Dvofaktorska analiza varijance s primjenama u psihologiji

Pavlek, Ivana

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Matematički odsjek

Ivana Pavlek

Dvofaktorska analiza varijance s primjenama u psihologiji

Diplomski rad

Voditelj rada:
doc.dr.sc. Miljenko Huzak

Zagreb, lipanj 2012.

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. ______________________________

Sadržaj

Uvod

U ovom radu će se obraditi matematički model dvofaktorske analize varijance i navesti će se dva primjera primjene u psihologiji. Od toga će jedan primjer biti samo za ilustraciju (3. poglavlje), a u drugom će biti obrađena stvarna situacija pomoću koje će se prikazati korištenje dvofaktorske analize varijance u praksi (5. poglavlje). Egzaktne matematičke definicije koristiti će se u 2. i 4. poglavlju, dok će se intuitivan pristup dvofaktorskoj analizi varijance obraditi u 3. poglavlju.

Gotovo svi bolji statistički programi računaju analizu varijance kao poseban slučaj regresije, te će se taj model obraditi egzaktno u 4. poglavlju. Osnovne značajke regresije koje će se koristiti biti će navedene u 2. poglavlju kao i osnovne definicije. Kod primjera će se koristiti programski jezik SAS.

Oznake će biti uglavnom uobičajene. Za označavanje skalara koristiti će se mala slova normalne debljine (kao što su x i β), za slučajne varijable velika (kao što su X i Y). Nešto deblja mala slova označavati će vektore (kao što su $\boldsymbol{x}$ i $\boldsymbol{\beta}$), dok će veća deblja slova označavati matrice (kao što je $\boldsymbol{X}$).

Ovom prilikom bih se željela bih se zahvaliti mentoru prof. Miljenku Huzaku koji mi je egzaktnije definirao temu koju sam željela i jer me strpljivo vodio tijekom cijelog pisanja rada. Na kraju bih se još zahvalila svojim roditeljima jer su mi omogućili da studiram što želim, te sestri Mariji koja mi je bila velika podrška u teškim danima studiranja.

1 Osnovni pojmovi

1.1 Osnovne definicije

Definicija 1.1. *Neprekidna slučajna varijabla X je* **normalna slučajna varijabla** *s parametrima μ i σ ako joj je funkcija gustoće*

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

Piše se: $X \sim N(\mu, \sigma^2)$.

Definicija 1.2. *Kažemo da je* $\mathbf{Y}$ **n-dimenzionalni normalni slučajni vektor** *s matematičkim očekivanjem μ i kovarijacijskom matricom Σ ako postoji m-dimenzionalni ($m \leq n$) slučajni vektor* $\mathbf{Z}$ *čije su komponente nezavisne $N(0,1)$ slučajne varijable i neslučajna $m \times n$ matrica* $\mathbf{A}$ *td. vrijedi:*

$$\mathbf{Y} \stackrel{D}{=} \mu + \mathbf{AZ} \quad i \quad \Sigma = \mathbf{AA}^T.$$

Piše se: $\mathbf{X} \sim N_m(\mu, \Sigma)$, ili kraće $\mathbf{X} \sim N(\mu, \Sigma)$.

Definicija 1.3. *Neprekidna slučajna varijabla T ima* **Studentovu** *ili* **t-distribuciju s n stupnjeva slobode** *ako postoje nezavisne slučajne varijable X i Y takve da $X \sim N(0,1)$, $Y \sim N(0,1)$ i*

$$T \stackrel{D}{=} \frac{X}{\sqrt{\frac{Y}{n}}}.$$

Piše se: $T \sim t(n)$.

Definicija 1.4. *Neka su $X_1, \ldots, X_n \sim N(0,1)$ nezavisne slučajne varijable. Tada $\mathbf{H} := X_1^2 + \cdots + X_n^2$ ima χ^2* **distribuciju s n stupnjeva slobode***. Piše se: $\mathbf{H} \sim \chi^2(n)$.*

Definicija 1.5. *Kažemo da slučajna varijabla V ima* **F-distribuciju** *(također poznatu kao* **Snedecorova F-distribucija** *ili* **Fisher-Snedecor distribucija***) s (m,n) stupnjeva slobode ako postoje nezavisne slučajne varijable $X \sim \chi^2(m)$ i $Y \sim \chi^2(n)$ i vrijedi*

$$V \stackrel{D}{=} \frac{X/m}{Y/n}.$$

Piše se: $V \sim F(m,n)$.

1.2 Ortogonalni Projektori

Definicija 1.6. *Kažemo da je* **H** **ortogonalni projektor** *ili* **ortogonalna projekcijska matrica** *na vektorski prostor* V *ako vrijedi*

a) $\boldsymbol{v} \in V \;\Rightarrow\; \mathbf{H}\boldsymbol{v} = \boldsymbol{v}$

b) $\boldsymbol{w} \perp V \;\Rightarrow\; \mathbf{H}\boldsymbol{w} = \boldsymbol{0}$.

Teorem 1.7 (Karakterizacija ortogonalne projekcijske matrice)**.** *Kvadratna matrica* **H** *je ortogonalna projekcijska matrica ako i samo ako je simetrična i idempotentna, to jest ako vrijedi*

a) $\mathbf{H} = \mathbf{H}^T$

b) $\mathbf{H} = \mathbf{H}^2$.

Definicija 1.8. *Neka je* **Z** n*-dimenzionalan slučajni vektor i* **A** *simetrična kvadratna matrica dimenzije* $n \times n$*. Tada je* kvadratna forma *od* **Z** *u odnosu na* **A** *slučajna varijabla* $\mathbf{Q} = \mathbf{Z}^T\mathbf{A}\mathbf{Z}$.

Teorem 1.9. *Ako je* **P** *neslučajna ortogonalna projekcijska matrica i* **Z** *multivarijantan slučajni vektor takav da* $\mathbf{Z} \sim N_n(\mathbf{0}, I)$*, tada kvadratna forma* $\mathbf{Z}^T\mathbf{P}\mathbf{Z} \sim \chi^2(r(P))$*, za* $r(\mathbf{P})$ *rang od* **P**.

Teorem 1.10. *Neka je* $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 I)$*. Tada vrijedi:*

$\mathbf{A}$ *i* $\mathbf{B}$ *simetrične matrice td.* $\mathbf{AB} = \mathbf{0} \Rightarrow$

$$\Rightarrow \mathbf{Y}^T\mathbf{A}\mathbf{Y} \text{ i } \mathbf{Y}^T\mathbf{B}\mathbf{Y} \text{ su nezavisne slučajne varijable.}$$

Teorem 1.11 (Teorem o projekciji)**.** *Neka je* M *potprostor euklidskog prostora* $\mathbb{R}^n$ *i* $\boldsymbol{x} \in \mathbb{R}^n$ *proizvoljan vektor. Tada:*

a) $(\,\exists!\, \hat{\boldsymbol{x}} \in \mathrm{M}) \qquad \|\boldsymbol{x} - \hat{\boldsymbol{x}}\| = \inf_{\boldsymbol{y} \in \mathrm{M}} \|\boldsymbol{x} - \mathbf{y}\|$

b) $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| = \min_{\boldsymbol{y} \in \mathrm{M}} \|\boldsymbol{x} - \mathbf{y}\| \;\wedge\; \hat{\boldsymbol{x}} \in \mathrm{M} \iff \boldsymbol{x} - \hat{\boldsymbol{x}} \in \mathrm{M}^\perp \;\wedge\; \hat{\boldsymbol{x}} \in \mathrm{M}$

gdje je $\mathrm{M}^\perp$ *ortogonalni komplement od* M *u* $\mathbb{R}^n$.

Propozicija 1.12. *Neka je* $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{V})$ *i neka je* **A** *proizvoljna* $n \times n$ *matrica. Tada vrijedi*

$$\mathrm{E}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}) = tr(\mathbf{A}\mathbf{V}) + \mu^T\mathbf{A}\mu\,.$$

2 Regresijska analiza

Regresijski model je povezan s modelom analize varijanci, te je stoga od interesa razumjeti njegovu interpretaciju projekcijama, na koju će se poslije često pozivati.

2.1 Model

Općeniti linearni model[1] je oblika

$$Y_i = \sum_{j=1}^{p} \theta_j x_{ij} + \varepsilon_i,\ i = 1, \ldots, n \tag{2.1.1}$$

gdje se Y_i naziva **zavisna varijabla** i-tog subjekta, a x_{ij} j-ta **nezavisna varijabla** i-tog subjekta. Varijable ε_i $(i = 1, \ldots, n)$ su slučajne varijable i nazivaju se **pogreške**, dok su θ_i $(i = 1, \ldots, n)$ neki nepoznati parametri.

Za regresijski model oblika (2.1.1) vrijedi da su pogreške međusobno nezavisne i normalne, da im je očekivanje $\mathrm{E}(\varepsilon_i) = 0$ i varijanca $\mathrm{Var}(\varepsilon_i) = \sigma^2$, to jest $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$.

Napomena 2.1. *Također se pretpostavlja da su $x_{1j}, ..., x_{nj}$ linearno nezavisni za svaki $j = 1, \ldots, p$.*

Eksplicitno korištenje (2.1.1) u teoriji nije praktično, pa se više koristi u matričnom zapisu

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{2.1.2}$$

gdje je $\mathbf{Y}$ slučajni vektor čije su komponente Y_i, $\mathbf{X}$ matrica čije su komponente x_{ij}, $\boldsymbol{\theta}$ vektor čije su komponente θ_j, te gdje je $\boldsymbol{\varepsilon}$ slučajni vektor čije su komponente ε_i. Važno je primijetiti da su $\mathbf{Y}$ i $\boldsymbol{\varepsilon}$ dimenzije n, ali je $\boldsymbol{\theta}$ dimenzije p. Matrica $\mathbf{X}$ dimenzije $n \times p$ se naziva **dizajn-matrica** ili **matrica modela**. Zbog prethodne napomene 2.1 dizajn-matrica mora biti punog ranga.

U ovakvom modelu je uobičajeno poznate nezavisne varijable smatrati neslučajnima, to jest fiksnima. Stoga je $\mathbf{X}$ fiksna matrica kao što je i $\boldsymbol{\theta}$ fiksni vektor. Jedine slučajne veličine u (2.1.2) su $\boldsymbol{\varepsilon}$ i $\mathbf{Y}$. Uvjet na vektor pogrešaka zapisan u matričnom obliku glasi

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I),$$

što povlači

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 I), \tag{2.1.3}$$

[1] Naziv linearnog modela ne dolazi od linearnosti po varijablama x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$ već po parametrima θ_j, $j = 1, \ldots, p$.

gdje je

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\theta}. \tag{2.1.4}$$

Vektor $\boldsymbol{\mu}$ zovemo **vektor očekivanja**.

Sada se početni model (2.1.1) može zapisati

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}. \tag{2.1.5}$$

Ovaj model se naziva **normalni linearni model** zbog pretpostavke na distribuciju u (2.1.3).

2.2 Metoda najmanjih kvadrata

Ako je $\mathbf{z}$ n-dimenzionalan vektor s komponentama z_i, tada je duljina od $\mathbf{z}$ dana s

$$\|\mathbf{z}\| = \sqrt{\sum_{i=1}^{n} z_i^2}\,.$$

Suma se može zapisati kompaktnije koristeći matričnu notaciju

$$\|\mathbf{z}\| = \sqrt{\mathbf{z}^T\mathbf{z}}. \tag{◑}$$

Procjenitelj najmanjih kvadrata nepoznatog parametra $\boldsymbol{\theta}$ je vektor $\hat{\boldsymbol{\theta}}$ koji minimizira **rezidualnu sumu kvadrata** danu izrazom

$$Q_1(\boldsymbol{\theta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}\|^2 \overset{◑}{=} (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})\,.$$

Minimum se postiže kada je gradijent (vektor prvih parcijalnih derivacija)

$$\nabla Q_1(\boldsymbol{\theta}) = -2\mathbf{X}^T(\mathbf{X}\boldsymbol{\theta})$$

jednak nula, iz čega slijedi

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\theta}}. \tag{2.1.6}$$

Ove linearne jednadžbe koje se moraju riješiti za $\hat{\boldsymbol{\theta}}$ nazivaju se **normalne jednadžbe**. Ako je $\mathbf{X}^T\mathbf{X}$ invertibilna matrica tada je rješenje jednadžbe (2.1.6) jedinstveno i dano s

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.1.7}$$

Ako inverz ne postoji, tada rješenje nije jedinstveno i za $\hat{\boldsymbol{\theta}}$ se može uzeti bilo koje rješenje od (2.1.6). Da bi inverz postojao, dovoljno je da je dizajn-matrica punog ranga.

2.3 Regresija kao projekcija

Ako se uvedu oznake $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ za stupce matrice $\mathbf{X}$, tada vrijedi

$$\mathbf{X}\boldsymbol{\theta} = \sum_{j=1}^{p} \theta_j \boldsymbol{x}_j \,.$$

Dakle, svaki mogući vektor očekivanja $\boldsymbol{\mu}$ u (2.1.4) je linearna kombinacija stupaca od $\mathbf{X}$, a skup svih mogućih vektora očekivanja

$$C(\mathbf{X}) = \{\mathbf{X}\boldsymbol{\theta} \colon \boldsymbol{\theta} \in \mathbb{R}^p\} \tag{2.1.8}$$

naziva se **prostor stupaca** od $\mathbf{X}$ zato jer je razapet stupcima od $\mathbf{X}$, to jest vrijedi da je $C(\mathbf{X})$ linearna ljuska $[\{x_1, \ldots, x_p\}]$, što je po konstrukciji vektorski prostor.

Problem najmanjih kvadrata sada se može preformulirati. Procjenitelj najmanjih kvadrata nepoznatog parametra $\boldsymbol{\mu}$ danog u (2.1.4) je vektor $\hat{\boldsymbol{\mu}}$ koji minimizira rezidualnu sumu kvadrata

$$Q_2(\boldsymbol{\mu}) = \|\mathbf{Y} - \boldsymbol{\mu}\|^2 = (\mathbf{Y} - \boldsymbol{\mu})^T(\mathbf{Y} - \boldsymbol{\mu})$$

gdje $\boldsymbol{\mu}$ prolazi vektorskim prostorom $C(\mathbf{X})$ danim u (2.1.8). Metoda najmanjih kvadrata pronalazi točku $\hat{\boldsymbol{\mu}}$ najbližu $\mathbf{Y}$ koja se nalazi u $C(\mathbf{X})$. Takva točka uvijek postoji i jedinstvena je.[2] Ako je $\mathbf{X}^T\mathbf{X}$ invertibilna, što se događa ako i samo ako su stupci od $\mathbf{X}$ linearno nezavisni vektori, tada je rješenje dano s

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \,.$$

Uobičajeno je pisati $\hat{\mathbf{Y}}$ umjesto $\hat{\boldsymbol{\mu}}$, te vrijedi

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \tag{2.1.9}$$

gdje je

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{2.1.10}$$

ortogonalna projekcijska matrica na vektorski prostor $C(\mathbf{X})$, koja se neformalno naziva *kapa matrica* (od engl. *hat matrix*) jer stavlja znak kapa ($\hat{}$) na vektor $\mathbf{Y}$. Komponente od vektora $\hat{\mathbf{Y}}$ nazivaju se **prediktne vrijednosti**.

Često se promatra i ortogonalna projekcijska matrica $\mathbf{M}$ na $C^{\perp}(\mathbf{X})$. Ta matrica je jednaka

$$\mathbf{M} = I - \mathbf{H}$$

[2] To je posljedica Teorema 1.10

za $\mathbf{H}$ kapa matricu. Dakle, za matricu $\mathbf{M}$ vrijedi da je simetrična i idempotentna, što se vidi iz simetričnosti i idempotentnosti kapa matrice:

$$(I-\mathbf{H})^T(I-\mathbf{H}) = (I-\mathbf{H})\,(I-\mathbf{H}) = I-\mathbf{H}-\mathbf{H}+\mathbf{H}^2 =$$
$$= I-\mathbf{H}-\mathbf{H}+\mathbf{H} = I-\mathbf{H} = \mathbf{M}. \tag{2.1.11}$$

2.4 Reziduali

Zbog toga što je $\mathbf{H}$ ortogonalna projekcijska matrica na $C(\mathbf{X})$ vrijedi

$$\mathbf{HX} = \mathbf{X} \tag{2.1.12}$$

iz čega slijedi da

$$\mathbf{HX}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\theta},$$

to jest

$$\mathbf{H}\boldsymbol{\mu} = \boldsymbol{\mu}. \tag{2.1.13}$$

Vektor pogrešaka je $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\theta}$, a vektor reziduala $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = \mathbf{MY}$, gdje se pod **rezidualima** podrazumijevaju procijenjene vrijednosti pogrešaka.

Lema 2.2. *Za vektor reziduala vrijedi* $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}\boldsymbol{\varepsilon}$.

Dokaz. Vrijedi

$$\mathbf{MX}\boldsymbol{\theta} = (I-\mathbf{H})\mathbf{X}\boldsymbol{\theta} = (\mathbf{X}-\mathbf{HX})\boldsymbol{\theta} \overset{(2.1.12)}{=} \mathbf{0}\,.$$

Sada iz toga lako slijedi tvrdnja, jer

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{MY} \overset{(2.1.2)}{=} \mathbf{MX}\boldsymbol{\theta} + \mathbf{M}\boldsymbol{\varepsilon} = \mathbf{M}\boldsymbol{\varepsilon}\,.$$

□

Lema 2.3. *Za rezidualnu sumu kvadrata vrijedi*

$$\|\hat{\boldsymbol{\varepsilon}}\|^2 = \boldsymbol{\varepsilon}^T\mathbf{M}\,\boldsymbol{\varepsilon}. \tag{2.1.14}$$

Dokaz. Tvrdnja slijedi iz svojstva idempotentnosti matrice $\mathbf{M}$ i dva puta primjene leme 2.1:

$$\|\hat{\boldsymbol{\varepsilon}}\|^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}^T\mathbf{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T\mathbf{M}^2\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T\mathbf{M}\boldsymbol{\varepsilon}.$$

□

2.5 Usporedba modela

Neka su zadani

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I) \tag{2.1.15}$$

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I) \tag{2.1.16}$$

normalni linearni modeli gdje je $\mathbf{X}$ $n \times p$ dizajn-matrica ranga r, a $\mathbf{X}_0$ dizajn-matrica čiji je broj stupaca jednak $p_0 < p$. Kažemo da su modeli (2.1.15) i (2.1.16) **ugniježđeni** ako je prostor stupaca jednog modela potprostor drugog, to jest[3]

$$C(\mathbf{X}_0) \subset C(\mathbf{X}).$$

Model (2.1.16) čiji je vektorski prostor potprostor drugog naziva se **restringirani model**, dok se (2.1.15) naziva **puni model**.

Ako se uvedu oznake $\mathbf{H}$ i $\mathbf{H}_0$ za redom kapa matricu punog i restringiranog modela, tada

$$C(\mathbf{X}_0) = C(\mathbf{H}_0) \text{ i } C(\mathbf{X}) = C(\mathbf{H}). \tag{2.1.17}$$

Nadalje, vrijedi

$$C(\mathbf{H}_0) \subset C(\mathbf{H}) \Rightarrow \mathbf{H}\mathbf{H}_0 = \mathbf{H}_0, \tag{2.1.18}$$

jer za proizvoljni $\boldsymbol{y} \in \mathbb{R}^p$, $C(\mathbf{H}_0) \subset C(\mathbf{H})$ povlači $\mathbf{H}_0\boldsymbol{y} \in C(\mathbf{H})$. Sada primjenom Definicije 1.6 a) za ortogonalnu projekcijsku matricu $\mathbf{H}$ na vektorski prostor $C(\mathbf{H})$ je $\mathbf{H}\boldsymbol{v} = \boldsymbol{v}$, $\boldsymbol{v} \in C(\mathbf{H})$ iz čega slijedi $\mathbf{H}\mathbf{H}_0\boldsymbol{y} = \mathbf{H}_0\boldsymbol{y}$.

Često se postavlja pitanje je li restringirani model (2.1.16) istinit ako je puni model (2.1.15) istinit, to jest može li se puni model reducirati na jednostavniji model. Može se pokazati da ukoliko je restringirani model istinit, i puni model je istinit. To slijedi iz

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}_0\boldsymbol{\gamma} \in C(\mathbf{X}_0) \subset C(\mathbf{X})$$

pa je $\mathrm{E}(\mathbf{Y})$ element i od $C(\mathbf{X})$, što znači da vrijedi $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$ za neki vektor θ.

Ono što se nalazi u pozadini testiranja istinitosti restringiranog modela je sljedeće:

1. Ako je restringirani model (2.1.16) istinit, to znači da je $\mathrm{E}(\mathbf{Y})$ jednako u oba modela, to jest da je $\mathbf{H}_0\mathbf{Y}$ blizu $\mathbf{H}\mathbf{Y}$ što znači da je $(\mathbf{H} - \mathbf{H}_0)\mathbf{Y}$ približno nul-vektor.

[3] Sam uvjet $p_0 < p$ da ima manje stupaca u matrici dizajna restringiranog modela nije dovoljan za ugniježđene modele.

2. S druge strane, velika razlika između $\mathbf{HY}$ i $\mathbf{H}_0\mathbf{Y}$ ukazuje na neistinitost restringiranog modela.

3. Odluka o istinitosti restringiranog modela ovisi o tome je li $(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}$ "dovoljno veliko".

4. Kako bi se izmjerila "veličina" od $(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}$ koristi se kvadrirana duljina tog vektora:

$$\|(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}\|^2 = \mathbf{Y}^T(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}. \tag{2.1.19}$$

   Prethodna jednakost slijedi iz činjenice da je $\mathbf{H}-\mathbf{H}_0$ također ortogonalna projekcijska matrica:

$$\begin{aligned}(\mathbf{H}-\mathbf{H}_0)^T(\mathbf{H}-\mathbf{H}_0) &= (\mathbf{H}-\mathbf{H}_0)\ (\mathbf{H}-\mathbf{H}_0)\\ &= \mathbf{H}^2-\mathbf{H}\mathbf{H}_0-\mathbf{H}_0\mathbf{H}+\mathbf{H}_0^2\\ &\overset{(2.1.18)}{=} \mathbf{H}^2-\mathbf{H}_0-(\mathbf{H}\mathbf{H}_0)^T+\mathbf{H}_0^2\\ &\overset{(2.1.18)}{=} \mathbf{H}^2-\mathbf{H}_0-\mathbf{H}_0+\mathbf{H}_0^2\\ &= \mathbf{H}-\mathbf{H}_0 \end{aligned} \tag{2.1.20}$$

   gdje su prva i zadnja jednakost posljedica idempotentnosti i simetričnosti kapa matrica $\mathbf{H}$ i $\mathbf{H}_0$.

5. Ako je restringirani model točan tada vrijedi

$$\mathrm{E}\Big(\frac{\mathbf{Y}^T(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}}{r(\mathbf{H}-\mathbf{H}_0)}\Big) = \sigma^2$$

   što slijedi iz propozicije 1.12 uzimanjem $\mathbf{A}=\mathbf{H}-\mathbf{H}_0$, koristeći da zbog idempotentnosti matrice $\mathbf{H}-\mathbf{H}_0$ vrijedi $\mathrm{tr}(\mathbf{H}-\mathbf{H}_0)=\mathrm{r}(\mathbf{H}-\mathbf{H}_0)$, te da za oba modela vrijedi (2.1.13).

6. Intuitivno, ako je $\mathbf{Y}^T(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}/r(\mathbf{H}-\mathbf{H}_0)$ blizu σ^2 to bi sugeriralo na istinitost restringiranog modela, a ako je ta statistika puno veća od σ^2, na neistinitost.

7. Zbog toga što σ^2 nije poznata umjesto nje uzima se njezin nepristrani procjenitelj:

$$MSE = \frac{(\mathbf{Y}-\mathbf{HY})^T(\mathbf{Y}-\mathbf{HY})}{\mathrm{r}(I-\mathbf{H})}.$$

To je nepristrani procjenitelj jer vrijedi

$$\begin{aligned}
&\mathrm{E}[(\mathbf{Y}-\mathbf{HY})^T(\mathbf{Y}-\mathbf{HY})] = \\
&\quad = \mathrm{E}[((I-\mathbf{H})\mathbf{Y})^T((I-\mathbf{H})\mathbf{Y})] = \mathrm{E}[\mathbf{Y}^T(I-\mathbf{H})^T(I-\mathbf{H})\mathbf{Y})] = \\
&\quad = \mathrm{E}[\mathbf{Y}^T(I-\mathbf{H})\mathbf{Y})] \overset{\text{(a)}}{=} \mathrm{E}[\mathrm{tr}(\mathbf{Y}^T(I-\mathbf{H})\mathbf{Y})] \overset{\text{(b)}}{=} \\
&\quad \overset{\text{(b)}}{=} \mathrm{E}[\mathrm{tr}((I-\mathbf{H})\mathbf{Y}\mathbf{Y}^T)] \overset{\text{(c)}}{=} \mathrm{tr}(\mathrm{E}[(I-\mathbf{H})\mathbf{Y}\mathbf{Y}^T]) = \\
&\quad = \mathrm{tr}((I-\mathbf{H})\mathrm{E}[\mathbf{Y}\mathbf{Y}^T])) = \mathrm{tr}((I-\mathbf{H})(\mathrm{Var}(\mathbf{Y}) + \mathrm{E}[\mathbf{Y}]\mathrm{E}[\mathbf{Y}]^T)\overset{(2.1.4)}{=} \\
&\quad \overset{(2.1.4)}{=} \mathrm{tr}((I-\mathbf{H})(I\sigma^2 + \boldsymbol{\mu}\boldsymbol{\mu}^T)) = \mathrm{tr}(I\sigma^2 + \boldsymbol{\mu}\boldsymbol{\mu}^T - \mathbf{H}(I\sigma^2 + \boldsymbol{\mu}\boldsymbol{\mu}^T))\overset{(2.1.13)}{=} \\
&\quad \overset{(2.1.13)}{=} \sigma^2\mathrm{tr}(I-\mathbf{H}) = \sigma^2\mathrm{r}(I-\mathbf{H})
\end{aligned}$$

gdje prvih nekoliko jednakosti slijedi iz svojstava idempotentnosti i simetričnosti matrice $I-\mathbf{H}$, te poznatih svojstava od kojih će se izdvojiti neka svojstva traga:

(a) $\mathrm{tr}(c) = c$, za neki skalar c, pa se to može primijeniti na $\mathbf{Y}^T(I-\mathbf{H})\mathbf{Y}$ jer je to skalar

(b) $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$ za neke matrice $\mathbf{A}$ i $\mathbf{B}$, pa se to može primijeniti za $\mathbf{A} = \mathbf{Y}^T$ i $\mathbf{B} = (I-\mathbf{H})\mathbf{Y}$

(c) očekivanje i trag komutiraju, jer su linearni operatori.

Primijetimo da se ovo moglo dobiti i kao direktna posljedica propozicije 1.12, jer vrijedi:

$$\mathrm{E}[(\mathbf{Y}-\mathbf{HY})^T(\mathbf{Y}-\mathbf{HY})] = \mathrm{E}[\mathbf{Y}^T\mathbf{M}\mathbf{Y})] \overset{\text{Prop1.12}}{=} \mathrm{tr}(\sigma^2\mathbf{M}I) = \sigma^2\mathrm{r}(M).$$

8. Iz svega navedenog prikladna test-statistika je

$$F = \frac{\mathbf{Y}^T(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}}{MSE\cdot\mathrm{r}(\mathbf{H}-\mathbf{H}_0)}. \tag{2.1.21}$$

Teorem 2.4. *Ako su modeli ugniježđeni i ako su oba istinita, tada*

a) $$\frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\sigma^2} \sim \chi^2(r(I-\mathbf{H}))$$

b) $$\frac{\|(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}\|^2}{\sigma^2} \sim \chi^2(r(\mathbf{H}-\mathbf{H}_0)).$$

Dokaz. Neka je $\mathbf{z} := \boldsymbol{\varepsilon}/\sigma$. Iz poznate distribucije od $\boldsymbol{\varepsilon}$ slijedi $\mathbf{z} \sim N(\mathbf{0}, I)$. Primjenom Leme 2.2, Teorema 1.9 i prethodne definicije od $\mathbf{z}$ vrijedi

$$\frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\sigma^2} = \frac{\boldsymbol{\varepsilon}^T \mathbf{M}\, \boldsymbol{\varepsilon}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}^T (I - \mathbf{H})\, \boldsymbol{\varepsilon}}{\sigma^2} = \mathbf{z}^T (I - \mathbf{H})\, \mathbf{z}$$

iz čega slijedi tvrdnja pod *a)* to jest

$$\mathbf{z}^T (I - \mathbf{H})\, \mathbf{z} \sim \chi^2(\mathrm{r}\,(I - \mathbf{H}))\,.$$

Za dokaz tvrdnje *b)* potrebno je primijetiti da iz pretpostavke točnosti oba modela slijedi

$$\mathbf{H}_0\, \boldsymbol{\mu} \overset{(2.1.13)}{=} \boldsymbol{\mu} \overset{(2.1.13)}{=} \mathbf{H}\boldsymbol{\mu}. \qquad (\text{🍃}_2)$$

Primjenom zadnje tvrdnje je

$$\begin{aligned}
\mathbf{Y}^T (\mathbf{H} - \mathbf{H}_0)\, \mathbf{Y} &\overset{(2.1.5)}{=} (\boldsymbol{\mu}^T + \boldsymbol{\varepsilon}^T)(\mathbf{H} - \mathbf{H}_0)\, (\boldsymbol{\mu} + \boldsymbol{\varepsilon}) = \\
&= (\boldsymbol{\mu}^T (\mathbf{H} - \mathbf{H}_0) + \boldsymbol{\varepsilon}^T (\mathbf{H} - \mathbf{H}_0))\, (\boldsymbol{\mu} + \boldsymbol{\varepsilon}) \overset{\text{🍃}_2}{\underset{\text{tm1.7.a}}{=}} \\
&\overset{\text{🍃}_2}{\underset{\text{tm1.7.a}}{=}} (\boldsymbol{\varepsilon}^T (\mathbf{H} - \mathbf{H}_0))\, (\boldsymbol{\mu} + \boldsymbol{\varepsilon}) = \\
&= \boldsymbol{\varepsilon}^T (\mathbf{H} - \mathbf{H}_0)\boldsymbol{\mu} + \boldsymbol{\varepsilon}^T (\mathbf{H} - \mathbf{H}_0)\boldsymbol{\varepsilon} \overset{\text{🍃}_2}{=} \boldsymbol{\varepsilon}^T (\mathbf{H} - \mathbf{H}_0)\boldsymbol{\varepsilon}. \qquad (\text{🍃})
\end{aligned}$$

Prema tome, budući da je $\mathbf{H} - \mathbf{H}_0$ ortogonalna projekcijska matrica jasno je da vrijedi *b)* jer iz

$$\frac{\|(\mathbf{H} - \mathbf{H}_0)\mathbf{Y}\|^2}{\sigma^2} \overset{\text{◑}}{\underset{\text{tm1.7.b}}{=}} \frac{\mathbf{Y}^T (\mathbf{H} - \mathbf{H}_0)\, \mathbf{Y}}{\sigma^2} \overset{\text{🍃}}{=} \frac{\boldsymbol{\varepsilon}^T (\mathbf{H} - \mathbf{H}_0)\boldsymbol{\varepsilon}}{\sigma^2}$$

primjenom Teorema 1.9 slijedi

$$\frac{\boldsymbol{\varepsilon}^T (\mathbf{H} - \mathbf{H}_0)\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2(\mathrm{r}\,(\mathbf{H} - \mathbf{H}_0))\,.$$

□

Lema 2.5. *Vrijedi*

$$\frac{\|(I - \mathbf{H})\mathbf{Y}\|^2}{\sigma^2} \sim \chi^2(r(I - \mathbf{H})).$$

Dokaz. Ovo je direktna posljedica Teorema 2.3 pod *a)*, jer vrijedi

$$\begin{aligned}
&\|(I - \mathbf{H})\mathbf{Y}\|^2 \overset{\text{◑}}{\underset{\text{tm1.7.b}}{=}} \mathbf{Y}^T (I - \mathbf{H})\, \mathbf{Y} \overset{(2.1.5)}{=} (\boldsymbol{\mu}^T + \boldsymbol{\varepsilon}^T)(I - \mathbf{H})\, (\boldsymbol{\mu} + \boldsymbol{\varepsilon}) = \\
&= (\boldsymbol{\mu}^T (I - \mathbf{H}) + \boldsymbol{\varepsilon}^T (I - \mathbf{H}))\, (\boldsymbol{\mu} + \boldsymbol{\varepsilon}) \overset{\text{🍃}_2}{\underset{\text{tm1.7.a}}{=}} (\boldsymbol{\varepsilon}^T (I - \mathbf{H}))\, (\boldsymbol{\mu} + \boldsymbol{\varepsilon}) = \\
&= \boldsymbol{\varepsilon}^T (I - \mathbf{H})\boldsymbol{\mu} + \boldsymbol{\varepsilon}^T (I - \mathbf{H})\boldsymbol{\varepsilon} \overset{\text{🍃}_2}{=} \boldsymbol{\varepsilon}^T (I - \mathbf{H})\boldsymbol{\varepsilon} = \|\hat{\boldsymbol{\varepsilon}}\|^2.
\end{aligned}$$

□

3 Uvod u dvofaktorsku ANOVA-u

3.1 Motivacija

Analizu varijance, često zvanu skraćeno ANOVA,[4] ponekad i F-test, razvio je Ronald A. Fischer. Mnoge praktične situacije zahtjevaju da se promatra utjecaj varijabli poticaja (ovdje zvanih *faktori* jer nisu neprekidne nego kategorijske) na ishod određene pojave (varijable odziva)[5], što će se slično kao kod regresije i ovdje promatrati. Svaki od faktora obično ima k ($k > 1$) različitih vrijednosti koje se nazivaju *razine*. Dakle, uzima se slučajni uzorak subjekata iz populacije i podijeli prema tome kojoj razini pripadaju za svaki od faktora (tu se dobiju različite grupe koje se često nazivaju *ćelije*).

Općenito ANOVA ima više varijacija. One mogu biti u odnosu na broj faktora koji utječu na varijablu odziva (jednofaktorska, dvofaktorska i općenito multifaktorska ANOVA), u odnosu na broj zavisnih varijabli, te testira li se više puta na istim subjektima ili samo jednom (dizajn unutar subjekata ili dizajn ponovljenih mjerenja i dizajn između subjekata).

U ovom radu će se razmatrati dizajn između subjekata dvofaktorske ANOVA-e s jednom varijablom odziva.[6] Pretpostavljat će se da je uzorak izabran na slučajan način, te da su dodatno i subjekti unutar svake ćelije izabrani na slučajan način. Zbog toga što se takav dizajn može dosta iskomplicirati, biti će riječ o jednom slučaju: balansiranom ili ortogonalnom dizajnu analize varijance. Za njega se dodatno pretpostavlja da ćelije imaju jednak broj elemenata. Taj slučaj se najviše koristi jer ima mnoga lijepa svojstva.

3.2 Poveznica s t-testom

Analiza Varijance je usko vezana uz t-test. Glavna razlika je da t-test mjeri razliku očekivanja dvaju grupa, dok ANOVA testira razliku očekivanja dvaju ili više grupa (razina). Dakle, u slučaju kada jednofaktorska ANOVA ima dvije razine faktora rezultat je jednak kao kod t-testa.

Teoretski, t-test bi se mogao koristiti i za više grupa, no problem je što dolazi do pogreške tipa I: odbacivanja nul-hipoteze kada je ona istinita. Naime, kada bi bile četiri grupe koje treba usporediti, trebalo bi šest puta koristiti t test ($1-2$, $1-3$, $1-4$, $2-3$, $2-4$ i $3-4$). Uz razinu značajnosti $\alpha = 0.05$ za

[4] od engl. *ANalysis Of VAriance*

[5] Pojam faktora se kod jednofaktorske analize varijance često naziva i nezavisna varijabla, a varijable odziva zavisna varijabla, no mi ćemo taj naziv izbjegavati jer se javlja u više različitih konteksta.

[6] Za više vidi [4].

svaku usporedbu ukupna vjerojatnost da dođe do pogreške tipa I je približno $6 \times 0.05 = 0.30$, što je prevelika greška za bilo koji model.

3.3 Poveznica s regresijom

Statički modeli su obično definirani u terminima parametara, te je čest slučaj da se dva modela mogu napisati u naizgled različitim formama a zapravo biti *ekvivalentni*. To je upravo slučaj s regresijom i ANOVA-om.

Ugrubo govoreći, ANOVA je statistički model koji se koristi za predviđanje neprekidne varijable, ukoliko je poznata jedna ili više kategorijskih (to jest nominalnih) varijabli. Regresija je također statistički model koji se koristi za predviđanje neprekidne varijable, ali uz pomoć jedne ili više neprekidnih varijabli. Neki smatraju da je ANOVA poseban slučaj regresije, jer se na dosta intuitivan način iz regresije može doći do ANOVA-e uvođenjem takozvanih *indikatorskih varijabli*. Indikatorska varijabla je jednaka jedan za točno jednu razinu svake kategorijske varijable, a u ostalima je jednaka nula. Primjerice, ukoliko postoji kategorijska varijabla s k razina u jednofaktorskoj ANOVA-i, tada će biti $k-1$ indikatorskih varijabli (zadnji indikator je uvijek višak jer je suma indikatorskih varijabli za određenu razinu uvijek jednaka jedan) u regresiji i efektivno se dobije jednak rezultat kao u ANOVA modelu.

Napomena 3.1. *Ovdje se pod* ekvivalentnim *modelima misli u smislu da iako određene procjene parametara nisu nužno jednake, suma kvadrata pogrešaka ta dva modela će uvijek biti jednaka.*

3.4 Primjer

Da bi se lakše i jasnije shvatila problematika i kasnije teorijske konstrukcije dvofaktorskog modela analize varijance najprije će se razmotriti tipičan primjer (vidi [5]) s primjenom u psihologiji.

Primjer 3.2. *Profesor statistike provodi eksperiment kako bi usporedio koja od metoda predavanja je efektivnija za nastavu. Metoda A je uobičajna metoda predavanja: nastava, domaće zadaće i završni ispit. Metoda B jednaka je metodi A, samo što profesor svaki tjedan predaje dodatni sat kako bi ilustrirao probleme na primjerima.*
Profesor također želi vidjeti na koji način metode djeluju na studente različitih matematičkih sposobnosti, te ih podijeli na izvanredne, prosječne i slabije. Dakle, u ovom eksperimentu su faktori matematičke sposobnosti i metode predavanja koji će se radi jednostavnosti redom nazivati faktor I i faktor II. Faktor I ima tri razine: izvanredni, prosječni i slabiji studenti, dok faktor II

ima dvije: metoda A i metoda B. Pojava koju se promatra su bodovi na završnom ispitu, to jest želi se saznati da li na bodove utječu spomenuti faktori. Može se promatrati nekoliko hipoteza (kasnije nul-hipoteza):

1. *Različite razine matematičkih sposobnosti ne utječu na razliku u bodovima završnog ispita.*
2. *Različite razine metoda ne utječu na razliku u bodovima završnog ispita.*
3. *Ne postoji interakcija između faktora I i II.*

Na slučajan način je izabrano pet studenata iz svake grupe. Na kraju semestra su zabilježeni sljedeći rezultati (od mogućih 50 bodova) iz završnog ispita:

	Metoda predavanja (faktor II)				
Mat. sposobnost (faktor I)	*Metoda A (1)*		*Metoda B (2)*		*Srednja vrijednost po retcima*
Izvanredni (A)	39 48 44	41 42	49 47 43	47 48	44.8
Prosječni (B)	43 40 42	36 35	38 45 42	46 44	41.1
Slabiji (C)	30 29 37	33 36	37 34 40	41 33	35.0
Srednja vrijednost po stupcima	$38.3\dot{3}$		$42.2\dot{6}$		40.3

Razlog zbog kojeg se grupe (u ovom slučaju od po pet studenata) u dvofaktorskoj ANOVA-i nazivaju *ćelije*, dolazi upravo iz prikaza podataka u tablici. Ukoliko se u svakoj ćeliji nalazi jednak broj podataka (kao što je u ovom primjeru, što je i bilo najavljeno da će se pretpostaviti) govori se o *balansiranoj ANOVA-i*.

Iz formulacije problema je jasna sličnost s regresijom jer se *izlazna* numerička vrijednost, često zvana *opservacija* (ovdje rezultat završnog ispita) neprekidne varijable shvaća kao posljedica određene razine kategorijske varijable.

Sada će se napraviti intuitivna analiza prethodnog primjera pomoću koje će se pokušati odrediti istinitost hipoteza. Prvo će se analizirati nešto jednostavniji slučaj tog primjera kada je broj faktora jednak jedan, a zatim će naglasak biti na dvofaktorskoj analizi, koja je i cilj promatranja.

3.4.1 Glavni efekti i interakcija

Pojam efekta je veoma bitan u analizi varijance. On se općenito odnosi na stupanj promjene varijable odziva pod utjecajem faktora. Pojmovi vezani uz efekte objasniti će se intuitivno i ilustrirati na primjeru 3.2, dok će se za njihove definicije koristiti jednadžbe koje će se izvesti nakon uvođenja modela. Važno je napomenuti da će se kod modela uvesti parametri, dok će sada biti riječ samo o procjenama tih parametara (dobivenih na temelju opservacija iz tablice primjera 3.2) pomoću kojih će se pokušati intuitivno objasniti glavni koncepti.

Glavni efekt faktora je utjecaj tog pojedinog faktora na varijablu odziva, bez prisutnosti drugih utjecaja. Primijetimo da tada model dvofaktorske ANOVA-e ima dva glavna efekta jer postoje dva faktora, dok jednofaktorska ima jedan.

Kada bismo željeli procijeniti glavni efekt faktora II promatrali bi ćelije koje odgovaraju Metodi A i uspoređivali ih s ćelijama koje odgovaraju Metodi B. To je ekvivalentno uspoređivanju srednjih vrijednosti stupaca (bez procjene opće srednje vrijednosti[7]). Dakle, procjena glavnog efekt faktora I iz podataka odgovara razlici srednjih vrijednosti redaka (bez procjene opće srednje vrijednosti), dok za faktor II odgovara razlici srednjih vrijednosti stupaca (bez procjene opće srednje vrijednosti). Na primjer, ako se gleda faktor II (jer je za dvije razine faktora intuitivnije jasniji pojam razlike) procjena glavnog efekta je razlika dviju srednjih vrijednosti redaka ($38.3\dot{3}$ i $42.2\dot{6}$) koja iznosi $3.9\dot{3}$.

Ukoliko bi parametri srednjih vrijednosti svih razina nekog faktora bili jednaki, govorilo bi se o **nepostojanju** (glavnog) **efekta faktora**, odnosno ukoliko bi se testiranjem pokazalo da razlike srednjih vrijednosti svih razina nekog faktora **nisu značajno** jednake nuli o (glavnom) efektu koji nije značajan (to se obično testira nul-hipotezom).

U istim terminima bi se moglo govoriti o **efektu** određene **razine** promatranog faktora što je razlika opće srednje vrijednosti i pripadne srednje

[7]Ukupna aritmetička sredina podataka ovdje će se nazivati *procjena opće srednje vrijednosti* (zadnji element zadnjeg stupca ili retka tablice u primjeru), gdje se pojam opće srednje vrijednosti odnosi na parametar u populaciji.

vrijednosti te razine.[8]

Za razliku od jednofaktorskog modela, u dvofaktorskom modelu ANOVA-e pojavljuje se i pojam interakcije. Ona postoji ukoliko glavni efekti faktora nisu nezavisni. U ovisnosti o interakciji govoriti će se o različitim modelima: o **dvofaktorskom aditivnom modelu** ukoliko ne postoji interakcija, a inače o **općem dvofaktorskom modelu.**

Ako postoji interakcija, to zapravo znači da utjecaj jednog faktora ovisi o razinama drugog faktora. Važno je primijetiti da interakcija ili postoji ili ne postoji, odnosno ukoliko utjecaj jednog faktora ovisi o razinama drugog faktora, tada i utjecaj drugog faktora ovisi o razinama prvog faktora. Također se može primijetiti da nije moguće govoriti o interakciji pojedinih razina, već samo o interakciji faktora, jer se ona odnosi isključivo na faktore.

Pojam interakcije se može bolje shvatiti uvođenjem pojma jednostavnog glavnog efekta, koji se kao i pojam interakcije pojavljuje u dvofaktorskoj analizi varijance.

Jednostavni glavni efekt faktora je glavni efekt tog faktora na određenoj razini drugog faktora. Na primjer, procjena jednostavnog glavnog efekta faktora II po razini I_1 (Izvanredni studenti) je 4 $(42.8 - 46.8)$, po razini I_2 je 3.8 $(39.2 - 43.0)$, a po razini I_3 je 4 $(37.0 - 33.0)$.[9]

Sada se može reći da su u slučaju postojanja interakcije jednostavni glavni efekti jednog faktora različiti po razinama drugog faktora, odnosno da je glavni efekt određene razine različit od jednostavnih glavnih efektata te razine. Zbog toga što nema prevelike razlike između jednostavnog glavnog efekta faktora II po različitim razinama faktora I (dobilo se 4, 3.8 i 4) to bi moglo ukazivati na nepostojanje interakcije. To je također vidljivo i uspoređivanjem s glavnim efektom faktora II gdje se dobilo $3.9\dot{3}$.

U slučaju dvofaktorskog aditivnog modela bez utjecaja jednog od faktora u primjeru 3.2 (na primjer faktora I), riječ je o jednofaktorskom modelu analize varijanci. To je najjednostavniji model i najbolje je prvo njega analizirati kako bi bilo jasan način na koji se dolazi do općenitijeg modela.

3.4.2 Jednofaktorska analiza varijance

Općenito, nul-hipoteza jednofaktorske analize varijance tvrdi da različite razine faktora imaju isti utjecaj na promatranu pojavu to jest da su razlike srednjih vrijednosti uzorka posljedica oscilacija radi slučajnosti uzorka, dok je alternativna hipoteza da su razlike srednjih vrijednosti uzorka posljedica

[8] Pokazuje se da je promatranje utjecaja glavnog efekta faktora usko vezano uz promatranje utjecaja pojedinih razina faktora.

[9] U zagradama su dane srednje vrijednosti ćelija koje će biti prikazane nešto kasnije u tablici 3

stvarne razlike u očekivanjima populacija.

To je tipičan problem analize varijance i cijela teorija se uglavnom sastoji od objašnjavanja postupaka za njegovo rješavanje. Budući da se ta teorija uglavnom bavi analizom odstupanja opservacija od srednjih vrijednosti, teorija je i dobila naziv *analiza varijance*.

Nimalo iznenađujuće, pretpostavke koje trebaju za korištenje ANOVA-e jednake su onima koje su trebale za t-test (to je jasno iz pretpostavki dizajna modela). Treba se pretpostaviti:

1) slučajni, nezavisni uzorci iz populacija
2) normalna distribuiranost populacija
3) jednakost varijanci populacija (tzv. *homogenost*).

Kod primjera su populacije svi studenti pojedine razine, a uzorci su dane opservacije koje su izabrane na slučajan način (pa zadovoljen uvjet 1)). Ovdje će se gledati pojednostavljeni primjer 3.2, gdje će se smatrati da samo jedan faktor (primjerice faktor I) djelujuje na zavisnu varijablu (rezultat završnog ispita). Radi jednostavnosti skraćeno će se o razinama govoriti kao o razini A, B i C umjesto redom Izvanredni, Prosječni i Slabiji.

Zbog sličnosti s modelom regresije, model se po (2.1.4) može napisati kao

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

to jest po elementima kao

$$\begin{aligned} Y_{iA} &= \mu_A + \varepsilon_{iA}, \quad i = 1, \dots, 10 \\ Y_{iB} &= \mu_B + \varepsilon_{iB}, \quad i = 1, \dots, 10 \\ Y_{iC} &= \mu_C + \varepsilon_{iC}, \quad i = 1, \dots, 10, \end{aligned} \tag{3.4.1}$$

za μ_A, μ_B i μ_C parametre srednjih vrijednosti razina.

U slučaju opservacija u primjeru, treći broj prvog retka 49 može shvatiti kao zbroj srednje vrijednosti tog retka (to jest 44.8) i ostatka do tog broja. Pretpostavlja se da ti ostaci dolaze od slučajnih varijabli koje se nazivaju slučajne pogreške. U ovakvim modelima je uobičajeno pretpostaviti da su te slučajne varijable, koje možemo označiti $\boldsymbol{\varepsilon_A}$, $\boldsymbol{\varepsilon_B}$ i $\boldsymbol{\varepsilon_C}$, nezavisne slučajne varijable sa zajedničkom normalnom razdiobom $N(0, \sigma^2)$, to jest slučajni vektor pogrešaka je $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$ (čime su zadovoljeni uvjeti 2) i 3)).

Sada se nul-hipoteza o nepostojanju značajne razlike u rezultatima završnog ispita formulira kao

$$H_0: \quad \mu_A = \mu_B = \mu_C,$$

što znači da ne postoji efekt faktora.

Sada se mogu preciznije defnirati efekti razina:

$$\delta_A = \mu_{opća} - \mu_A, \quad \delta_B = \mu_{opća} - \mu_B, \quad \delta_C = \mu_{opća} - \mu_C, \tag{3.4.2}$$

a nul-hipoteza se može zapisati kao:

$$H_0: \quad \delta_A = \delta_B = \delta_C = 0,$$

za po formulama $\mu_{opća} = \frac{1}{n}(n_A\mu_A + n_B\mu_B + n_C\mu_C)$ i $n = n_A + n_B + n_C$, ako n_A, n_B i n_C brojevi elemenata po redom razinama A, B i C (u primjeru 3.2 su se za te brojeve uzeli $n_A = n_B = n_C = 10$).

Iz formule $\mu_{opća}$ i definicija (3.4.2) slijedi:

$$\delta_A + \delta_B + \delta_C = 0. \tag{3.4.3}$$

Sada se model (3.4.1) može zapisati kao (i ti modeli su ekvivalentni što će biti pokazano u općenitijem slučaju kasnije):

$$\begin{aligned} Y_{Ai} &= \mu_{opća} + \delta_A + \varepsilon_{Ai}, \quad i = 1, \ldots, 10 \\ Y_{Bi} &= \mu_{opća} + \delta_B + \varepsilon_{Bi}, \quad i = 1, \ldots, 10 \\ Y_{Ci} &= \mu_{opća} + \delta_C + \varepsilon_{Ci}, \quad i = 1, \ldots, 10. \end{aligned} \tag{3.4.4}$$

Model (3.4.4) se u općenitijem obliku (za proizvoljnih n razina faktora) često naziva i *model efekata* ([3]).

Kako bi se odredilo jesu li opservacije različitih razina iz iste distribucije ili nisu, treba primijetiti da se ukupno odstupanje (u primjeru rezultata završnog ispita) može rastaviti na odstupanje opservacija od srednje vrijednosti njihove razine i odstupanje srednjih vrijednosti razina od opće srednje vrijednosti (to jest efekata razina). Na primjer za razinu A, ako je $\hat{\mu}_{opća}$ procjena opće srednje vrijednosti, $\hat{\mu}_A$ procjena srednje vrijednosti razine A, te y_{ij} opservacije u tablici za koje se pretpostavlja da su nastale mjerenjem slučajne varijable Y_{ij}, $i \in \{A, B, C\}$, $j = 1, \ldots, 10$:

$$y_{Ai} - \hat{\mu}_{opća} \quad = \quad y_{Ai} - \hat{\mu}_A \quad + \quad \hat{\mu}_A - \hat{\mu}_{opća}, \quad i = 1, \ldots, 10$$

Kada bi se prosumiralo po svim razinama dobio bi se rastav[10]

$$SS_T \quad = \quad SS_W \quad + \quad SS_B,$$

[10] Ovaj rastav ne bi bio moguć bez pretpostavke o balansiranom dizajnu, to jest jednakom broju elemenata u svakoj ćeliji (vidi [3]).

gdje se SS_T (engl. *total sum of squares*) naziva *ukupna suma kvadrata*, SS_W (engl. *within-groups sum of squares*) *suma kvadrata unutar grupa* (to jest razina) ili *suma kvadrata pogrešaka* (engl. *sum of square error*) i SS_B (engl. *between-groups sum of squares*) *suma kvadrata između grupa.*[11] Upravo to što se ukupna varijabilnost podataka može na ovako jednostavan način rastaviti je bitno svojstvo ANOVA-e.

Formule za opservacije primjera 3.2 su sljedeće:

$$\begin{aligned}
SS_T &= \sum_{ij} \left(y_{ij} - \hat{\mu}_{opća} \right)^2 \\
SS_W &= SS_{A_{razina}} + SS_{B_{razina}} + SS_{C_{razina}} \\
&= \sum_i \left(y_{Ai} - \hat{\mu}_A \right)^2 + \sum_i \left(y_{Bi} - \hat{\mu}_B \right)^2 + \sum_i \left(y_{Ci} - \hat{\mu}_C \right)^2 \\
SS_B &= n \left(\left(\hat{\mu}_A - \hat{\mu}_{opća} \right)^2 + \left(\hat{\mu}_B - \hat{\mu}_{opća} \right)^2 + \left(\hat{\mu}_C - \hat{\mu}_{opća} \right)^2 \right). \qquad (3.4.5)
\end{aligned}$$

Važno je primijetiti da SS_W ne ukazuje na to postoji li efekt faktora, jer se svaka opservacija gleda u odnosu na srednju vrijednost svoje grupe. S druge strane, SS_B ukazuje na postojanje efekta faktora, jer što je veći efekt faktora, to se više međusobno razlikuju srednje vrijednosti razina, što znači da će se više razlikovati od opće srednje vrijednosti.

Nadalje, uvode se:

$$s_W^2 = \frac{SS_W}{df_W} \qquad s_B^2 = \frac{SS_B}{df_B},$$

koje se redom nazivaju *procjena varijance unutar grupa*[12] i *procjena varijance između grupa*,[13] za $df_W = n - k$ *broj stupnjeva slobode unutar grupa* (ukupan broj opservacija minus broj grupa), $df_B = k - 1$ *broj stupanjeva slobode između grupa* (broj grupa minus jedan), te k broj razina (u ovom primjeru $k = 3$, a $n = 15$). Intuitivno su s_W^2 i s_B^2 redom odstupanje srednjih vrijednosti razina od opće srednje vrijednosti i odstupanje opservacija određene razine od prosjeka te razine. Iz njihove formule i prethodne napomene, jasno je da s_B^2 ukazuje na postojanje ili nepostojanje efekta faktora, dok s_W^2 ne ukazuje.

[11]Primijetimo da su sve ove oznake dane za procjenitelje, odnosno *statistike*, a ne parametre.

[12]Često se označava i sa MS_B od engl. *mean square between.*

[13]Često se označava i sa MS_W od engl. *mean square within* ili sa MS_E od engl. *mean square error.*

To znači da statistika koja se u literaturi naziva *empirijski F-omjer* (ili kraće *F-omjer*) i označava

$$F_{obt} = s_B^2 / s_W^2$$

je dobar indikator za odlučivanje o tome treba li odbaciti nul-hipotezu ili ne. Dakle, jer se s_B^2 povećava s efektom faktora, a s_W^2 ostaje isti, što je veći F-omjer to je manje vjerojatna nul-hipoteza.

Kao što je prethodno rečeno, dosta veliki F-omjer bi trebao nagovijestiti odbacivacivanje nul-hipoteze. Egzaktnije, ako je vrijednost statistike F_{obt} veća od F_{crit} koja se obično iščitava iz tablica, odbacuje se H_0, to jest ukratko:

$$\text{Ako } F_{obt} \geq F_{crit} \text{ odbija se hipoteza } H_0\,.$$

$$\text{Ako } F_{obt} \leq F_{crit} \text{ ne odbija se hipoteza } H_0\,. \tag{3.4.6}$$

U danom primjeru dobije se (uobičajeno je da se cjelokupni postupak prikazuje u tzv. *ANOVA tablici jednofaktorskog modela*)[14]:

The ANOVA Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	66.7720000	33.3860000	7.14	0.0091
Error	12	56.1280000	4.6773333		
Corrected Total	14	122.9000000			

Slika 1: ANOVA tablica za jednofaktorski model

Za nivo značajnosti $\alpha = 0.05$ je $F_{crit} = 3.89$ za $df_B = 2$ i $df_W = 12$. Dakle, $F_{crit} \leq 7.14 = F_{obt}$ što znači da se odbija hipoteza H_0.

Zaključak je da postoji značajna razlika u rezultatima završnog ispita između razina, to jest razine nisu iz iste populacije, i ekvivalentno u terminima efekata: efekt faktora I (matematička sposobnost) značajno utječe na rezultat završnog ispita.

To se moglo i direktno zaključiti iz tablice promotrivši zadnji stupac, to jest p-vrijednost. Niska p-vrijednost ukazuje na to da je malo vjerojatno da su grupe iz iste populacije. U ovom slučaju nivo značajnosti je $\alpha = 0.05$ pa pošto $0.0091 < 0.05$ to znači da treba odbaciti hipotezu H_0.

U originalnom problemu primjera 3.2 postoje dva faktora i iako je ovo

[14] Ovdje je korištena procedura `proc anova` programskog jezika SAS, ali se postupak mogao izvesti i u bilo kojem drugom programskom jeziku.

		Faktor II					
		II_1	$\dots$	II_j	$\dots$	II_{m_2}	
Faktor I	I_1	μ_{11}	$\dots$	μ_{1j}	$\dots$	μ_{1m_2}	$\mu_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	I_i	μ_{i1}	$\dots$	μ_{ij}	$\dots$	μ_{im_2}	$\mu_{i\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	I_{m_1}	μ_{m_11}	$\dots$	μ_{m_1j}	$\dots$	$\mu_{m_1m_2}$	$\mu_{m_1\cdot}$
		$\mu_{\cdot 1}$	$\dots$	$\mu_{\cdot j}$	$\dots$	$\mu_{\cdot m_2}$	μ

Tablica 1: Tablica parametara srednjih vrijednosti

temelj razmišljanja doći će do bitnih promjena, pogotovo u slučaju općeg modela s interakcijom.

3.4.3 Uvod u dvofaktorski model

Ovaj model je nešto kompliciraniji nego jednofaktorski, upravo zbog više faktora i mogućnosti interakcije. Iz tog razloga će se prvo napraviti analiza za općenit slučaj, a zatim će se to primijeniti na primjer 3.2.

Neka faktor I općenito ima m_1 ($m_1 \geq 2$), a faktor II m_2 ($m_2 \geq 2$) razina, te neka u svakoj ćeliji ima l ($l \geq 1$) opservacija, što znači da se raspolaže s ukupno $n = m_1 m_2 l$ opservacija Y_{ijk} ($i = 1, \dots, m_1, j = 1, \dots, m_2, k = 1, \dots, l$) slučajne varijable Y. Sada su bitne srednje vrijednosti za svaku ćeliju koje će biti označene u skladu s tablicom 1, gdje su

$$\mu_{i\cdot} = \frac{1}{m_2} \sum_{j=1}^{m_2} \mu_{ij}, \quad i = 1, \dots, m_1$$

$$\mu_{\cdot j} = \frac{1}{m_1} \sum_{i=1}^{m_1} \mu_{ij}, \quad j = 1, \dots, m_2$$

parametri srednje vrijednosti i-tog retka, odnosno j-tog stupca, dok je μ parametar opće srednje vrijednosti kao i prije, to jest vrijedi

$$\mu = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mu_{ij} \,.$$

Tada se model može postaviti kao

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, m_1, \quad j = 1, \ldots, m_2, \quad k = 1, \ldots, l \qquad (3.4.7)$$

za μ_{ij} parametri srednjih vrijednosti ćelija iz tablice 1.

Kao i prije pretpostavlja se da su ε_{ijk} nezavisne slučajne varijable sa zajedničkom normalnom razdiobom $N(0, \sigma^2)$, iz čega je $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$.

Nul-hipoteze za ovaj općenit slučaj (u skladu s nul-hipotezama primjera):

$$H_{12}: \quad (\mu_{ij} - \mu) - (\mu_{i\cdot} - \mu) - (\mu_{\cdot j} - \mu) = 0, \; i = 1, \ldots, m_1, \; j = 1, \ldots, m_2$$

$$H_{10}: \quad \mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{i\cdot} = \ldots = \mu_{m_1\cdot}$$

$$H_{02}: \quad \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot j} = \ldots = \mu_{\cdot m_2},$$

gdje intuitivno H_{12} znači da ne postoji interakcija između faktora I i II, H_{10} da različite razine faktora I ne utječu na promatranu izlaznu veličinu, a H_{02} da različite razine faktora II ne utječu na promatranu varijablu odziva.

Kako bi se odredio model efekata (slično kao u (3.4.4) jednofaktorskog modela, samo za općeniti slučaj) uvode se sljedeće definicije za *glavni efekt i-te razine faktora* I, te za $\delta_{\cdot j}$ *glavni efekt j-te razine faktora* II:

$$\delta_{i\cdot} = \mu_{i\cdot} - \mu, \quad i = 1, \ldots, m_1$$

$$\delta_{\cdot j} = \mu_{\cdot j} - \mu, \quad j = 1, \ldots, m_2 \,.$$

Analogno kao u jednofaktorskom slučaju u (3.4.3) vrijedi:

$$\sum_{i=1}^{m_1} \delta_{i\cdot} = \sum_{j=1}^{m_2} \delta_{\cdot j} = 0 \,.$$

No sada postoji i efekt δ_{ij} koji karakterizira doprinos međusobne interakcije i-te razine faktora I i j-te razine faktora II. *Interakcija faktora* I *i faktora* II, definira se

$$\delta_{ij} = \mu_{ij} - (\mu + \delta_{i\cdot} + \delta_{\cdot j}), \quad i = 1, \ldots, m_1, \quad j = 1, \ldots, m_2, \qquad (3.4.8)$$

što je ekvivalentno

$$\delta_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu, \quad i = 1, \ldots, m_1, \quad j = 1, \ldots, m_2 \,.$$

Iz prethodne definicije δ_{ij} i definicije srednjih vrijednosti pokazuje se da također vrijedi:

$$\sum_{i=1}^{m_1} \delta_{ij} = \sum_{j=1}^{m_2} \delta_{ij} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \delta_{ij} = 0. \qquad (3.4.9)$$

Sada se iz (3.4.7) i (3.4.8) može postaviti model efekata

$$Y_{ijk} = \mu + \delta_{i\cdot} + \delta_{\cdot j} + \delta_{ij} + \varepsilon_{ijk}, \tag{3.4.10}$$

za svaki $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, $k = 1, \ldots, l$, koji je ekvivalentan modelu (3.4.7). Na prvi pogled nije jasno zbog čega bi ta dva modela bila ekvivalentna, jer nemaju jednak broj parametara. Model u jednadžbi (3.4.7) ima $m_1 m_2$ parametra, a model u jednadžbi (3.4.10) $1 + m_1 + m_2 + m_1 m_2$ parametara. No, iz jednadžbi (3.4.8) i (3.4.9) se vidi da parametri modela efekata nisu svi linearno nezavisni, već se jedni mogu prikazati preko drugih. U skladu s tim, zapravo u jednadžbi (3.4.10) ima $1 + (m_1 - 1) + (m_2 - 1) + (m_1 - 1)(m_2 - 1) = m_1 m_2$ parametara, što je isto kao za (3.4.7).

Nul-hipoteze modela efekata su:

$$\begin{aligned} H_{12}: &\quad \delta_{ij} = 0, \quad i = 1, \ldots, m_1, \quad j = 1, \ldots, m_2 \\ H_{10}: &\quad \delta_{1\cdot} = \delta_{2\cdot} = \ldots = \delta_{m_1\cdot} = 0 \\ H_{02}: &\quad \delta_{\cdot 1} = \delta_{\cdot 2} = \ldots = \delta_{\cdot m_2} = 0, \end{aligned}$$

pri čemu se za svaku od njih uzima kao alternativna da barem na jednom mjestu vrijedi znak nejednakosti. Na H_{12} se može gledati kao na hipotezu da ne postoji interakcija faktora I i II, na H_{10} da ne postoji glavni efekt faktora I, a na H_{02} da ne postoji glavni efekt faktora II. Važno je naglasiti da je promatranje hipoteza H_{10} i H_{02} od interesa jedino ako se testiranjem H_{12} utvrdi da interakcija ne postoji, jer inače utjecaj glavnih faktora ovisi i o interakciji. To znači da je svaka od tih hipoteza ekvivalentna hipotezi koja se promatrala u jednofaktorskom modelu, jer se zapravo odnose na aditivni model.

Kako bi se odlučilo treba li odbaciti nul-hipotezu ili ne, u jednofaktorskom modelu je bilo važno odrediti dvije statistike varijance: s_B^2 i s_W^2. Do zaključka o odbacivanju nul-hipoteze (za pojednostavljenje primjera 3.2) došlo se testirajući jesu li te statistike procjene istog parametra. Ako jesu (to jest ako nema efekta faktora, pa s_B^2 ostaje nepromjenjen) tada se ne odbacuje nul-hipoteza.

U analizi dvofaktorske ANOVA-e potrebno je umjesto dvije odrediti četiri statistike varijance: s_R^2, s_C^2, s_{RC}^2 i s_W^2. Statistika s_W^2 odgovara procjenitelju varijance unutar grupa u jednofaktorskom modelu koja je neosjetljiva na utjecaj efekata. U sve tri hipoteze u dvofaktorskom modelu ona će se uspoređivati s preostalima kako bi se odlučilo o tome treba li te hipoteze odbaciti. Ostale statistike biti će osjetljive na efekte faktora.

Statistika s_R^2 se naziva *procjenitelj varijance redaka*. Ona se bazira na varijabilnosti srednjih vrijednosti redaka, te je stoga osjetljiva na glavni efekt faktora II. Statistika s_C^2 se naziva *procjenitelj varijance stupaca.* Ona se bazira na varijabilnosti srednjih vrijednosti stupaca, te je stoga osjetljiva na glavni efekt faktora I. Statistika s_{RC}^2 se naziva *procjenitelj varijance redaka po stupcima.* Ona se bazira na varijabilnosti srednjih vrijednosti ćelija, te je stoga osjetljiva na interakciju faktora I i II.

Ukoliko ne postoji glavni efekt faktora I (to jest ako je H_{10} istinita hipoteza), s_R^2 je nepristrani procjenitelj od σ^2.[15] To je intuitivno jasno jer ako H_{10} nije istinita postojati će veće razlike između srednjih vrijednosti razina faktora I (to jest srednjih vrijednosti redaka), pa procjenitelj varijance redaka neće biti dobar procjenitelj varijance populacije. Također, ako ne postoji glavni efekt faktora II (H_{02} istinita), s_C^2 je nepristrani procjenitelj od σ^2, a ako nema interakcije između faktora I i II (H_{12} istinita), s_{RC}^2 je nepristrani procjenitelj od σ^2. Dakle, mogu se formirati F-omjeri:

$$\begin{array}{ll} \text{Za faktor II,} & F_{obt} = s_R^2/s_W^2 \\ \text{Za faktor I,} & F_{obt} = s_C^2/s_W^2 \\ \text{Za interakciju faktora I i II,} & F_{obt} = s_{RC}^2/s_W^2. \end{array}$$

Kao i kod jednofaktorske analize varijance, svaka vrijednost F_{obt} se gleda u odnosu na F_{crit}. Za slučaj po retcima, ako je $F_{obt} \geq F_{crit}$, to znači da postoji glavni efekt faktora II, za slučaj po stupcima da postoji glavni efekt faktora I, a za retke po stupcima da postoji interakcija.

Kako bi se došlo do konkretnih formula za statistike s_R^2, s_C^2, s_{RC}^2 i s_W^2 potrebna su slična razmatranja kao za jednofaktorski model.

3.4.4 Procjenitelj varijance unutar grupa (s_W^2)

Ovaj procjenitelj je izveden iz varijabilnosti opservacija unutar svake ćelije. Zbog toga što na podatke u ćeliji utječu jednake razine faktora I i II, varijabilnost razina nije prisutna. Njegova formula glasi:

$$s_W^2 = \frac{SS_W}{df_W},$$

za SS_W sumu kvadrata unutar grupa, a df_W broj stupnjeva slobode unutar grupa. Suma kvadrata unutar grupa (SS_W) je zbroj suma kvadrata za svaku ćeliju, to jest vrijedi

$$SS_W = SS_{11} + SS_{12} + \cdots + SS_{m_1 m_2},$$

[15] Kažemo da je statistika koja se koristi za procjenu parametra nepristrani procjenitelj ukoliko je srednja vrijednost uzoračke distribucije te statistike jednaka vrijednosti tog parametra.

gdje SS_{11} = suma kvadrata ćelije određene presjekom 1. retka i 1. stupca
SS_{12} = suma kvadrata ćelije određene presjekom 1. retka i 2. stupca
$SS_{m_1 m_2}$ = suma kvadrata ćelije određene presjekom m_1-tog retka i m_2-tog stupca.

To je konceptualna formula, no puno je praktičnija:

$$SS_W = \sum_{ijk} (y_{ijk} - \hat{\mu}_{ij})^2 =$$

$$= \sum_{ijk} y_{ijk}^2 - \frac{\left(\sum\limits_{k=1}^{l} y_{11k}\right)^2 + \left(\sum\limits_{k=1}^{l} y_{12k}\right)^2 + \cdots + \left(\sum\limits_{k=1}^{l} y_{m_1 m_2 k}\right)^2}{l},$$

za kao i ranije podatke y_{ijk}, za koje se pretpostavlja da su nastali mjerenjem slučajne varijable Y_{ijk}, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, $k = 1, \ldots, l$. Ovo je ekvivalentno formuli od SS_W u (3.4.5.) za jednofaktorski model, razlika je jedino u tome što sume umjesto po razinama idu po ćelijama.

Zbog toga što svaka ćelija doprinosi $l - 1$ stupnjem slobode, a ukupno je $m_1 m_2$ ćelija, ukupan broj stupnjeva slobode unutar grupa je $m_1 m_2 (l - 1)$, to jest $df_W = m_1 m_2 (l - 1)$.

3.4.5 Procjenitelj varijance redaka (s_R^2)

Ovaj procjenitelj je baziran na razlici srednjih vrijednosti redaka. Analogan je s_B^2, procjenitelju varijance između grupa za jednofaktorski model. Iz jednofaktorskog modela je poznato da je to procjenitelj varijance populacije plus efekt faktora (ukoliko nul-hipoteza nije istinita). Na sličan način, procjenitelj varijance redaka s_R^2 u dvofaktorskoj ANOVA-i je procjenitelj varijance populacije plus efekt faktora I (ukoliko H_{10} nije istinita). Ako ne postoji glavni efekt faktora I, tada su parametri srednjih vrijednosti redaka jednaki ($\mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{m_1\cdot}$) i razlike između uzoračkih srednjih vrijednosti redaka biti će uzrokovane slučajnošću uzorka identičnih populacija. U tom slučaju, s_R^2 je nepristran procjenitelj od populacijske varijance σ^2.

Formula je:

$$s_R^2 = \frac{SS_R}{df_R},$$

za SS_R sumu kvadrata redaka, a df_R broj stupnjeva slobode redaka. Suma kvadrata redaka (SS_R) je slična formuli sume kvadrata između grupa SS_W u (3.4.5.). Jedina razlika je da se umjesto srednjih vrijednosti razina koriste srednje vrijednosti redaka. Konceptualna formula je

$$SS_R = n_{redak} \left((\hat{\mu}_{1\cdot} - \hat{\mu})^2 + (\hat{\mu}_{2\cdot} - \hat{\mu})^2 + \cdots + (\hat{\mu}_{m_1\cdot} - \hat{\mu})^2 \right),$$

za n_{redak} broj opservacija u jednom retku, to jest $n_{redak} = m_1 l$. Povećanjem glavnog efekta faktora I, srednje vrijednosti redaka su sve udaljenije jedna od druge, zbog čega se $(\hat{\mu}_{1\cdot} - \hat{\mu})^2, (\hat{\mu}_{2\cdot} - \hat{\mu})^2, \ldots, (\hat{\mu}_{m_1\cdot} - \hat{\mu})^2$ povećavaju. Dakle, SS_R se povećava s povećanjem glavnog efekta faktora I, a time se povećava i s_R^2.

Praktičnija formula bi bila:

$$SS_R = \left[\frac{\left(\sum_{jk} y_{1jk}\right)^2 + \left(\sum_{jk} y_{2jk}\right)^2 + \cdots + \left(\sum_{jk} y_{m_1jk}\right)^2}{n_{redak}} \right] - \frac{\left(\sum_{ijk} y_{ijk}\right)^2}{n}.$$

Zbog toga što retci doprinose $m_1 - 1$ stupnjem slobode, ukupan broj stupnjeva slobode redaka je $df_R = m_1 - 1$. Ovo je slično kao u slučaju jednofaktorske ANOVA-e za broj stupnjeva slobode između grupa df_B kada se umjesto broja redaka gledao broj razina.

3.4.6 Procjenitelj varijance stupaca (s_C^2)

Procjenitelj varijance stupaca se definira po sličnom postupku kao procjenitelj varijance redaka, samo što je baziran na razlici srednjih vrijednosti stupaca, te se umjesto srednjih vrijednosti redaka koriste srednje vrijednosti stupaca.

Dakle, procjenitelj varijance stupaca s_C^2 je procjenitelj varijance populacije plus efekt faktora II (ukoliko H_{02} nije istinita). Također, ako ne postoji glavni efekt faktora II, tada su parametri srednjih vrijednosti stupaca jednaki ($\mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot m_2}$) i razlike između uzoračkih srednjih vrijednosti stupaca biti će uzrokovane slučajnošću uzorka identičnih populacija, što znači da je s_C^2 nepristran procjenitelj od σ^2.

Formula je:

$$s_C^2 = \frac{SS_C}{df_C},$$

za SS_C sumu kvadrata stupaca, a df_C broj stupnjeva slobode stupaca. Konceptualna formula za SS_C je

$$SS_C = n_{stupac} \left((\hat{\mu}_{\cdot 1} - \mu)^2 + (\hat{\mu}_{\cdot 2} - \hat{\mu})^2 + \cdots + (\hat{\mu}_{\cdot m_2} - \hat{\mu})^2 \right),$$

za n_{stupac} broj elemenata u jednom stupcu, to jest $n_{stupac} = m_2 l$. Analogno kao u SS_R, iz prethodne formule je vidljivo da se povećanjem glavnog efekta faktora II povećava suma kvadrata stupaca SS_C a time i s_C^2, što znači da je s_C^2 dobar indikator istinitosti hipoteze H_{02}.

Praktična formula:

$$SS_C = \left[\frac{\left(\sum\limits_{ik} y_{i1k}\right)^2 + \left(\sum\limits_{ik} y_{i2k}\right)^2 + \cdots + \left(\sum\limits_{ik} y_{im_2k}\right)^2}{n_{stupac}} \right] - \frac{\left(\sum\limits_{ijk} y_{ijk}\right)^2}{n}.$$

Analogno kao s $df_R = m_1 - 1$ vrijedi da je broj stupnjeva slobode stupaca $df_C = m_2 - 1$. Dakle, s_R^2 i s_C^2 imaju dosta sličnosti s s_B^2 u jednofaktorskom modelu.

3.4.7 Procjenitelj varijance redaka po stupcima (s_{RC}^2)

Kao što je ranije rečeno, interakcija postoji kada su jednostavni glavni efekti jednog faktora različiti po razinama drugog faktora. Drugi način na koji se to može reći je da interakcija postoji ako je glavni efekt nekog faktora različit od (jednog od) jednostavnih glavnih efekata tog faktora.
Procjenitelj varijance redaka po stupcima s_{RC}^2 se koristi kako bi se procijenila interakcija između faktora. On je procjenitelj varijance populacije plus interakcija od faktora (ukoliko H_{12} nije istinita). Ako ne postoji interakcija, tada su parametri srednjih vrijednosti ćelija jednake ($\mu_{11} = \mu_{12} = \ldots = \mu_{m_1m_2}$) i razlike između uzoračkih srednjih vrijednosti ćelija biti će uzrokovane slučajnošću uzorka identičnih populacija. Tada je s_{RC}^2 nepristran procjenitelj populacijske varijance σ^2.

Formula je:

$$s_{RC}^2 = \frac{SS_{RC}}{df_{RC}},$$

za SS_{RC} sumu kvadrata redaka po stupcima, a df_{RC} broj stupnjeva slobode redaka po stupcima. Konceptualna formula za SS_{RC} je

$$SS_{RC} = l\left((\hat{\mu}_{11} - \hat{\mu})^2 + (\hat{\mu}_{12} - \hat{\mu})^2 + \cdots + (\hat{\mu}_{m_1m_2} - \hat{\mu})^2\right) - SS_R - SS_C,$$

a praktičnija za računati je

$$SS_{RC} = \left[\frac{(\sum_{k=1}^{l} y_{11k})^2 + (\sum_{k=1}^{l} y_{12k})^2 + \cdots + (\sum_{k=1}^{l} y_{m_1m_2k})^2}{l} \right] - \frac{(\sum\limits_{ijk} y_{ijk})^2}{n}.$$

Broj stupnjeva slobode za procjenu varijance redaka po stupcima iznosi $df_{RC} = (m_1 - 1)(m_2 - 1)$.

3.4.8 Ukupna varijabilnost

Kao i kod jednofaktorske ANOVA-e i ovdje se pokazuje da se ukupna varijabilnost može rastaviti na zbroj jednostavnijih varijabilnosti

$$SS_T = SS_R + SS_C + SS_{RC} + SS_W, \tag{3.4.11}$$

gdje je

$$SS_T = \sum_{ijk}\Big(y_{ijk} - \hat{\mu}\Big)^2 = \sum_{ijk}\left(y_{ijk} - \frac{\sum\limits_{ijk} y_{ijk}}{n}\right)^2 = \sum_{ijk}\Big(y_{ijk}\Big)^2 + \frac{\Big(\sum\limits_{ijk} y_{ijk}\Big)^2}{n}.$$

Taj rastav je na jednostavan način prikazan na slici 2.

3.4.9 Računanje F-omjera

Intuitivno se računaju tri F-omjera:

1. Za testiranje glavnog efekta faktora I:

$$F_{obt} = \frac{s_R^2}{s_W^2} \approx \frac{\sigma^2 + \text{glavni efekt faktora I}}{\sigma^2}.$$

2. Za testiranje glavnog efekta faktora II:

$$F_{obt} = \frac{s_C^2}{s_W^2} \approx \frac{\sigma^2 + \text{glavni efekt faktora II}}{\sigma^2}.$$

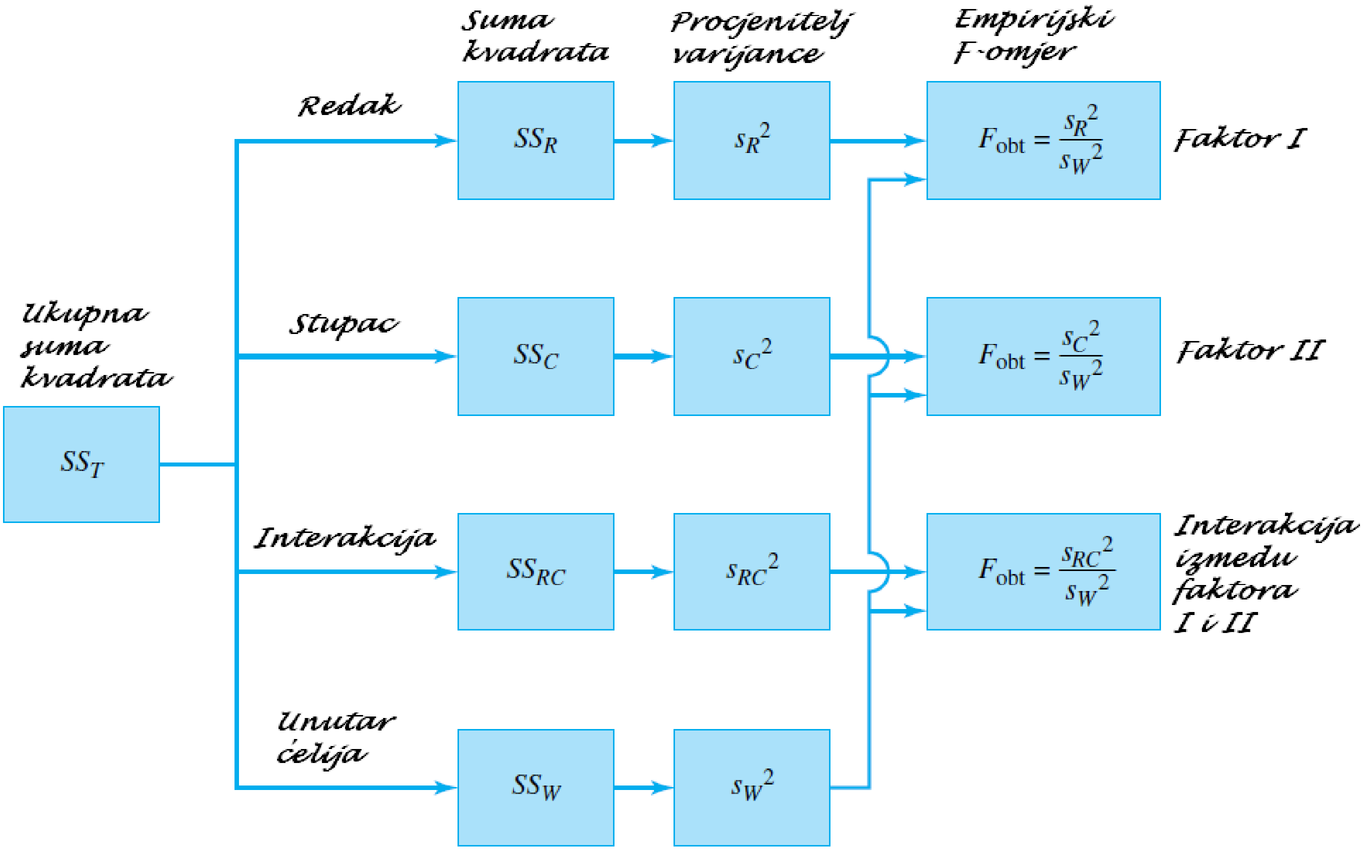
3. Za testiranje interakcije između faktora I i II:

$$F_{obt} = \frac{s_{RC}^2}{s_W^2} \approx \frac{\sigma^2 + \text{interakcija faktora I i II}}{\sigma^2}.$$

Glavni efekt svakog faktora, kao i interakcija faktora testira se kao i u jednofaktorskoj ANOVA-i uspoređivanjem F_{obt} s odgovarajućim F_{crit}. Pravilo odlučivanja o istinitosti svih hipoteza je isto kao u (3.4.6):

$$\text{Ako } F_{obt} \geq F_{crit} \text{ odbija se hipoteza } H_0.$$
$$\text{Ako } F_{obt} \leq F_{crit} \text{ ne odbija se hipoteza } H_0.$$

S tim da se za F_{crit} gleda broj stupnjeva slobode brojnika i nazivnika. Na primjer za testiranje H_{10} o neznačajnosti utjecaja faktora I, za broj stupnjeva slobode se uzimaju df_C i df_W. Cijeli do sada opisan postupak prikazan je na slici 2.



Slika 2: Postupak ukratko

3.4.10 Rješenje primjera

Početne hipoteze koje su se promatrale u primjeru 3.2 su zapravo testirale redom postojanje glavnog efekta faktora I, postojanje glavnog efekta faktora II i postojanje interakcije između faktora I i II, to jest ukoliko se podaci iz primjera zapišu u skladu s tablicom srednjih vrijednosti 1:

$$
\begin{aligned}
H_{10}: &\quad \mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot} \\
H_{02}: &\quad \mu_{\cdot 1} = \mu_{\cdot 2} \\
H_{12}: &\quad (\mu_{ij} - \mu) - (\mu_{i\cdot} - \mu) - (\mu_{\cdot j} - \mu) = 0, \quad i = 1,2,3,\ j = 1,2\,.
\end{aligned}
$$

Tablica 1 za slučaj opservacija primjera 3.2 prikazana je u tablicama 2 i 3. Primijetimo da se za faktor II hipotezom H_{02} testira značajnost razlike srednjih vrijednosti ćelija prvog stupca i srednjih vrijednosti ćelija drugog

		Faktor II		
		II_1	II_2	
Faktor I	I_1	$\hat{\mu}_{11}$	$\hat{\mu}_{12}$	$\hat{\mu}_{1\cdot}$
	I_2	$\hat{\mu}_{21}$	$\hat{\mu}_{22}$	$\hat{\mu}_{2\cdot}$
	I_3	$\hat{\mu}_{31}$	$\hat{\mu}_{32}$	$\hat{\mu}_{3\cdot}$
		$\hat{\mu}_{\cdot 1}$	$\hat{\mu}_{\cdot 2}$	$\hat{\mu}$

Tablica 2: Tablica procjena srednjih vrijednosti primjera

		Faktor II		
		II_1	II_2	
Faktor I	I_1	42.8	46.8	44.8
	I_2	39.2	43.0	41.1
	I_3	33.0	37.0	35.0
		$38.3\dot{3}$	$42.2\dot{6}$	40.3

Tablica 3: Tablica brojeva srednjih vrijednosti primjera

stupca, a za faktor I hipotezom H_{10} se testira značajnost razlike srednjih vrijednosti ćelija prva tri retka. Također je poznato da se hipotezom H_{12} testira je li glavni efekt nekog faktora jednak jednostavnom glavnom efektu određene razine.

Postupak rješavanja se egzaktno provodi po formulama koje su navedene i dođe se do F_{obt} kao što je naznačeno na slici 2. Kao i u jednofaktorskoj ANOVA-i ovdje je korišten programski jezik SAS, čiji su rezultati prikazani na slici 3. Iz nje se može iščitati:

$$SS_R = 489.800, \quad SS_C = 116.033, \quad SS_{RC} = 0.067, \quad SS_W = 258.4$$

$$s_R^2 = 489.800, \quad s_C^2 = 116.033, \quad s_{RC}^2 = 0.067, \quad s_W^2 = 258.4$$

$$F_{obt} = \frac{s_R^2}{s_W^2} = 22.75, \quad F_{obt} = \frac{s_C^2}{s_W^2} = 10.78, \quad F_{obt} = \frac{s_{RC}^2}{s_W^2} = 0.00$$

i p-vrijednosti za redom H_{10}, H_{02} i H_{12}:

$$pv_R < .001, \quad pv_C = 0.0031, \quad pv_{RC} = 0.9969\,.$$

Obično se prvo gleda je li interakcija značajna, jer ako *je*, to znači da glavni efekti sami za sebe ne objašnjavaju potpuni utjecaj faktora na varijablu odziva. To je zbog toga što je u razlici jednostavnih efekata sadržan i glavni efekt.

Iz dobivenih rezultata je jasno da interakcija ne postoji[16] jer se ne može odbaciti nul-hipoteza H_{12} ($pv_{RC} = 0.9969 > 0.05$), a to se može vidjeti i iz grafa dobivenog pomoću programskog jezika SAS na slici 4. Interpretacija

[16] To se i intuitivno zaključilo analiziranjem efekata u poglavlju neposredno nakon primjera.

The SAS System

The GLM Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	605.9000000	121.1800000	11.26	<.0001
Error	24	258.4000000	10.7666667		
Corrected Total	29	864.3000000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
faktorI	2	489.8000000	244.9000000	22.75	<.0001
faktorII	1	116.0333333	116.0333333	10.78	0.0031
faktorI*faktorII	2	0.0666667	0.0333333	0.00	0.9969

Slika 3: ANOVA tablica za dvofaktorski model

interakcije pomoću grafa je veoma jedostavna: ukoliko su linije paralelne ne postoji interakcija (što je slučaj u ovom primjeru), a ako značajno odstupaju od paralelnosti postoji.

Kada bi interakcija bila značajna testirala bi se značajnost svakog pojedinog jednostavnog glavnog efekta. To je zbog toga što interakcija postoji ako glavni efekt jednog faktora ovisi o razinama drugog faktora, pa je dalje pitanje u kojim razinama drugog faktora dolazi do značajne razlike u glavnom efektu. Time bi se odredilo točno koje razine faktora doprinose interakciji, to jest što točno utječe na varijablu odziva.[17]

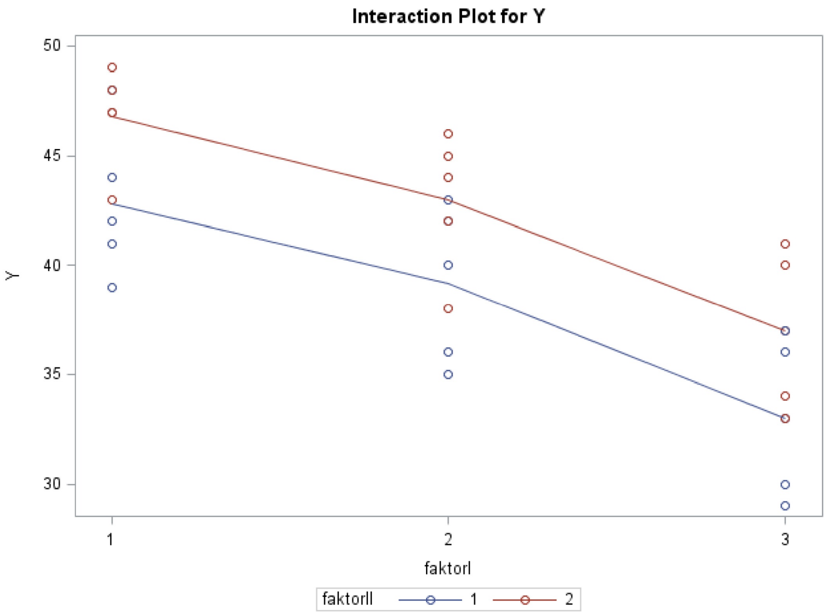
U ovom slučaju nema interakcije, pa se rezultati mogu interpretirati samo pomoću glavnih efekata faktora. To se moglo vidjeti ako se izračuna da je F_{crit} za redom utjecaj faktora I, faktora II i interakcije jednak 3.40, 4.26 i 3.40.

Iz rezultata se vidi da se H_{10} i H_{02} odbijaju (jer $pv_R < .001 < 0.05$ i $pv_C = 0.0031 < 0.05$), što znači da je utjecaj faktora I i II značajan. Treba nadodati da je interpretacija tog rezultata jednostavna u slučaju faktora II kada postoje samo dvije razine: metoda A i metoda B, jer to znači da postoji značajna razlika u rezultatima završnog ispita između metode A i metode B. Kod faktora I, pošto imaju tri razine, ne zna se odmah koje dvije od tih

[17] O tome će biti više govora u praktičnom primjeru u 5. poglavlju.

razina se značajno razlikuju, te je potrebno provesti takozvani *post-hoc test*.

Dakle, zaključak je sljedeći: ne postoji interakcija između faktora, što znači da je efekt jednog faktora bio jednak po svim razinama drugog faktora. Nadalje, razlike između očekivanja različitih razina faktora I (odnosno II) su prevelike da bi se to moglo pripisati slučajnom uzorku, to jest metode predavanja utječu na rezultate završnog ispita, kao i matematičke sposobnosti. Iz slike 4 se može vidjeti da se po razinama faktora I (1-Izvanredni, 2-Prosječni, 3-Slabiji) smanjuju bodovi na završnom ispitu, te da je metoda B (2-crvena linija) superiornija metodi A (1-plava linija).



Slika 4: Prikaz interakcije za primjer 3.2

4 Model dvofaktorske ANOVA-e

U ovom poglavlju promotrit će se egzaktni model analize varijance koji će objasniti matematičku pozadinu modela iz 3. poglavlja.

Ovaj model će se nadovezati na model regresije u 2. poglavlju jer postoji način da se prikaže u obliku

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}\,. \tag{4.1}$$

Kao i prije će se prvo promotriti jednostavniji slučaj jednofaktorskog modela.

4.1 Jednofaktorski model kao regresijski model

Njegov model efekata je općeniti slučaj modela (3.4.4) za pojednostavljeni primjer 3.2

$$Y_{i1} = \mu + \delta_i + \varepsilon_{i1}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m$$

gdje n broj razina faktora I, m broj razina faktora II, μ opću srednju vrijednost i δ_i efekt razine i. Ako je $i = 4$ broj razina faktora i u svakoj razini jedna vrijednost, model izgleda:

$$\begin{aligned}
Y_{11} &= \mu + \delta_1 + \varepsilon_{11} \\
Y_{21} &= \mu + \delta_2 + \varepsilon_{21} \\
Y_{31} &= \mu + \delta_3 + \varepsilon_{31} \\
Y_{41} &= \mu + \delta_4 + \varepsilon_{41}\,.
\end{aligned}$$

To se iz prikaza

$$\begin{aligned}
Y_{11} &= 1\mu + 1\delta_1 + 0\delta_2 + 0\delta_3 + 0\delta_4 + 1\varepsilon_{11} \\
Y_{21} &= 1\mu + 0\delta_1 + 1\delta_2 + 0\delta_3 + 0\delta_4 + 1\varepsilon_{21} \\
Y_{31} &= 1\mu + 0\delta_1 + 0\delta_2 + 1\delta_3 + 0\delta_4 + 1\varepsilon_{31} \\
Y_{41} &= 1\mu + 0\delta_1 + 0\delta_2 + 0\delta_3 + 1\delta_4 + 1\varepsilon_{41}
\end{aligned}$$

može formirati u vektorski zapis, gdje je na primjer

$$Y_{11} = [\,1\ \ 1\ \ 0\ \ 0\ \ 0\,]\,[\,\mu\ \ \delta_1\ \ \delta_2\ \ \delta_3\ \ \delta_4\,]^T + \varepsilon_{11}\,.$$

Iz toga slijedi matrični zapis

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{41} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{31} \\ \varepsilon_{41} \end{bmatrix}. \tag{4.2}$$

Lako je za vidjeti da slično vrijedi za općeniti slučaj, to jest dobije se upravo izraz oblika (4.1.). Ako ćelija sadržava više od jedne opservacije, sve opservacije u istoj grupi imati će isti model, to jest isti redak matrice $\mathbf{X}$. Zbog toga je uobičajno da se model zapisuje

$$\begin{aligned} Y_{1\cdot} &= \mu + \delta_1 + \varepsilon_{1\cdot} \\ Y_{2\cdot} &= \mu + \delta_2 + \varepsilon_{2\cdot} \\ Y_{3\cdot} &= \mu + \delta_3 + \varepsilon_{3\cdot} \\ Y_{4\cdot} &= \mu + \delta_4 + \varepsilon_{4\cdot}, \end{aligned}$$

što znači da je određen samo u terminima razina, a ne opservacija. Ako je s l označen broj elemenata u svakoj grupi, tada vrijedi $Y_{i\cdot} = \sum_{j=1}^{l} Y_{ij}/l$ i $\varepsilon_{i\cdot} = \sum_{j=1}^{l} \varepsilon_{ij}/l$, gdje je za ovaj slučaj $i = 1, \ldots, 4$. Sada iz toga i od prije poznate pretpostavke $\varepsilon_{ij} \sim N(0, \sigma^2)$ slijedi $\varepsilon_{i\cdot} \sim N(0, \sigma^2/l), i = 1, \ldots, 4,\ j = 1, \ldots, l$.

Kako bi se model dalje opisao u općenitom linearnom obliku

$$Y_i = \sum_{j=1}^{p} \theta_j x_{ij} + \varepsilon_i,\ i = 1, \ldots, n$$

uvode se dihotomne varijable koje se često nazivaju *indikatorske varijable* (ili kraće *indikatori*). Prvo će se taj model izvesti za slučaj kada broj razina faktora $i = 2$. Tada vrijedi

$$\begin{bmatrix} Y_{1\cdot} \\ Y_{2\cdot} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1\cdot} \\ \varepsilon_{2\cdot} \end{bmatrix}.$$

Bitno je za primijetiti da dizajn-matrica ovog modela nije punog ranga, što je problem za regresijski model. Kako bi se to ispravilo primijenjuju se elementarne transformacije po stupcima:

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

Dakle, parametar δ_2 je viška (što je i logično jer za glavne efekte vrijedi $\delta_2 + \delta_1 = 0$ pa se jedna može iskazati preko drugoga). Dakle, dovoljno je promatrati

$$\begin{bmatrix} Y_{1\cdot} \\ Y_{2\cdot} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1\cdot} \\ \varepsilon_{2\cdot} \end{bmatrix}$$

to jest

$$\begin{bmatrix} Y_{1\cdot} \\ Y_{2\cdot} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \theta_0 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \theta_1 + \begin{bmatrix} \varepsilon_{1\cdot} \\ \varepsilon_{2\cdot} \end{bmatrix}$$

ili još jednostavnije

$$\begin{aligned} Y_{1\cdot} &= \theta_0 + \varepsilon_{1\cdot} \\ Y_{2\cdot} &= \theta_0 + \theta_1 + \varepsilon_{2\cdot}\,, \end{aligned}$$

uz oznake $\theta_0 = \mu + \delta_1$ i $\theta_0 + \theta_1 = \mu + \delta_2$.

Iz toga se dobije da je model oblika

$$Y_{i\cdot} = \theta_0 + \theta_1 x + \varepsilon_{i\cdot}, \quad i = 1, 2$$

za $\theta_0 = \mu$, $\theta_1 = \delta_1$ i x indikator takav da

$$x = \begin{cases} 1, & \text{ako postoji efekt razine 2} \\ 0, & \text{ako ne postoji efekt razine 2} \end{cases}$$

promatranog faktora.

Ukoliko je $\theta_1 = 0$ to znači da faktor ne utječe značajno na odstupanja opservacija, to jest da razlike između efekata razine 1 i razine 2 nisu značajne. Moglo bi se, dakle, uzeti da vrijednost 1 od indikatora x označava postojanje utjecaja razine 2, a vrijednost 0 postojanje utjecaja razine 1.

Nadalje, za slučaj kada je broj razina $i = 4$ u modelu (4.2.) je iz sličnih razloga jedna varijabla viška, pa se model može zapisati kao

$$\begin{bmatrix} Y_{1\cdot} \\ Y_{2\cdot} \\ Y_{3\cdot} \\ Y_{4\cdot} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1\cdot} \\ \varepsilon_{2\cdot} \\ \varepsilon_{3\cdot} \\ \varepsilon_{4\cdot} \end{bmatrix}.$$

Pitanje koje se tu postavlja je kako definirati indikator. U slučaju u kojem su postojale dvije razine to je bilo lako jer su bile i dvije oznake: 0 i 1. Kako bi se dalo odgovor na to pitanje treba primijetiti da je broj potrebnih indikatora x kako bi se opisalo k razina faktora jednak $k - 1$ (koliki je i broj stupnjeva slobode jednofaktorske ANOVA-e). Za ovaj primjer bilo bi dovoljno definirati

$$x_1 = \begin{cases} 1, \text{ ako je razina 1} \\ 0, \text{ inače} \end{cases}, \; x_2 = \begin{cases} 1, \text{ ako je razina 2} \\ 0, \text{ inače} \end{cases} \; i \; x_3 = \begin{cases} 1, \text{ ako je razina 3} \\ 0, \text{ inače.} \end{cases}$$

Ukoliko je $x_1 = x_2 = x_3 = 0$ tada je jasno da je riječ o utjecaju efekta 4. razine. Sada se može napisati regresijska jednadžba ovog modela kao

$$Y_{i\cdot} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \varepsilon_{i\cdot} \quad i = 1, \dots, 4\,.$$

Sličnim razmišljanjem dolazi se do općenite regresijske jednadžbe jednofaktorskog modela.

4.2 Dvofaktorski model kao regresijski model

Sada će se raspisati kako se dolazi do regresijske jednadžbe dvofaktorskog modela efekata, to jest modela

$$Y_{ijk} = \mu + \delta_{i\cdot} + \delta_{\cdot j} + \delta_{ij} + \varepsilon_{ijk}\,.$$

Kada je slučaj kao u primjeru 3.2, to jest faktor I ima tri razine, faktor II dvije, a u svakoj ćeliji je pet opservacija tada je model

$$\begin{aligned}
Y_{11\cdot} &= \mu + \delta_{1\cdot} + \delta_{\cdot 1} + \delta_{11} + \varepsilon_{11\cdot}\\
Y_{12\cdot} &= \mu + \delta_{1\cdot} + \delta_{\cdot 2} + \delta_{12} + \varepsilon_{12\cdot}\\
Y_{21\cdot} &= \mu + \delta_{2\cdot} + \delta_{\cdot 1} + \delta_{21} + \varepsilon_{21\cdot}\\
Y_{22\cdot} &= \mu + \delta_{2\cdot} + \delta_{\cdot 2} + \delta_{22} + \varepsilon_{22\cdot}\\
Y_{31\cdot} &= \mu + \delta_{3\cdot} + \delta_{\cdot 1} + \delta_{31} + \varepsilon_{31\cdot}\\
Y_{32\cdot} &= \mu + \delta_{3\cdot} + \delta_{\cdot 2} + \delta_{32} + \varepsilon_{32\cdot},
\end{aligned}$$

a od njega je matrični zapis

$$\begin{bmatrix} Y_{11\cdot}\\ Y_{12\cdot}\\ Y_{21\cdot}\\ Y_{22\cdot}\\ Y_{31\cdot}\\ Y_{32\cdot} \end{bmatrix} = \begin{bmatrix} 1&1&0&0&1&0&1&0&0&0&0&0\\ 1&1&0&0&0&1&0&1&0&0&0&0\\ 1&0&1&0&1&0&0&0&1&0&0&0\\ 1&0&1&0&0&1&0&0&0&1&0&0\\ 1&0&0&1&1&0&0&0&0&0&1&0\\ 1&0&0&1&0&1&0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} \mu\\ \delta_{1\cdot}\\ \delta_{2\cdot}\\ \delta_{3\cdot}\\ \delta_{\cdot 1}\\ \delta_{\cdot 2}\\ \delta_{11}\\ \delta_{12}\\ \delta_{21}\\ \delta_{22}\\ \delta_{31}\\ \delta_{32} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11\cdot}\\ \varepsilon_{12\cdot}\\ \varepsilon_{21\cdot}\\ \varepsilon_{22\cdot}\\ \varepsilon_{31\cdot}\\ \varepsilon_{32\cdot} \end{bmatrix}. \tag{4.3}$$

Ovo se može poopćiti, te se dobije traženi oblik:

$$\mathbf{Y}. = \mathbf{X}.\boldsymbol{\theta} + \boldsymbol{\varepsilon}., \tag{4.4.a}$$

gdje je slično kao u jednofaktorskom slučaju $\mathbf{Y}.$ p-dimenzionalni slučajni vektor srednjih vrijednosti ćelija, za p broj ćelija (=broj razina faktora I $\times$ broj razina faktora II, što je $p = 6$ za prethodni slučaj), $\boldsymbol{\varepsilon}.$ $p \times 1$ slučajni vektor srednjih vrijednosti pogrešaka i $\mathbf{X}.$ $p \times m$ dizajn-matrica, za m broj parametara (u slučaju primjera $m = 12$ gdje su parametri po stupcima redom: 1. stupac srednja vrijednost, 2.-4. stupac utjecaj faktora I, 5. i 6. stupac utjecaj faktora II, 7.-12. stupac interakcija faktora I i II). Koeficijenti ili efekti koji se obično procjenjuju u regresiji sadržani su u $m \times 1$ vektoru $\boldsymbol{\theta}$.

Također se model (4.4.a) može zapisati i po svakoj opservaciji

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{4.4.b}$$

gdje je $\mathbf{Y}$ n-dimenzionalni slučajni vektor, za $n = m_1 m_2 l$ ukupan broj opservacija, $\boldsymbol{\varepsilon}$ $n \times 1$ slučajni vektor i $\mathbf{X}$ $n \times m$ dizajn-matrica, za m broj parametara.

Kao i u jednofaktorskom slučaju, niti ova dizajn-matrica nije punog ranga. U (4.3) se računanjem dobije da je rang jednak 6, pa se neki stupci zanemaruju. Od svakog faktora se zanemaruje efekt zadnje razine, to jest to po stupcima izgleda $1 + (3-1) + (2-1) + (3-1)(2-1) = 6$.

Sličnom logikom se dobije i općenita regresijska jednadžba dvofaktorskog modela po elementima modela (4.4.b):

$$Y_{ij\cdot} = \theta_0 + \sum_{i=1}^{m_1\text{-}1} \theta_{i\cdot} x_{i\cdot} + \sum_{j=1}^{m_2\text{-}1} \theta_{\cdot j} x_{\cdot j} + \sum_{i=1}^{m_1\text{-}1} \sum_{j=1}^{m_2\text{-}1} \theta_{ij} x_{ij} + \varepsilon_{ij\cdot} \tag{4.5}$$

za $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, gdje su $x_{i\cdot}$ oznake za indikatore razina faktora I, $x_{\cdot j}$ oznake za indikatore razina faktora II i x_{ij} oznake za indikatore interakcije između faktora I i II, to jest

$$x_{i\cdot} := \begin{cases} 1, & \text{ako je razina } \mathrm{I}_i \text{ faktora I} \\ 0, & \text{inače} \end{cases} \qquad x_{\cdot j} := \begin{cases} 1, & \text{ako je razina } \mathrm{II}_j \text{ faktora II} \\ 0, & \text{inače} \end{cases}$$

$$x_{ij} := \begin{cases} 1, & \text{ako je ćelija koja odgovara } \mathrm{I}_i \ i \ \mathrm{II}_j \\ 0, & \text{inače.} \end{cases}$$

Dio modela (4.5) s interakcijom

$$\sum_{i=1}^{m_1\text{-}1} \sum_{j=1}^{m_2\text{-}1} \theta_{ij} x_{ij}$$

zapravo govori da postoje dva efekta (po retcima i po stupcima) za koje vrijedi da niti jedan nema utjecaj na varijablu odziva bez prisutnosti drugoga. To znači da se oznaka x_{ij} odnosi na samo jednu varijablu, što je i logično jer nema smisla govoriti o interakciji jednog faktora.

Međutim, kao što je i prije napomenuto, češće se koristi matrični oblik modela (4.5), te će se model (4.4.b) sada detaljnije raspisati. To neće biti problem, jer je taj model sličan modelu (4.5), samo što će umjesto varijabli biti riječ o vektorima. Na primjer za model u (4.3) po svim opservacijama i s matricom punog ranga uvedene su sljedeće oznake po stupcima:

$$\mathbf{Y} = \begin{bmatrix} \mathbb{1}_{30} & \boldsymbol{x}_{1\cdot} & \boldsymbol{x}_{2\cdot} & \boldsymbol{x}_{\cdot 1} & \boldsymbol{x}_{11} & \boldsymbol{x}_{21} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_{1\cdot} \\ \theta_{2\cdot} \\ \theta_{\cdot 1} \\ \theta_{11} \\ \theta_{21} \end{bmatrix} + \boldsymbol{\varepsilon}\,.$$

To se poopćuje pa se sada matrični model (4.4.b) može zapisati:

$$\mathbf{Y} = \theta_0 \mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1} \theta_{i\cdot} \boldsymbol{x}_{i\cdot} + \sum_{j=1}^{m_2\text{-}1} \theta_{\cdot j} \boldsymbol{x}_{\cdot j} + \sum_{i=1}^{m_1\text{-}1} \sum_{j=1}^{m_2\text{-}1} \theta_{ij} \boldsymbol{x}_{ij} + \boldsymbol{\varepsilon}\,. \tag{4.6}$$

Dakle, gotovo sve oznake su slične kao one u modelima (4.4.b) i (4.5), samo su uvedene dodatne oznake za stupce dizajn-matrice $\mathbf{X}$.

4.3 Opći dvofaktorski model

Nul-hipoteze koje treba testirati za dvofaktorsku analizu varijance su da ne postoji interakcija, da ne postoji glavni efekt faktora I i da ne postoji glavni efekt faktora II. Te hipoteze se sada mogu izraziti na sljedeći način:

$$\begin{aligned} &H_{12} : \theta_{ij} = 0,\ i = 1, \dots, m_1 - 1,\ j = 1, \dots, m_2 - 1 \\ &H_{10} : \theta_{i\cdot} = 0,\ i = 1, \dots, m_1 - 1 \\ &H_{02} : \theta_{\cdot j} = 0,\ j = 1, \dots, m_2 - 1. \end{aligned}$$

Također se može ispitati i postoji li općenito neki efekt faktora:

$$H_0 : \theta_{ij} = \theta_{i\cdot} = \theta_{\cdot j} = 0,\ i = 1, \dots, m_1 - 1,\ j = 1, \dots, m_2 - 1\,.$$

Važno je primijetiti da je primjerice testiranje hipoteze H_{01} ekvivalentno testiranju istinitosti modela nul-hipoteze:

$$\mathbf{Y} = \theta_0 \mathbb{1}_n + \sum_{j=1}^{m_2\text{-}1} \theta_{\cdot j} \boldsymbol{x}_{\cdot j} + \sum_{i=1}^{m_1\text{-}1} \sum_{j=1}^{m_2\text{-}1} \theta_{ij} \boldsymbol{x}_{ij} + \boldsymbol{\varepsilon}\,, \tag{4.7}$$

ako se pretpostavlja da je model (4.6) istinit.

Iz 2. poglavlja slijedi da postoji kapa matrica tog modela, oznake $\mathbf{P}_{10}$ za koju vrijedi

$$\mathbf{P}_{10}\mathbf{Y} = \hat{\mathbf{Y}} = \mathbf{X}_0\hat{\gamma} = \hat{\theta}_0\mathbb{1}_n + \sum_{j=1}^{m_2\text{-}1}\hat{\theta}_{\cdot j}\boldsymbol{x}_{\cdot j} + \sum_{i=1}^{m_1\text{-}1}\sum_{j=1}^{m_2\text{-}1}\hat{\theta}_{ij}\boldsymbol{x}_{ij}$$

gdje je $\hat{\gamma}$ procjenitelj najmanjih kvadrata u matričnom modelu $\mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ restringiranog modela (4.7). Kako ne bi došlo do miješanja oznaka ovdje se uvodi $\hat{\mathbf{Y}}_{10} := \hat{\mathbf{Y}}$.

Slično, testiranje hipoteze H_{02} je ekvivalentno testiranju istinitosti modela nul-hipoteze:

$$\mathbf{Y} = \theta_0\mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1}\theta_{i\cdot}\boldsymbol{x}_{i\cdot} + \sum_{i=1}^{m_1\text{-}1}\sum_{j=1}^{m_2\text{-}1}\theta_{ij}\boldsymbol{x}_{ij} + \boldsymbol{\varepsilon}. \tag{4.8}$$

za kojeg postoji kapa matrica oznake $\mathbf{P}_{02}$ takva da

$$\hat{\mathbf{Y}}_{02} := \mathbf{P}_{02}\mathbf{Y} = \hat{\theta}_0\mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1}\hat{\theta}_{i\cdot}\boldsymbol{x}_{i\cdot} + \sum_{i=1}^{m_1\text{-}1}\sum_{j=1}^{m_2\text{-}1}\hat{\theta}_{ij}\boldsymbol{x}_{ij},$$

dok je:

$$\mathbf{Y} = \theta_0\mathbb{1}_n + \boldsymbol{\varepsilon}. \tag{4.9}$$

ekvivalentno testiranju istinitosti modela nul-hipoteze H_0, za kojeg postoji kapa matrica oznake $\mathbf{P}_0$ takva da

$$\hat{\mathbf{Y}}_0 := \mathbf{P}_0\mathbf{Y} = \hat{\theta}_0\mathbb{1}_n,$$

a hipoteze H_{12}:

$$\mathbf{Y} = \theta_0\mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1}\hat{\theta}_{i\cdot}\boldsymbol{x}_{i\cdot} + \sum_{j=1}^{m_2\text{-}1}\hat{\theta}_{\cdot j}\boldsymbol{x}_{\cdot j} + \boldsymbol{\varepsilon}. \tag{4.10}$$

za kojeg postoji kapa matrica oznake $\mathbf{P}_{12}$ takva da

$$\hat{\mathbf{Y}}_{12} := \mathbf{P}_{12}\mathbf{Y} = \hat{\theta}_0\mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1}\hat{\theta}_{i\cdot}\boldsymbol{x}_{i\cdot} + \sum_{j=1}^{m_2\text{-}1}\hat{\theta}_{\cdot j}\boldsymbol{x}_{\cdot j}\,.$$

U svim ovim slučajevima se pretpostavlja da je je model (4.6) istinit. Primijetimo da vrijedi rastav:

$$I - \mathbf{P}_0 = (\mathbf{P}_{10} - \mathbf{P}_0) + (\mathbf{P}_{02} - \mathbf{P}_0) + (\mathbf{P}_{00} - \mathbf{P}_{10} - \mathbf{P}_{02} + \mathbf{P}_0) + (I - \mathbf{P}_{00}). \tag{4.11}$$

Zbog toga što su $\mathbf{P}_0$, $\mathbf{P}_{10}$, $\mathbf{P}_{02}$, $\mathbf{P}_{00}$ i $\mathbf{P}_{12}$ ortogonalne projekcijske matrice, tada su po (2.1.11) i (2.1.20) i $I-\mathbf{P}_0$, $\mathbf{P}_{10}-\mathbf{P}_0$, $\mathbf{P}_{02}-\mathbf{P}_0$ i $I-\mathbf{P}_{00}$ ortogonalne projekcijske matrice. Također, jer su $\mathbf{P}_{00} - \mathbf{P}_{10}$ i $\mathbf{P}_{02} - \mathbf{P}_0$ ortogonalne projekcijske matrice, tada je iz istog razloga i $(\mathbf{P}_{00}-\mathbf{P}_{10})-(\mathbf{P}_{02}-\mathbf{P}_0)=\mathbf{P}_{00}-\mathbf{P}_{10}-\mathbf{P}_{02}+\mathbf{P}_0$ ortogonalna projekcijska matrica.
Također vrijedi da su $\mathbf{P}_0$, $\mathbf{P}_{10}$, $\mathbf{P}_{02}$, $\mathbf{P}_{00}$, $\mathbf{P}_{00}-\mathbf{P}_{10}-\mathbf{P}_{02}+\mathbf{P}_0$ i $\mathbf{P}_{12}$ međusobno okomiti projektori.
Kako bi se to lako vidjelo uvode se oznake

$\mathcal{M}_{R\times C} := C(\mathbf{X})$ prostor stupaca općeg modela (4.6)

$\mathcal{M}_{R+C} := [\{\mathbb{1}_n, \boldsymbol{x}_{1\cdot}, \ldots, \boldsymbol{x}_{m_1-1\cdot}, \boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot m_2-1}\}]$ prostor stupaca aditivnog modela (4.10)

$\mathcal{M}_{RI} := [\{\mathbb{1}_n, \boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot m_2-1}, \boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{m_1-1\, m_2-1}\}]$ prostor stupaca modela (4.7)

$\mathcal{M}_{CI} := [\{\mathbb{1}_n, \boldsymbol{x}_{1\cdot}, \ldots, \boldsymbol{x}_{m_1-1\cdot}, \boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{m_1-1\, m_2-1}\}]$ prostor stupaca modela (4.8)

$\mathcal{M}_{\emptyset} := [\{\mathbb{1}_n\}]$ prostor stupaca modela (4.9)

$$\begin{tikzcd}
 & & \mathcal{M}_{RI} \arrow[dl] & \\
\mathcal{M}_{R+C} \arrow[r] & \mathcal{M}_{R\times C} & & \mathcal{M}_{\emptyset} \arrow[ul] \arrow[dl] \\
 & & \mathcal{M}_{CI} \arrow[ul] &
\end{tikzcd}$$

Slika 5: Odnos prostora stupaca različitih prostora za opći model

očito je da vrijedi odnos kao na slici 5 (s tim da na primjer $\mathcal{M}_{R+C} \longrightarrow \mathcal{M}_{R\times C}$ označava $\mathcal{M}_{R+C} \subset \mathcal{M}_{R\times C}$). Stoga na primjer vrijedi

$$(\mathbf{P}_{10}-\mathbf{P}_0)(\mathbf{P}_{02}-\mathbf{P}_0) = \mathbf{P}_{10}\mathbf{P}_{02}+\mathbf{P}_{10}\mathbf{P}_0+\mathbf{P}_0\mathbf{P}_{02}-\mathbf{P}_0^2 = \mathbf{P}_0+\mathbf{P}_0\mathbf{P}_0+\mathbf{P}_0-\mathbf{P}_0^2 = 0$$

gdje druga jednakost slijedi iz (2.1.18), jer $\mathcal{M}_{\emptyset} \subset \mathcal{M}_{RI}$ i $\mathcal{M}_{\emptyset} \subset \mathcal{M}_{CI}$, a ostali slično. Sada se može primijeniti Teorema 1.10 (Teorem o projekciji) na rastav (4.11) pomnožen zdesna s $\mathbf{Y}$:

$$\begin{aligned}\|(I-\mathbf{P}_0)\mathbf{Y}\|^2 = {} & \|(\mathbf{P}_{10}-\mathbf{P}_0)\mathbf{Y}\|^2 + \|(\mathbf{P}_{02}-\mathbf{P}_0)\mathbf{Y}\|^2 + \\ & + \|(\mathbf{P}_{00}-\mathbf{P}_{10}-\mathbf{P}_{02}+\mathbf{P}_0)\mathbf{Y}\|^2 + \|(I-\mathbf{P}_{00})\mathbf{Y}\|^2.\end{aligned}$$

Sada uvođenjem nove oznake za opću srednju vrijednost $\overline{\mathbf{Y}} = \hat{\mu}$ i jer za model (4.8) iz formula (2.1.7) i (2.1.10) slijedi (ako se model promatra u obliku (2.1.2)):

$$\hat{\theta}_0 = (\mathbb{1}_n^T\mathbb{1}_n)^{-1}\mathbb{1}_n^T\mathbf{Y} = \frac{1}{n}\sum_{ijk} Y_{ijk} = \overline{\mathbf{Y}}$$

$$\Rightarrow \mathbf{P}_0\mathbf{Y} = \mathbb{1}_n(\mathbb{1}_n^T\mathbb{1}_n)^{-1}\mathbb{1}_n^T\mathbf{Y} = \mathbb{1}_n\hat{\theta}_0 = \hat{\theta}_0\mathbb{1}_n = \overline{\mathbf{Y}}\mathbb{1}_n\,. \tag{4.12}$$

Primijetimo da također vrijedi

$$\hat{\mathbf{Y}}_{12} - \hat{\mathbf{Y}}_{00} = \hat{\mathbf{Y}}_{00} - \hat{\mathbf{Y}}_{10} - \hat{\mathbf{Y}}_{02} + \hat{\mathbf{Y}}_0\,,$$

pa se primjenom toga i (4.12) jednadžba može pojednostaviti na:

$$\|\mathbf{Y} - \overline{\mathbf{Y}}\mathbb{1}_n\|^2 = \|(\mathbf{P}_{10} - \mathbf{P}_0)\mathbf{Y}\|^2 + \|(\mathbf{P}_{02} - \mathbf{P}_0)\mathbf{Y}\|^2 +$$
$$+ \|(\mathbf{P}_{12} - \mathbf{P}_{00})\mathbf{Y}\|^2 + \|(I - \mathbf{P}_{00})\mathbf{Y}\|^2.$$

Može se pokazati da je prethodni izraz ekvivalentan (3.4.11). Primjerice, ukupna varijabilnost se podudara jer uz oznaku $\mathbb{J}_n$ za $n \times n$ matricu jedinica i iz (4.12) slijedi

$$\mathbf{P}_0 = \mathbb{1}_n(\mathbb{1}_n^T\mathbb{1}_n)^{-1}\mathbb{1}_n^T = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T = \frac{1}{n}\mathbb{J}_n$$

iz čega

$$\begin{aligned}\|\mathbf{Y} - \overline{\mathbf{Y}}\mathbb{1}_n\|^2 &= \|(I - \mathbf{P}_0)\mathbf{Y}\|^2 = ((I - \mathbf{P}_0)\mathbf{Y})^T((I - \mathbf{P}_0)\mathbf{Y}) = \mathbf{Y}^T(I - \mathbf{P}_0)^2\mathbf{Y} = \\ &= \mathbf{Y}^T(I - \mathbf{P}_0)\mathbf{Y} = \mathbf{Y}^T(I - \frac{1}{n}\mathbb{J}_n)\mathbf{Y} = \mathbf{Y}^T\mathbf{Y} - \frac{1}{n}\mathbf{Y}^T\mathbb{J}_n\mathbf{Y} = \\ &= \mathbf{Y}^T\mathbf{Y} - \frac{1}{n}(\mathbf{Y}^T\mathbf{Y})(\mathbf{Y}^T\mathbf{Y}) = \sum_{ijk}\Big(Y_{ijk}\Big)^2 + \frac{\Big(\sum_{ijk} Y_{ijk}\Big)^2}{n} = SS_T\end{aligned}$$

a za ostale slično.

Dakle, ako se želi testirati hipoteza

$$H_{12} : \theta_{ij} = 0,\ i = 1, \ldots, m_1 - 1,\ j = 1, \ldots, m_2 - 1$$

to je ekvivalentno testiranju istinitosti modela (4.10) ako je model (4.6) istinit. Zbog toga što vrijedi $\mathcal{M}_{R+C} \subset \mathcal{M}_{R\times C}$ ovo su ugniježdeni modeli, te se sada može primijeniti test-statistika u (2.1.21) koja u ovom slučaju izgleda

$$F_{12} = \frac{\dfrac{\mathbf{Y}^T(\mathbf{P}_{00} - \mathbf{P}_{12})\mathbf{Y}}{\mathrm{r}(\mathbf{P}_{00} - \mathbf{P}_{12})}}{\dfrac{\mathbf{Y}^T(I - \mathbf{P}_{00})\mathbf{Y}}{\mathrm{r}(I - \mathbf{P}_{00})}}$$

gdje

$$
\begin{aligned}
\mathrm{r}(\mathbf{P}_{00}-\mathbf{P}_{12}) &= \mathrm{r}(\mathbf{P}_{00})-\mathrm{r}(\mathbf{P}_{12}) = \dim(C(\mathbf{P}_{00}))-\dim(C(\mathbf{P}_{12})) \overset{(2.1.16)}{=} \\
&= \dim(\mathcal{M}_{R\times C})-\dim(\mathcal{M}_{R+C}) = m_1m_2-(m_1+m_2-1) = \\
&= m_2(m_1-1)-(m_1-1) = (m_1-1)(m_2-1) \\
\mathrm{r}(I-\mathbf{P}_{00}) &= \mathrm{r}(I)-\mathrm{r}(\mathbf{P}_{00}) = n-\dim(\mathcal{M}_{R\times C}) = m_1m_2l-m_1m_2 = m_1m_2(l-1)
\end{aligned}
$$

Zbog toga što

$$(\mathbf{P}_{00}-\mathbf{P}_{12})(I-\mathbf{P}_{00}) = \mathbf{P}_{00}-\mathbf{P}_{00}^2-\mathbf{P}_{12}+\mathbf{P}_{12}\mathbf{P}_{00} \underset{\mathrm{tm1.7.b}}{\overset{(2.1.17)}{=}} \mathbf{P}_{00}-\mathbf{P}_{00}-\mathbf{P}_{12}+\mathbf{P}_{12} = 0$$

iz Teorema 1.9 slijedi da su brojnik i nazivnik od F_{12} nezavisni.

Nadalje, uz pretpostavku istinitosti oba modela iz Teorema 2.3 slijedi

$$\frac{\|(\mathbf{P}_{00}-\mathbf{P}_{12})\mathbf{Y}\|^2}{\sigma^2} \sim \chi^2((m_1-1)(m_2-1))$$

a Leme 2.4

$$\frac{\|(I-\mathbf{P}_{00})\,\boldsymbol{Y}\|^2}{\sigma^2} \sim \chi^2(m_1m_2(l-1))$$

iz čega po Definiciji 1.5 F-distribucije slijedi

$$F_{12} \sim F((m_1-1)(m_2-1), m_1m_2(l-1)). \qquad (❀_{12})$$

Postupak ide slično za testiranje hipoteze

$$H_{10} : \theta_{i\cdot} = 0,\ i = 1,\dots,m_1-1$$

koja je ekvivalentna testiranju istinitosti modela (4.7) uz pretpostavku istinitosti početnog modela (4.6). Zbog toga što vrijedi $\mathcal{M}_{RI} \subset \mathcal{M}_{R\times C}$ ovo su ugniježdeni modeli. U ovom slučaju je test-statistika:

$$F_{10} = \frac{\dfrac{\mathbf{Y}^T(\mathbf{P}_{00}-\mathbf{P}_{10})\mathbf{Y}}{\mathrm{r}(\mathbf{P}_{00}-\mathbf{P}_{10})}}{\dfrac{\mathbf{Y}^T(I-\mathbf{P}_{00})\mathbf{Y}}{\mathrm{r}(I-\mathbf{P}_{00})}},$$

gdje je

$$
\begin{aligned}
&\mathrm{r}(\mathbf{P}_{00}-\mathbf{P}_{10}) = \mathrm{r}(\mathbf{P}_{00})-\mathrm{r}(\mathbf{P}_{10}) = \dim(C(\mathbf{P}_{00}))-\dim(C(\mathbf{P}_{10})) \overset{(2.1.17)}{=} \\
&\quad = \dim(\mathcal{M}_{R\times C})-\dim(\mathcal{M}_{RI}) = m_1m_2-1-(m_2-1)-(m_1-1)(m_2-1) = \\
&\quad = [1+m_1-1+m_2-1+(m_1-1)(m_2-1)]-1-(m_2-1)-(m_1-1)(m_2-1) = \\
&\quad = m_1-1 \\
&\mathrm{r}(I-\mathbf{P}_{00}) = m_1m_2(l-1).
\end{aligned}
$$

Zbog toga što vrijedi

$$(\mathbf{P}_{00}-\mathbf{P}_{10})(I-\mathbf{P}_{00})=\mathbf{P}_{00}-\mathbf{P}_{00}\mathbf{P}_{00}-\mathbf{P}_{10}+\mathbf{P}_{10}\mathbf{P}_{00}\underset{\text{tm1.7.b}}{\overset{(2.1.18)}{=}}\mathbf{P}_{00}-\mathbf{P}_{00}-\mathbf{P}_{10}+\mathbf{P}_{10}=0$$

iz Teorema 1.9 slijedi da su brojnik i nazivnik od F_{10} nezavisni.

Nadalje, uz pretpostavku istinitosti modela iz Teorema 2.3 i Leme 2.4 slijedi

$$\frac{\|(\mathbf{P}_{00}-\mathbf{P}_{10})\mathbf{Y}\|^2}{\sigma^2}\sim\chi^2(m_1-1)$$
$$\frac{\|(I-\mathbf{P}_{00})\,\mathbf{Y}\|^2}{\sigma^2}\sim\chi^2(m_1m_2(l-1))$$

iz čega po Definiciji 1.5 F-distribucije slijedi

$$F_{10}\sim F(m_1-1,m_1m_2(l-1)). \qquad (❀_1)$$

Sada za testiranje hipoteze

$$H_{02}:\theta_{\cdot j}=0,\ j=1,\ldots,m_2-1$$

koja je ekvivalentna testiranju istinitosti modela (4.8) uz pretpostavku istinitosti modela (4.6). Zbog toga što vrijedi $\mathcal{M}_{CI}\subset\mathcal{M}_{R\times C}$ ovo su ugniježdeni modeli. Za njega je test-statistika

$$F_{02}=\frac{\dfrac{\mathbf{Y}^T(\mathbf{P}_{00}-\mathbf{P}_{02})\mathbf{Y}}{\mathrm{r}(\mathbf{P}_{00}-\mathbf{P}_{02})}}{\dfrac{\mathbf{Y}^T(I-\mathbf{P}_{00})\mathbf{Y}}{\mathrm{r}(I-\mathbf{P}_{00})}},$$

gdje je

$$\begin{aligned}\mathrm{r}(\mathbf{P}_{00}-\mathbf{P}_{02})&=\mathrm{r}(\mathbf{P}_{00})-\mathrm{r}(\mathbf{P}_{02})=\dim(C\,(\mathbf{P}_{00}))-\dim(C\,(\mathbf{P}_{02}))\overset{(2.1.17)}{=}\\&=\dim(\mathcal{M}_{R\times C})-\dim(\mathcal{M}_{CI})=m_1m_2-1-(m_1-1)-(m_1-1)(m_2-1)=\\&=[1+m_1-1+m_2-1+(m_1-1)(m_2-1)]-1-(m_1-1)-(m_1-1)(m_2-1)=\\&=m_2-1\\\mathrm{r}(I-\mathbf{P}_{00})&=m_1m_2(l-1).\end{aligned}$$

Zbog toga što vrijedi

$$(\mathbf{P}_{00}-\mathbf{P}_{02})(I-\mathbf{P}_{00})=\mathbf{P}_{00}-\mathbf{P}_{00}\mathbf{P}_{00}-\mathbf{P}_{02}+\mathbf{P}_{02}\mathbf{P}_{00}\underset{\text{tm1.7.b}}{\overset{(2.1.18)}{=}}\mathbf{P}_{00}-\mathbf{P}_{00}-\mathbf{P}_{02}+\mathbf{P}_{02}=0$$

iz Teorema 1.9 slijedi da su brojnik i nazivnik od F_{02} nezavisni.

Ponovo uz pretpostavku istinitosti modela iz Teorema 2.3 i Leme 2.4 slijedi

$$\frac{\|(\mathbf{P}_{00}-\mathbf{P}_{02})\mathbf{Y}\|^2}{\sigma^2}\sim\chi^2(m_2-1)$$

$$\frac{\|(I-\mathbf{P}_{00})\,\boldsymbol{Y}\|^2}{\sigma^2} \sim \chi^2(m_1 m_2(l-1))$$

iz čega po Definiciji 1.5 F-distribucije slijedi

$$F_{02} \sim F(m_2-1, m_1m_2(l-1)). \qquad (❀_2)$$

Dakle, ovim je opravdano korištenje F-distribucije u 3. poglavlju. Iako u 3. poglavlju nije bilo toliko govora o aditivnom modelu, on je zapravo indirektno spomenut pri rješavanju primjera 3.2 jer je testiranjem ustanovljeno da ne postoji interakcija. Sada će se na sličan način opisati i nešto jednostavniji aditivni model, kako bi se dobila potpunija slika o dvofaktorskom modelu varijance.

4.4 Dvofaktorski aditivni model

Nul-hipoteze koje treba testirati za aditivni model slične su hipotezama H_{10} i H_{02}, samo što se dodatno pretpostavlja da je član interakcije jednak nula:

$$H_{10}^A: \ \theta_{i\cdot} = 0, \theta_{ij} = 0, \ i = 1, \ldots, m_1-1, \ j = 1, \ldots, m_2 - 1$$
$$H_{02}^A: \ \theta_{\cdot j} = 0, \theta_{ij} = 0, \ i = 1, \ldots, m_1-1, \ j = 1, \ldots, m_2 - 1\,.$$

Matrični regresijski prikaz aditivnog modela je oblika (4.10), odnosno

$$\mathbf{Y} = \theta_0 \mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1} \theta_{i\cdot}\boldsymbol{x}_{i\cdot} + \sum_{j=1}^{m_2\text{-}1} \theta_{\cdot j}\boldsymbol{x}_{\cdot j} + \boldsymbol{\varepsilon}\,.$$

Slično kao u općem modelu, testiranje hipoteze H_{10}^A ekvivalentno je testiranju istinitosti modela nul-hipoteze

$$\mathbf{Y} = \theta_0 \mathbb{1}_n + \sum_{j=1}^{m_2\text{-}1} \theta_{\cdot j}\boldsymbol{x}_{\cdot j} + \boldsymbol{\varepsilon} \qquad (4.13)$$

ako se pretpostavlja da je model (4.10) istinit.

Iz 2. poglavlja slijedi da postoji kapa matrica tog modela, oznake $\mathbf{P}_{10}^A$ za koju vrijedi

$$\hat{\mathbf{Y}}_{10}^A := \mathbf{P}_{10}^A \mathbf{Y} = \hat{\theta}_0 \mathbb{1}_n + \sum_{j=1}^{m_2\text{-}1} \hat{\theta}_{\cdot j}\boldsymbol{x}_{\cdot j}\,.$$

Testiranje hipoteze H_{02}^A je ekvivalentno testiranju istinitosti modela nulhipoteze

$$\mathbf{Y} = \theta_0 \mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1} \theta_{i\cdot}\boldsymbol{x}_{i\cdot} + \boldsymbol{\varepsilon}. \qquad (4.14)$$

za kojeg postoji kapa matrica oznake $\mathbf{P}_{02}^{A}$ takva da

$$\hat{\mathbf{Y}}_{02}^{A} := \mathbf{P}_{02}^{A}\mathbf{Y} = \hat{\theta}_0\mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1} \hat{\theta}_{i\cdot}\boldsymbol{x}_{i\cdot}\,,$$

gdje se pretpostavlja da je model (4.10) istinit.

Uz prethodno navedene oznake za prostore stupaca, za potrebe aditivnog modela, uvesti će se i oznake

$\mathcal{M}_R := [\{\mathbb{1}_n, \boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot m_2-1}\}]$ prostor stupaca modela (4.13)

$\mathcal{M}_C := [\{\mathbb{1}_n, \boldsymbol{x}_{1\cdot}, \ldots, \boldsymbol{x}_{m_1-1\cdot}\}]$ prostor stupaca modela (4.14).

Za aditivni model će biti važan odnos prostora prikazan na slici 6.

$$\begin{tikzcd}
 & \mathcal{M}_R \arrow[dl] & \\
\mathcal{M}_{R+C} & & \mathcal{M}_{\emptyset} \arrow[ul] \arrow[dl] \\
 & \mathcal{M}_C \arrow[ul] &
\end{tikzcd}$$

Slika 6: Odnos prostora stupaca različitih prostora za aditivni model

I za aditivni model vrijedi rastav varijabilnosti na sumu kvadrata koji je karakterističan za ANOVA-u. Prvo primijetimo da vrijedi

$$I - \mathbf{P}_0 = (\mathbf{P}_{10}^{A} - \mathbf{P}_0) + (\mathbf{P}_{02}^{A} - \mathbf{P}_0) + (I - \mathbf{P}_{10}^{A} - \mathbf{P}_{02}^{A} + \mathbf{P}_0). \tag{4.16}$$

Sada zbog toga što su $\mathbf{P}_0$, $\mathbf{P}_{10}^{A}$ i $\mathbf{P}_{02}^{A}$ ortogonalne projekcijske matrice, tada su po (2.1.11) i (2.1.20) i $I - \mathbf{P}_0$, $\mathbf{P}_{10}^{A} - \mathbf{P}_0$ i $\mathbf{P}_{02}^{A} - \mathbf{P}_0$ ortogonalne projekcijske matrice. Također, jer su $I - \mathbf{P}_{10}^{A}$ i $\mathbf{P}_{02}^{A} - \mathbf{P}_0$ ortogonalne projekcijske matrice, tada je i $(I - \mathbf{P}_{10}^{A}) - (\mathbf{P}_{02}^{A} - \mathbf{P}_0) = I - \mathbf{P}_{10}^{A} - \mathbf{P}_{02}^{A} + \mathbf{P}_0$ po (2.1.20) ortogonalna projekcijska matrica.

Kao i prije, lako se pokaže da su $I - \mathbf{P}_0$, $\mathbf{P}_{10}^{A} - \mathbf{P}_0$, $\mathbf{P}_{02}^{A} - \mathbf{P}_0$ i $I - \mathbf{P}_{10}^{A} - \mathbf{P}_{02}^{A} + \mathbf{P}_0$ međusobno okomiti projektori, te se može primijeniti Teorem o projekciji na rastav (4.11) pomnožen zdesna s $\mathbf{Y}$:

$$\|(I - \mathbf{P}_0)\mathbf{Y}\|^2 = \|(\mathbf{P}_{10}^{A} - \mathbf{P}_0)\mathbf{Y}\|^2 + \|(\mathbf{P}_{02}^{A} - \mathbf{P}_0)\mathbf{Y}\|^2 + \|(I - \mathbf{P}_{10}^{A} - \mathbf{P}_{02}^{A} + \mathbf{P}_0)\mathbf{Y}\|^2.$$

Primijetimo da vrijedi

$$\hat{\mathbf{Y}}_{12} = \hat{\theta}_0\mathbb{1}_n + \sum_{i=1}^{m_1\text{-}1} \hat{\theta}_{i\cdot}\boldsymbol{x}_{i\cdot} + \sum_{j=1}^{m_2\text{-}1} \hat{\theta}_{\cdot j}\boldsymbol{x}_{\cdot j} + \hat{\theta}_0\mathbb{1}_n - \hat{\theta}_0\mathbb{1}_n = (\mathbf{P}_{10}^{A} + \mathbf{P}_{02}^{A} - \mathbf{P}_0)\mathbf{Y}$$

iz čega slijedi da je prethodno uvedena kapa matrica $\mathbf{P}_{12}$ aditivnog modela (4.10) linearna kombinacija ostalih kapa matrica, to jest vrijedi:

$$\mathbf{P}_{12} = \mathbf{P}^A_{10} + \mathbf{P}^A_{02} - \mathbf{P}_0. \tag{4.11}$$

Sada se primjenom toga i (4.12) rastav se može pojednostaviti na

$$\|\mathbf{Y} - \overline{\mathbf{Y}}\mathbb{1}_n\|^2 = \|(\mathbf{P}^A_{10} - \mathbf{P}_0)\mathbf{Y}\|^2 + \|(\mathbf{P}^A_{02} - \mathbf{P}_0)\mathbf{Y}\|^2 + \|(I - \mathbf{P}_{12})\mathbf{Y}\|^2,$$

što je poznati rastav u literaturi za aditivni model.

Testirajmo prvo hipotezu

$$H^A_{10} : \ \theta_{i\cdot} = 0, \ \theta_{ij} = 0, \ i = 1, \ldots, m_1 - 1, \ j = 1, \ldots, m_2 - 1,$$

koja je ekvivalentna testiranju istinitosti modela (4.13) uz pretpostavku istinitosti aditivnog modela (4.10). Zbog toga što vrijedi $\mathcal{M}_R \subset \mathcal{M}_{R+C}$ ovo su ugniježdeni modeli. Test-statistika u (2.1.21) je sada:

$$F^A_{10} = \frac{\dfrac{\mathbf{Y}^T(\mathbf{P}_{12} - \mathbf{P}^A_{10})\mathbf{Y}}{\mathrm{r}(\mathbf{P}_{12} - \mathbf{P}^A_{10})}}{\dfrac{\mathbf{Y}^T(I - \mathbf{P}_{12})\mathbf{Y}}{\mathrm{r}(I - \mathbf{P}_{12})}},$$

gdje je

$$\begin{aligned}
\mathrm{r}(\mathbf{P}_{12} - \mathbf{P}^A_{10}) &= \mathrm{r}(\mathbf{P}_{12}) - \mathrm{r}(\mathbf{P}^A_{10}) = \dim(C(\mathbf{P}_{12})) - \dim(C(\mathbf{P}^A_{10})) \overset{(2.1.17)}{=} \\
&= \dim(\mathcal{M}_{R+C}) - \dim(\mathcal{M}_R) = 1 + m_1 - 1 + m_2 - 1 - (1 + m_2 - 1) = \\
&= m_1 - 1 \\
\mathrm{r}(I - \mathbf{P}_{12}) &= m_1 m_2 l - (1 + m_1 - 1 + m_2 - 1) = m_1 m_2 l - m_1 - m_2 + 1.
\end{aligned}$$

Zbog toga što vrijedi

$$(\mathbf{P}_{12} - \mathbf{P}^A_{10})(I - \mathbf{P}_{12}) = \mathbf{P}_{12} - \mathbf{P}^2_{12} - \mathbf{P}^A_{10} + \mathbf{P}^A_{10}\mathbf{P}_{12} \underset{\text{tm1.7.b}}{\overset{(2.1.18)}{=}} \mathbf{P}_{12} - \mathbf{P}_{12} - \mathbf{P}^A_{10} + \mathbf{P}^A_{10} = 0$$

iz Teorema 1.9 slijedi da su brojnik i nazivnik od F^A_{10} nezavisni.

Nadalje, uz pretpostavku istinitosti modela iz Teorema 2.3 i Leme 2.4 slijedi

$$\frac{\|(\mathbf{P}_{12} - \mathbf{P}^A_{10})\mathbf{Y}\|^2}{\sigma^2} \sim \chi^2(m_1 - 1)$$

$$\frac{\|(I - \mathbf{P}_{12})\,\boldsymbol{Y}\|^2}{\sigma^2} \sim \chi^2(m_1 m_2 (l - 1))$$

iz čega po Definiciji 1.5 F-distribucije slijedi

$$F_{10}^A \sim F(m_1 - 1, m_1 m_2 l - m_1 - m_2 + 1). \qquad (❀_{A1})$$

Nadalje, testirajmo hipotezu

$$H_{02}^A : \theta_{\cdot j} = 0,\ \theta_{ij} = 0,\ i = 1, \ldots, m_1 - 1,\ j = 1, \ldots, m_2 - 1,$$

koja je ekvivalentna testiranju istinitosti modela (4.14) uz pretpostavku istinitosti aditivnog modela (4.10). Zbog toga što vrijedi $\mathcal{M}_C \subset \mathcal{M}_{R+C}$ ovo su ugniježđeni modeli. Za njega je test-statistika

$$F_{02}^A = \frac{\dfrac{\mathbf{Y}^T(\mathbf{P}_{12} - \mathbf{P}_{02}^A)\mathbf{Y}}{\mathrm{r}(\mathbf{P}_{12} - \mathbf{P}_{02}^A)}}{\dfrac{\mathbf{Y}^T(I - \mathbf{P}_{12})\mathbf{Y}}{\mathrm{r}(I - \mathbf{P}_{12})}},$$

gdje je

$$\begin{aligned}
\mathrm{r}(\mathbf{P}_{12} - \mathbf{P}_{02}^A) &= \mathrm{r}(\mathbf{P}_{12}) - \mathrm{r}(\mathbf{P}_{02}^A) = \dim(C(\mathbf{P}_{12})) - \dim(C(\mathbf{P}_{02}^A)) \overset{(2.1.17)}{=} \\
&= \dim(\mathcal{M}_{R+C}) - \dim(\mathcal{M}_C) = 1 + m_1 - 1 + m_2 - 1 - (1 + m_1 - 1) = \\
&= m_2 - 1 \\
\mathrm{r}(I - \mathbf{P}_{12}) &= m_1 m_2 l - m_1 - m_2 + 1.
\end{aligned}$$

Zbog toga što vrijedi

$$(\mathbf{P}_{12} - \mathbf{P}_{02}^A)(I - \mathbf{P}_{12}) = \mathbf{P}_{12} - \mathbf{P}_{12}^2 - \mathbf{P}_{02}^A + \mathbf{P}_{02}^A \mathbf{P}_{12} \underset{\mathrm{tm1.7.b}}{\overset{(2.1.18)}{=}} \mathbf{P}_{12} - \mathbf{P}_{12} - \mathbf{P}_{02}^A + \mathbf{P}_{02}^A = 0$$

iz Teorema 1.9 slijedi da su brojnik i nazivnik od F_{02}^A nezavisni.

Uz pretpostavku istinitosti modela iz Teorema 2.3 i Leme 2.4 slijedi

$$\frac{\|(\mathbf{P}_{12} - \mathbf{P}_{02}^A)\mathbf{Y}\|^2}{\sigma^2} \sim \chi^2(m_1 - 1)$$

$$\frac{\|(I - \mathbf{P}_{12})\mathbf{Y}\|^2}{\sigma^2} \sim \chi^2(m_1 m_2 (l - 1)),$$

iz čega po Definiciji 1.5 F-distribucije slijedi

$$F_{02}^A \sim F(m_2 - 1, m_1 m_2 l - m_1 - m_2 + 1). \qquad (❀_{A2})$$

Primijetimo još da ukoliko je broj elemenata u svakoj ćeliji jednak jedan, to jest $l = 1$, tada su dobivene F-distribucije:

$$F_{10}^A \sim F(m_1 - 1, (m_1 - 1)(m_2 - 1))$$

$$F_{02}^A \sim F(m_2 - 1, (m_1 - 1)(m_2 - 1)).$$

5 Praktični primjer

Ovaj primjer rađen je na temelju članka J. D. Austin, *An experimental study of the effects of three instructional methods in basic probability and statistics*. Pošto u njemu postoji dosta stvari koje nemaju veze s primjenom ANOVA-e one ovdje neće biti spomenute, dok će neke važnije za shvaćanje primjera biti spomenute, ali ne detaljno.

5.1 Opis eksperimenta

Provedeno je istraživanje kojim se željelo ustanoviti koja je metoda predavanja vjerojatnosti i statistike[18] uspješnija: P (grafički engl. *pictorial*), MP (eksperimenti i grafovi engl. *manipulative-pictorial*) ili S (simbolički engl. *symbolic*). [Ovo je tzv. Brunerov model - za više vidi [1].]

Subjekti istraživanja su bili pretežno studenti prve ili druge godine studija, kojih je na početku bilo ukupno 80, dok je u tablici 4 zapisano koliko ih je završilo eksperiment.

Tablica 4: Broj studenata koji su završili eksperiment

	Termini predavanja (faktor II)		
Metoda predavanja (faktor I)	MWF	TTS	ukupno
P	10	12	22
MP	11	10	21
S	16	12	28
Ukupno	37	34	71

Mogli su se upisati u jedan od dva moguća termina predavanja: MWF i TTS.[19] Prije nego što je eksperiment započeo, studenti su rangirani prema prethodnim ocjenama iz matematike. Prva tri najuspješnija studenta na slučajan su način dodijeljena različitim metodama P, MP i S, te tako do kraja. Za svaku metodu bilo je ukupno 12 predavanja, nakon kojih je uslijedio završni ispit koji je bio za sve jednak.

[18] To je bio jedan kolegij na sveučilištu Purdue (Indiana, SAD).

[19] *MWF* je engleska skraćenica za *Monday*, *Wednesday*, *Friday*, a *TTS* za *Tuesday*, *Thursday*, *Saturday*.

5.2 Statistička interpretacija

Primijetimo da je cilj ovog istraživanja zaključiti postoji li značajna razlika u rezultatima završnog ispita obzirom na različite metode. Iz uzorka studenata danog kolegija nastoji se zaključiti koji bi princip predavanja bio općenito bolji. Iako se čini da bi se mogla koristiti jednofaktorska ANOVA kojoj je faktor metoda, ovdje će se koristiti dvofaktorska ANOVA. To je zbog toga što se uzima u obzir da je za studente napravljena još jedna podjela pri upisu u različite termine (MWF i TTS).

Dakle, imamo dva faktora: metodu i termin predavanja. Od toga metoda ima tri, a termin dvije razine.

Važno je primijetiti da u tablici 4 broj subjekata u ćelijama nije jednak. Zbog toga što se naša cijela teorija zasnivala na pretpostavci jednakosti subjekata u svakoj ćeliji, te jer većina programa koristi tu istu teoriju ovdje će se uzeti da je l jednak njihovoj srednjoj vrijednosti. U ovom slučaju to je $l = (10 + 12 + 11 + 10 + 16 + 12)/6 = 11.8\dot{3}$.

5.3 Rezultati pomoćnog testiranja

Prije glavnog testiranja na rezultatima završnog ispita, napravljen je pomoćni test dvofaktorske ANOVA-e na prethodnim ocjenama iz matematike. To je zbog toga kako bi se uvjerili da u pojedinim ćelijama ili razredima nema značajne razlike u subjektima.

Slika 7: Rezultati testiranja prethodnih ocjena iz matematike

PREVIOUS MATHEMATICS GRADE: ANALYSIS AND CELL DATA

Source	*df*	*MS*	*F*	*p*
Treatment	2	1.82	1.50	.24
Classes	1	.41	.34	>.50
Interaction	2	.11	.09	>.50
Error	65	1.22		

Treatment	Class	Cell $\bar{X}_{ij}$	Cell s_{ij}^2	Treatment $\bar{X}_i$
	MWF	2.75	.87	
S	TTS	2.83	1.61	2.79
	MWF	2.10	.77	
P	TTS	2.42	1.17	2.27
	MWF	2.64	2.05	
MP	TTS	2.70	.90	2.67

Note.—Scale: A = 1, B = 2, C = 3, D = 4, and F = 5.

Ako se žele raditi testiranja ANOVA-e, prethodno se trebaju provjeriti njezine pretpostavke. Ovdje je korišten Bartlettov test za testiranje homogenosti varijanci, a za provjeru normalnosti se obično ili nacrta uzoračka distribucija ili napravi tablica statistika. Pošto su ovdje bili zadovoljeni uvjeti rezultati se mogu uzeti kao vjerodostojni.

Rezultati primjene dvofaktorske ANOVA-e prikazani su na slici 7 gdje se nalazi originalna tablica rezultata iz spomenutog članka.

Komentirajmo prvo gornji dio tablice. Tu su testirane prethodno spomenute nul hipoteze glavnih efekata i interakcije za ocjene iz matematike. Za rezultate testiranja hipoteze interakcije je dan treći redak pod engl. nazivom *interaction*. Vidimo da je naznačena formula dana u ❀$_{12}$ ovdje primijenjena, jer $(m_1-1)(m_2-1) = (3-1)(2-1) = 2$ i $m_1 m_2 (l-1) = 2 \cdot 3 \cdot (11.8\dot{3}-1) = 65$. Primijetimo da je drugi broj stavljen posebno pod *Error* jer će se pojaviti i za testiranje hipoteza glavnih efekata faktora. Brojevi koji se u 3.poglavlju označavali s s^2_{RC} i F_{crit} ovdje su dani redom $MS = 0.11$ i $F = 0.09$. Za interakciju p-vrijednost iznosi nešto više od 0.50, pa se može zaključiti da ne postoji interakcija između faktora (u eksperimentu se pretpostavlja razina značajnosti $\alpha = 0.05$, no to se moglo zaključiti i za ostale uobičajene razine značajnosti). Sada se rezultati za ostale hipoteze mogu interpretirati kao za aditivni model.

U prvom retku su analogno dobiveni brojevi za testiranje utjecaja glavog efekta faktora I (metode, ovdje engl. *Treatment*) pomoću formule ❀$_1$ (isti rezultat se dobije i primjenom formule ❀$_{A1}$ jer interakcija nije značajna) gdje $m_1 - 1 = 3 - 1 = 2$, te također $s^2_R = 1.82$ i $F_{crit} = 1.50$. P-vrijednost iznosi 0.24, što znači da na razini značajnosti od 5% ne možemo odbaciti nul-hipotezu o jednakosti između metoda P, MP i S. Također, u drugom retku je testiran utjecaj glavog efekta faktora II (termina, odnosno ovdje engl. *Class*) pomoću formule ❀$_2$ (odnosno ❀$_{A2}$) gdje $m_2 - 1 = 2 - 1 = 1$, te također $s^2_C = 0.41$ i $F_{crit} = 0.34$. P-vrijednost iznosi nešto više od 0.50, što znači da se na razini značajnosti od 5% ne može odbaciti nul-hipoteza o jednakosti termina MWF i TTS. Dakle, ništa nije ispalo značajno, što smo i željeli kako bi mogli prijeći na glavno testiranje.

Tablica ispod ove daje pregled osnovnih statistika po faktorima, i to redom srednje vrijednosti ćelija, standardne devijacije ćelija i srednje vrijednosti za različite metode. Ispod tablice su navedene oznake za ocjene koje su za nas obrnute od uobičajenog: 1 (Odličan), 2 (Vrlo dobar), 3 (Dobar), 4 (Dovoljan) i 5 (Nedovoljan). Radi usporedbe mogu se pogledati različite srednje vrijednosti metoda P, MP i S koje iznose redom: 2.27, 2.79 i 2.67. Primjećujemo da su metode MP i S slične po rezultatima i nešto lošije od metode P koja je imala najbolje rezultate. Testiranjem ta razlika nije ispala značajna da bi se mogla poopćiti na cijelu populaciju, ali ovo je samo ilustracije radi.

5.4 Rezultati glavnog testiranja

Ovdje je napravljeno testiranje dvofaktorske ANOVA-e za ukupne bodove završnog ispita. Uz to, pošto se završni ispit sastojao od 4 dijela, istraživači su željeli saznati postoji li značajna razlika u nekim dijelovima rješavanja ispita.

TABLE 4
CELL DATA FOR DEPENDENT VARIABLES

Variable	Treatment	Class	Cell $\bar{X}_{ij}$	Cell s^2_{ij}	Treatment $\bar{X}_i$	Bartlett's Test[a]
Comprehension	S	MWF	27.75	35.93	29.04	5.57
		TTS	30.75	96.20		
	P	MWF	34.30	81.79	34.68	
		TTS	35.00	32.73		
	MP	MWF	32.00	76.60	32.38	
		TTS	32.80	48.62		
Computation	S	MWF	12.06	19.80	13.39	16.46**
		TTS	15.17	2.52		
	P	MWF	14.00	11.78	14.00	
		TTS	14.00	9.64		
	MP	MWF	12.91	23.09	13.48	
		TTS	14.10	4.32		
Application	S	MWF	25.00	316.53	25.11	5.79
		TTS	25.25	84.39		
	P	MWF	40.40	213.48	37.32	
		TTS	34.75	230.02		
	MP	MWF	32.64	341.65	37.10	
		TTS	42.00	187.56		
Analysis	S	MWF	10.06	69.26	11.25	3.79
		TTS	12.83	35.61		
	P	MWF	18.90	112.32	17.82	
		TTS	16.92	82.63		
	MP	MWF	16.27	50.02	17.33	
		TTS	18.50	77.39		
Total	S	MWF	74.87	983.18	78.79	3.23
		TTS	84.00	448.91		
	P	MWF	107.60	1126.93	103.82	
		TTS	100.67	844.24		
	MP	MWF	93.82	1228.76	100.29	
		TTS	107.40	668.93		

Note.—Maximum possible score: Comprehension 44, Computation 16, Application 64, Analysis 36, and Total 160.
[a] Theoretical distribution is Chi-square with 5 degrees of freedom.
* Significant at the .05 level.
** Significant at the .01 level.

Slika 8: Osnovne statistike testiranja i Bartlettov test

Različiti dijelovi ispita su bili: razumijevanje (engl. *comprehension*), računanje (engl. *computation*), primjena (engl. *application*) i analiza (engl.

analysis).[20]

Originalne statistike članka, kao i Bartlettov test homogenosti dani su na slici 8. Ovdje se kod dijela *Computation* (računanja) javlja problem, jer na razini značajnosti od 1% nije zadovoljen uvjet homogenosti varijanci.

TABLE 5
ANALYSIS-OF-VARIANCE TABLES

						Pooling			
Variance	Source	*df*	*MS*	*F*	*p*	*df*	*MS*	*F*	*p*
Total	Trmt.	2	4474.82	5.07**	.01	2	4673.02	5.37**	.01
	Class	1	478.40	.54	>.50				
	Inter.	2	637.71	.72	>.50				
	Error	65	883.43			68	870.92		
Comprehension	Trmt.	2	181.39	3.03	.06	2	201.85	3.46*	.04
	Class	1	38.94	.65	>.50				
	Inter.	2	10.60	.18	>.50				
	Error	65	59.95			68	58.30		
Application	Trmt.	2	1269.47	5.42**	.01	2	1242.10	5.32**	.01
	Class	1	30.21	.13	>.50				
	Inter.	2	305.58	1.30	.29				
	Error	65	234.33			68	233.31		
Analysis	Trmt.	2	323.14	4.62*	.02	2	341.15	4.99**	.01
	Class	1	17.48	.25	>.50				
	Inter.	2	38.82	.56	>.50				
	Error	65	69.96			68	68.34		
Computation MWF	Trmt.	2	11.62	.62	>.50				
	Error	34	18.64						
Computation TTS	Trmt.	2	4.94	.89	>.50				
	Error	31	5.57						

* Significant at the .05 level.
** Significant at the .01 level.

Slika 9: Glavno testiranje

Na slici 9 su dani rezultati iz članka u kojima se jedino nije provela ANOVA za dio računanja (*Computation*) jer nisu bili zadovoljeni uvjeti. Umjesto toga, za taj dio provedena je jednofaktorska ANOVA posebno za svaki od termina MWF i TTS, te možemo vidjeti da utjecaj faktora metode niti u jednom slučaju nije bio značajan. Također, interakcija i faktor termina niti za jedan dio ispita nisu ispali značajni. Još se može uočiti da je faktor metode značajan za ukupne bodove na završnom ispitu i za dio primjene na

[20] Ovo je tzv. Bloomova klasifikacija.

razini značajnosti od 1%, dok je za dio analize na razini značajnosti od 5%.

Dio tablice označen sa *Pooling* odnosi se na provođenje testa na reduciranom modelu. To znači da ako se testiranjem dobilo da neki efekti nisu značajni, oni su izbačeni iz modela.[21] Time se jasnije vidi koliko su zapravo značajni efekti koji imaju utjecaja na varijablu odziva. Ono što se tada dobilo dodatno je da je faktor metode za dio primjene postao značajan za razinu značajnosti od 1%, te da je za dio razumijevanja također postao značajan za razinu značajnosti od 5%.

TABLE 6
PAIRWISE ANALYSIS OF TREATMENT AVERAGES

Variable	Treatment Pair	Difference	Scheffé's Test[a]	P
Total	P—S	25.03[b]	4.43*	.02
	MP—S	21.50	3.18*	.05
(s^2 = 870.92)	P—MP	3.53	.08	>.50
Comprehension	P—S	5.65	3.37*	.04
	MP—S	3.35	1.15	.34
(s^2 = 58.30)	P—MP	2.30	.49	>.50
Application	P—S	12.21	3.94*	.02
	MP—S	11.99	3.70*	.03
(s^2 = 233.31)	P—MP	.22	.00	>.50
Analysis	P—S	6.57	3.89*	.03
	MP—S	6.08	3.25*	.05
(s^2 = 68.34)	P—MP	.49	.02	>.50

[a] The computed value has a $F(2, 68)$ distribution.
[b] Here $\bar{x}_P - \bar{x}_S = 25.03$ or the average score of treatment group P was 25.03 points higher than the average score of treatment group S on the (Total) variable.
* Significant at the .05 level.
** Significant at the .01 level.

Slika 10: Scheffeov test za svaku razinu

Zbog toga što postoje tri razine faktora metode nije jasno koji od njih prave razliku. Rezultati Scheffeovog testa koji testira razliku svake razine faktora dani su na slici 10. Tu vidimo da kod svih postoji značajna razlika između P i S razine, dok razlika između MP i S nije značajna jedino u dijelu razumijevanja.

[21] Ovdje je to napravljeno prema Bozivichovom principu, što znači da su se odbacili oni efekti čija je p-vrijednost bila veća od 0.25.

5.5 Zaključak

Testiranjem se pokazalo se da su se ukupni bodovi, analiza i primjena na završnom ispitu značajno povećali kod metoda P i MP u odnosu na metodu S. Za razumijevanje su se povećali rezultati kod P metode u odnosu na S metodu, dok je za MP metodu *moguće* da su se povećali u odnosu na S metodu. Kod računanja nije bilo bitne razlike u metodama.

Literatura

[1] J. D. Austin, *An experimental study of the effects of three instructional methods in basic probability and statistics*, Journal for Research in Mathematical Education, Vol. 5, No. 3 (1974), 146-154

[2] R. Christensen, *Plane Answers to Complex Questions*, 4th Edition, Springer, United States of America, 2011.

[3] S. E. Maxwell, Harold D. Delaney, *Designing Experiments and Analyzing Data - A Model Comparison Perspective*, 2nd Edition, Lawrence Erlbaum Associates, United States of America, 2004.

[4] D. C. Montgomery, *Design and Analysis of Experiments*, 5th Edition, United States of America, Wiley, 2001.

[5] R. R. Pagano, *Understanding Statistics in the Behavioral Sciences*, Wadsworth, Canada, 2009.

[6] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, 1993.

[7] N. H. Timm, *Applied Multivariate Analysis*, Springer, United States of America, 2002.

Sažetak

U ovom radu je obrađen model dvofaktorske analize varijance, te je također ilustriran na primjerima iz psihologije. Dvofaktorska ANOVA se koristi pri testiranju razlike između dvaju ili više grupa. Egzaktni dvofaktorski model ANOVA-e uveden je pomoću regresijskog modela koji je baziran na projekcijama.

U prvom poglavlju su dani od prije znani rezultati i definicije. U drugom poglavlju su uvedena osnovna svojstva regresijskog modela koja će se koristiti.

Dvofaktorska ANOVA je jak statistički test iz kojeg se lako mogu donijeti potrebni zaključci, no njegovi rezultati su često u tablicama koje se treba naučiti interpretirati. Tome je posvećeno 3. poglavlje s radnim primjerom iz psihologije. To je uvod u dvofaktorsku ANOVA-u u kojem je sadržan intuitivan pristup njezinim osnovnim svojstvima. Osnovni pojmovi su faktori, glavni efekti, jednostavni glavni efekti, interakcija, balansirani dizajn i drugi.

U četvrtom poglavlju je uveden dvofaktorski ANOVA model. Posebno je napravljen za opći i aditivni model.

U petom poglavlju je dan praktični primjer u psihologiji. Baziran je na *An experimental study of the effects of three instructional methods in basic probability and statistics*. To je članak objavljen u *Journal for Research in Mathematical Education* 1974. godine.

Summary

In this thesis is analyzed two-way ANOVA model and there are also given some examples in psychology. Two-way ANOVA is used when you need to test the difference between two or more groups. The exact two way ANOVA model is here built with applications of regression model which is based on projections.

In the first chapter are given well known results and definitions. In the second chapter are basic regression properties which will be used.

The Two-way ANOVA is a strong statistical test which easily can be used to draw conclusions, but the output of this test includes more tables which you first have to become acquainted with. This is done in the third chapter with illustrative example in psychology. It is an introduction to two-way ANOVA and it contains an intuitive approach to it's basic features. The main concepts are factors, main effects, simple effects, interaction, orthogonal design and others.

In the fourth chapter is introduced exact two way ANOVA model. It is done separately for saturated and additive model.

In the fifth chapter is given pratical example in psychology. Example is based on *An experimental study of the effects of three instructional methods in basic probability and statistics*. It is an 1974 article published in *Journal for Research in Mathematical Education*.

Životopis

Rođena sam 22. travnja 1988. godine u Vukovaru. U Zagrebu sam završila osnovnu školu Pavleka Miškine i prirodoslovno-matematičku gimnaziju Lucijana Vranjanina.

2007. godine upisala sam Preddiplomski studij matematike na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu, dok sam ga 2010. godine završila i upisala Diplomski studij Matematičke statistike.