplomski rad

Voditelj rada:  
prof. dr. sc. Anamarija Jazbec

Zagreb, srpanj, 2014

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. \_\_\_\_\_

*Zaručnici Zrinki, koja je s puno razumijevanja, bila velika potpora i pomoć, osobito u trenucima kada me je trebalo pogurati i ohrabriti. Kolegi i nadasve prijatelju Hrvoju koji me strpljivo i nesebično, čak i nakon što je diplomirao, pratio i pomagao tijekom studija. Roditeljima koji su stoički sve podnosili i nisu gubili vjeru u mene u trenucima kada je ja nisam imao. Sestrama Ani, Maji, Petri i Antoniji i posebno bratu Luki koji su uvijek imali lijepu riječ za mene. Samome sebi kao daljnji poticaj i pokazatelj da mogu i znam. Bogu koji sve započinje, svime upravlja te sve dovršava.*

# Sadržaj

|  |           |
|--|-----------|
| <b>Sadržaj</b>   | <b>iv</b> |
| <b>Uvod</b>  | <b>1</b>  |
| <b>1 Testiranje hipoteza o jednakosti sredina</b>                      | <b>3</b>  |
| 1.1 Uvod . . . . .   | 3         |
| 1.2 Studentov $t$ test za dva uzorka . . . . .                         | 7         |
| 1.3 Primjer . . . . .  | 8         |
| <b>2 Univariatna analiza varijance</b>                                 | <b>13</b> |
| 2.1 Uvod u problematiku . . . . .                                      | 13        |
| 2.2 Jednofaktorski model analize varijance . . . . .                   | 14        |
| 2.3 Dvofaktorski model analize varijance . . . . .                     | 18        |
| 2.4 Primjer - dvofaktorska analiza varijance . . . . .                 | 23        |
| <b>3 Analiza varijance ponovljenih mjerena</b>                         | <b>27</b> |
| 3.1 Uvod . . . . .   | 27        |
| 3.2 Jednofaktorski model . . . . .                                     | 28        |
| 3.3 Primjer - jednofaktorski model . . . . .                           | 33        |
| 3.4 Dvofaktorski model analize varijance ponovljenih mjerena . . . . . | 36        |
| 3.5 Primjer - dvofaktorski model . . . . .                             | 39        |
| <b>Bibliografija</b>   | <b>45</b> |

# Uvod

Statistika je grana primjenjene matematike koja se bavi prikupljanjem i analizom podataka te interpretacijom rezultata analize uz uporabu dobro definiranih metoda. Na diplomskom studiju Matematičke statistike, osim teorije, naglasak je bio na metodama analize i njezinim primjenama. Jedna od čestih i popularnih metoda je Analiza varijance (skraćeno ANOVA) koju ćemo u ovome radu obraditi te staviti naglasak na njezinu generalizaciju, Analizu varijance ponovljenih mjerena. Metodu (ANOVA), kao određeni matematički model i praktičnu tehniku za istraživanje nekih bioloških fenomena, prvi je razvio i dao joj ime (eng. analysis of variance) poznati engleski statističar R. A. Fisher (1890-1962)[2]. ANOVA je specijalni slučaj linijske regresije, koja je opet specijalni slučaj generaliziranih linijskih modela kojima je zajedničko da minimiziraju grešku modela.

U prvom poglavlju ćemo obraditi Studentov t test koji nam je potreban zbog boljeg razumijevanja ANOVE te uvesti osnove statističke definicije i pojmove koji će nam trebati u dalnjem radu. Proći ćemo pretpostavke za t test; postavljanje hipoteza; t test za dva uzorka i njegovu interpretaciju. U drugom poglavlju ćemo preko gore spomenutih generaliziranih linijskih modela doći do modela ANOVE. Osim uvođenja osnovnih pojmovova za shvaćanje analize varijance tu će biti govora o jednofaktorskome i dvofaktorskome modelu; postavljanju hipoteza i njihovu testiranju te testiranju hipoteza o adekvatnosti modela. U trećem poglavlju nadogradit ćemo model ANOVE dodajući ponovljena mjerena kao faktor koristeći prethodno stečena znanja. Time ćemo obraditi naš glavni model i temu ovoga rada, analizu varijance ponovljenih mjerena.

Svako poglavlje započet ćemo problemom, odnosno pitanjem koje će nam poslužiti kao daljnja motivacija. Na kraju poglavlja metodu ćemo ilustrirati kroz konkretni primjer riješen u programskom sustavu SAS. Ispis iz SAS-a ćemo povezati s pojmovima iz poglavlja, a rezultate interpretirati.

Cilj statistike je bolje razumijevanje svijeta, odnosno traženje odgovora na razlike (varijabilitet) među prikupljenim podacima. Da bi iz podataka mogli učiti potrebno ih je pravilno čitati. U današnjem svijetu da bi mogli pravilno čitati podatke potrebni su nam brzi i kvalitetni programski sustavi zbog velike količine podataka. Zbog toga koristimo programski sustav SAS kojeg ćemo kroz primjere bolje upoznati.



# Poglavlje 1

## Testiranje hipoteza o jednakosti sredina

### 1.1 Uvod

#### Motivacija

Prepostavimo da nas zanima jesu li muškarci u Hrvatskoj prosječno viši od muškaraca u Kini. U tom slučaju imamo dvije populacije, prva populacija su muškarci u Hrvatskoj, a druga muškarci u Kini. Kada bismo mogli izmjeriti sve muškarce u Hrvatskoj i sve muškarce u Kini u vrlo kratkom periodu, statistički testovi nam tada ne bi bili potrebni. Jednostavno bismo usporedili prosjeke jedne i druge populacije i zaključili gdje su viši muškarci. Kako nam je nemoguće izmjeriti sve muškarce u Hrvatskoj i Kini, moramo smisliti nešto drugo. Izabrat ćemo uzorak iz jedne i iz druge populacije koji će populaciju dobro predstavljati. I tu dolazi na red statistika. Mi se nećemo baviti izborom uzorka nego ćemo pretpostaviti da imamo dva velika uzorka iz jedne i druge populacije. Sada ih moramo usporediti. Iako nam prosjek uzorka može sugerirati tko je viši nepravilno je na temelju njega zaključiti da su muškarci u jednoj zemlji viši. Zašto je tomu tako? Jer uzorak može biti loš, odnosno nereprezentativan. Tu nam treba statistički test da potvrdi je li razlika koja postoji između prosjeka dvaju uzorka značajna ili ne.

#### Definicije i osnovni pojmovi

Neka je  $\Omega$  skup elementarnih događaja.

**Definicija 1.1.1.** *Uređena trojka  $(\Omega, (\mathcal{F}), P)$ , gdje je  $\mathcal{F}$   $\sigma$ -algebra na  $\omega$  i  $P$  vjerojatnost na  $\mathcal{F}$ , zove se vjerojatnosni prostor.*

Vjerojatnosni prostor osnovni je objekt u teoriji vjerojatnosti. Vjerojatnosni prostor može biti diskretni vjerojatnosni prostor kod kojega je  $\Omega$  konačan ili prebrojiv skup ili opći

vjerojatnosni prostor gdje slučajna varijabla odnosno veličina koja se mjeri u vezi s nekim slučajnim pokusom, može primiti sve realne vrijednosti ili sve realne vrijednosti iz nekog intervala. Definirat ćemo slučajnu varijablu na općem vjerojatnosnom prostoru.

Neka je  $\mathbb{R}$  skup realnih brojeva. Sa  $\mathcal{B}$  označimo  $\sigma$ -algebru generiranu familijom svih otvorenih skupova na  $\mathbb{R}$ .  $\mathcal{B}$  zovemo  $\sigma$ -algebra Borelovih skupova na  $\mathbb{R}$ , a elemente  $\sigma$ -algebре  $\mathcal{B}$  zovemo Borelovi skupovi. Iz definicije slijedi (vidi N. Sarapa, poglavlje 8 [3]) da je svaki otvoreni, zatvoreni, poluotvoreni, poluzatvoreni interval Borelov skup. Također neograničeni intervali, jednočlani skupovi, prebrojivi podskupovi, skup svih racionalnih, iracionalnih brojeva su Borelovi skupovi. Ipak može se dokazati da postoje skupovi na  $\mathbb{R}$  koji nisu Borelovi. Dakle,  $\mathcal{B} \neq \mathcal{P}(\mathbb{R})$ .

**Definicija 1.1.2.** Neka je  $(\Omega, (F), P)$  vjerojatnosni prostor. Funkcija  $X : \Omega \rightarrow \mathbb{R}$  jest slučajna varijabla (na  $\Omega$ ) ako je  $X^{-1}(B) \in \mathcal{F}$  za proizvoljno  $B \in \mathcal{B}$ , tj.  $X^{-1}(\mathcal{B}) \subset \mathcal{F}$ .

Jedan od osnovnih pojmova u teoriji vjerojatnosti jest pojam funkcije distribucije slučajne varijable.

**Definicija 1.1.3.** Neka je  $X$  slučajna varijabla na  $\Omega$ . Funkcija distribucije od  $X$  jest funkcija  $F_x : \mathbb{R} \rightarrow [0, 1]$  definirana sa

$$F_x(x) = P\{\omega \in \Omega : X(\omega) \leq x\} = P(X \leq x), \quad x \in \mathbb{R}. \quad (1.1)$$

Koristit ćemo oznaku  $F_x = F$ , ako je jasno o kojoj se slučajnoj varijabli radi. Usko uz pojam funkcije distribucije vezan je i pojam funkcije gustoće.

**Definicija 1.1.4.** Neka je  $X$  slučajna varijabla na  $(\Omega, (F), P)$  i neka je  $F_x$  njezina funkcija distribucije. Kažemo da je  $X$  absolutno neprekidna, ili, kraće, neprekidna slučajna varijabla ako postoji nenegativna realna Borelova funkcija  $f$  na  $\mathbb{R}$  ( $f : \mathbb{R} \rightarrow \mathbb{R}_+$ ) takva da je

$$F_x(x) = \int_{-\infty}^x f(t) d\lambda(t), \quad x \in \mathbb{R}. \quad (1.2)$$

Integral u (1.2) je Lebesgueov integral funkcije  $f$  u odnosu na Lebesgueovu mjeru  $\lambda$ . Za funkciju distribucije  $F_x$  neprekidne slučajne varijable  $X$ , dakle za funkciju u (1.2) kažemo da je absolutno neprekidna funkcija distribucije.

Ako je  $X$  neprekidna slučajna varijabla, tada se funkcija  $f$  iz (1.2) zove funkcija gustoće vjerojatnosti od  $X$ , tj. od njezine funkcije distribucije  $F_x$  ili, kraće, gustoća od  $X$ .

Iz (1.2) i poglavlja 9.2. iz [3] slijedi da ako znamo gustoću neprekidne slučajne varijable  $X$ , znamo vjerojatnosti svih događaja koji su u vezi s tom slučajnom varijablom. Također slijedi da je funkcija distribucije neprekidne slučajne varijable  $X$  u potpunosti određena njezinom gustoćom.

Sada ćemo definirati distribucije koje ćemo kasnije koristiti.

*Normalna distribucija*

Neka su  $\mu, \sigma \in \mathcal{R}$ ,  $\sigma > 0$ . Neprekidna slučajna varijabla  $X$  ima normalnu distribuciju s parametrima  $\mu$  i  $\sigma^2$  ako joj je gustoća dana sa

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}. \quad (1.3)$$

To ćemo označavati  $X \sim N(\mu, \sigma^2)$ .

$X$  ima jediničnu normlnu distribuciju ako je  $X \sim N(0, 1)$ , dakle

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}. \quad (1.4)$$

*Studentova distribucija*

Neka je  $n \in \mathbb{N}$ . Neprekidna slučajna varijabla  $X$  ima Studentovu  $t$ -distribuciju sa  $n$  stupnjeva slobode ako joj je gustoća  $f$  dana sa

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}. \quad (1.5)$$

gdje je  $\Gamma$  gama funkcija. Danu distribuciju označavamo  $X \sim t(n)$ .

Dalje uvodimo pojam očekivanja i varijance za neprekidnu slučajnu varijablu  $X$  u oznaci  $EX$  i  $VarX$ .

Neka je  $X$  neprekidna slučajna varijabla s gustoćom  $f_x$ . Tada je s

$$EX = \int_{-\infty}^{\infty} xf_x(x) d\lambda(x). \quad (1.6)$$

dana formula za računanje očekivanja slučajne varijable  $X$ , a s

$$VarX = \int_{-\infty}^{\infty} (x - EX)^2 f_x(x) d\lambda(x). \quad (1.7)$$

je dana formula za računanje varijance slučajne varijable  $X$ .

Pozitivan drugi korijen iz varijance zovemo standardna devijacija od  $X$  i označujemo  $\sigma_x$ .

Definirat ćemo slučajan uzorak te osnovne statistike uzorka.

**Definicija 1.1.5.** Neka je  $(\Omega, (F), P)$  vjerojatnosni prostor i neka je  $X$  slučajna varijabla na  $\Omega$  s funkcijom distribucije  $F$ . Kažemo da  $X_1, X_2, \dots, X_n$  čine slučajan uzorak duljine  $n$  iz populacije s funkcijom distribucije  $F$  ako su  $X_1, X_2, \dots, X_n$  nezavisne i jednako distribuirane slučajne varijable sa zajedničkom funkcijom distribucije  $F$ .

Ako funkcija  $F$  ima gustoću  $f$ , tj. ako je  $X$  neprekidna ili diskretna slučajna varijabla s gustoćom  $f$ , kažemo da je uzorak uzet iz gustoće  $f$ . Intuitivno slučajan uzorak duljine  $n$  odgovara nizu od  $n$  nezavisnih mjerena slučajnog svojstva nekog statističkog skupa (populacije), i to slučajnog svojstva koje se opisuje slučajnom varijablom  $X$ .

**Definicija 1.1.6.** Neka je  $X = (X_1, \dots, X_n)$  slučajan uzorak iz funkcije distribucije  $F$  i  $g : \mathcal{R}^n \rightarrow \mathcal{R}$  Borelova funkcija. Slučajnu varijablu  $T = g(X)$  zovemo statistika.

U primjenama se najčešće promatraju sljedeće dvije statistike:

$$\bar{X} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad (1.8)$$

koju zovemo aritmetička sredina ili očekivanje uzorka i

$$S^2 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (1.9)$$

koju zovemo varijanca uzorka.

Definicija uzorka (Def. 1.1.5.) i Propozicija 11.4 iz N. Sarapa poglavlje 11.3 [3] sugeriraju da očekivanje uzorka može poslužiti kao "dobra" procjena za očekivanje populacije, ako je to očekivanje nepoznato.

## Testiranje hipoteza

Statistička hipoteza je tvrdnja o veličini parametra  $\mu$  ili o obliku distribucije osnovnog skupa čija se vjerodostojnost ispituje pomoću slučajnog uzorka. Postupak ili pravilo kojim se donosi odluka o prihvaćanju ili neprihvaćanju tvrdnje na temelju podataka iz slučajnog uzorka naziva se testiranjem statističkih hipoteza[5]. Statistički testovi dijele se na parametarske i neparametarske. Pri provođenju parametarskih testova polazi se od danog oblika i karakteristika distribucije numeričke varijable u osnovnom skupu.  $T$  test je parametarski test.

Svaki postupak testiranja polazi od *nulte hipoteze* i *alternativne hipoteze*. Sadržaj alternativne hipoteze uvijek proturijeći sadržaju nulte hipoteze. Sadržaj hipoteza određujemo mi sami. Sud koji izvire iz odluke o prihvaćanju ili neprihvaćanju nulte hipoteze nije kategoričan jer se odluka donosi na temelju vrijednosti iz slučajnog uzorka, odnosno dijela podataka. Stoga se u postupku odluke mogu pojaviti dvije vrste pogrešaka. Pogreška tipa I. i pogreška tipa II. Pogreška tipa I. je kada se odbaci istinita nulta hipoteza. Pogreška tipa II. je kada se prihvati nulta hipoteza iako je lažna. Tablični prikaz je u tablici 1.1.

Pogreška tipa I. predočuje se vjerojatnošću odbacivanja nulte hipoteze  $\alpha$ . Još se naziva razinom značajnosti. S  $\beta$  označujemo vjerojatnost da se prihvati lažna nulta hipoteza,

odnosno da se učini greška tipa  $II$ . Vjerojatnost odbacivanja lažne nulte hipoteze naziva se snagom statističkog testa. Ta je vjerojatnost jednaka  $(1 - \beta)$ .

| Odluka                  | $H_0$ je istinita   | $H_0$ je lažna       |
|-------------------------|---------------------|----------------------|
| Prihvati nultu hipotezu | odluka ispravna     | pogreška tipa $II$ . |
| Odbaci nultu hipotezu   | pogreška tipa $I$ . | odluka ispravna      |

Tablica 1.1: Vrste grešaka

Hipotezu  $H_0$  ćemo odbaciti ukoliko nam je testna statistika upala u kritično područje odnosno ukoliko nam je  $p$  vrijednost (vjerojatnost odbacivanja istinite nulte hipoteze) manja od razine značajnosti  $\alpha$  koju ćemo u primjerima standardno uzimati da bude 0.05.

## 1.2 Studentov $t$ test za dva uzorka

### Povijest $t$ -testa

T distribuciju otkrio je je 1908. godine William Sealy Gosset, kemičar sa Oxforda radeći u Guinness-ovojoj pivovari u Dublinu. Otkrio ju je tražeći jeftin način za kontrolu proizvodnje piva. Kako je politika Guinnessa bila da zaposlenici ne smiju objavljivati svoja otkrića, Gosset je objavio svoje otkriće pod pseudonimom "Student". Iz toga razloga se ta distribucija danas zove Studentova distribucija, odnosno pripadni test, Studentov  $t$ -test. Rad je objavljen u matematičkom časopisu *Biometrika* čiji je urednik bio poznati matematičar Karl Pearson.

### Prepostavke i testna statistika

Neka su iz dvije populacije sa sredinama  $\mu_1$  i  $\mu_2$ , i nepoznatim varijancama  $\sigma_1^2$  i  $\sigma_2^2$  izvučena dva slučajna uzorka veličine  $n_1 < 30$  i  $n_2 < 30$  respektivno. Na uzorcima je izmjerena zavisna varijabla  $Y$  uz prepostavke da su

$$Y_i \sim N(\mu_i, \sigma_i^2) \quad \text{i} \quad \sigma_1^2 = \sigma_2^2 \quad \text{gdje je } i = 1, 2.$$

Hipoteze koje hoćemo testirati su sljedeće:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Iz dva izvučena uzorka računaju se sredine  $\bar{y}_1, \bar{y}_2$  i varijance  $s_1^2, s_2^2$  koje su nam procjene za  $\mu_i, \sigma_i^2$ . Označimo sa  $d$  razliku sredina uzoraka  $d = \bar{y}_1 - \bar{y}_2$ . Tada je *sampling distribucija*<sup>1</sup>

---

<sup>1</sup>distribucija dobivena iz uzorka

za  $d$  uz istinitu  $H_0$ , gdje su uzorci nezavisni normalna sa sredinom nula i varijancom  $\sigma^2 * (\frac{1}{n_1} + \frac{1}{n_2})$ . Testna statistika je tada oblika s pripadnom distribucijom:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \stackrel{H_0}{\sim} t_{n_1+n_2-2} \quad (1.10)$$

gdje je  $s^2$  procjenitelj za  $\sigma^2$  oblika

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (1.11)$$

Ova test statistika se naziva *pooled t* statistika.

U dvosmjernom testu  $H_0$  se odbacuje za  $t < t_{\frac{\alpha}{2}}$ ,  $t > t_{1-\frac{\alpha}{2}}$  gdje su  $t_{\frac{\alpha}{2}}$  i  $t_{1-\frac{\alpha}{2}}$  granice kritičnog područja.

Odstupanje od gore navedenih klasičnih pretpostavki zahtijeva prilagodbe ovisno o tome od koje se pretpostavke odstupa. Ukoliko nemamo normalnost možemo probati napraviti transformaciju na podacima, za odstupanje od jednakosti varijance koristimo drugu test statistiku (*satterthwaite t* statistika), za zavisnost između uzoraka koristimo  $t$  test za zavisne uzorke (koristimo drugu test statistiku).

## 1.3 Primjer

### Programski sustav SAS

Programski sustav SAS je modularni, integrirani aplikacijski sustav koji na jednostavan i fleksibilan način omogućuje kako elementarnu, tako i sofisticiranu analizu podataka uporabom *"point and click"* tehnike rada preko grafičkih korisničkih sučelja (Grafical User Interface -GUI), gotovih programa - SAS procedura, ili programiranjem.

Glavna područja primjene programskog sustava SAS: rukovanje, deskriptivna i grafička analiza podataka; statistička analiza i podrška u odlučivanju, prezentacija rezultata i razvoj aplikacija.

Prednost SAS-a je da ima jednaku sintaksu i sučelje na mirkoračunalima, mini računalima i velikim računalima te se SAS programi mogu izvoditi bez ikakve promjene koda programa interaktivno ili u *batch*<sup>2</sup> modu.

Programski sustav SAS je nastao 1966. godine u SAD-u za potrebe poljoprivrednih istraživanja na sveučilištu North Carolina. Tvorci sustava su Jim Goodnight i John Sall.

---

<sup>2</sup>u pozadini omogućavajući neki drugi rad na računalu

## Zadatak

Imamo trideset i jednog ispitanika u dobi od 35 do 54 godine kojima je mjereno vrijeme da pretrče jednu i pol milju (2,41 kilometar)<sup>3</sup>. Ispitanike ćemo podijeliti u dvije grupe. Jedna grupa će sadržavati ispitanike koji imaju između 35 i 44 godine, a druga ispitanike koji imaju između 45 i 54 godine. Zanima nas postoji li razlika u sredinama između grupa odnosno kome treba manje da pretrči jednu i pol milju. Uzorci su nezavisni. Vrijeme je izraženo u minutama. Dio podataka je dan na slici 1.1. U prvom redu su označene variable i broj opservacije (*Obs*). Varijabla *age* nam označava broj godina, *runtime* potrebno vrijeme da se pretrči jedna milja, a varijabla *AgeGr* nam je grupna varijabla koja označava kojoj dobroj grupi opspervacija pripada.

| The SAS System |     |         |         |
|----------------|-----|---------|---------|
| Obs            | age | runtime | AgeGr   |
| 1              | 44  | 10.13   | _35-44_ |
| 2              | 44  | 11.37   | _35-44_ |
| 3              | 40  | 10.07   | _35-44_ |
| 4              | 44  | 8.65    | _35-44_ |
| 5              | 42  | 8.17    | _35-44_ |
| 6              | 38  | 9.22    | _35-44_ |
| 7              | 40  | 11.95   | _35-44_ |
| 8              | 43  | 10.85   | 35-44   |
| 9              | 44  | 13.08   | _35-44_ |
| 10             | 38  | 8.63    | _35-44_ |
| 11             | 57  | 12.63   | _45-54_ |
| 12             | 54  | 11.17   | _45-54_ |
| 13             | 52  | 9.63    | _45-54_ |
| ...            |     |         |         |

Slika 1.1: Dio podataka - Ispis iz SAS-a

---

<sup>3</sup>Podaci preuzeti s tečaja STAT3 u Srcu ak. godine 2013/14, voditelj tečaja: mr. sc. Vesna Hljuz Dobrić

## Kod

```
proc univariate data=fit normal;
var runtime;
class AgeGr;
run;

proc ttest data=fit plots=all;
var runtime;
class AgeGr;
run;
```

## Ispis

| The UNIVARIATE Procedure |            |                  |            |
|--------------------------|------------|------------------|------------|
| Variable: runtime        |            |                  |            |
| AgeGr = _35-44_          |            |                  |            |
| Moments                  |            |                  |            |
| N                        | 10         | Sum Weights      | 10         |
| Mean                     | 10.212     | Sum Observations | 102.12     |
| Std Deviation            | 1.60268386 | Variance         | 2.56859556 |
| Skewness                 | 0.45169976 | Kurtosis         | -0.7230033 |
| Uncorrected SS           | 1065.9668  | Corrected SS     | 23.11736   |
| Coeff Variation          | 15.6941232 | Std Error Mean   | 0.50681314 |

  

| The UNIVARIATE Procedure |            |                  |            |
|--------------------------|------------|------------------|------------|
| Variable: runtime        |            |                  |            |
| AgeGr = _45-54_          |            |                  |            |
| Moments                  |            |                  |            |
| N                        | 21         | Sum Weights      | 21         |
| Mean                     | 10.7642857 | Sum Observations | 226.05     |
| Std Deviation            | 1.27600773 | Variance         | 1.62819571 |
| Skewness                 | 0.90790835 | Kurtosis         | 0.96243037 |
| Uncorrected SS           | 2465.8307  | Corrected SS     | 32.5639143 |
| Coeff Variation          | 11.8540864 | Std Error Mean   | 0.27844771 |

Slika 1.2: Deskriptivna statistika za obje grupe - Ispis iz SAS-a

| The TTEST Procedure |    |         |         |         |         |         |
|---------------------|----|---------|---------|---------|---------|---------|
| Variable: runtime   |    |         |         |         |         |         |
| AgeGr               | N  | Mean    | Std Dev | Std Err | Minimum | Maximum |
| _35-44_             | 10 | 10.2120 | 1.6027  | 0.5068  | 8.1700  | 13.0800 |
| 45-54               | 21 | 10.7643 | 1.2760  | 0.2784  | 8.9200  | 14.0300 |
| Diff (1-2)          |    | -0.5523 | 1.3857  | 0.5324  |         |         |

  

| AgeGr                    | Method | Mean    | 95% CL Mean     | Std Dev | 95% CL        | Std Dev |
|--------------------------|--------|---------|-----------------|---------|---------------|---------|
| _35-44_                  |        | 10.2120 | 9.0655 11.3585  | 1.6027  | 1.1024 2.9259 |         |
| 45-54                    |        | 10.7643 | 10.1835 11.3451 | 1.2760  | 0.9762 1.8426 |         |
| Diff (1-2) Pooled        |        | -0.5523 | -1.6411 0.5366  | 1.3857  | 1.1035 1.8628 |         |
| Diff (1-2) Satterthwaite |        | -0.5523 | -1.7874 0.6828  |         |               |         |

  

| Method        | Variances | DF     | t Value | Pr >  t |
|---------------|-----------|--------|---------|---------|
| Pooled        | Equal     | 29     | -1.04   | 0.3081  |
| Satterthwaite | Unequal   | 14.653 | -0.96   | 0.3550  |

  

| Equality of Variances |        |        |         |        |  |
|-----------------------|--------|--------|---------|--------|--|
| Method                | Num DF | Den DF | F Value | Pr > F |  |
| Folded F              | 9      | 20     | 1.58    | 0.3783 |  |

Slika 1.3: Rezultati  $t$  testa - Ispis iz SAS-a

## Interpretacija

Prvo smo provjeravali normalnost uzoraka sa procedurom *univariate* koja nam daje testove za normalnost koji nam svi redom ne odbacuju  $H_0$  za obje grupe. Osim testova normalnosti procedura *univariate* nam daje i deskriptivnu statistiku za obje grupe u tablicama na slici 1.2. Sredina odnosno srednje vrijeme za prvu grupu nam je 10,21 minuta sa standardnom devijacijom 1,60, a za drugu grupu 10,76 minuta sa standardnom devijacijom 1,28. Već iz samih sredina vidimo da ćemo teško odbaciti nultu hipotezu.

Procedura *ttest* čiji su rezultati prikazani u tablici na slici 1.3 nam testira jednakost sredina. Osim toga automatski nam testira i pretpostavku o jednakosti varijanci sa metodom *Folded F* koja nam daje  $p$  vrijednost od 0,38 pa prihvaćamo nultu hipotezu da su varijance grupa jednake. Opcijom *plots=all* specificiramo da hoćemo sve grafove koje nam daje procedura *ttest*. Kako su nam varijance jednake gledamo *Pooled* metodu kod koje je testna statistika -1,04, a  $p$  vrijednost 0,3081 što je veće od  $\alpha$  (0,05), pa prema tome prihvaćamo nultu hipotezu u  $t$  testu odnosno prosječna vremena potrebna za pretrčati jednu i pol milju kod mlađe i starije grupe ispitanika značajno se ne razlikuju.



## Poglavlje 2

# Univariatna analiza varijance

### 2.1 Uvod u problematiku

Prepostavimo da tri različite tvornice automobila A, B i C proizvode, među ostalim, i tip automobila približno iste snage motora, pa se želi provjeriti hipoteza da potrošnja goriva ne ovisi o marki (tvornici) automobila.<sup>1</sup> Kako organizirati eksperiment koji će omogućiti donošenje odluke o prihvaćanju, odnosno odbacivanju postavljene hipoteze?

Odmah se nameće ideja da se uzme nekoliko automobila svake marke, proveze određeni broj kilometara sa svakim od njih, te izmjeri potrošnja goriva. No, svaki vozač zna da potrošnja goriva ovisi i o mnogim drugim faktorima (vrsta ceste, vozačko iskustvo, godišnje doba i sl.). Da bi se eliminirao utjecaj ceste, mogu se svi automobili voziti po istoj cesti. Želi li se eliminirati i utjecaj vozača, čini se razumnim slučajno izabrati vozače, tako da se dobivene vrijednosti potrošnje goriva mogu smatrati vrijednostima slučajnog uzorka. U svakom slučaju cilj nam je utvrditi utjecaj samo jednog faktora, faktora proizvođača, na potrošnju goriva.

Budući da je marka automobila nenumeričko obilježje izmjerene brojčane vrijednosti (litara/100 km) se mogu shvatiti kao vrijednosti izlazne varijable, uzrokovane odgovarajućim vrijednostima (A,B,C) nenumeričkog faktora - marke automobila. To nam pokazuje da će matematički model za opisivanje promatranog fenomena imati određene sličnosti sa regresijskim modelima, u smislu da se izlazna numerička vrijednost shvaća kao posljedica djelovanja nenumeričke vrijednosti (razine, tretmana, grupe) ulazne varijable (djelujućeg faktora), u ovom slučaju marke automobila, koja određuje srednju vrijednost izlazne varijable (potrošnje goriva) i kojoj se dodaje slučajna greška.

Čini se prilično logičnim da se na početku postavljeno pitanje o postojanju ili nepostojanju značajne razlike u potrošnji goriva formulirana kao problem testiranja nul hipoteze  $H_0 : \mu_a = \mu_b = \mu_c$ , prema alternativnoj hipotezi da se bar dvije sredine razlikuju.

---

<sup>1</sup>Motivacijski primjer i način obrade preuzeti iz [2]

To je tipični problem analize varijance i cijela teorija analize varijance se uglavnom sastoji od objašnjenja postupka za njegovo rješavanje. Budući da se ta teorija uglavnom bavi analizom rasipanja (varijabilnosti) izlaznih podataka, teorija je i dobila naziv *analiza varijance*.

Očigledno je da rasipanje aritmetičkih sredina može poslužiti kao indikator valjanosti hipoteze  $H_0$ . Da smo, recimo, dobili sve tri aritmetičke sredine međusobno jednake, onda bi njihova varijanca bila nula i u tom slučaju ne bismo odbacili nultu hipotezu. Osim rasipanja aritmetičkih sredina među grupama možemo računati i varijabilnost unutar svake grupe te dobiti ponderiranu sredinu tih varijanci. Pokazuje se da je omjer te dvije varijance (u brojniku je ponderirana varijanca među grupama, u nazivniku ponderirana varijanca unutar grupa) prikladan indikator za donošenje odluke o hipotezi  $H_0$ , jer očigledno da je prevelika vrijednost tog omjera (veliko rasipanje aritmetičkih sredina među grupama i malo rasipanje unutar grupa) upućuje na odbacivanje nulte hipoteze.

## 2.2 Jednofaktorski model analize varijance

Univarijatnu analizu varijance karakterizira jedna kriterijska - zavisna varijabla i jedna ili više eksperimentalnih - nezavisnih varijabli - faktora. Obzirom na broj faktora u modelu u univarijatnoj analizi varijance razlikuju se jednofaktorska i višefaktorska ANOVA. Obzirom na izbor nivoa faktora model može biti model s fiksnim efektima i model sa slučajnim efektima. Višefaktorski eksperiment može sadržavati glavne efekte i ukrštene (interakcijske) efekte.

Jednofaktorski, *potpuno slučajni dizajn* (CRD) najjednostavniji je eksperimentalni dizajn s jednom zavisnom varijablom kontinuiranog tipa i jednom faktorskom varijabom diskretnog tipa s konačnim brojem nivoa. Razlikujemo istraživačke - *sampling* studije (gdje faktor s  $m$  nivoa definira  $m$  populacija i iz svake populacije  $i$  se izabire po jedan reprezentativni uzorak s  $n_i$  opservacija - ukupno dakle  $m$  uzoraka ili grupa) i eksperimentalna studija ( $m$  nivoa faktora definira  $m$  tretmana i iz jedne populacije slučajno se izabire jedan uzorak te se jedinicama analize slučajno pridruži jedan od  $m$  tretmana, što definira  $m$  poduzoraka). Na osnovu sredina po uzorcima ili poduzorcima treba donijeti zaključak o očekivanjima populacija i zaključak o efikasnosti tretmana. Hipoteze koje se testiraju su  $H_0$  : očekivanja populacija / efekti tretmana su jednaki i  $H_1$  : postoje barem dvije populacije čija su očekivanja različita / postoje barem dva efekta koja različito utječu na sredine.

Klasične prepostavke za valjano testiranje hipoteza su:

- uzorci su nezavisni i slučajni (randomizacija) - opservacije u uzorcima nezavisne
- $m$  populacijskih varijanci su međusobno jednake - varijance zavisne varijable po uzorcima homogene

- svaka od  $m$  populacija ima normalnu distribuciju - zavisna varijabla po uzorcima aproksimativno normalna

Ako se  $H_0$  odbaci, možemo koristiti *post hoc* testove da utvrdimo koje se sredine uzorka razlikuju.

U gornjem primjeru i pripadnom tekstu istaknuli smo bitne momente, koji će nam olakšati shvaćanje općih apstraktnih pojmove koje ćemo sada definirati.

Prepostavlja se da je dano  $m$  ( $m \geq 3$ ) nizova podataka

$$\left\{ \begin{array}{l} y_{11}, y_{12}, \dots, y_{1n_1} \\ y_{21}, y_{22}, \dots, y_{2n_2} \\ \dots, \dots, \dots, \dots, \\ y_{m1}, y_{m2}, \dots, y_{mn_m} \end{array} \right., \quad n_1, \dots, n_m \in \mathbb{N}, \quad (2.1)$$

i da je  $i$ -ti ( $i = 1, \dots, m$ ) niz dobiven mjerjenjem slučajnog varijable  $Y_i \sim N(\mu_i, \sigma^2)$ , te da su  $Y_1, \dots, Y_m$  nezavisne slučajne varijable. To znači da se  $y_{ij}$  ( $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ ) može interpretirati kao vrijednost slučajne varijable

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad (2.2)$$

gdje su  $\epsilon_{ij} \sim N(0, \sigma^2)$  nezavisne slučajne varijable.

Možemo, dakle, reći da je  $Y_{ij}$  izlazna slučajna varijabla, čije vrijednosti  $y_{ij}$  nastaju djelovanjem  $i$ -te razine ( $\mu_i$ ) određenog faktora, uz dodatak slučajne greške ( $\epsilon_{ij}$ ). Djelujući faktor (ulazna varijabla) najčešće ima nenumeričko obilježje i u tome je glavna razlika u odnosu na model jednodimenzionalne (jednofaktorske) regresije.

U primjeru na početku poglavlja djelujući je faktor marka automobila i imamo tri ( $m = 3$ ) razine A, B i C toga faktora. Potrošnja goriva je izlazna varijabla  $y_{ij}$  i za svaki  $i$  imamo  $n_i$  vrijednosti izlazne varijable. Ukupno se raspolaže sa  $n = n_1 + n_2 + n_3$  podataka.

Općenito se stavlja

$$n = \sum_{i=1}^m n_i, \quad (2.3)$$

pri čemu  $n$  označuje ukupni broj podataka.

Sada možemo općenito definirati i glavni problem jednofaktorske analize varijance, koji se sastoji u određivanju postupka za testiranje nul hipoteze

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_m, \quad (2.4)$$

prema alternativnoj hipotezi da u (2.4) postoje  $i$  i  $j$  takvi da  $\mu_i \neq \mu_j$ . Drugim riječima, problem se sastoji u određivanju kritičnog područja, zadane razine značajnosti, pri testiranju hipoteze o jednakosti očekivanja  $m$  nezavisnih slučajnih varijabli normalnih razdioba zajedničke nepoznate varijance  $\sigma^2$ , na temelju  $m$  nizova podataka (2.1).

U praktičnim situacijama hipoteza  $H_0$  obično se iskazuje kao hipoteza da razlike razine djelujućeg faktora ne utječu na promatranu izlaznu veličinu, odnosno da tretman nema utjecaj na promatranu veličinu.

Stavimo

$$\mu = \frac{1}{n} \sum_{i=1}^m n_i \mu_i, \quad \delta_i = \mu_i - \mu, \quad i = 1, \dots, m. \quad (2.5)$$

Uobičajeno je  $\mu$  zvati *opća srednja vrijednost* dok se  $\delta_i$  zove *efekt*  $i$ -te razine djelujućeg faktora. U tom svjetlu jednadžba (2.2) može zapisati u obliku

$$Y_{ij} = \mu + \delta_i + \epsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n_i. \quad (2.6)$$

a hipoteza  $H_0$  iz (2.4)

$$H_0 = \delta_1 = \delta_2 = \dots = \delta_m = 0 \quad (2.7)$$

Jednadžba (2.6) se može protumačiti tako da se izlazna vrijednost  $y_{ij}$  shvati kao zbroj opće vrijednosti  $\mu$ , efekta  $\delta_i$   $i$ -te razine djelujućeg faktora i vrijednosti  $\epsilon_{ij}$  slučajne greške  $E_{ij}$ . Hipotezom  $H_0$ , zapisanom u obliku (2.7), postavlja se teza da su efekti beznačajni.

Da bi se definirala prikladna test statistika, pomoću koje će se odrediti kritično područje i pripadajuća  $p$  vrijednost zadane razine značajnosti  $\alpha$ , uvest ćemo sljedeće označke

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, \dots, m, \quad (2.8)$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^m n_i \bar{Y}_i \quad (2.9)$$

gdje je  $\bar{Y}_i$  aritmetička sredina  $i$ -te grupe (niza podataka), a  $\bar{Y}$  aritmetička sredina svih podataka.

Nadalje označimo sa  $SST$  sumu kvadrata zbog tretmana (kvadriranu razliku sredine pojedine grupe od opće srednje vrijednosti) i  $SSE$  sumu kvadrata pogrešaka (kvadriranu razliku pojedine opservacije od svoje sredine)

$$SST = \sum_{i=1}^m n_i (\bar{Y}_i - \bar{Y})^2 \quad (2.10)$$

$$SSE = \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2. \quad (2.11)$$

Zbrojimo li  $SST$  i  $SSE$  dobivamo  $SS$  odnosno kvadratnu sumu svih odstupanja od sredine

$$SS = \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2. \quad (2.12)$$

Podijelimo li svaki od zadnja tri izraza sa pripadajućim brojem nezavisnih podataka/grupa dobivamo sljedeće

$$MST = \frac{SST}{m-1} \quad (2.13)$$

$$MSE = \frac{SSE}{n-m} \quad (2.14)$$

$$MSS = \frac{SS}{n-1} \quad (2.15)$$

gdje je  $MST$  srednjekvadratno odstupanje zbog razlika među grupama,  $MSE$  srednjekvadratno odstupanje razlika unutar grupe, a  $MSS$  srednjekvadratno ukupno odstupanje.

Pod pretpostavkom istinitosti  $H_0$  vrijedi sljedeće:

$$Y_{ij} \sim N(\mu, \sigma^2), \quad i = 1, \dots, m, \quad j = 1, \dots, n_i. \quad (2.16)$$

$$Y_i \sim N(\mu, \frac{\sigma^2}{n_i}), \quad i = 1, \dots, m, \quad (2.17)$$

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n}) \quad (2.18)$$

pa na temelju Ž.Pauše IV.4. [2] i nezavisnosti od  $SST$  i  $SSE$  vrijedi

$$\frac{1}{\sigma^2} SST = \frac{m-1}{\sigma^2} MST \sim \chi(m-1), \quad (2.19)$$

$$\frac{1}{\sigma^2} SSR = \frac{n-m}{\sigma^2} MSE \sim \chi(n-m), \quad (2.20)$$

Primjeni li se Ž. Pauše V.6. [2] dobiva se

$$F = \frac{MST}{MSE} \stackrel{H_0}{\sim} F(m-1, n-m), \quad (2.21)$$

što će nam biti test statistika za našu analizu. Da je ta vrijednost prikladna za donošenje odluke o prihvaćanju ili odbacivanju  $H_0$  može se zaključiti iz činjenice da je

$$E[MST] = \sigma^2 + \frac{1}{m-1} \sum_{i=1}^m n_i \delta_i^2, \quad E[MSE] = \sigma^2, \quad (2.22)$$

bez obzira na  $H_0$ . Ako je hipoteza  $H_0$  istinit, onda je prvo očekivanje jednak nula zbog (2.7), pa se može očekivati da će vrijednost test statistike F biti oko jedan, a ako hipoteza  $H_0$  nije istinita, onda se može očekivati povećanje  $MSE$ , pa stoga i test statistike. S većom test statistikom smanjuje se vjerojatnost odbacivanja hipoteze  $H_0$ .

Kritično područje je oblika  $[f_\alpha(m - 1, n - m), +\infty]$ , gdje je  $f_\alpha$  kvantil pripadne  $F$  distribucije uz zadanu razinu značajnosti  $\alpha$ , a pripadna  $P$ -vrijednost  $p = \mathbb{P}(F > f|H_0)$  odnosno vjerojatnost da je testna statistika upala u kritično područje uz istinitost hipoteze  $H_0$ . Razina značajnosti  $\alpha$  i  $p$  vrijednost se mogu interpretirati kao površine ispod grafa funkcije gustoće za pripadne vrijednosti na osi  $x$ . Tako, ako je  $p < \alpha$  znači da je testna statistika desno od  $f_\alpha$  tj. upala je u kritično područje.

Zbog boljeg pregleda i veće jasnoće problema i njegovih rezultata, uobičajeno je da se sve navedene statistike prikazuju u obliku ANOVA tablice. U tablici 2.2 prikazan je opći oblik ANOVA tablice za jednofaktorski model analize varijance.

| Izvor varijabilnosti | Sume kvadrata  | Stupnjevi slobode | Varijanca               | Test statistika $F$   | $p$ vrijednost              |
|----------------------|--|-------------------|-------------------------|-----------------------|-----------------------------|
| Između grupa         | $SST = \sum_{i=1}^m n_i(\bar{Y}_i - \bar{Y})^2$              | $m - 1$           | $MST = \frac{SST}{m-1}$ | $F = \frac{MST}{MSE}$ | $p = \mathbb{P}(F > f H_0)$ |
| Unutar grupa         | $SSE = \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$ | $n - m$           | $MSE = \frac{SSE}{n-m}$ |                       |                             |
| Ukupno               | $SS = \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$    | $n - 1$           |                         |                       |                             |

Tablica 2.1: ANOVA tablica - jednofaktorski model

## 2.3 Dvofaktorski model analize varijance

Na početku poglavlja smo imali primjer u kojem smo htjeli provjeriti hipotezu da potrošnja goriva ne ovisi o marki odnosno tvornici automobila. U obzir smo uzeli samo jedan faktor, i to marku automobila. Prirodno bi se bilo pitati utječe li još neki faktor na potrošnju goriva. Odgovor je da ih utječe puno, mi ćemo na primjer uzeti još jedan faktor i to faktor iskustva koji također ima više razina. Sada se nameće pitanje o postojanju ili nepostojanju značajnog utjecaja na potrošnju goriva jednog i drugog faktora, te o eventualnom postojanju međusobne interakcije između ta dva faktora.<sup>2</sup>

<sup>2</sup>Motivacijski primjer i način obrade preuzeti iz [2]

Kategorizirajmo vozače u 5 grupa:

1. grupa - početnici s vozačkim iskustvom manjim od 2 godine
2. grupa - vozači s iskustvom od 2 do 5 godina
3. grupa - vozači s iskustvom od 5 do 10 godina
4. grupa - vozači s iskustvom od 10 do 20 godina
5. grupa - vozači s iskustvom većim od 20 godina

Sada iz svake grupe vozača slučajno biramo po devet vozača i trojica voze automobil marke A, trojica marke B i trojica marke C pri čemu se mjeri odgovarajuća potrošnja goriva (broj litara na 100 kilometara). Dakle ovaj model karakteriziraju dvije nezavisne varijable (faktori) diskretnog tipa: prvi faktor sa  $m_1$  razinama, drugi faktor sa  $m_2$  razinama. Nivoi faktora definiraju tretmane, kombinacije nivoa faktora definiraju  $m_1 m_2$  tretmana; te jedna zavisna varijabla  $Y$  kontinuiranog tipa.

Prepostavimo općenito da prvi faktor kao gore ima  $m_1 \geq 2$ , a drugi faktor ima  $m_2 \geq 2$  razina dok je svaki uzorak duljine  $l$ . Ukupno imamo  $m_1 m_2 l$  podataka  $y_{ijk}$  što možemo zapisati u obliku  $Y_{ijk} = \mu_{ij} + E_{ijk}$ , gdje je  $\mu_{ij}$  neslučajna veličina koja karakterizira djelovanje  $i$ -te razine I. faktora i  $j$ -te razine II. faktora, dok je  $E_{ijk}$  slučajna varijabla koja karakterizira grešku  $k$ -tog mjerjenja. Matematički opis dvofaktorskog modela analize varijance općenito je izražen jednadžbom

$$Y_{ijk} = \mu_{ij} + E_{ijk}, \quad i = 1, \dots, m_1 \quad j = 1, \dots, m_2, \quad k = 1, \dots, l, \quad (2.23)$$

gdje su  $E_{ijk}$  nezavisne slučajne varijable sa zajedničkom normalnom razdiobom  $N(0, \sigma^2)$ . Smatra se da je  $y_{ijk}$  rezultat djelovanja  $i$ -te razine I. faktora,  $j$ -te razine II. faktora, međusobne interakcije obaju faktora i slučajne greške.

Uvedimo oznake

$$\mu = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mu_{ij} \quad (2.24)$$

$$\mu'_i = \frac{1}{m_2} \sum_{j=1}^{m_2} \mu_{ij}, \quad \delta'_i = \mu'_i - \mu, \quad i = 1, \dots, m_1, \quad (2.25)$$

$$\mu''_i = \frac{1}{m_1} \sum_{i=1}^{m_1} \mu_{ij}, \quad \delta''_j = \mu''_j - \mu, \quad j = 1, \dots, m_2, \quad (2.26)$$

gdje je  $\mu$  opća srednja vrijednost svih opservacija,  $\mu'_i$  srednja vrijednost za fiksiranu  $i$ -tu razinu I. faktora, a  $\mu''_j$  srednja vrijednost za fiksiranu  $j$ -tu razinu II. faktora. Zato se  $\delta'_i$  zove glavni efekt  $i$ -te razine I. faktora, a  $\delta''_j$  glavni efekt  $j$ -te razine II. faktora.

Uvede li se zapis

$$\mu_{ij} = \mu + \delta'_i + \delta''_j + \delta_{ij} \quad (2.27)$$

može se reći da je očekivana vrijednost rastavljena na zbroj opće srednje vrijednosti, dva glavna efekta, dok je  $\delta_{ij}$  doprinos međusobne interakcije  $i$ -te razine I. faktora i  $j$ -te razine II. faktora odnosno interakcijski efekt. Stoga se kao nul hipoteze prirodno nameću sljedeće tri hipoteze

$$H_{01} : \delta'_1 = \delta'_2 = \dots = \delta'_{m_1} = 0, \quad (2.28)$$

$$H_{02} : \delta''_1 = \delta''_2 = \dots = \delta''_{m_2} = 0, \quad (2.29)$$

$$H_{12} : \delta_{ij} = 0, \quad \text{za } i = 1, \dots, m_1, \quad j = 1, \dots, m_2, \quad (2.30)$$

pri čemu se za svaku od njih, kao alternativna hipoteza, uzima da bar na jednom mjestu ne vrijedi znak jednakosti.

Prve dvije hipoteze odgovaraju na pitanje je li I. faktor značajan odnosno je li II. faktor značajan, dok se testiranje zadnje hipoteze želi dobiti odgovor na pitanje postoji li interakcija između I. i II. faktora, koja uzrokuje značajne promjene na izlaznim podacima.

Glavni je problem i u ovome modelu da se definiraju prikladne test statistike za testiranje hipoteza (2.28), (2.29) i (2.30). Kao i u jednofaktorskome modelu ukupno rasipanje podataka (varijabilitet) ćemo razdvojiti na više komponenti. U tu svrhu uvodimo sljedeće statistike

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^l Y_{ijk} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad n = m_1 m_2 l, \quad (2.31)$$

$$\bar{Y}'_i = \frac{1}{m_2 l} \sum_{j=1}^{m_2} \sum_{k=1}^l Y_{ijk} \sim N\left(\mu + \delta'_i, \frac{\sigma^2}{m_2 l}\right), \quad i = 1, \dots, m_1, \quad (2.32)$$

$$\bar{Y}''_j = \frac{1}{m_1 l} \sum_{i=1}^{m_1} \sum_{k=1}^l Y_{ijk} \sim N\left(\mu + \delta''_j, \frac{\sigma^2}{m_1 l}\right), \quad j = 1, \dots, m_2, \quad (2.33)$$

$$\bar{Y}_{ij} = \frac{1}{l} \sum_{k=1}^l Y_{ijk} \sim N\left(\mu + \delta'_i + \delta''_j + \delta_{ij}, \frac{\sigma^2}{l}\right), \quad i = 1, \dots, m_1, \quad j = 1, \dots, m_2 \quad (2.34)$$

$$SST_1 = m_2 l \sum_{i=1}^{m_1} (\bar{Y}'_i - \bar{Y})^2, \quad (2.35)$$

$$SST_2 = m_1 l \sum_{j=1}^{m_2} (\bar{Y}_j'' - \bar{Y})^2, \quad (2.36)$$

$$SST_{12} = l \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (\bar{Y}_{ij} - \bar{Y}'_i - \bar{Y}''_j + \bar{Y})^2, \quad (2.37)$$

$$SSE = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^l (Y_{ijk} - \bar{Y}_{ij})^2, \quad (2.38)$$

$$SS = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^l (Y_{ijk} - \bar{Y})^2. \quad (2.39)$$

Navedene statistike imaju sljedeću interpretaciju:

$\bar{Y}$  - aritmetička sredina svih mjerena

$\bar{Y}'_i$  - aritmetička sredina svih mjerena  $i$ -toga retka I. faktora

$\bar{Y}''_j$  - aritmetička sredina svih mjerena  $j$ -toga stupca II. faktora

$\bar{Y}_{ij}$  - aritmetička sredina svih mjerena iz  $(i, j)$ -toga polja

$SST_1$  - suma kvadrata odstupanja sredina redaka od zajedničke sredine

$SST_2$  - suma kvadrata odstupanja sredina stupaca od zajedničke sredine

$SST_{12}$  - interakcijska suma kvadrata

$SSE$  - suma kvadrata odstupanja mjerena od odgovarajućih sredina u polju

$SS$  - suma kvadrata odstupanja svih mjerena od njihove aritmetičke sredine

Sada vidimo da smo ukupan izvor varijabilnosti rastavili na sljedeći način:

$$SS = SST_1 + SST_2 + SST_{12} + SSE \quad (2.40)$$

slično kao i u jednofaktorskome modelu samo sada imamo više izvora varijabilnosti. Također pokazuje se prema [2] da su

$$MST_1 = \frac{1}{m_1 - 1} SST_1 \quad (2.41)$$

$$MST_2 = \frac{1}{m_2} SST_2 \quad (2.42)$$

$$MST_{12} = \frac{1}{(m_1 - 1)(m_2 - 1)} SST_{12} \quad (2.43)$$

nepristrani procjenitelji za nepoznati parametar  $\sigma^2$  samo ako su redom hipoteze (2.28), (2.29) i (2.30) istinite. Dok je

$$MSE = \frac{1}{m_1 m_2 (l - 1)} SSE \quad (2.44)$$

nepristrani procjenitelj za nepoznati parametar  $\sigma^2$  bez obzira na gore navedene hipoteze.

Ako su hipoteze (2.28), (2.29) i (2.30) doista neistinete, onda se mogu očekivati veće vrijednosti statistika  $SST_1$ ,  $SST_2$  i  $SST_{12}$  nego u slučaju stvarne istinitosti navedenih hipoteza. Stoga je kao i u jednofaktorskome modelu analize varijance kao test statistiku uzeti omjer odgovarajućih varijanci odnosno  $F$  test sa odgovarajućim stupnjevima slobode.

Za hipotezu da utjecaj I. faktora nije značajan,  $H_{01}$  test statistika je

$$F_1 = \frac{MST_1}{MSE} \sim F(m_1 - 1, m_1 m_2 (l - 1)). \quad (2.45)$$

Za hipotezu da utjecaj II. faktora nije značajan,  $H_{02}$  test statistika je

$$F_2 = \frac{MST_2}{MSE} \sim F(m_2 - 1, m_1 m_2 (l - 1)). \quad (2.46)$$

Za hipotezu da utjecaj interakcije nije značajan,  $H_{12}$  test statistika je

$$F_{12} = \frac{MST_{12}}{MSE} \sim F((m_1 - 1)(m_2 - 1), m_1 m_2 (l - 1)), \quad (2.47)$$

Dobije li se vrijednost navedenih test statistika mnogo veća od jedinice, to će nas uputiti na odbacivanje nultih hipoteza. Slično kao i u jednofaktorskome modelu navedene statistike se prikazuju u ANOVA tablici za dvofaktorski model.

| Izvor variabilnosti | Suma kvadrata | Stupnjevi slobode    | Varijanca  | Test statistika $F$             |
|---------------------|---------------|----------------------|------------|---------------------------------|
| I. faktor           | $SST_1$       | $m_1 - 1$            | $MST_1$    | $F_1 = \frac{MST_1}{MSE}$       |
| II. faktor          | $SST_2$       | $m_2 - 1$            | $MST_2$    | $F_2 = \frac{MST_2}{MSE}$       |
| Interakcija         | $SST_{12}$    | $(m_1 - 1)(m_2 - 1)$ | $MST_{12}$ | $F_{12} = \frac{MST_{12}}{MSE}$ |
| Greška              | $SSE$         | $m_1 m_2 (l - 1)$    | $MSE$      |                                 |
| Ukupno              | $SS$          | $m_1 m_2 l - 1$      |            |                                 |

Tablica 2.2: ANOVA tablica - dvofaktorski model

U tablici 2.3 nismo stavili pripadne  $p$  vrijednosti za svaku od hipoteza koje se dobiju analogno kao u jednofaktorskome modelu kao vjerojatnost da test statistika upadne u kritično područje uz istinitost nulte hipoteze.

## Ukupna značajnost modela

Osim testiranja hipoteza koje smo gore obradili ostaje nam pitanje značajnosti cjelokupnog modela odnosno sljedeće hipoteze:

$$H_{00} = \mu_{11} = \mu_{12} = \dots = \mu_{1m_2} = \mu_{21} = \dots = \mu_{2m_2} = \dots = \mu_{m_11} = \dots = \mu_{m_1m_2}, \quad (2.48)$$

kojom testiramo da su sve sredine po svakom faktoru jednake. Kada malo razmislimo to je i ono što nas najviše zanima. Prema gornjem primjeru sa markama automobila i duljinom iskustva zanima nas je li srednja potrošnja goriva jednaka bez obzira koju marku automobila vozimo i koliko imamo iskustva. Upravo to testira gornja hipoteza. Ukoliko se ona pokaže da nije značajna odnosno da razlika među sredinama nije značajna tada niti ne trebamo testirati druge hipoteze. Ukoliko je značajna testiramo kako bismo vidjeli gdje je ta varijabilnost odnosno koji faktor čini tu razliku značajnom. Suma kvadrata je jednaka

$$SSM = SST_1 + SST_2 + SST_{12} \quad (2.49)$$

odnosno zbroju sume kvadrata faktora I., II. i njihove interakcije. Korigirana varijanca je jednaka

$$MSM = \frac{1}{m_1 m_2 - 1} SSM. \quad (2.50)$$

Tada je test statistika kao i prije jednaka omjeru odgovarajućih varijanci te uz istinitost  $H_{00}$  ima distribuciju

$$F_m = \frac{MSM}{MSE} \sim F(m_1 m_2 - 1, m_1 m_2 l - 1). \quad (2.51)$$

Ovime smo obradili univarijantnu jednofaktorsku i dvofaktorsku analizu varijance sa fiksnim efektima te ćemo sada obraditi na primjeru s dva faktora.

## 2.4 Primjer - dvofaktorska analiza varijance

### Zadatak

Mjeren je rast pet vrsta trave tijekom četiri tjedna pod utjecajem tri metoda klijanja sjemenja<sup>3</sup>. Za svaku kombinaciju trave i vrste zasađeno je šest lonaca. Lonci su slučajno raspoređeni u prostoriji za rast. Nakon četiri tjedna trava je ošišana i za svaki je lonac izmjerena i zapisana količina dobivene trave. Pitanje je postoji li razlika u količini dobivene

---

<sup>3</sup>Podaci preuzeti s tečaja STAT3 u Srcu ak. godine 2013/14, voditelj tečaja: mr. sc. Vesna Hljuz Dobrić

trave u odnosu na metodu i vrstu trave? Dodatna pitanja su nam: je li promjena prinosa trave po vrsti trave različita za razne metode klijanja i je li promjena prinosa trave po metodama klijanja različita za razne vrste trave? Dakle imamo 2 faktora - vrstu trave i metodu klijanja. Za svaku kombinaciju imamo šest lonaca pa je to ukupno devedeset podataka. Dio podataka je dan u tablici na slici 2.1, a deskriptiva na slici 2.2. Podatke ćemo pripremiti za analizu te ćemo koristiti proceduru Anova. Varijabla *metoda* označava metodu klijanja a *vrsta* vrstu trave. Varijable *t1* do *t6* označavaju lonce trave za svaku metodu i vrstu. Metoda\*vrsta označava interakciju.

| Obs | Metoda | Vrsta | t1   | t2   | t3   | t4   | t5   | t6   | trt |
|-----|--------|-------|------|------|------|------|------|------|-----|
| 1   | A      |       | 22.1 | 24.1 | 19.1 | 22.1 | 25.1 | 18.1 | A1  |
| 2   | A      |       | 27.1 | 15.1 | 20.6 | 28.6 | 15.1 | 24.6 | A2  |
| 3   | A      |       | 22.3 | 25.8 | 22.8 | 28.3 | 21.3 | 18.3 | A3  |
| 4   | A      |       | 19.8 | 28.3 | 26.8 | 27.3 | 26.8 | 26.8 | A4  |
| 5   | A      |       | 20.0 | 17.0 | 24.0 | 22.5 | 28.0 | 22.5 | A5  |
| 6   | B      |       | 13.5 | 14.5 | 11.5 | 6.0  | 27.0 | 18.0 | B1  |
| 7   | B      |       | 16.9 | 17.4 | 10.4 | 19.4 | 11.9 | 15.4 | B2  |
| 8   | B      |       | 15.7 | 10.2 | 16.7 | 19.7 | 18.2 | 12.2 | B3  |
| 9   | B      |       | 15.1 | 6.5  | 17.1 | 7.6  | 13.6 | 21.1 | B4  |
| 10  | B      |       | 21.8 | 22.8 | 18.8 | 21.3 | 16.3 | 14.3 | B5  |
| 11  | C      |       | 19.0 | 22.0 | 20.0 | 14.5 | 19.0 | 16.0 | C1  |
|     |        |       | ...  | ...  |      |      |      |      |     |

Slika 2.1: Tablica djela nepripremljenih podataka - Ispis iz SAS-a

## Kod

```
PROC ANOVA data=trava plots=all;
class metoda vrsta;
model trava = metoda vrsta metoda*vrsta;
means metoda vrsta metoda*vrsta;
run;
quit;
```

| Level of Metoda |    | N          | Trava      |         | Trava      |            |
|-----------------|----|------------|------------|---------|------------|------------|
|                 |    |            | Mean       | Std Dev | Mean       | Std Dev    |
| A               | 30 | 23.0100000 | 3.98638632 |         | 21.7666667 | 2.73252020 |
| B               | 30 | 15.6966667 | 4.89344976 |         | 21.8500000 | 5.88854821 |
| C               | 30 | 16.6066667 | 4.92830903 |         | 23.1333333 | 3.50238014 |

  

| Level of Vrsta |    | N          | Trava      |         | Trava      |            |
|----------------|----|------------|------------|---------|------------|------------|
|                |    |            | Mean       | Std Dev | Mean       | Std Dev    |
| 1              | 18 | 18.4222222 | 5.18624375 |         | 15.0833333 | 7.05277723 |
| 2              | 18 | 19.0000000 | 5.01949142 |         | 15.2333333 | 3.44480285 |
| 3              | 18 | 18.6333333 | 4.95046047 |         | 15.4500000 | 3.61593695 |
| 4              | 18 | 18.1000000 | 7.39856902 |         | 13.5000000 | 5.60535458 |
| 5              | 18 | 18.0333333 | 5.78364835 |         | 19.2166667 | 3.36773910 |

  

| Level of Metoda |   | N | Trava      |            | Trava      |            |
|-----------------|---|---|------------|------------|------------|------------|
|                 |   |   | Mean       | Std Dev    | Mean       | Std Dev    |
| A               | 1 | 6 | 21.7666667 | 2.73252020 | 25.9666667 | 3.07679487 |
| A               | 2 | 6 | 21.8500000 | 5.88854821 | 22.3333333 | 3.71034590 |
| A               | 3 | 6 | 23.1333333 | 3.50238014 | 18.4166667 | 2.72794184 |
| A               | 4 | 6 | 25.9666667 | 3.07679487 | 19.9166667 | 3.36773910 |
| A               | 5 | 6 | 22.3333333 | 3.71034590 | 17.3166667 | 4.40927054 |
| B               | 1 | 6 | 15.0833333 | 7.05277723 | 14.8333333 | 5.72421756 |
| B               | 2 | 6 | 15.2333333 | 3.44480285 | 12.5500000 | 5.35490429 |
| B               | 3 | 6 | 15.4500000 | 3.61593695 |            |            |
| B               | 4 | 6 | 13.5000000 | 5.60535458 |            |            |
| B               | 5 | 6 | 19.2166667 | 3.36773910 |            |            |
| C               | 1 | 6 | 21.8500000 | 5.88854821 |            |            |
| C               | 2 | 6 | 23.1333333 | 3.50238014 |            |            |
| C               | 3 | 6 | 25.9666667 | 3.07679487 |            |            |
| C               | 4 | 6 | 22.3333333 | 3.71034590 |            |            |
| C               | 5 | 6 | 18.4166667 | 2.72794184 |            |            |

Slika 2.2: Deskriptivna statistika - Ispis iz SAS-a

## Interpretacija

Što se tiče koda proceduru koju smo koristili je *anova* što možemo jer imamo balansirani dizajn. Sa *data* definiramo *data set* na kojem radimo analizu a sa *plots=all* specificiramo da hoćemo sve grafove koje daje procedura. Naredbom *class* kažemo da imamo dva faktora a *model* kako nam izgleda model koji testiramo. Naredba *means* je dodatna koja nam ispisuje aritmetičke sredine za svaki faktor te interakciju što je u tablicama na slici 2.2.

Prva tablica na slici 2.3 nam testira sveukupnu značajnost modela te kako nam je *p* vrijednost manja od 0,0001 zaključujemo kako nam je model značajan. Sljedeće tablica nam daje deskriptivne statistike (R kvadrat, koeficijent varijacije, standardnu devijaciju i aritmetičku sredinu svih izmjerjenih vrijednosti). Zadnja tablica nam je naša Anova tablica (2.3) s time da je greška modela i *SS* dan u prvoj tablici. Iz tablice zaključujemo sljedeće: Postoji značajna razlika u prinosu trave po metodama kljanja za razne vrste trave (*p* vrijednost manja od 0,0001), ne postoji značajna razlika u prinosu trave po vrstama trave za razne metode (*p* vrijednost je 0,96) i postoji značajna razlika u količini dobivene trave u odnosu na metodu i vrstu trave (*p* vrijednost je 0,024) što znači da se vrsta trave ne ponaša približno isto po metodama.

## Ispis

| The ANOVA Procedure       |    |                |             |         |        |  |
|---------------------------|----|----------------|-------------|---------|--------|--|
| Dependent Variable: Trava |    |                |             |         |        |  |
| Source                    | DF | Sum of Squares | Mean Square | F Value | Pr > F |  |
| Model                     | 14 | 1339.024889    | 95.644635   | 4.87    | <.0001 |  |
| Error                     | 75 | 1473.766667    | 19.650222   |         |        |  |
| Corrected Total           | 89 | 2812.791556    |             |         |        |  |

  

| R-Square | Coeff Var | Root MSE | Trava Mean |  |
|----------|-----------|----------|------------|--|
| 0.476048 | 24.04225  | 4.432857 | 18.43778   |  |

  

| Source       | DF | Anova SS    | Mean Square | F Value | Pr > F |  |
|--------------|----|-------------|-------------|---------|--------|--|
| Metoda       | 2  | 953.1562222 | 476.5781111 | 24.25   | <.0001 |  |
| Vrsta        | 4  | 11.3804444  | 2.8451111   | 0.14    | 0.9648 |  |
| Metoda*Vrsta | 8  | 374.4882222 | 46.8110278  | 2.38    | 0.0241 |  |

Slika 2.3: Tablice rezultata - Ispis iz SAS-a

# Poglavlje 3

## Analiza varijance ponovljenih mjerena

### 3.1 Uvod

Dizajni analize varijance koje ćemo razmatrati u ovome poglavlju specijalni su slučajevi slučajnih potpunih blok dizajna (eng. *randomized complete-block*[4]) u kojima se svaki subjekt smatra blokom na koji su primjenjeni svi tretmani. Kada konstruiramo blokove, greška (varijabilnost) se smanjuje kada su subjekti više homogeni. Kako radimo ponovljena mjerena na istim subjektima ne možemo imati više homogeniji blok. Ovakav dizajn ima tri velike prednosti: imamo više podataka o svakom subjektu, grešku možemo dalje objasniti te trebamo manje subjekata za istu snagu testa i ovaj dizajn pokazuje se ekonomičan kada je teško doći do subjekta jer na svakom subjektu treba primjeniti sve tretmane. Najčešće primjene ovakvih dizajna su u psihologiji, farmaceutskoj industriji i agronomiji.

### Motivacija

Prije nego neki lijek izađe na tržište odnosno dobije odobrenje da se može prodavati mora proći nekoliko faza. Jedna od tih faza je i testiranje lijeka na ljudima. Recimo da se radi o lijeku za povećavanje apetita te želimo vidjeti koliko je i je li uopće efikasan. Iz nema zanimljive populacije izabrali bi na slučajan način  $n$  ljudi koji su nam voljni pomoći u eksperimentu. Nema smisla da im damo samo jedanput lijek i onda za tjedana dana izmjerimo njihovu tjelesnu masu jer tjedan dana je premali vremenski period, a postoje i mnogi drugi faktori koji mogu utjecati na povećanje/smanjenje tjelesne mase. Očito moraju uzimati lijek jedan duži vremenski period te im moramo redovito mjeriti njihovu tjelesnu masu. Sama grafička usporedba težine na početku eksperimenta i na kraju eksperimenta nam može ukazati hoćemo li reći da je lijek efikasan ili ne. Za malo precizniji odgovor potrebna nam je analiza varijance ponovljenih mjerena (eng. *whitin-subject design*). U

ovome primjeru zavisna varijabla nam je vrijeme. Osim vremena imamo i nezavisne varijable koje nam mogu biti različiti tretmani koje opet primjenjujemo nad istim subjektima. Npr. zanima nas utjecaj različitih lijekova za smirenje na obavljanje određenih fizičkih radnji. Svakom subjektu (čovjeku) ćemo dati sve tretmane (lijekove) te ćemo nakon toga mjeriti koliko mu treba vremena da obavi zadane radnje. Zavisna varijabla nam je u ovom slučaju različiti nivoi tretmana odnosno različiti lijekovi.

## Zavisnost

U modelima koje smo obradili u 2. poglavlju koristili smo pretpostavku nezavisnosti unutar svake grupe i između grupa. Nezavisnost unutar svake grupe, bilo da imamo jedan ili više faktora nam kaže da vrijednost jedne opservacije nema utjecaja na vrijednost druge opservacije unutar te iste grupe. U ponovljenim mjerenjima također imamo pretpostavku nezavisnosti unutar jednog ponavljanja, te pretpostavku nezavisnosti između ljudi dok je očito da će postojati korelacija unutar bloka odnosno da će različita mjerena nad istim subjektom biti povezana. Kako prepostavljamo da je distribucija multivarijatna normalna koreliranost nam je isto što i zavisnost. Prema tome opservacije unutar svakog bloka (čovjeka) će biti zavisne pa ćemo tu varijabilnost moći bolje objasniti.

## 3.2 Jednofaktorski model

### Aditivni model

Prepostavimo da imamo  $k$  tretmana i  $n$  osoba. U tablici 3.2 imamo sljedeći zapis odnosno notaciju.  $Y_{11}$  nam predstavlja mjerena nad prvom osobom pod tretmanom jedan,  $Y_{12}$  mjereno za prvu osobu pod tretmanom dva,  $Y_{1j}$  mjereno za prvu osobu pod tretmanom  $j$ . Generalno prvi indeks označuje osobu, dok drugi indeks označuje tretman (ili vrijeme) pod kojim je mjereno izvršeno.<sup>1</sup>

Simbol  $P_1$  reprezentira zbroj svih tretmana nad osobom jedan,  $P_2$  zbroj svih tretmana nad osobom 2,  $P_i$  zbroj svih tretmana nad osobom  $i$  tj.

$$P_i = \sum_{j=1}^k Y_{ij}. \quad (3.1)$$

Stoga je aritmetička sredina svih opservacija na osobi  $i$  dana sa

$$\bar{P}_i = \frac{P_i}{k}. \quad (3.2)$$

---

<sup>1</sup>notacija djelomično preuzeta iz [4]

Analogno po recima možemo gledati zbroj i aritmetičku sredinu po stupcima. Sa  $T_1$  označit ćemo zbroj svih  $n$  opservacija pod prvim tretmanom, sa  $T_2$  zbroj svih  $n$  opservacija pod drugim tretmanom, sa  $T_j$  zbroj svih  $n$  opservacija pod tretmanom  $j$  tj.

$$T_j = \sum_{i=1}^n Y_{ij}. \quad (3.3)$$

Stoga je aritmetička sredina svih opservacija u tretmanu  $j$  dana sa

$$\bar{T}_j = \frac{T_j}{n}. \quad (3.4)$$

| Osoba \ Tretman | 1           | 2           | ... | $j$         | ... | $k$         | Zbroj    | Sredina     |
|-----------------|-------------|-------------|-----|-------------|-----|-------------|----------|-------------|
| 1               | $Y_{11}$    | $Y_{12}$    |     | $Y_{1j}$    |     | $Y_{1k}$    | $P_1$    | $\bar{P}_1$ |
| 2               | $Y_{21}$    | $Y_{22}$    |     | $Y_{2j}$    |     | $Y_{2k}$    | $P_2$    | $\bar{P}_2$ |
| $\vdots$        | $\vdots$    | $\vdots$    |     | $\vdots$    |     | $\vdots$    | $\vdots$ | $\vdots$    |
| i               | $Y_{i1}$    | $Y_{i2}$    |     | $Y_{ij}$    |     | $Y_{ik}$    | $P_i$    | $\bar{P}_i$ |
| $\vdots$        | $\vdots$    | $\vdots$    |     | $\vdots$    |     | $\vdots$    | $\vdots$ | $\vdots$    |
| $n$             | $Y_{n1}$    | $Y_{n2}$    |     | $Y_{nj}$    |     | $Y_{nk}$    | $P_n$    | $\bar{P}_n$ |
| Zbroj           | $T_1$       | $T_2$       | ... | $T_j$       | ... | $T_k$       | $G$      |             |
| Sredina         | $\bar{T}_1$ | $\bar{T}_2$ | ... | $\bar{T}_j$ | ... | $\bar{T}_k$ |          | $\bar{G}$   |

Tablica 3.1: Notacija

Zbroj svih opservacija ćemo označiti sa  $G$  i on je jednak

$$G = \sum_{i=1}^n P_i = \sum_{j=1}^k T_j = \sum_{i=1}^n \sum_{j=1}^k Y_{ij}, \quad (3.5)$$

dok ćemo sa  $\bar{G}$  označiti aritmetičku sredinu svih opservacija

$$\bar{G} = \frac{G}{kn} = \frac{\sum_{i=1}^n P_i}{n} = \frac{\sum_{j=1}^k T_j}{k}. \quad (3.6)$$

Ukupna varijabilnost u dizajnu je suma kvadrata odstupanja svih opservacija od aritmetičke sredine  $\bar{G}$

$$SS = \sum_{i=1}^n \sum_{j=1}^k (Y_{ij} - \bar{G})^2. \quad (3.7)$$

Stupnjevi slobode za  $SS$  su  $kn - 1$ . Sada tu varijabilnost kao i u Poglavlju 2 idemo rastaviti na izvore varijabilnosti. Postoje dva izvora varijabilnosti. Jedan izvor vrijabilnosti je varijabilnost između ljudi odnosno varijabilnost između subjekata (eng. *between-subject*). Da povučemo paralelu sa Poglavljem 2 u tablici 2.2 jednofaktorskog modela to je  $SST$ . Ovdje ćemo tu varijabilnost označiti sa  $SSB$  te ona glasi:

$$SSB = k \sum_{i=1}^n (\bar{P}_i - \bar{G})^2. \quad (3.8)$$

Ovaj izvor varijabilnosti dolazi od toga da su ljudi odnosno promatrani subjekti različiti. Kako imamo  $n$  aritmetičkih sredina, ovaj izvor varijabilnosti ima  $n - 1$  stupnjeva slobode. Drugi izvor varijabilnosti jest varijabilnost unutar subjekta odnosno ljudi (eng. *within-subject*) i definira se na sljedeći način: za svaki subjekt  $i$  možemo izračunati njegovu kvadratnu udaljenost od njegove aritmetičke sredine  $P_i$ . Suma svih odstupanja za sve subjekte jest naša preostala varijabilnost. U 2.2 to nam je  $SSE$ . Ovdje ćemo to označiti sa  $SSW$ . Zapisano formulom to je

$$SSW = \sum_{i=1}^n \sum_{j=1}^k (Y_{ij} - \bar{P}_i)^2. \quad (3.9)$$

Stupnjevi slobode za varijabilnost unutar pojedinog subjekata su  $k - 1$  pa su stupnjevi slobode za zbroj varijabilnosti unutar subjekta ( $SSW$ )  $n(k - 1)$ . Sada smo ukupnu varijabilnost rastavili na 2 dijela ( $SS = SSB + SSE$ ) sa pripadnim stupnjevima slobode koji također odgovaraju ( $kn - 1 = n - 1 + n(k - 1)$ ). Kao što vidimo naš  $SSW$  nam je jednak  $SSE$  u modelu univarijatne analize varijance s jednim faktorom. Razlika u odnosu na taj model je što  $SSW$  možemo dalje rastavljati. Još nismo iskoristili varijabilnost zbog različitog tretmana na istim subjektima odnosno efekt ponavljanja. Označimo tu varijabilnost sa  $SST_b$  te je ona dana formulom

$$SST_b = n \sum_{j=1}^k (\bar{T}_j - \bar{G})^2, \quad (3.10)$$

sa  $k - 1$  stupnjem slobode. Označimo sada sa  $SSE_b$  preostalu, neobjasnjenu varijabilnost te je ona dana sa formulom

$$SSE_b = SSW - SST_b = \sum_{i=1}^n \sum_{j=1}^k ((Y_{ij} - \bar{G}) - (\bar{P}_i - \bar{G}) - (\bar{T}_j - \bar{G}))^2. \quad (3.11)$$

Broj stupnjeva slobode za  $SSE_b$  se dobije kada od broja stupnjeva slobode za  $SSW$  oduzmemmo stupnjeve slobode za  $SST_b$ . Odnosno  $n(k - 1) - (k - 1) = (k - 1)(n - 1)$ . Prema

tome stupnjevi slobode za neobjašnjenu varijabilnost su  $(k - 1)(n - 1)$ . Sada smo dakle rastavili ukupno varijabilnost na tri dijela. Dakle  $SS = SSB + SSW = SSB + SST_b + SSE$ . Kao i u poglavlju 2. varijance dobijemo kada sume kvadrata podijelimo sa pripadajućim stupnjevima slobode. Nulta hipoteza koju u ovom modelu testiramo je da razlike među tretmanima (ili vremenu) nije značajna nasuprot alternativnoj hipotezi da se barem dvije sredine značajno razlikuju. Priklada statistika za tu hipotezu nam je kao i u poglavlju 2,  $F$  statistika koja u brojniku ima varijancu  $MST_b$ , a u nazivniku varijancu  $MSE$ .

Sve gore navedeno možemo prikazati u tablici 3.2.

| Izvor varijabilnosti | Suma kvadrata | Stupnjevi slobode | Varijanca                   | Test statistika $F$     |
|----------------------|---------------|-------------------|-----------------------------|-------------------------|
| Između subjekata     | $SSB$         | $n - 1$           | $MSB = \frac{SSB}{n-1}$     | $F = \frac{MSB}{MSE}$   |
| Unutar subjekata     | $SSW$         | $n(k - 1)$        | $MSE = \frac{SSW}{n(k-1)}$  |                         |
| Efekt tretmana       | $SST_b$       | $k - 1$           | $MST_b = \frac{SST_b}{k-1}$ | $F = \frac{MST_b}{MSE}$ |
| Greška               | $SSE_b$       | $(n - 1)(k - 1)$  | $MSE$                       |                         |
| Ukupno               | $SS$          | $nk - 1$          |                             |                         |

Tablica 3.2: ANOVA tablica - jednofaktorski model s ponavljanim mjeranjima

## Model i prepostavke

Matematički model iz prethodnog poglavlja možemo zapisati kao i u jednofaktorskom modelu analize varijance bez ponavljanja (2.6) na sljedeći način:

$$X_{ij} = \mu + \pi_i + \delta_j + \epsilon_{ij}. \quad (3.12)$$

Kao i u (2.6)  $\mu$  je ukupna aritmetička sredina,  $\delta_j$  efekt  $j$ -tog tretmana, a  $\epsilon_{ij}$  greška koja je normalno distribuirana te nezavisna. Jedino što je novo je  $\pi_i$  koja je vezana uz osobu  $i$  i koja je po prepostavci normalno distribuirana. Tu dodatnu varijablu dobijemo jer imamo više mjerena nad istim subjektom.

Prepostavke koje testiramo su sljedeće: da razlike u tretmanima nisu značajne, da razlike između subjekata nisu značajne i sveukupnu značajnost modela. Sve  $F$  statistike se dobiju analogno kao u Poglavlju 2.

Neke smo prepostavke već spomenuli u uvodu te kroz obradu modela a sada ćemo ih nabrojati te neke objasniti.

### Univarijatne pretpostavke

- Normalnost - opservacije unutar svakog nivoa tretmana su normalno distribuirane
- Sferičnost (eng. *sphericity*) - jednakost varijanci
- Nezavisnost - opservacije po subjektima su nezavisne jedan od druge, subjekti su slučajni

### Multivarijatne pretpostavke

- Normalnost - Nivoi tretmana su imaju multivarijatnu normalnu distribuciju
- Nezavisnost - razlike među sredinama po nivoima tretmana su nezavisne jedna od druge

### Sferičnost

Neka je  $\Sigma_y$  kovarijacijska matrica modela. Uvjet cirkularnosti za  $\Sigma_y$  glasi:

$$\sigma_{jj} - \sigma_{j'j'} - 2\sigma_{jj'} = 2\lambda, \quad \text{za sve } j \neq j' \quad (3.13)$$

za neki  $\lambda > 0$ .

Huynh and Feldt (1970)[1] su pokazali da ukoliko vrijedi sljedeći uvjet

$$\sigma_{Y_j - Y_{j'}}^2 = 2\lambda, \quad \text{za sve } j \neq j', \quad (3.14)$$

tada je opravdano koristiti uobičajenu analizu varijance ponovljenih mjerena. Ovaj uvjet se još zove Huynh-Feldtov uvjet. Zapravo nam kaže da je razlika varijanca između svaka dva  $Y$ -a konstantna.

Uvjet cirkularnosti povlači sferičnost matrice  $\Sigma_x$ , koju definiramo kao

$$\Sigma_x = M^* \Sigma_y M^{*\prime}, \quad (3.15)$$

gdje je su varijable  $X$  normirane ortogonalne transformacije originalnih varijabli  $Y$ . Matrica  $M^*$  je ortonormirana matrica reda  $(k-1) \times k$  čiji su redovi normirani i međusobno ortogonalni.

Dakle da bismo provjerili uvjet (3.13) moramo provjeriti uvjet sferičnosti. U SAS-u koristimo Mauchly-jev (1940)[1] test za sferičnost. Kada Huynh-Feldtov uvjet nije zadovoljen uzimamo prilagođenu  $p$  vrijednost tako da odgovarajuće stupnjeve slobode množimo sa skalarom  $\epsilon$ , koji je funkcija elemenata u kovarijacijskoj matrici ponovljenih mjerena. Načine izračuna  $\epsilon$ -a dali su Huynh i Feldt (1976) te Greenhouse i Geisser (1959). Huynh-feldt-ova prilagodba stupnjeva slobode je malo konzervativnija nego Greenhouse-ova i Geisser-ova.

### 3.3 Primjer - jednofaktorski model

#### Zadatak

Imamo 5 subjekata koji su u ovom slučaju ljudi. Svakome subjektu su dana 4 različita lijeka u odgovarajućim vremenskim razmacima. Nakon što im je dan lijek mjereno je koliko je vremena potrebno da se obavi niz standardiziranih fizičkih radnji koje su im prije objašnjene. Podaci se mogu vidjeti na slici 3.1. Varijabla *Osoba* označava osobu, a varijabla *Lijek* koji je lijek primila. U varijabli *Time* je vrijeme potrebno da se izvede zadana fizička radnja. Tako npr. deseta opservacija nam kaže da je osobi 3 kada je dobila lijek 2 trebalo 20 sekundi da izvrši zadano. Imamo 5 osoba i 4 lijeka. Ukupno 20 podataka. Dizajn je balansiran.

#### Kod

```

DATA anova_1;
    Osoba+1;
    DO Lijek =1  to  4;
        INPUT Time @;
    OUTPUT;
END;
    DATALINES;
30 28 16 34
14 18 10 22
24 20 18 30
38 34 20 44
26 28 14 30
;

proc print data=anova_1;
run;

proc glm data=anova_1;
class osoba lijek;
model Time=osoba lijek;
means lijek;
run;

```

| Obs | Osoba | Lijek | Time |
|-----|-------|-------|------|
| 1   | 1     | 1     | 30   |
| 2   | 1     | 2     | 28   |
| 3   | 1     | 3     | 16   |
| 4   | 1     | 4     | 34   |
| 5   | 2     | 1     | 14   |
| 6   | 2     | 2     | 18   |
| 7   | 2     | 3     | 10   |
| 8   | 2     | 4     | 22   |
| 9   | 3     | 1     | 24   |
| 10  | 3     | 2     | 20   |
| 11  | 3     | 3     | 18   |
| 12  | 3     | 4     | 30   |
| 13  | 4     | 1     | 38   |
| 14  | 4     | 2     | 34   |
| 15  | 4     | 3     | 20   |
| 16  | 4     | 4     | 44   |
| 17  | 5     | 1     | 26   |
| 18  | 5     | 2     | 28   |
| 19  | 5     | 3     | 14   |
| 20  | 5     | 4     | 30   |

Slika 3.1: Tablica podataka - Ispis iz SAS-a

## Ispis

Prvi dio koda nam učitava podatake odnosno priprema ih za analizu. Sa procedurom *print* dobijemo ispis naših podataka odnosno podatke na slici 3.1. Procedurom *glm*, što je skraćeno od *General linear models*, radimo analizu. Specificiramo data set na kojem vršimo analizu, grupne varijable su nam osoba i lijek te u naredbi *model* definiramo naš model. Naredba *means lijek* je dodatna naredba koja nam daje deskriptivnu analizu zavisne varijable *Time* po lijekovima koja je u tablici na slici 3.2. Ispis na slici 3.3 sadrži izvore varijabilnosti, stupnjeve slobode, sume kvadrata, varijance i vrijednosti testnih statistika i pripadajućih *p* vrijednosti.

| Level of Lijek | N | Time      |            |
|----------------|---|-----------|------------|
|                |   | Mean      | Std Dev    |
| 1              | 5 | 26.400000 | 8.76356092 |
| 2              | 5 | 25.600000 | 6.54217089 |
| 3              | 5 | 15.600000 | 3.84707681 |
| 4              | 5 | 32.000000 | 8.00000000 |

Slika 3.2: Deskriptivna statistika - Ispis iz SAS-a

| The GLM Procedure        |    |                |             |         |        |  |
|--------------------------|----|----------------|-------------|---------|--------|--|
| Dependent Variable: Time |    |                |             |         |        |  |
| Source                   | DF | Sum of Squares | Mean Square | F Value | Pr > F |  |
| Model                    | 7  | 1379.000000    | 197.000000  | 20.96   | <.0001 |  |
| Error                    | 12 | 112.800000     | 9.400000    |         |        |  |
| Corrected Total          | 19 | 1491.800000    |             |         |        |  |

  

| R-Square | Coeff Var | Root MSE | Time Mean |
|----------|-----------|----------|-----------|
| 0.924387 | 12.31302  | 3.065942 | 24.90000  |

  

| Source | DF | Type I SS   | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| Osoba  | 4  | 680.8000000 | 170.2000000 | 18.11   | <.0001 |
| Lijek  | 3  | 698.2000000 | 232.7333333 | 24.76   | <.0001 |

  

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| Osoba  | 4  | 680.8000000 | 170.2000000 | 18.11   | <.0001 |
| Lijek  | 3  | 698.2000000 | 232.7333333 | 24.76   | <.0001 |

Slika 3.3: Rezultati - Ispis iz SAS-a

## Interpretacija

U prvoj tablici na slici 3.3 testiramo sveukupno značajnost modela. Pod *Model* nam je objašnjena varijabilnost ( $SS_B + SST_b$ ) a pod *Error* neobjašnjena varijabilnost ( $SSE_b$ ) sa pripadnim stupnjevima slobode te sumama kvadrata. *F* statistiku dobijemo kao omjer dvaju varijanci (197 u brojniku i 9,4 u nazivniku), te iznosi 20,96 s pripadnom *p* vrijednosti koja je manja od 0,0001, pa prema tome odbacujemo nultu hipotezu odnosno model nam je značajan što znači da smo modelom uspjeli objasniti više varijabilnosti nego što je

ostalo neobjašnjeno. *Corrected Total* predstavlja ukupnu sumu kvadrata odnosno ukupnu varijabilnost.

U drugoj tablici su nam dane standardne statistike koje procedura ispisuje poput  $R$  kvadrata koji nam govori koliko smo postotak varijabilnosti objasnili te koeficijent varijacije, standardnu devijaciju i sredinu svih vremena. U našem slučaju  $R^2$  nam je preko 92% što je jako dobro.

Sljedeće dvije tablice su skoro pa identične osim što se razlikuju u načinu izračuna sume kvadrata što u ovom slučaju nije važno jer na oba načina dobivamo iste sume. U tim tablicama *Osoba* predstavlja varijabilnost između subjekata ( $SS_B$ ), a *Lijek* varijabilnost efekta tretmana ( $SS_{T_b}$ ). Obje nulte hipoteze koje tu testiramo odbacujemo na razini značajnosti od 5% jer su nam pripadne  $p$  vrijednosti manje od 0,0001. Dakle postoji značajna razlika među sredinama subjekata i razlika među tretmanima. Usporedba se može napraviti sa tablicom 3.2.

### 3.4 Dvofaktorski model analize varijance ponovljenih mjerena

Kao što i sam naziv kaže, u dvofaktorskem modelu imamo dva faktora. Modelska jednadžba nam je ista dvofaktorskome modelu analize varijance bez ponovljenih mjerena ((2.23) i (2.27)) samo će nam rastav suma kvadrata biti različit. Dakle izlazna vrijednost nam je rezultat očekivane vrijednosti i greške koja je normalno distribuirana. Očekivana vrijednost nam je suma očekivanih vrijednosti I. faktora (grupe), II. faktora (ponavljanja) i njihove međusobne interakcije. Navedeno ima sljedeći matematički zapis:

$$Y_{ij} = \mu + \pi_i + \delta_j + \pi\delta_{ij} + \epsilon_{ij}. \quad (3.16)$$

Prepostavimo da imamo  $n$  ispitanika na kojima testiramo novi lijek za snižavanje krvnog tlaka te ih podijelimo u 3 skupine ovisno o dobi. Na one koji imaju manje od 25 godina, one koji imaju između 25 i 45 godina, te one koji imaju preko 45 godina. Na svakome ispitaniku mjerimo tlak  $k$  puta. Tada model dan formulom (3.16) možemo interpretirati na sljedeći način.  $\delta_j$  je efekt  $j$ -tog ponavljanja, a  $\pi_i$  efekt  $i$ -te grupe (mladi, srednja dob, stari), dok se njihova interakcija  $\pi\delta_{ij}$  interpretira kao "ponašaju li se grupe tijekom vremena jednako"? Podaci su dani u tablici 3.4, dok tablica 3.4 predstavlja standardnu Anova tablicu sa sumama kvadrata, stupnjevima slobode, varijancama i  $F$  statistikama.

U tablici (3.4) je prikazan slučaj sa dvije grupe odnosno tretmana. Općenito, neka je  $t$  broj grupe odnosno tretmana, neka je  $k$  broj ponovljenih mjerena i neka u svakoj grupi imamo  $l$  subjekata. Prema tome u svakome ponovljenom mjerenu za pojedinu grupu imamo  $l$  podataka.  $Y_{ijm}$  predstavlja  $m$ -to mjerene za  $i$ -tu osobu pod tretmanom  $j$ -ot. Tada ukupnu varijabilnost  $SS$  (ukupno kvadratno odstupanje od aritmetičke sredine) možemo

### 3.4. DVOFAKTORSKI MODEL ANALIZE VARIJANCE PONOVLJENIH MJERENJA

| Grupa    | Osoba \ Ponavljanje | 1         | 2         | ... | $j$       | ... | $k$       |
|----------|---------------------|-----------|-----------|-----|-----------|-----|-----------|
| A        | 1                   | $Y_{111}$ | $Y_{112}$ |     | $Y_{11j}$ |     | $Y_{11k}$ |
| A        | 2                   | $Y_{211}$ | $Y_{212}$ |     | $Y_{21j}$ |     | $Y_{21k}$ |
| $\vdots$ | $\vdots$            | $\vdots$  | $\vdots$  |     | $\vdots$  |     | $\vdots$  |
| A        | $l$                 | $Y_{l11}$ | $Y_{l12}$ |     | $Y_{l1j}$ |     | $Y_{l1k}$ |
| B        | 1                   | $Y_{121}$ | $Y_{122}$ |     | $Y_{12j}$ |     | $Y_{12k}$ |
| $\vdots$ | $\vdots$            | $\vdots$  | $\vdots$  |     | $\vdots$  |     | $\vdots$  |
| B        | $l$                 | $Y_{l21}$ | $Y_{l22}$ |     | $Y_{l2j}$ |     | $Y_{l2k}$ |

Tablica 3.3: Notacija

podijeliti na dvije sume kvadrata: sumu kvadrata između subjekata ( $SSB$ ) i sumu kvadrata unutar subjekata ( $SSW$ ) koje definiramo na isti način kao i u jednofaktorskome modelu analize varijance ponovljenih mjerena ((3.8) i (3.9) redom). Broj stupnjeva slobode za  $SSB$  je ukupan broj subjekata minus jedan, a za  $SSW$  broj subjekata što množi broj ponavljanja minus jedan procijenjeni parametar za svaki subjekt. I jednu i drugu sumu kvadrata možemo dalje podijeliti na sljedeći način:

$$SSB = SST_1 + SSE_1 \quad (3.17)$$

i

$$SSW = SST_2 + SST_{12} + SSE_2. \quad (3.18)$$

U (3.18) varijabilnost između subjekata možemo djelomično objasniti zbog različitih tretmana ( $SST_1$ ). Ono što nismo uspjeli objasniti jest prirodna varijabilnost zbog različitih tretmana ( $SSE_1$ ). U (3.19) varijabilnost unutar subjekata djelomično možemo objasniti ponovljenim mjeranjima ( $SST_2$ ), djelomično interakcijom tretmana i vremena ( $SST_{12}$ ), dok je neobjašnjena varijabilnost ( $SSE_2$ ) uzrokovana različitim tretmanima i ponovljenim mjeranjima. Uspoređujući s jednofaktorskim modelom ponovljenih mjerena vidimo da smo dodajući još jedan faktor podijelili varijabilnost između subjekata na dvije sume kvadrata te varijabilnost unutar umjesto na dvije podijelili na tri sume kvadrata (tablica 3.2). Ukoliko usporedimo ovaj model s dvofaktorskim modelom bez ponovljenim mjerena (tablica 2.3) možemo primjetiti da smo "samo" podijelili grešku odnosno neobjašnjenu varijabilnost ( $SSE$ ) na dvije neobjašnjene varijabilnosti ( $SSE_1$  i  $SSE_2$ ) koje proizlaze iz toga da su mjerena zavisna. Greška  $SSE_1$  pripada varijabilnosti između, a  $SSE_2$  varijabilnosti unutar subjekata. Prema tome  $SST_1$ ,  $SST_2$  i  $SST_{12}$  se definiraju analogno (uz zamjenu  $m_1 = t$  i  $m_2 = k$ ) kao u Poglavlju 2 gdje smo obradili dvofaktorski model bez ponavljanja ((2.35), (2.36) i (2.37) redom). Neobjašnjene varijance možemo protumačiti na sljedeći

način:  $SSE_1$  je greška subjekata unutar tretmana, dok je  $SSE_2$  greška efekta ponavljanja i subjekata unutar tretmana.

| Izvor varijabilnosti          | Suma kvadrata | Stupnjevi slobode | Korigirana varijanca | Test statistika $F$            |
|-------------------------------|---------------|-------------------|----------------------|--------------------------------|
| <b>Između subjekata</b>       | $SSB$         | $tl - 1$          |                      |                                |
| Efekt tretmana                | $SST_1$       | $t - 1$           | $MST_1$              | $F_1 = \frac{MST_1}{MSE_1}$    |
| Greška tretmana               | $SSE_1$       | $t(l - 1)$        | $MSE_1$              |                                |
| <b>Unutar subjekata</b>       | $SSW$         | $tl(k - 1)$       |                      |                                |
| Efekt ponavljanja             | $SST_2$       | $k - 1$           | $MST_2$              | $F_2 = \frac{MST_2}{MSE_2}$    |
| Interakcija                   | $SST_{12}$    | $(t - 1)(k - 1)$  | $MST_{12}$           | $F_3 = \frac{MST_{12}}{MSE_2}$ |
| Greška ponavljanja i tretmana | $SSE_2$       | $t(l - 1)(k - 1)$ | $MSE_2$              |                                |
| Ukupno                        | $SS$          | $tkl - 1$         |                      |                                |

Tablica 3.4: ANOVA tablica - dvofaktorski model s ponovljenim mjeranjima

## Hipoteze i ukupna značajnost modela

Možemo testirati tri hipoteze. Hipotezu da je efekt grupe (tretmana) beznačajan, efekt ponavljanja beznačajan i interakcija beznačajna. Prva nulta hipoteza jest da su sredine grupa jednake nasuprot alternativnoj hipotezi da se barem dvije razlikuju. Testna statistika nam je sljedeća:

$$F_1 = \frac{MST_1}{MSE_1} \sim F(t - 1, t(l - 1)). \quad (3.19)$$

Druga nulta hipoteza je da su sredine u svakom ponavljanju jednake nasuprot alternativnoj da se barem dvije razlikuju. Testna statistika je

$$F_2 = \frac{MST_2}{MSE_2} \sim F(k - 1, t(l - 1)(k - 1)). \quad (3.20)$$

Treća nulta hipoteza je da su sredine svakog tretmana u svakom ponavljanju jednake nasuprot alternativnoj da se barem dvije značajno razlikuju. Test statistika je

$$F_3 = \frac{MST_{12}}{MSE_2} \sim F((t - 1)(k - 1), t(l - 1)(k - 1)). \quad (3.21)$$

Testirat ćemo i sveukupnu značajnost modela gdje nam je testna statistika sljedeća:

$$F = \frac{MST_1 + MST_2 + MST_{12}}{MSE_1 + MSE_2} \sim F(tk - 1, tk(l - 1)). \quad (3.22)$$

Kao i u prijašnjim poglavljima velik brojnik i mali nazivnik sugerirat će odbacivanje nultih hipoteza.

## 3.5 Primjer - dvofaktorski model

### Zadatak

Podatke sam preuzeo iz [1]. Bez obzira na to što su podaci djelomični analiza se može napraviti. Ukupan broj subjekata je trideset koji su podijeljeni u tri grupe. Subjekti su krave iz Australije. U originalnom eksperimentu imamo sedamdeset i dvije krave. Grupe su vrste dijeta. Imamo sljedeće dijete kojima su podvrgnute krave: *Barkley*, *Mixed*, *Lupins*. *Mixed* dijeta je mješavina *Barkley* i *Lupins* dijeta. Svakoj dijeti je podvrgnuto deset krava pa prema tome imamo balansirani dizajn. Svrha eksperimenta je bila da vidimo kako pojedina dijeta djeluje na količinu proteina u mlijeku. Razina proteina je mjerena tjedno iz uzorka mlijeka za svaku kravu. Podaci u tablici na slici 3.4 prikazuju dio podataka. Imamo jednu nezavisnu varijablu dijeta koja ima 3 razine, i pet zavisnih varijabli: week1, week2, week3, week4 i week5.

### Kod

```
proc glm data=krave;
  class dijeta;
  model week1-week5=dijeta;
  repeated time 5 polynomial / summary printe;
  means dijeta;
run;
```

Kao i prije koristimo proceduru *glm*. Naredbom *class* kažemo da nam je varijabla dijeta grupna varijabla. Naredbom *model* definiramo model na kojem ćemo raditi analizu. S lijeve strane su nam zavisne varijable, s desne nezavisne. Naredbom *repeated* kažemo SAS-u da imamo ponovljena mjerjenja. Iza te naredbe moramo dati ime zavisnoj varijabli koje je *time*. Broj 5 označava koliko imamo ponovljenih mjerjenja. Nakon toga smo dodali opciju *polynomial* te iza kose crte *summary* i *printe*. Opcija *polynomial* kreira četiri nove varijable iz pet zavisnih varijabli. Prva je linearni efekt vremena, druga kvadratni, treća kubični... Opcijom *summary* dobivamo rezultate za četiri transformirane varijable, dok opcijom *printe* ispisujemo matrice kojima provjeravamo prepostavke.

| Obs | dijeta  | Krava | week1 | week2 | week3 | week4 | week5 |
|-----|---------|-------|-------|-------|-------|-------|-------|
| 1   | Barkley | 1     | 3.63  | 3.57  | 3.47  | 3.65  | 3.89  |
| 2   | Barkley | 2     | 3.24  | 3.25  | 3.29  | 3.09  | 3.38  |
| 3   | Barkley | 3     | 3.98  | 3.60  | 3.43  | 3.30  | 3.29  |
| 4   | Barkley | 4     | 3.66  | 3.50  | 3.05  | 2.90  | 2.72  |
| 5   | Barkley | 5     | 4.34  | 3.76  | 3.68  | 3.51  | 3.45  |
| 6   | Barkley | 6     | 4.36  | 3.71  | 3.42  | 3.95  | 4.06  |
| 7   | Barkley | 7     | 4.17  | 3.60  | 3.52  | 3.10  | 3.78  |
| 8   | Barkley | 8     | 4.40  | 3.86  | 3.56  | 3.32  | 3.64  |
| 9   | Barkley | 9     | 3.40  | 3.42  | 3.51  | 3.39  | 3.35  |
| 10  | Barkley | 10    | 3.75  | 3.89  | 3.65  | 3.42  | 3.32  |
| 11  | Mixed   | 11    | 3.38  | 3.38  | 3.10  | 3.90  | 3.15  |
| 12  | Mixed   | 12    | 3.80  | 3.51  | 3.19  | 3.11  | 3.35  |
| 13  | Mixed   | 13    | 4.17  | 3.71  | 3.32  | 3.10  | 3.07  |
|     |         | ...   |       |       |       |       |       |

Slika 3.4: Tablica djela podataka - Ispis iz SAS-a

## Ispis i interpretacija

Prvo što nam SAS daje jest univarijatna analizu varijance za svaku zavisnu varijablu *week* po grupnoj varijabli *dijeta* s pripadnim box-plotovima. Taj ispis možemo spriječiti dodajući opciju *nouni* iza kose crte u naredbi model. Sljedeći dio ispisa jest zbog naredbe *repeated* i objasnit ćemo sada dio po dio.

| The GLM Procedure           |        |                      |
|-----------------------------|--------|----------------------|
| Class Level Information     |        |                      |
| Class                       | Levels | Values               |
| dijeta                      | 3      | Barkley Lupins Mixed |
| Number of Observations Read |        | 30                   |
| Number of Observations Used |        | 30                   |

Slika 3.5: Općeniti podaci - Ispis iz SAS-a

| The GLM Procedure<br>Repeated Measures Analysis of Variance |   |   |   |   |   |  |  |  |  |
|---|---|---|---|---|---|--|--|--|--|
| Repeated Measures Level Information                         |   |   |   |   |   |  |  |  |  |
| Dependent Variable week1 week2 week3 week4 week5            |   |   |   |   |   |  |  |  |  |
| Level of time   | 1 | 2 | 3 | 4 | 5 |  |  |  |  |

Slika 3.6: Tablica nivoa zavisne varijable - Ispis iz SAS-a

| Level of dijeta | N  | week1 |      | week2 |      | week3 |      | week4 |      | week5 |      |
|-----------------|----|-------|------|-------|------|-------|------|-------|------|-------|------|
|                 |    | Mean  | Std  |
| Barkley         | 10 | 3.89  | 0.41 | 3.61  | 0.19 | 3.45  | 0.18 | 3.36  | 0.30 | 3.48  | 0.37 |
| Lupins          | 10 | 3.74  | 0.45 | 3.42  | 0.34 | 3.29  | 0.38 | 3.24  | 0.30 | 3.11  | 0.37 |
| Mixed           | 10 | 3.92  | 0.39 | 3.52  | 0.24 | 3.30  | 0.26 | 3.29  | 0.32 | 3.39  | 0.29 |

Slika 3.7: Deskriptivne statistike - Ispis iz SAS-a

U tablici na slici 3.5 dani su nam opći podaci o broju opservacija i grupnoj varijabli, na 3.6 dani su nam nivoi zavisne varijable. Vidimo da nam je zavisna varijabla time i da ima pet mjerena. Vrlo je važna ova tablica da vidimo jesmo li postavili dobro model. Nakon općih podataka slijede deskriptivne statistike na u tablici na slici 3.7. Sljedeći dio ispisa je dan u matrici na slici 3.8. To je korelacijska matrica koja odgovara na pitanje: Ako kontroliramo efekt dijete, do koje mjere tada razina proteina u jednom trenutku predviđa razinu proteina u drugom trenutku? Ispod korelacije je  $p$  vrijednost za par vremena. Većina ovih korelacija bi trebala biti značajna. Vidimo da u našem slučaju samo week1 i week4 nisu značajna ( $p$  vrijednost=0,2207).

Sljedeći dio ispisa daje korelacijsku matricu za transformirane varijable i Manova testove i njih možemo zanemariti. Nakon toga testiramo pretpostavku sferičnosti u tablici na slici 3.9 za transformirane podatke i na ortogonalnim komponentama to jest nekoreliranim. Važno je da uvijek pogledamo test za ortogonalne komponente. Ukoliko nije značajan znači da je uvjet sferičnosti zadovoljen. U našam slučaju  $p$  vrijednost je 0,2701 što je veće od 0,05 pa je uvjet sferičnosti zadovoljen.

Sada dolazimo do glavnih testova i statistika koje smo obradili u poglavljiju. Tablica na slici 3.10 testira razlike između subjekata (eng. *Between Subjects Effects*). Njome testiramo da dijeta nema efekt na razinu proteina u mlijeku krave. Izvore varijabilnosti možemo usporediti sa tablicom 3.4. Oznaka za sumu kvadrata za izvor varijabilnosti grupe nam je  $SST_1$  (1,0054) sa pripadnim stupnjevima slobode ( $t - 1 = 3 - 1 = 2$ ), a za neobjašnjenu varijabilnost  $SSE_1$  (8,6243) sa stupnjevima slobode ( $t(l - 1) = 3(10 - 1) = 27$ ).

| Partial Correlation Coefficients from the Error SSCP Matrix / Prob >  r |          |          |          |          |          |
|---|----------|----------|----------|----------|----------|
| DF = 27   | week1    | week2    | week3    | week4    | week5    |
| <b>week1</b>  | 1.000000 | 0.613004 | 0.548948 | 0.238948 | 0.495486 |
|   |          | 0.0005   | 0.0025   | 0.2207   | 0.0073   |
| <b>week2</b>  | 0.613004 | 1.000000 | 0.505814 | 0.434938 | 0.538662 |
|   | 0.0005   |          | 0.0060   | 0.0207   | 0.0031   |
| <b>week3</b>  | 0.548948 | 0.505814 | 1.000000 | 0.382738 | 0.564099 |
|   | 0.0025   | 0.0060   |          | 0.0444   | 0.0018   |
| <b>week4</b>  | 0.238948 | 0.434938 | 0.382738 | 1.000000 | 0.534734 |
|   | 0.2207   | 0.0207   | 0.0444   |          | 0.0034   |
| <b>week5</b>  | 0.495486 | 0.538662 | 0.564099 | 0.534734 | 1.000000 |
|   | 0.0073   | 0.0031   | 0.0018   | 0.0034   |          |

Slika 3.8: Korelacijska matrica - Ispis iz SAS-a

| Sphericity Tests      |    |                     |            |            |  |
|-----------------------|----|---------------------|------------|------------|--|
| Variables             | DF | Mauchly's Criterion | Chi-Square | Pr > ChiSq |  |
| Transformed Variates  | 9  | 0.6466045           | 11.082188  | 0.2701     |  |
| Orthogonal Components | 9  | 0.6466045           | 11.082188  | 0.2701     |  |

Slika 3.9: Test sferičnosti - Ispis iz SAS-a

Vrijednost  $F_1$  statistike je 1,57, a  $p$  vrijednosti 0,2257 pa prema tome ne odbacujemo nultu hipotezu. Odnosno zaključujemo da razlike među sredinama grupa nisu značajne. Druga važna tablica je na slici 3.11 (eng. *Within Subjects Effects*). U njoj testiramo druge dvije hipoteze koje smo obradili. Prvi izvor varijabilnosti nam je *time* a njegovu sumu kvadrata u tablici 3.4 smo označili sa  $SST_2$  (6,2699) sa 4 stupnja slobode ( $k - 1 = 5 - 1 = 4$ ). Sljedeći izvor varijabilnosti nam je interakcija  $SST_{12}$  (0,3464) *time\*dijeta* sa 8 stupnjeva slobode ( $(t - 1)(k - 1) = (3 - 1)(5 - 1) = 8$ ). Na kraju je neobjašnjena varijabilnost  $SS E_2$  (6,3974) koja ima 108 stupnjeva slobode ( $t(l - 1)(k - 1) = 3(10 - 1)(5 - 1) = 108$ ). Vrijednost  $F_2$  statistike nam je 26,46 a pripadna  $p$  vrijednost nam je manja od 0,0001 pa nam je vrijeme značajno. Vrijednost  $F_3$  statistike nam je 0,73 a pripadne  $p$  vrijednosti 0,6637 pa nam interakcija nije značajna. Dva zadnja redu u tablici su nam prilagođene  $p$  vrijednosti koje gledamo kada prepostavka sferičnosti nije zadovoljena. G-G označava

Greenhouse-Geisserovu prilagodbu, a H-F-L Huynh-Feldt-Lecoutrovu prilagodbu  $p$  vrijednosti. Bez obzira što ih ne gledamo možemo komentirati da obje prilagodbe ne bih radile razliku u zaključivanju. U donjoj tablici su na slici 3.11 su Greenhouse-Geisserov i Huynh-Feldt-Lecoutrovu epsiloni koji se koriste za prilagodbu stupnjeva slobode koji onda produciraju spomenute prilagođene  $p$  vrijednosti. Iz ovoga primjera zaključujemo sljedeće: različita vrsta dijete značajno utječe na razinu proteina u kravini mljeku, vrijeme također značajno utječe na razinu proteina u krvi dok interakcija vremena i dijete nije značajna, to jest različite dijete se tijekom vremena ponašaju slično.

| The GLM Procedure<br>Repeated Measures Analysis of Variance<br>Tests of Hypotheses for Between Subjects Effects |    |             |             |         |        |  |
|---|----|-------------|-------------|---------|--------|--|
| Source  | DF | Type III SS | Mean Square | F Value | Pr > F |  |
| dijeta  | 2  | 1.00540933  | 0.50270467  | 1.57    | 0.2257 |  |
| Error   | 27 | 8.62435400  | 0.31942052  |         |        |  |

Slika 3.10: Tablica između subjekata - Ispis iz SAS-a

| The GLM Procedure<br>Repeated Measures Analysis of Variance<br>Univariate Tests of Hypotheses for Within Subject Effects |     |             |             |         |                |                  |
|--|-----|-------------|-------------|---------|----------------|------------------|
| Source   | DF  | Type III SS | Mean Square | F Value | Pr > F         | Adj Pr > F       |
|  |     |             |             |         | G - G    H-F-L |                  |
| time   | 4   | 6.26993333  | 1.56748333  | 26.46   | <.0001         | <.0001    <.0001 |
| time*dijeta  | 8   | 0.34649067  | 0.04331133  | 0.73    | 0.6637         | 0.6336    0.6518 |
| Error(time)  | 108 | 6.39749600  | 0.05923607  |         |                |                  |

  

|                              |        |
|------------------------------|--------|
| Greenhouse-Geisser Epsilon   | 0.7959 |
| Huynh-Feldt-Lecoutre Epsilon | 0.9147 |

Slika 3.11: Tablica unutar subjekata - Ispis iz SAS-a



# Bibliografija

- [1] J. Lawson, *Design and Analysis of Experiments with SAS*, Chapman & Hall/CRC Press, 2010.
- [2] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, 1993.
- [3] N. Sarapa, *Teorija vjerojatnosti*, sv. 3, Školska knjiga, 2002.
- [4] B. J. Winer, D. R. Brown i K. M. Michels, *Statistical Principles in Experimental Design*, sv. 3, McGraw-Hill.
- [5] I. Šošić, *Statistical Principles in Experimental Design*, sv. 3, McGraw-Hill.



# Sažetak

U ovome radu koji je podijeljen u tri poglavlja glavna tema nam je bila analiza varijance ponovljenih mjerena. Analiza varijance ponovljenih mjerena je statistička metoda kojom testiramo jednakost sredina na podacima čija mjerena se ponavljaju. U prvom poglavlju, tako reći uvodnome, djelomično je obrađen  $t$  test. Točnije obrađen je  $t$  test za dva uzorka kao primjer testiranja jednakosti sredina. U drugom poglavlju proširili smo to testiranje na više uzoraka. Obrađena je univariatna analiza varijance s jednim i s dva faktora. Objasnjena je varijabilnost rastavom na različite sume kvadrata, dan je matematički model, a sve statistike sa sumama kvadrata i pripadnim stupnjevima slobode su onda sažete u standardnim Anova tablicama. Nakon što je dobro pripremljen teren, obrađena je i analiza varijance ponovljenih mjerena s jednim faktorom i sa dva faktora. Nadograđujući već obrađene modele bilo je lakše objasniti i razumjeti daljnje rastavljenje definiranih suma kvadrata. Kao glavni razlog dodatnih suma kvadrata pokazala su se dodatna mjerena na istim subjektima, kojim homogeniziramo uzorak i smanjujemo neobjašnjenu varijabilnost. Rečeno je koje su prednosti, dan je model, navedene su i obrađene prepostavke te su opet sve statistike dane u Anova tablicama. Na dva zanimljiva i stvarna primjera su numerički obrađena oba modela.

Kroz primjere, kojih ima četiri:  $t$  test za 2 uzorka, dvofaktorski model analize varijance, jednofaktorski model analize varijance ponovljenih mjerena i dvofaktorski model analize varijance ponovljenih mjerena, pokazali smo konkretnu primjenu obrađenih metoda. Time smo diplomski rad učinili ne samo teoretskim nego i primjenjenim. Također smo se kroz primjere bolje upoznali sa programskim sustavom SAS, jednim od najpopularnijih alata za analizu podataka kako u akademskim krugovima tako i u poslovnom svijetu.

Iako se analiza varijance već duže vrijeme primjenjuje za razne statističke probleme nije zastarjela. Štoviše, jedna je od najkorištenijih i najaktualnijih statističkih metoda koja se i dalje aktivno razvija.



# Summary

In this thesis which is divided in three chapters the main theme was analysis of variance of repeated measures. Analysis of variance of repeated measures is a statistical method which tests the equality of means having repeated measurements on same subjects. In chapter one, which I would call introduction, partly was elaborated *t* test. To be correct it was elaborated *t* test for two samples as an example of testing the equality of means. In chapter two we expanded testing on more samples. It was elaborated univariate analysis of variance with one and two factors. The variability was explained by dividing on various sum of squares, it was given the mathematical model and all statistics with sum of squares and corresponding degrees of freedom are given in standard Anova tables. After good elaboration of analysis of variance was made, it was elaborated analysis of variance of repeated measures with one and two factors. Already elaborated models were upgraded and easier to explain and understand further dividing defined sum of squares. It was shown that a main cause of extra sum of squares were additional or more measurements on same subjects, which caused less unexplained variability as a result of more homogenized sample. The advantages were discussed, mathematical model was given, assumptions were named and elaborated and again all relevant statistics were given in standard Anova tables. On two interesting and real examples numerically were discussed both models.

Throughout examples, which we had four: *t* test for two samples, two factor analysis of variance, one factor and two factor analysis of variance of repeated measures, concrete examples of elaborated models were shown. By that, this final thesis is not just theoretical but applied as well. Also, throughout examples program system SAS was introduced, which is one of the most popular tools for data analysis in academic and business world.

Although analysis of variance is used for already longer period of time for diverse statistical problems it hasn't become old fashioned. On contrary, it is one of the most used and up to date statistical method which is still actively developing.



# Životopis

Rođen sam 08.05.1988. godine u Zagrebu. Drugo sam od šestero djece. Cijeli život živim u Zagrebu. Pohađao sam osnovnu školu Ive Andrića. Srednješkolsko sam obrazovanje stekao u općoj, I. gimnaziji koju sam završio 2007. godine te sam iste godine upisao Pred-diplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu. Pred-diplomski studij završio sam sa zimskim semestrom akademske godine 2011./12. te na jesen iste godine upisujem diplomski studij Matematička statistika. Još kao desetogodišnjak počinjem trenirati nogomet te ga treniram sve do prve godine na fakultetu. Tijekom razdoblja na fakultetu igram mali nogomet za selekciju PMF-a. Zadnju godinu na fakultetu radim puno radno vrijeme u jednoj farmaceutskoj kompaniji kao student. Zaručio sam se u jesen 2013. godine te se planiram oženiti u ljeto 2014. godine sa Zrinkom koju sam upoznao na matematici i koja je prije godinu dana diplomirala.