

**SNJEŽANA PIVAC**

ISBN 978-953-281-033-2

**STATISTIČKE METODE**  
(predavanja, diplomski studij,  
kolegij "Statističke metode")

**e-nastavni materijal**

**Split, 2010.**



**Prof. dr. sc. Snježana Pivac**  
**Ekonomski fakultet u Splitu**

**ISBN 978-953-281-033-2**

# **STATISTIČKE METODE**

**e-nastavni materijal**

**Split, 2010.**

**Prof. dr. sc. Snježana Pivac**  
**Ekonomski fakultet u Splitu**

**ISBN 978-953-281-033-2**

**Statističke metode**  
**3/2011**  
**e-nastavni materijal**

Recenzenti:

Prof. dr. sc. Mirjana Čižmešija, Sveučilište u Zagrebu, Ekonomski fakultet  
Zagreb

Dr. sc. Branka Marasović, Sveučilište u Splitu, Ekonomski fakultet Split

Tehnički urednik:

Prof. dr. sc. Snježana Pivac

## SADRŽAJ

<b>PREDGOVOR.....</b>	<b>9</b>
<b>1 DEFINIRANJE VARIJABLI I NJIHOVO MJERENJE .....</b>	<b>11</b>
1.1 STATISTIKA .....	11
1.2 PROGRAMSKA POTPORA ZA PRIMJENU STATISTIČKIH METODA U KONKRETNIM ANALIZAMA.....	13
1.3 IZVORI PODATAKA I METODE NJIHOVA PRIBAVLJANJA .....	15
1.4 PRIPREMA PODATAKA ZA STATISTIČKU ANALIZU .....	18
1.5 UZORAK.....	18
1.6 UNOŠENJE VARIJABLI U BAZU PODATAKA U RAČUNALU.....	21
1.7 UREĐIVANJE I PRIKAZIVANJE PODATAKA.....	29
1.8 SREDNJE VRIJEDNOSTI .....	51
1.9 DISPERZIJA.....	55
1.10 ASIMETRIJA .....	59
1.11 ZAobljenost .....	61
1.12 SLOŽENO PRIKAZIVANJE STATISTIČKIH PODATAKA .....	72
<b>2 VJEROJATNOST I PROCJENA PROSJEČNE VRIJEDNOSTI.....</b>	<b>83</b>
2.1 VJEROJATNOST .....	83
2.2 DISKONTINUIRANA SLUČAJNA VARIJABLA.....	89
2.3 KONTINUIRANA SLUČAJNA VARIJABLA .....	107
2.4 PROCJENA PROSJEČNE VRIJEDNOSTI.....	112
<b>3 TESTIRANJE HIPOTEZA SA ZAVISNIM I NEZAVISNIM UZORCIMA.....</b>	<b>121</b>
3.1 ZNANSTVENE I STATISTIČKE HIPOTEZE .....	121
3.2 TESTIRANJE HIPOTEZE O PROSJEČNOJ VRIJEDNOSTI JEDNOG OSNOVNOG SKUPA .....	123
3.3 TESTIRANJE HIPOTEZE O RAZLICI PROSJEČNIH VRIJEDNOSTI DVAJU NEZAVISNIH OSNOVNIH SKUPOVA .....	125
3.4 TESTIRANJE HIPOTEZE O RAZLICI PROSJEČNIH VRIJEDNOSTI DVAJU ZAVISNIH OSNOVNIH SKUPOVA.....	135
3.5 TESTIRANJE HIPOTEZE O NEZAVISNOSTI DVAJU KVALITATIVNIH OBILJEŽJA ELEMENTA OSNOVNOG SKUPA.....	148
3.6 TESTIRANJE HIPOTEZE DA DISTRIBUCIJA IMA ODREĐENI OBLIK.....	164
3.6.1 Kolmogorov - Smirnov test.....	166

3.7	TESTIRANJE HIPOTEZA SA ZAVISNIM UZORCIMA .....	173
3.7.1	<i>McNemarov test za dva zavisna uzorka .....</i>	173
3.7.2	<i>Friedman test za više od dva zavisna uzorka .....</i>	178
3.8	TESTIRANJE HIPOTEZA S NEZAVISNIM UZORCIMA .....	186
3.8.1	<i>Mann-Whitney U-test za dva nezavisna uzorka (Wilcoxon T-test ili test zbroja rangova).....</i>	186
3.8.2	<i>Kruskal-Wallis test za više od dva nezavisna uzorka.....</i>	192
3.9	ANALIZA UTJECAJA PROMJENJIVOG/IH FAKTORA NA KRETANJE SLUČAJNE VARIJABLE .....	200
3.9.1	<i>Djelovanje jednog promjenjivog faktora na kretanje slučajne varijable (nezavisni uzorci) .....</i>	201
3.9.2	<i>Djelovanje dva promjenjiva faktora na kretanje slučajne varijable.....</i>	210
3.10	MULTIVARIANTNA CLUSTER ANALIZA.....	218
<b>4</b>	<b>ISPITIVANJE OVISNOSTI IZMEĐU NUMERIČKIH I NEKIH NENUMERIČKIH (DUMMY) VARIJABLI .....</b>	<b>233</b>
4.1	DIJAGRAM RASIPANJA .....	233
4.2	KOEFICIJENT LINEARNE KORELACIJE .....	238
4.3	KOEFICIJENT KORELACIJE RANGA .....	242
4.4	KOEFICIJENT PARCIJALNE KORELACIJE .....	246
4.5	KENDALLOV KOEFICIJENT KORELACIJE RANGA .....	249
4.6	REGRESIJSKA ANALIZA .....	253
4.7	VIŠESTRUKA (MULTIPLA) REGRESIJA .....	275
4.7.1	<i>Metode odabira varijabli u modelu višestruke regresije.....</i>	278
4.7.2	<i>Problem međuovisnosti regresorskih varijabli .....</i>	279
4.8	REGRESIJSKO MODELIRANJE U UVJETIMA NARUŠENIH OSNOVNIH PRETPOSTAVKI .....	301
4.8.1	<i>Problem heteroskedastičnosti varijance reziduala .....</i>	301
4.8.2	<i>Problem autokorelacije reziduala .....</i>	306
4.9	NENUMERIČKE DUMMY VARIJABLE I POSLOVNE PROGNOZE.....	336
4.9.1	<i>Dummy varijable konstantnog člana.....</i>	337
4.9.2	<i>Dummy varijable za promjene u nagibu.....</i>	339
4.9.3	<i>Sezonske dummy varijable .....</i>	340
<b>5</b>	<b>TABLICE ODABRANIH STATISTIČKIH DISTRIBUCIJA .....</b>	<b>347</b>
A	POVRŠINE ISPOD NORMALNE KRIVULJE .....	347
B	KRITIČNE VRIJEDNOSTI T, STUDENTOVE DISTRIBUCIJE .....	348

C1	KRITIČNE VRIJEDNOSTI HI-KVADRAT DISTRIBUCIJE (ZA $P \leq 0,05$ ).....	349
C2	KRITIČNE VRIJEDNOSTI HI-KVADRAT DISTRIBUCIJE (ZA $P > 0,05$ ).....	350
D1	KRITIČNE VRIJEDNOSTI F-DISTRIBUCIJE (ZA $P = 0,05$ ) .....	351
D2	KRITIČNE VRIJEDNOSTI F-DISTRIBUCIJE (ZA $P = 0,01$ ) .....	352
E1	KRITIČNE VRIJEDNOSTI DURBIN-WATSONOVOG POKAZATELJA (ZA $P = 0,05$ ) .....	353
E2	KRITIČNE VRIJEDNOSTI DURBIN-WATSONOVOG POKAZATELJA (ZA $P = 0,01$ ) .....	354
<b>6</b>	<b>LITERATURA .....</b>	<b>355</b>





## PREDGOVOR

Integrirani tekst predavanja "Statističke metode" namijenjen je prvenstveno studentima diplomskog studija Ekonomskog fakulteta u Splitu i prati nastavni plan i program kolegija "Statističke metode". Predavanjima su obuhvaćene teorijske osnove i objašnjenja za svako obrađeno i uključeno područje statistike. Kroz rješavanje mnoštva izvornih konkretnih primjera daju se objašnjenja dobivenih rezultata i njihovo kritičko vrednovanje.

Kao posebnu vrijednost potrebno je istaknuti da su gotovo svi primjeri i zadaci riješeni uz upotrebu statističkog programa za računala *SPSS 16.0* (Statistical Package for the Social Sciences), koji je upravo prilagođen društvenim, pa tako i ekonomskim istraživanjima. Ovaj program pruža mnoštvo mogućnosti za provođenje statičkih metoda i tehnika na konkretnim analizama. Svaki primjer riješen u *SPSS*-u u integriranom tekstu ima oznaku u obliku slike *CD*-a. Takvi primjeri su priloženi u elektroničkom obliku u dokumentu *SPSS* i svaki od njih ima i svoje rješenje u posebnom dokumentu *Output*-a programa *SPSS*. Sve je to na *CD*-u koji prati ova predavanja i vježbe kolegija Statističke metode. Na taj način studentima je omogućena i kontrola valjanosti usvojenoga gradiva. Tekst na taj način može osposobiti čitatelja da samostalno statistički analizira određene pojave na stručno zadovoljavajući način i da uz upoznavanje odabrane statističke metodologije savlada rad i rješavanje statističkih problema na računalu. Naime, upotrebom statističkih paketa, počevši već od pripremne faze statističkog istraživanja, znatno se skraćuje i pojednostavljuje vrijeme potrebno za primjenu statističkih tehnika. Na taj se način statistički postupci približavaju mnogim korisnicima, a u nastavu se uvode suvremene metode rada koristeći statističke pakete.

Studenti koji nastave obrazovanje na poslijediplomskim studijima bit će pripremljeni za samostalni stručni i znanstveni rad u svojim istraživanjima (detaljna obrada ankete i/ili istraživanje međuovisnosti između različitih ekonomskih veličina), gdje se ekonomski problemi statistički rješavaju upotrebom računala. Dio studenata koji će se nakon diplomiranja neposredno uključiti u poslovnu praksu, također će imati koristi, zbog mogućnosti upotrebe usvojenog znanja statističke teorijske i programske potpore pri poslovnim i ekonomskim analizama.

Potrebno je napomenuti da je program *SPSS* kompatibilan s *Microsoft Excelom*, pa se već postojeći podaci iz jednog mogu jednostavno kopirati u drugi program. To je važna činjenica, s obzirom da je poznato da je *Microsoft Excel* jedan

od najraširenijih programa za obradu velikih baza podataka te je lako dostupan većini korisnika.

U integriranom tekstu predavanja korištene su oznake i simboli, koji su preuzeti iz standardne statističke literature. Ako se, pak, u literaturi koriste različite oznake, upotrijebljena je češće spominjana verzija.

Ovaj rukopis je nastao kao rezultat višegodišnjeg iskustva autorice kod primjene statističkih metoda u izradi brojnih znanstvenih i stručnih radova, studija i analiza te kao rezultat predavačkog iskustva pri prenošenju znanja iz područja statistike na mnoge generacije studenata Ekonomskog i drugih fakulteta.

Split, ožujak 2010.

*Autorica*

# 1 DEFINIRANJE VARIJABLI I NJIHOVO MJERENJE

## 1.1 Statistika

**Statistika je posebna znanstvena disciplina** koja u svrhu realizacije postavljenih ciljeva istraživanja na organiziran način prikuplja, odabire, grupira, prezentira i vrši analizu informacija ili podataka, te interpretira rezultate provedene analize.

**Statistika se kao znanstvena disciplina može podijeliti na:**

1. deskriptivnu statistiku i
2. inferencijalnu statistiku.

**Deskriptivna ili opisna statistika temelji se na potpunom obuhvatu statističkog skupa, čiju masu podataka organizirano prikuplja, odabire, grupira, prezentira i interpretira dobivene rezultate analize.** Na taj način se, izračunavanjem različitih karakteristika statističkog skupa, sirova statistička građa svodi na lakše razumljivu i jednostavniju formu. Ako se statističke metode i tehnike primjenjuju na **čitav statistički skup**, dakle ako su istraživanjem obuhvaćeni svi elementi skupa oni tvore **statističku populaciju**.

**Inferencijalna statistika temelji se na dijelu (uzorku) jedinica izabranih iz cjelovitog statističkog skupa, pomoću kojeg se uz primjenu odgovarajućih statističkih metoda i tehnika donose zaključci o čitavom statističkom skupu.** Uvijek je prisutan odgovarajući stupanj rizika kada se koriste rezultati iz uzorka, za kojeg je poželjno da bude izabran na slučajan način i da bude reprezentativan. Inferencijalna statistika pripada skupini **induktivnih metoda**, kojima se izvode zaključci polazeći **od posebnoga prema općem**.

**Statističke metode temelj su za provođenje statističke analize društvenih i prirodnih pojava.**

**Predmet proučavanja statistike su određene zakonitosti koje se javljaju u masovnim pojavama. Zadaća statistike je da uoči zakonitosti u masovnim i slučajnim pojavama, te da ih iskaže brojčano.**

**Masovne pojave su skupine istovrsnih elemenata, koji imaju jedno ili više zajedničkih svojstava. Takvu skupinu nazivamo statističkom masom ili statističkim skupom.**

**Statistički skup potrebno je definirati pojmovno, prostorno i vremenski.**

**Pojmovno** odrediti statistički skup podrazumijeva odrediti pojam ili svojstvo svakog elementa promatranog skupa.

**Prostorno** odrediti statistički skup znači odrediti prostor na koji se odnosi ili kojemu pripadaju elementi statističkog skupa.

**Vremenski** odrediti statistički skup znači odrediti vremenski trenutak ili razdoblje kojim će se obuhvatiti svi elementi koji ulaze u statistički skup.

**Primjer 1.1:**

**Statistički skup**, "studenti 1. godine diplomskog studija Ekonomskog fakulteta u Splitu u Republici Hrvatskoj u akademskoj godini 2008./09, na dan 24.10.2008. godine", je vrlo precizno određen i iz takve njegove definicije mogu se odrediti svi njegovi elementi.

Svojstvo svakog elementa odnosno studenta definirano je najprije **pojmovno**, tj. jasno je da se radi o studentima 1. godine diplomskog studija Ekonomskog fakulteta.

Elementi ovog skupa određeni su i **prostorno**, tj. nalaze se u Splitu na području Republike Hrvatske.

**Vremenska** definicija ovog statističkog skupa upućuje na vremenski interval s kojim su elementi skupa obuhvaćeni tj. akademsku godinu 2008./09. i vremenski trenutak u kojem je izvršena selekcija studenata tj. 24.10.2008. godine.

**Statistička obilježja (varijable) su opće karakteristike elemenata statističkog skupa, po kojima su ti elementi međusobno slični i po kojima se međusobno razlikuju.**

**Primjer 1.2:**

**Statistički skup**, "studenti 1. godine diplomskog studija Ekonomskog fakulteta u Splitu u Republici Hrvatskoj u akademskoj godini 2008./09, na dan 24.10.2008. godine", promatran je prema **obilježju** "spol".

Sasvim je jasno da je spol jedna od karakteristika svakog studenta. Postoji muški i ženski spol. Neki studenti su jednakog spola, a neki se razlikuju po spolu.

## 1.2 Programska potpora za primjenu statističkih metoda u konkretnim analizama

Prikupljeni statistički podaci predstavljaju "**sirovu**" statističku građu koju je potrebno na odgovarajući način urediti i pripremiti za analizu.

Danas se statistički podaci uglavnom obrađuju prikladnim programima za računalo. U praksi postoje suvremeni statistički programski paketi koji omogućavaju, uz vrlo jednostavno rukovanje, unos, pripremu i obradu prikupljenih statističkih podataka.

Primjeri suvremenih statističkih paketa su: STATISTICA, SAS, SPSS. U ovoj knjizi primjeri i zadaci obrađeni su u programu **SPSS 16.0**, čija kratica odgovara engleskom nazivu programa *Statistical Package for the Social Sciences*, koji je upravo prilagođen društvenim, pa tako i ekonomskim istraživanjima. Statistički paketi se koriste za složenije statističke analize.

Potrebno je napomenuti da je jedan od najpopularnijih softwera (programskih jezika) u provođenju različitih aspekata statističke analize posebno za velike baze podataka i Microsoft Excel. **MS Excel** je u funkciji svih bitnih faza procesa statističke analize od formiranja baze podataka, sređivanja i grupiranja podataka, grafičkog prikazivanja statističkih nizova, izračuna temeljnih karakteristika statističkog niza pa sve do složenijih statističkih analiza i procedura i analize vremenskih nizova. Ovaj program je **kompatibilan sa SPSS-om**, pa se već postojeći podaci iz jednog programa mogu jednostavno kopirati u drugi.

Upotrebom statističkih paketa, počevši već od pripremne faze statističkog istraživanja, znatno se skraćuje i pojednostavljuje vrijeme potrebno za primjenu statističkih metoda i tehnika. Na taj se način statistički postupci približavaju mnogim korisnicima.

Prikupljene statističke podatke potrebno je unijeti i pohraniti u odgovarajuću datoteku odabranog statističkog paketa. Kada se unose podaci kvalitativnog ili opisnog karaktera, potrebno je izvršiti šifriranje ili kodiranje vrijednosti takvih obilježja.

Potrebno je napomenuti da je u programskom paketu **SPSS** oznaka za statistička obilježja: *Variable Type*.

Općenito se statistička obilježja mogu podijeliti na:

1. **kvalitativna statistička obilježja**, koja mogu poprimiti različite oblike, ali se *izražavaju opisno*. Ako se modalitetima ovog obilježja slučajno pridruže

brojevi, s njima *nisu dopuštene nikakve računske operacije* (u **SPSS-u**: *Variable Type: String*).

2. **kvantitativna statistička obilježja**, koja *se izražavaju bročano* (u **SPSS-u**: *Variable Type: Numeric*).

**Kvalitativna statistička obilježja** se mogu podijeliti na: a) **nominalna statistička obilježja** i b) **redoslijedna statistička obilježja**.

- a) **Nominalna statistička obilježja** *se izražavaju opisno*.
- b) **Redoslijedna statistička obilježja** *se mijenjaju prema intenzitetu ili rangu*. Na primjer: uspjeh na ispitu iz predmeta "Statističke metode" (ovdje se javljaju modaliteti obilježja od 1 do 5, ali se s njima *ne vrše nikakve računske operacije*, već se pomoću njih elementi skupa, odnosno studenti, mogu rangirati).

**Kvantitativna statistička obilježja** se još nazivaju i **numerička** statistička obilježja, a mogu se podijeliti na: a) **neprekidna ili kontinuirana statistička obilježja** i b) **prekidna ili diskontinuirana statistička obilježja**.

- a) **Neprekidna ili kontinuirana statistička obilježja** su takva numerička obilježja koja mogu poprimiti neprebrojivo beskonačno mnogo vrijednosti. Primjeri takvog obilježja su: visina, težina, duljina, starost itd.
- b) **Prekidna ili diskontinuirana statistička obilježja** su takva numerička obilježja koja mogu poprimiti prebrojivo beskonačno mnogo vrijednosti. Primjeri takvog obilježja su: broj djece, broj učenika u razredu, navršene godine starosti, visina plaće u kunama itd.

### 1.3 Izvori podataka i metode njihova pribavljanja

Osnovne faze statističkog istraživanja su:

1. statističko promatranje,
2. grupiranje (tabelarno i grafičko prikazivanje statističkih podataka),
3. statistička analiza i interpretacija rezultata provedene analize.

**Statističko promatranje** je organizirano prikupljanje statističkih podataka. Pojedinačne metode statističkog promatranja su:

- a) mjerenje,
- b) brojanje,
- c) ocjenjivanje,
- d) evidentiranje i
- e) anketiranje.

Jedan od načina statističkog promatranja i prikupljanja podataka je **mjerenje**. Mjeri se na primjer urod pšenice po jedinici poljoprivredne površine. Može se mjeriti težina proizvoda, kao i visina stanovništva nekog područja.

**Brojanjem** se može doći do podataka o broju zaposlenih u pojedinim organizacijskim jedinicama poduzeća, o broju upisanih učenika u srednje škole. Broje se i noćenja turista u turističkoj sezoni, pa se na taj način dobivaju podaci relevantni za analizu uspješnosti sezone.

**Ocjenjivanjem** kao metodom prikupljanja statističkih podataka određuje se kvaliteta provođenja određenih radnji, pa se stoga taj postupak obično veže za redoslijedna obilježja statističkog skupa. Ocjenjuju se i rangiraju studenti na testu iz statistike, ocjenjuje se usluga nekog hotelskog poduzeća i slično.

**Evidentiranje** podataka podrazumijeva kontinuirano praćenje kretanja neke pojave u duljem ili kraćem vremenskom razdoblju. Za potrebe evidencije često se uređuju odgovarajući obrasci. Na primjer pri promatranju izostanaka i broja ostvarenih radnih sati djelatnika u nekom poduzeću postoji obrazac za evidenciju gdje zaposlenici upisuju: ime i prezime, vrijeme dolaska i odlaska s radnog mjesta, vrijeme izostanka, razlog izostanka. Evidentiranje se može vršiti i tehničkim uređajima. To su umrežena računala, optički čitači, uređaji za brojenje npr. putnika i slično.

**Anketa i/ili intervju** je metoda kojom se prikupljaju podaci uz pomoć unaprijed pripremljenih upitnika, na kojima ispitanici svojim odgovorima daju

informacije o promatranim obilježjima statističkog skupa. Da bi anketa uspjela potrebno je veliku pozornost obratiti sastavljanju upitnika. Sastavlja je statističar, koji se često konzultira i s psihologom. Upiti moraju biti kratki, precizni i jasni. Moraju biti postavljeni tako da ne sugeriraju odgovor. Broj pitanja ne smije biti velik da ne zamara one koji odgovaraju. Pri provođenju ankete pristup ispitaniku, odnosno jedinici statističkog skupa može izravan i neizravan.

**Izravan pristup** ostvaruje se kada osoba ili tim koji provodi anketu izlaze na teren i u direktnom kontaktu s ispitanicima prikuplja odgovore na pitanja iz upitnika. **Neizravan pristup** ostvaruje se putem pošte, telefonom, elektroničkom poštom i web-om. Na ovaj način smanjuju se troškovi prikupljanja podataka (npr. putni troškovi osoba koje provode anketu tj. anketara). Iako se i na ovaj način ispitanicima prezentira tko provodi i koja je svrha istraživanja, praksa je pokazala da je ovim neizravnim pristupom anketiranja prisutan velik postotak neodaziva, kao i često velik postotak nevaljano i nepotpuno popunjenih upitnika. Naravno, razloge treba tražiti u činjenici da ispitaniku nije na raspolaganju osoba koja će mu pojasniti nejasna pitanja.

Ovisno o obimu istraživanja **organizaciju prikupljanja podataka može provoditi jedna osoba, skupina istraživača ili osoblje specijalizirane ustanove**, kojoj je to osnovna djelatnost. Ako se radi o manjem istraživanju, u prikupljanju podataka će sudjelovati manji broj istraživača, dok će veće istraživanje zahtijevati rad većeg broja istraživača.

Nakon precizne definicije zadatka, cilja i predmeta istraživanja tj. statističkog skupa pristupa se organiziranom prikupljanju statističkih podataka. Uspješnost i objektivnost ovog prvog koraka uvjetuje kvalitetu rezultata ostalih faza statističkog istraživanja. Nepotpune i neistinite prikupljene informacije do kojih bi se došlo u ovoj fazi značile bi da konačan rezultat statističkog istraživanja sadrži grešku. Pri tom **greška može biti sistematska i slučajna**.

**Sistematsku grešku** je lakše uočiti jer se ona ponavlja (npr. neispravnost određenog mjernog instrumenta, neistinito izjašnjavanje ispitanika).

**Slučajnu grešku** je teško precizno identificirati, jer se ona ne javlja kod svakog mjerenja i ne javlja se istim intenzitetom. Stoga se kod slučajne greške često veže pretpostavka o poništavanju njenog utjecaja na globalnoj razini promatranja.

U ovisnosti o **karakteru izvora podataka**, statistički podaci se dijele na:

- a) sekundarne podatke i
- b) primarne podatke.



**Sekundarni podaci** su oni koji se pribavljaju iz već postojećih baza podataka različitih državnih ustanova. Takvi se podaci prikupljaju sustavno na odgovarajući način, a njihov opseg ne ovisi o donošenju neke poslovne odluke ili zadanom cilju nekakvog istraživanja.

Takve podatke u Hrvatskoj prikupljaju: Državni zavod za statistiku, Hrvatska narodna banka, Hrvatska gospodarska komora, te neke druge specijalizirane agencije. Jedan od najčešće korištenih sekundarnih izvora podataka u Hrvatskoj je Statistički ljetopis Hrvatske u izdanju Hrvatskog zavoda za statistiku. U svjetskim okvirima poznat je World Statistical Yearbook, a putem Internet-a su danas dostupne mnoge baze sekundarnih podataka (na primjer: Eurostat, U.S: Census Bureau i slično). Na razini poduzeća, raznovrsna specifična izvješća o poslovanju i zaposlenicima imaju sekundarni karakter.

Sekundarni podaci su uglavnom brojevi. Predloženi su tablicama, a vrlo često i grafičkim prikazima.

**Primarni podaci** prikupljaju se neposrednim promatranjem svojstava elemenata statističkog skupa u skladu s unaprijed definiranim ciljevima statističkog istraživanja. Prikupljanje ovih podataka zahtjeva definiranje statističkog skupa, izbor obilježja koja se žele istražiti, određivanje modaliteta promatranog obilježja, pripremanje anketnih upitnika i/ili pratećih formulara te organiziranje i provođenje samog prikupljanja podataka.

Ako se podaci prikupljaju za sve članove nekog skupa, takovo promatranje naziva se **census**. Primjer takvog prikupljanja je popis stanovništva. Vrlo često su takva istraživanja skupa, pa se podaci prikupljaju za podskup osnovnog skupa, odnosno za uzorak. Takvo promatranje se naziva **reprezentativno promatranje**.

Prema vremenu promatranje se može podijeliti na **jednokratno, periodično i tekuće**. *Jednokratno promatranje* provodi se jednom i nema ponavljanja. *Periodično promatranje* se ponavlja nakon jednakih vremenskih perioda (npr. bilanca uspjeha u poduzeću). Za *tekuće promatranje* mjerenje je kontinuirano (npr. proizvodnja mlijeka).

Prikupljanje primarnih podataka može se vršiti i pomoću **statističkih pokusa**. Statističkim pokusom se vrši mjerenje vrijednosti obilježja nastalih u kontroliranim uvjetima. Vrlo čestu primjenu statistički pokus ima u marketinškom istraživanju tržišta. Na primjer, želi se istražiti kako boja pakovanja deterdženta za rublje ima utjecaja na kupnju. U takvom slučaju istraživač će u pokusu odrediti različite boje pakiranja (vrijednosti kontroliranog obilježja) i u njima će izložiti proizvod. U određenom vremenu vršiti će se mjerenje opsega prodaje. Na taj način dobiveni statistički podaci su primarni.

## 1.4 Priprema podataka za statističku analizu

Prije početka obrade prikupljenih podataka potrebno je izvršiti kontrolu sirove statističke građe. Kontrola se može vršiti u tijeku ili na kraju postupka prikupljanja podataka, što ovisi i o različitim metodama prikupljanja.

**Preventivna kontrola** obavlja se već u tijeku samog postupka prikupljanja statističkih podataka. Pri provođenju ankete to podrazumijeva kontrolu upitnika pri njegovom preuzimanju od ispitanika. Kontrolira se da li su dani odgovori na sva pitanja i da li su pravilno popunjena predviđena mjesta za tražene odgovore.

**Naknadna kontrola** prikupljenih podataka obavlja se nakon postupka prikupljanja. **Formalnom kontrolom** se uspoređuje realizirani broj prikupljenih podataka s onim planiranim obuhvatom statističkog skupa. Ako je anketa vršena neizravnim putem, na primjer poštom, uspoređuje se broj odaslanih upitnika s brojem vraćenih upitnika. **Materijalnom kontrolom** ispituje se potpunost i točnost sadržaja prikupljenih podataka. Na primjeru ankete to podrazumijeva kontrolu potpunosti i logičnosti danih odgovora.

## 1.5 Uzorak

**Uzorak** je podskup osnovnog statističkog skupa, a uzima se u svrhu ispitivanja obilježja elemenata osnovnog skupa (ili populacije).

Poželjno je da uzorak bude **što veći**, ali u konkretnim istraživanjima može se dogoditi da:

- povećavanjem uzorka istraživanje postaje sve skuplje,
- ponekad se uzorci uništavaju (npr. kemijska analiza nekih prehrambenih artikala).

Kad god je moguće poželjno je izabrati **slučajan uzorak**. U takvom uzorku svaka jedinica populacije (osnovnog skupa) ima jednaku vjerojatnost da bude izabrana. Uzorak **treba biti reprezentativan**, a **ne selekcioniran** (npr. potrebno je obuhvatiti istraživanjem i gradsko i seosko stanovništvo, zatim u nekom medicinskom istraživanju nije cilj samo analizirati i donositi zaključke za "dobrovoljce").

Ako neki članovi populacije imaju veću šansu od drugih da budu izabrani, takav uzorak se naziva **pristran uzorak** (biased sample).

Slučajan uzorak sastavlja se prema određenim principima, koji odgovaraju zakonu slučaja. Najbolji način je upotreba "tablice slučajnih brojeva" ili korištenje kompjuterskog sustava slučajnog izbora (generator slučajnih brojeva).

Pri izradi **tablice slučajnih brojeva** koristi se deseterostrana prizma s brojevima od 0 do 9.

### **Primjer 1.3.**

Nakon npr. 40 bacanja deseterostrane prizme dobije se na slučajan način 40 brojeva; skupljeni u skupine od po 4:

7766 7520 1607 6048 2771 4733 8558 8681 5204 3806

Zadatak je izabrati uzorak veličine 350 iz osnovnog skupa veličine 790. (Prije početka odabira svakoj osobi iz osnovnog skupa dodjeljuje se broj od 1 do 790!).

### **Rješenje 1.3.**

Najprije je potrebno otvoriti tablicu slučajnih brojeva *slučajno*! Pretpostavka je da se tablica otvorila na stranicu na kojoj je u prvom retku prikazani niz od 40 brojeva.

Kako je u osnovnom skupu ukupno 790 osoba, potrebno je čitati troznamenaste brojeve iz tablice (u redovima, u stupcima ili dijagonalno).

Ako se brojevi čitaju po redovima, prvi troznamenasti broj iz prezentiranog retka, tj. prva odabrana osoba je ona označena brojem 776, zatim 675, pa 201 itd.

Ako se naiđe na broj veći od 790, on se odbacuje, jer u osnovnom skupu nema više od 790 osoba. Ako se dogodi da se neki broj ponavlja, on se opet odbacuje, jer se u uzorak jedna osoba ne može odabrati dva puta.

**Uzorak** može biti i **sistematski**, ako se jedinice iz osnovnog skupa biraju sistematski. Na primjer, ako se po redu u uzorak bira svaki 10.-ti element iz osnovnog skupa.

**Stratificirani ili slojeviti uzorak** je takav uzorak koji se dobije na način da se populacija podijeli u slojeve ili stratumne prema nekim karakteristikama, te da se iz svake od grupa uzme slučajni uzorak. Na primjer u društvenim istraživanjima stratumi se mogu birati prema dobnim skupinama.

**Klaster uzorci** su lošija varijanta slučajnog uzorka i koriste se u velikim tržišnim, ekonomskim ili političkim istraživanjima. Na primjer, pri ispitivanju mišljenja stanovnika nekog grada o nekoj problematici, grad se može prema planu podijeliti na 50-ak blokova, odnosno kvartova. Tada se na slučaj biraju neki blokovi u

kojima onda anketari *detaljno* intervjuiraju sve stanovnike. Tada se čak vraćaju na adrese dok ne dobiju intervju od svakog stanovnika u odabranim blokovima.

**Kvotni uzorci** su još lošiji, jer predstavljaju neslučajni stratificirani uzorak. Koriste se kod "ad hoc" organiziranih istraživanja za potrebe tržišta, za prikupljanje mišljenja građana o nekom pitanju i slično. Istraživač unaprijed izabere broj ljudi (kvotu) svakog pojedinog stratumu koje mora intervjuirati. Stoga se ovi uzorci nazivaju kvotni.

**Prigodni uzorak** je onaj koji se "nađe pri ruci", jer je drugi nedostupan. Na primjer dostupni bolesnici na odjelu u bolnici, prisutni studenti neke godine studija i slično.

**Pri istraživanju** se mogu dogoditi **pogreške**, koje mogu biti:

- a) **pogreške izvan uzorka** (zbog odabrane pogrešne metode istraživanja, zbog pogrešne obrade rezultata i slično),
- b) **pogreške uzorka**.

Veličinu uzorka nije moguće općenito definirati, jer to ovisi o varijabilnosti pojave koja se mjeri i o preciznosti kojom se pojava želi izmjeriti. Za uzorak je najvažnije da bude reprezentativan, tj. kad god je to moguće slučajno sakupljen.

**Frakcija odabiranja** ( $f$ ) je omjer jedinica u uzorku i broja jedinica u osnovnom skupu:

$$f = \frac{n}{N}, \quad (1.1)$$

gdje je :

$n$  - broj jedinica u uzorku,

$N$  - broj jedinica u osnovnom skupu.

**Korak izbora** je recipročna vrijednost frakcije odabiranja ( $1/f$ ) i upotrebljava se kod sistematskog izbora jedinica u uzorak. To znači da ako je korak izbora jednak 20, da se u uzorak iz osnovnog skupa odabire svaki 20. element.

**Broj svih mogućih uzoraka** (bez ponavljanja) veličine  $n$  iz osnovnog skupa veličine  $N$  jednak je broju kombinacija bez ponavljanja  $n$ -tog razreda:  $K = \binom{N}{n}$ .

Pomoću uzorka vrši se procjena određenih parametara osnovnog skupa i testiraju se hipoteze o nepoznatim parametrima osnovnog skupa.

## 1.6 Unošenje varijabli u bazu podataka u računalu

U svim statističkom programskim paketima, pa tako i u **SPSS**-u potrebno je za svaku varijablu odabrati odgovarajuću skalu mjerenja.

U **SPSS**-u oznaka za skalu mjerenja je **Measure** i postoje 3 vrste skale mjerenja:

- **Nominal:** *nominalne skale*, za koje se umjesto imena predmeta navodi njegov broj. Brojevi služe samo za identifikaciju.
- **Ordinal:** *ordinalne skale*, koje služe za označavanje redoslijeda (određuju je li nešto veće ili manje od drugoga).
- **Scale:** *numerička statistička obilježja*.

### Primjer 1.4:

Na odabranom području izvršena je analiza zainteresiranosti učenika i studenata za poduke iz različitih predmeta upitnikom u kojem su između ostalih bila sljedeća pitanja:

- |   |                            |
|---|----------------------------|
| 1. Iz kojih predmeta ste pohađali poduke? | 2. Željeli biste pohađati: |
| a) Matematika                             | a) Individualne poduke     |
| b) Fizika                                 | b) Grupne poduke           |
| c) Engleski jezik                         | c) Svejedno mi je          |
| d) Hrvatski jezik                         | 3. Spol                    |
| e) Ostalo                                 | a) žensko                  |
|   | b) muško                   |

Potrebno je izvršiti kodiranje odabranih pitanja!

Kako se može vidjeti u tablici 1.1 obilježje - varijabla "poduka iz željenog predmeta" može poprimiti 5 oblika: Matematika, Fizika, Engleski jezik, Hrvatski jezik i ostalo. Svakom obliku obilježja dodijeljena je šifra ili kod: A, B, C, D i E. Na taj način, pri unosu podataka, umjesto da se pišu čitave riječi, upisuje se šifra, s čim se postiže velika ušteda vremena.

**Tablica 1.1.**

**Kodna lista**

OBILJEŽJE (varijabla)	OBLICI OBILJEŽJA	KOD
poduke iz željenog predmeta (V1)	Matematika	A
	Fizika	B
	Engleski jezik	C
	Hrvatski jezik	D
	Ostalo	E
željena vrsta poduka (V2)	Individualne poduke	A
	Grupne poduke	B
	Nije bitno	C
spol (V3)	ženski	1
	muški	2

*Izvor: Simulirani podaci.*

Kodovi ne moraju biti slova, pa se kod obilježja spol, za "ženski" unosi 1, a za "muški" 2.

Moderni statistički paketi imaju mogućnosti kreiranja i grupiranja unesenih podataka u različitim željenim varijantama u tablice, što opet ovisi o vrsti istraživanja i postavljenim ciljevima statističkog istraživanja.

**Primjer 1.5:**

Kodirana pitanja iz primjera 3 potrebno je za 10 ispitanika unijeti u tablicu!

**Tablica 1.2.**

**Tablica s kodiranim podacima (ili matrica podataka)**

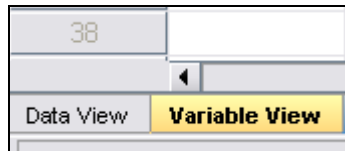
Ispitanici	V1	V2	V3
1	A	A	1
2	C	A	2
3	A	B	1
4	B	C	1
5	A	A	2
6	A	B	1
7	D	A	2
8	E	C	1
9	B	A	2
10	A	A	2

*Izvor: Simulirani podaci.*

U tablici 1.2 uneseni su podaci o 3 varijable (3 anketna pitanja) za 10 ispitanika.

Slika 1.1.

**Variable view za definiranje varijabli u programu SPSS**



Izvor: Iz programa SPSS.

Na slici 1.1 prikazan je aktiviran radni list **Variable view** u programskom paketu **SPSS**, na kojem se definiraju varijable, tj. pitanja iz anketnog upitnika. Na tom listu vrši se i kodiranje odgovarajućih kategorija nominalnih varijabli. Slika 1.2 prikazuje definirane varijable na listu **Variable view**. Ovdje se radi o nominalnom obilježju spol (**Name: v1**) i numeričkom obilježju visina u cm (**Name: v2**)<sup>1</sup>.

Slika 1.2.

**Prikaz definiranih varijabli u programu SPSS**

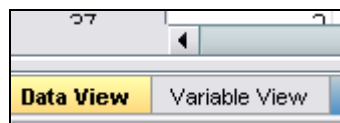
Vježbe. 27.10.08.sav [DataSet1] - SPSS Data Editor										
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help										
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	v1	Numeric	3	0	Spol	{1, musko}...	None	8	Right	Nominal
2	v2	Numeric	3	0	Visina u cm	None	None	8	Right	Scale

Izvor: Simulirani podaci.

Na slici 1.3 prikazan je aktiviran radni list **Data view** u programskom paketu **SPSS**, u kojem se unose podatci, tj. odgovori iz anketnog upitnika.

Slika 1.3.

**Data view za unošenje podataka u programu SPSS**



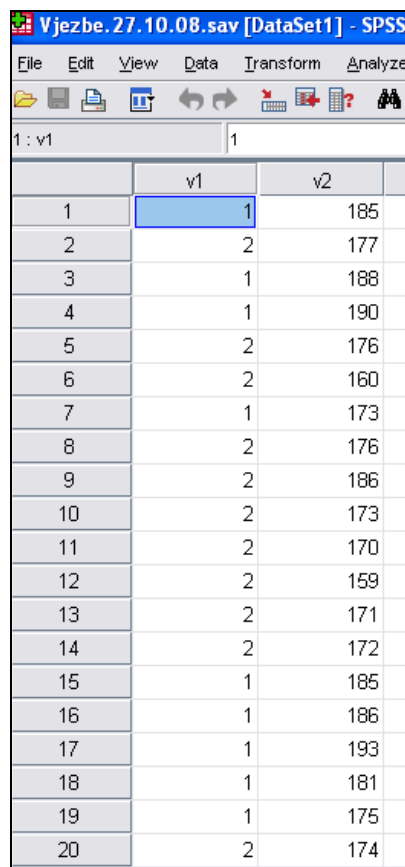
Izvor: Iz programa SPSS.

<sup>1</sup> Potrebno je napomenuti da su zbog lakše obrade podataka obje varijable označene da su **Type: Numeric**, iako je jasno da je spol nominalna varijabla i trebala bi biti označena **Type: String**.

Slika 1.4 prikazuje podatke u radnom listu **Data view** u programskom paketu **SPSS**. Ovdje su uneseni odgovori za 20 ispitanika na pitanja kratko nazvana v1 i v2. Svaki redak predstavlja jednog ispitanika, a svaki stupac predstavlja pitanje u anketnom upitniku. Unos i formiranje baze podataka moguće je organizirati na dva načina. Jedan način je da se za svakog ispitanika po redu unese najprije 1. pitanje, zatim za svakog ispitanika istim redom unese 2. pitanje itd. Tako se u **SPSS**-u popunjava stupac po stupac. Drugi način je da se za svakog ispitanika unesu sva pitanja. U **SPSS**-u se tada podacima popunjava redak po redak. Ovaj drugi način unosa podataka je možda jednostavniji i češći u praksi, jer se odjednom unosi čitav anketni listić i manja je mogućnost pogreške.

**Slika 1.4.**

**Prikaz podataka o varijablama v1 i v2 za 20 ispitanika u programu SPSS**



Vježbe. 27. 10. 08. sav [DataSet1] - SPSS

	v1	v2
1	1	185
2	2	177
3	1	188
4	1	190
5	2	176
6	2	160
7	1	173
8	2	176
9	2	186
10	2	173
11	2	170
12	2	159
13	2	171
14	2	172
15	1	185
16	1	186
17	1	193
18	1	181
19	1	175
20	2	174

*Izvor: Simulirani podaci.*



Takav način obrade podataka putem računala olakšava rad s masom statističkih podataka i omogućava veliku uštedu vremena.



### Primjer 1.6:

Na odabranom području izvršena je analiza stanovništva upitnikom u kojem su između ostalih bila sljedeća pitanja:

1. Spol:
  - a) Muški
  - b) Ženski

2. Visina:

.....

**Tablica 1.3.**

**Kodna lista**

OBILJEŽJE	OBLICI OBILJEŽJA	KOD
spol (v1)	Muški	1
	Ženski	2
visina u cm (v2)	.....	.....

*Izvor: Simulirani podaci.*

- a) Kodirana pitanja iz tablice 1.3 potrebno je unijeti u dokument programa SPSS!
- b) Za 20 ispitanika potrebno je u SPSS unijeti konkretne podatke o spolu i visini u cm prema kodiranim podacima u tablici 1.4!

**Tablica 1.4.**

**Podaci o spolu i visini za 20 ispitanika**

Ispit.	v1	v2	Ispit.	v1	v2
1.	1	185	11.	2	170
2.	2	177	12.	2	159
3.	1	188	13.	2	171
4.	1	190	14.	2	172
5.	2	176	15.	1	185
6.	2	160	16.	1	186
7.	1	173	17.	1	193
8.	2	176	18.	1	181
9.	2	186	19.	1	175
10.	2	173	20.	2	174

*Izvor: Simulirani podaci.*



### Rješenje 1.6:

a) U programu **SPSS** pitanja, odnosno varijable se definiraju u dijelu **Variable view** koji se bira u donjem lijevom kutu ekrana (slika 1.1.)

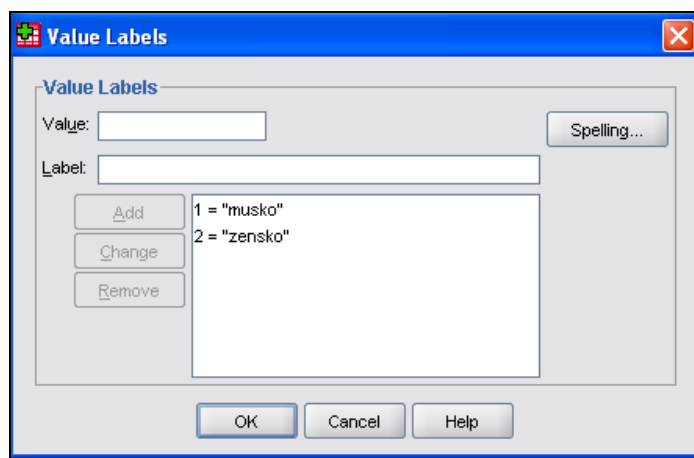
Prema slici 1.2 može se vidjeti da se varijable mogu nazvati (**Name**) v1 i v2. Radi lakšeg rada s podacima naznačena je vrsta podataka (**Type**) **Numeric**. U stupcu **Label** označeno je da je varijabla v1: Spol, a varijabla v2: Visina u cm. Zadnji stupac definira skalu mjerenja (**Measure**), gdje je za v1-Spol označena opisna, tj. **Nominal** skala, a za v2-Visinu u cm numerička, tj. **Scale** skala.

U stupcu **Values** za varijablu v1-Spol uneseni su kodovi prema tablici 1.3. Klikom miša u prvo polje stupca **Values** otvara se prozor **Value Labels**, u kojem se definira konkretno kodiranje. To je za ovaj primjer prikazano na slici 1.5.

Na slici 1.5 vidi se da muški spol ima kod 1, a ženski spol kod 2. Potrebno je napomenuti da pri unosu objašnjenja i vrijednosti treba izbjegavati slova koja ne pripadaju engleskoj abecedi jer ih **SPSS** u nekim varijantama neće prepoznati.

Slika 1.5.

### Prikaz definiranih varijabli u programu SPSS



Izvor: Prema podacima autora.

b) U programu **SPSS** podaci se prema definiranim varijablama unose u dijelu **Data view** koji se bira u donjem lijevom kutu ekrana (slika 1.3).

Naravno, nakon što su se podaci unijeli u program **SPSS**, potrebno ih je snimiti na standardan način kako se to radi kod većine klasičnih programa za računalu. Na

glavnom izborniku odabire se **File**, a u njegovu padajućem izborniku **Save As** te se uneseni podaci snimaju pod željenim imenom.



### Primjer 1.7:

Zadatak je za sve ispitanike od numeričke varijable džeparac (v2) formirati novu varijablu grupirani džeparac (v3) po principu koji je prikazan u tablici.

**Tablica 1.5.**

#### Princip formiranja nove varijable grupirani džeparac (v3)

Grupe	Visina džeparca	Nova vrijednost
1	0 - 200	200
2	201 - 400	400
3	401 - 600	600
4	601 - 800	800
5	801 - 1000	1000
6	1001 - 3500	3500

*Izvor: Simulirani podaci.*



### Rješenje 1.7:

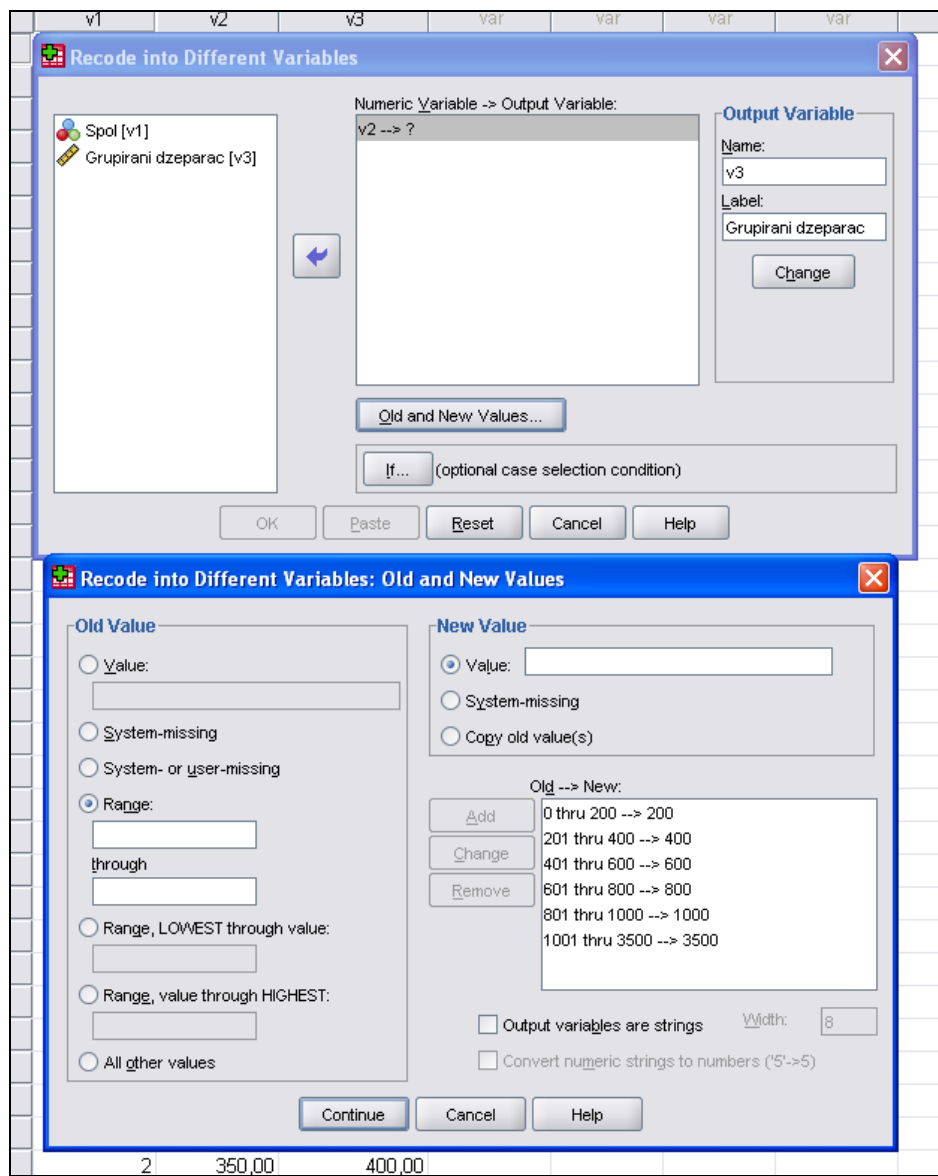
Da bi se pomoću varijable v2 - džeparac formirala nova varijabla grupirani džeparac po principu koji je prikazan u tablici 1.5, potrebno je na glavnom izborniku odabrati **Transform**, a u njegovu padajućem izborniku odabrati **Recode into Different Variables** (jer bi se odabirom **Recode into Same Variables** izgubila postojeća varijabla). U polje **Numeric Variable** bira se v2 - "Džeparac", a u **Output Variable, Name** je v3, a **Label** Grupirani džeparac.

Odabirom ikone **Old and New Values** otvara se novi prozor **Recode into Different Variables: Old and New Values**. Aktiviranjem opcije **Range** upisuju se odgovarajuće grupe visine džeparca, kako je zadano u tablici 1.5. Nakon upisivanja svakog pojedinog **Range**, upisuje se odgovarajuća **New Value**, koja se svaki put kao nova grupa dodaje klikom na ikonu **Add**. Nakon što su unesene sve grupe ikonom **Continue**, zatim **Change** i **OK**, u novom stupcu **Data View** kreirat će se nova varijabla Grupirani džeparac, odnosno v3.

Na slici 1.6 prikazani su prozori **Recode into Different Variables** i **Old and New Values**.

Slika 1.6.

Prozori Recode into Different Variables pomoću kojih se kreira nova varijabla  
v3 - Grupirani džeparac



Izvor: Simulirani podaci.

## 1.7 Uređivanje i prikazivanje podataka

### 1.7.1 Grupiranje podataka

Ako se urede prikupljeni statistički podaci prema nekom obilježju ili karakteristici dobiva se statistički niz.

Grupiranje statističkih podataka je postupak diobe statističkog skupa na određeni broj podskupova prema prethodno utvrđenim modalitetima promatranog obilježja i uz poštivanje principa isključivosti i iscrpnosti.

Princip isključivosti podrazumijeva da svaki element osnovnog skupa istovremeno može pripadati samo jednoj grupi tj. podskupu. Princip iscrpnosti podrazumijeva da postupkom grupiranja trebaju biti obuhvaćeni svi elementi osnovnog skupa.

### 1.7.2 Tabeliranje

Tabeliranje je postupak svrstavanja grupiranih prikupljenih statističkih podataka u tablice.

Statističke tablice kao jedan od oblika prikazivanja statističkih podataka prisutne su u literaturi svuda oko nas.

Tablica nastaje crtanjem okomitih i vodoravnih linija prema određenim pravilima. Svaka **statistička tablica mora imati: naslov, broj tablice (ako ih ima više), tekstualni dio, numerički ili brojčani dio i izvor podataka i po potrebi napomenu.**

### 1.7.3 Relativni brojevi strukture

Relativni brojevi su "**neimenovani**", te se stoga pomoću njih mogu uspoređivati i analizirati pojave koje imaju različitu jedinicu mjere ili različit broj elemenata. Na taj način dobije se relativna važnost dijela ili cjeline statističkog niza.

**Relativni brojevi nastaju dijeljenjem dviju veličina. Veličina s kojom se dijeli zove se osnova relativnog broja i po njoj se relativni brojevi međusobno razlikuju.**

**Relativni brojevi strukture pokazuju odnos dijela prema cjelini**, i njima se olakšava analiza rasporeda podataka prema modalitetima obilježja u jednom statističkom nizu, odnosno njihova struktura. Najčešće se izražavaju u postocima, a mogu i u promilima.

Ako relativni brojevi strukture ( $P_i$ ) pokazuju odnos apsolutnih frekvencija prema opsegu statističkog skupa (tj. ukupnom broju elemenata statističkog skupa) tada se zovu relativne frekvencije ( $fr_i$ ).

$$P_i = \frac{dio}{cjelina}; \quad fr_i = \frac{f_i}{\sum_{i=1}^n f_i} \quad (1.2)$$

$$P_i = \frac{dio}{cjelina} \cdot 100 \text{ (u \%)}; \quad fr_i = \frac{f_i}{\sum_{i=1}^n f_i} \cdot 100 \text{ (u \%)} \quad (1.3)$$

$$P_i = \frac{dio}{cjelina} \cdot 1000 \text{ (u \%)}; \quad fr_i = \frac{f_i}{\sum_{i=1}^n f_i} \cdot 1000 \text{ (u \%) } \quad (1.4)$$

Zbroj svih relativnih brojeva u jednom statističkom nizu je 1, ili 100, ili 1000, u ovisnosti o tome kako je relativni broj izražen.

Relativni brojevi strukture se **grafički** mogu prikazivati pomoću strukturnih stupaca, strukturnih krugova, polukrugova, ili nekim drugim geometrijskim likom. Pri tom se za usporedbu dvaju ili više statističkih nizova konstruiraju geometrijski likovi jednakih površina, jer je zbroj relativnih frekvencija uvijek isti.

#### 1.7.4 Grafičko prikazivanje statističkih nizova

Uz statističke tablice, pomoćno sredstvo u analizi statističkih nizova su grafički prikazi.

**Grafikonima se na jednostavan i pregledan način uz pomoć različitih geometrijskih likova prezentiraju osnovne karakteristike statističkih nizova.**

Grafički prikazi statističkih podataka su pregledniji i razumljiviji u odnosu na njihovo prikazivanje statističkom tablicom. Grafikoni omogućuju jednostavnije uočavanje glavnih karakteristika promatranih pojava, ali vrlo često ta preglednost ide na štetu preciznosti statističkih informacija. Stoga je poželjno uz grafički prikaz prezentirati i tablicu s originalnim vrijednostima statističkog niza. Suvremeni statistički programski paketi, a između njih i SPSS, naravno u skladu sa statističkom teorijom, imaju mnoštvo mogućnosti kreiranja grafičkih prikaza. Pomoću njih se mogu odabrati različite boje, oblici i linije na grafikonu, što omogućuje još zorniji prikaz promatrane pojave.

Oznake na grafikonu moraju biti takve da onaj tko čita grafikon može jasno raspoznati koji su elementi i koja je pojava prikazana. Stoga i grafikon mora imati **naslov, jedinice mjere promatranog obilježja, oznake modaliteta obilježja, izvor podataka i po potrebi kazalo ili tumač oznaka.**

Postoje tri skupine grafičkih prikaza:

- 1) površinski grafikoni,
- 2) linijski grafikoni, i
- 3) kartogrami.



#### Primjer 1.8.

Zadani su podaci o spolu i džeparcu uzorka studentske populacije na jednom sveučilištu.

Zadatak je formirati i komentirati apsolutne frekvencije varijabli:

- spol,
- grupirani džeparac.



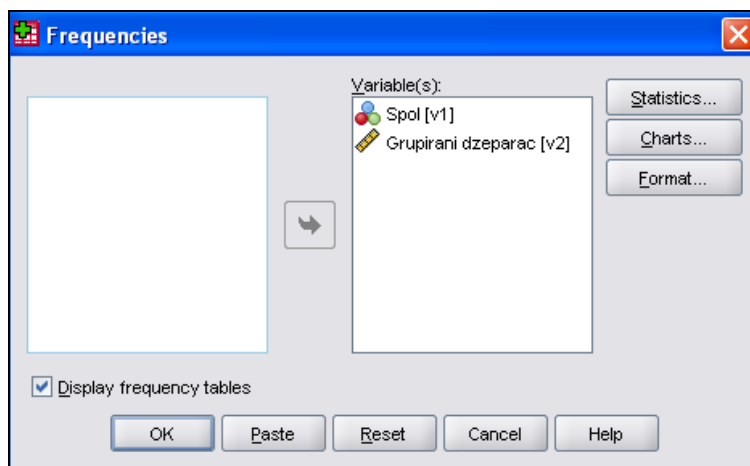
#### Rješenje 1.8.

Potrebno je na glavnom izborniku odabrati *Analyze*, a na njegovu padajućem izborniku odabrati *Descriptive Statistics*. Dalje se bira *Frequencies*. U ovom slučaju se mogu obje varijable v1 - Spol i v2 - Grupirani džeparac prebaciti u polje *Variable(s)*. Konačni izgled prozora *Frequencies* s odabranim varijablama prikazan je na slici 1.7.

Klikom na **OK** u *Outputu* programa *SPSS* dobije se tablica frekvencija uzorka prema odabranim obilježjima, odnosno varijablama. Rezultat je prikazan u tablicama 1.6 i 1.7.

Slika 1.7.

Prozor "Frequencies" iz izbornika "Descriptive Statistics" s odabranim varijablama v1 i v2



Izvor: Simulirani podaci.

Tablica 1.6.

Ispitanici prema obilježju:

Spol					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Musko	94	39,3	39,5	39,5
	Zensko	144	60,3	60,5	100,0
	Total	238	99,6	100,0	
Missing	System	1	,4		
Total		239	100,0		

Izvor: Simulirani podaci.

Prva frekvencija u stupcu **Frequency** u tablici 1.6 znači da je među ispitanicima bilo 94 studenta. Druga frekvencija istog stupca znači da je među ispitanicima bilo 144 studentice. Na postavljeno pitanje je odgovorilo 238 ispitanika, dok frekvencija 1 iz retka **Missing System** znači da 1 ispitanik nije dao odgovor na pitanje o njegovom spolu. Stupac **Percent** prikazuje relativne frekvencije u postotcima za valjane podatke uključujući i one koji nedostaju. Stupac **Valid Percent** prikazuje originalne valjane frekvencije u postotcima. Prvi podatak iz tog stupca pokazuje da je među onima koji su odgovorili na pitanje o spolu bilo 39,5% muškog spola, a 60,5% ženskog spola.



Tablica 1.7.

## Ispitanici prema obilježju:

Grupirani džeparac					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	200,00	73	30,5	42,2	42,2
	400,00	52	21,8	30,1	72,3
	600,00	30	12,6	17,3	89,6
	800,00	6	2,5	3,5	93,1
	1000,00	7	2,9	4,0	97,1
	3500,00	5	2,1	2,9	100,0
	Total	173	72,4	100,0	
Missing	System	66	27,6		
Total		239	100,0		

Izvor: Simulirani podaci.

Prva frekvencija u stupcu **Frequency** u tablici 1.7 znači da je među ispitanicima bilo 73 studenta s džeparcem od 200 kn ili manje. Druga frekvencija istog stupca znači da je među ispitanicima bilo 52 studenta s džeparcem između 201 kn i 400 kn itd. Na postavljeno pitanje je odgovorilo 173 ispitanika, dok frekvencija 66 iz retka **Missing System** znači da 66 ispitanika nije dalo odgovor na pitanje o njihovu džeparcu. Stupac **Percent** prikazuje relativne frekvencije u postocima za valjane podatke uključujući i one koji nedostaju. Stupac **Valid Percent** prikazuje originalne valjane frekvencije u postocima. Prvi podatak iz tog stupca pokazuje da je među onima koji su odgovorili na pitanje o džeparcu bilo 42,2% studenata s džeparcem od 200 kn ili manje itd.



## Primjer 1.9.

Zadatak je za sve ispitanike od varijable "Prosječna ocjena na I. godini studija" (v1) formirati novu varijablu "Grupirana prosječna ocjena na I. godini studija" (v2) po principu koji je prikazan u tablici 1.8.

Tablica 1.8.

## Princip formiranja nove varijable grupirana prosj. ocjena na I. god. studija (v2)

Grupe	Prosje. ocjena na I. god. stud.	Nova vrijednost
1	2,00-2,49	2
2	2,50-3,49	3
3	3,50-4,49	4
4	4,50-5,00	5

Izvor: Simulirani podaci.



### Rješenje 1.9.

Da bi se pomoću varijable v1 - "Prosječna ocjena na I. godini studija" formirala nova varijabla v2 - "Grupirana ocjena na I. godini studija" po principu koji je prikazan u tablici 1.8, potrebno je na glavnom izborniku odabrati **Transform**, a u njegovu padajućem izborniku odabrati **Recode into Different Variables**. U polje **Numeric Variable** bira se v1 - "Prosječna ocjena na I. godini studija", a u **Output Variable, Name** je v2, a **Label** "Grupirana ocjena na I. godini studija".

Slika 1.8.

Prozor "Recode into Different Variables: Old and New Values", pomoću kojih se kreira nova varijabla v2 - Grupirana ocjena u srednjoj školi

Izvor: Simulirani podaci.

Odabirom ikone **Old and New Values** otvara se novi prozor **Recode into Different Variables**. Aktiviranjem opcije **Range** upisuju se odgovarajuće grupe ocjena, kako je zadano u tablici 1.8. Nakon upisivanja svakog pojedinog **Range**, upisuje se odgovarajuća **New Value**, koja se svaki put kao nova grupa dodaje klikom na ikonu **Add**. Nakon što su unesene sve grupe ikonom **Continue**, zatim **Change** i **OK**, u novom stupcu **Data View**, kreirat će se nova varijabla "Grupirana ocjena na I. godini

studija", odnosno v2. Na slici 1.8 prikazan je prozor *Recode into Different Variables: Old and New Values*.



#### Primjer 1.10.

Zadatak je komentirati apsolutne i relativne frekvencije varijable: Grupirana ocjena na I. godini studija (v2).

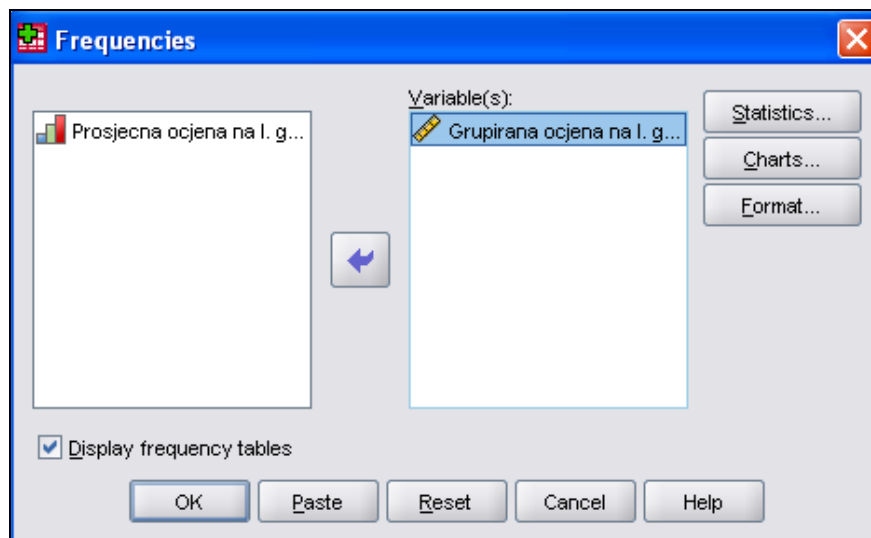


#### Rješenje 1.10.

Potrebno je na glavnom izborniku odabrati *Analyze*, a na njegovu padajućem izborniku odabrati *Descriptive Statistics*. Dalje se bira *Frequencies*, gdje se u dio prozora *Variable(s)* klikom na strelicu prebaci varijabla v2 - Grupirana ocjena na I. godini studija. Konačni izgled prozora *Frequencies* s odabranom varijablom prikazan je na slici 1.9.

Slika 1.9.

Prozor "Frequencies" iz izbornika "Descriptive Statistics" s odabranom varijablom v2



Izvor: Simulirani podaci.

Klikom na **OK** u **Outputu** programa **SPSS** dobije se tablica frekvencija uzorka prema odabranom obilježju, odnosno varijabli. Rezultat je prikazan u tablici 1.9.

Tablica 1.9.

## Ispitanici prema obilježju:

Grupirana ocjena na I. godini studija					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2,00	1	,4	,4	,4
	3,00	34	14,2	14,3	14,7
	4,00	113	47,3	47,5	62,2
	5,00	90	37,7	37,8	100,0
Total		238	99,6	100,0	
Missing	System	1	,4		
Total		239	100,0		

Izvor: Simulirani podaci.

Prva frekvencija u stupcu **Frequency** u tablici 1.9 znači da je među ispitanicima bio 1 student s ocjenom 2 na I. godini studija. Druga frekvencija istog stupca znači da je među ispitanicima bilo 34 studenta s ocjenom 3, ... itd. Na postavljeno pitanje je odgovorilo 238 ispitanika, dok frekvencija 1 iz retka **Missing System** znači da 1 ispitanik nije dao odgovor na pitanje o ocjeni na I. godini studija. Stupac **Percent** prikazuje relativne frekvencije u postotcima za valjane podatke uključujući i one koji nedostaju. Stupac **Valid Percent** prikazuje originalne valjane frekvencije u postotcima. Treći podatak iz tog stupca pokazuje da je 47,5% studenata imalo prosječnu ocjenu na I. godini studija 4.



## Primjer 1.11.

Zadatak je između svih ispitanika odabrati samo one koji su ženskog spola.



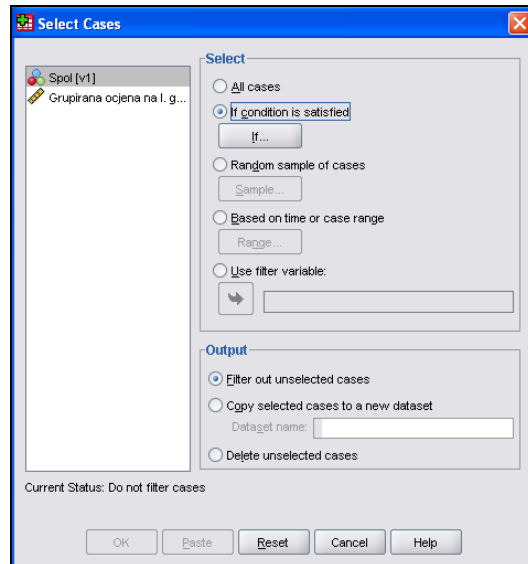
## Rješenje 1.11.

U glavnom izborniku bira se **Data**, a na njegovu padajućem izborniku **Select Cases**, gdje je potrebno aktivirati opciju **If condition is satisfied** kako je prikazano na slici 1.10.

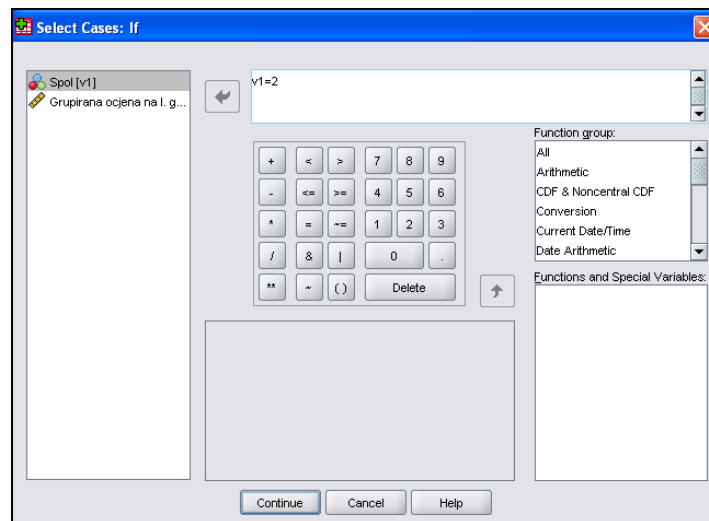
Klikom na ikonu **If...** otvara se novi prozor **Select cases: If**, kako je prikazano na slici 1.11.

Slika 1.10.

Prozor "Select Cases" s aktiviranom opcijom "If condition is satisfied"

*Izvor: Simulirani podaci.*

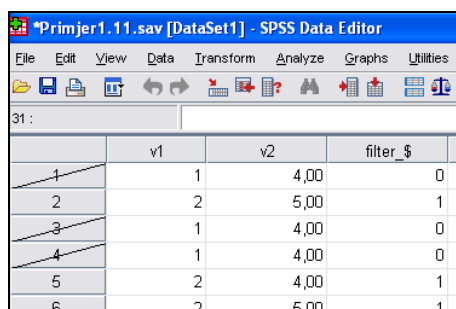
Slika 1.11.

Prozor "Select Cases: If" s naznačenim uvjetom  $v1=2$ *Izvor: Simulirani podaci.*

Strelicom se varijabla v1 prebaci u odgovarajuće polje i izjednači s 2, jer je uvjet da se biraju samo oni ispitanici ženskog spola, a ženski spol je pri kodiranju označen s 2. Klikom na **Continue** i **OK**, u dodatnom stupcu **Data View** javlja se nova varijabla **filter\_\$**, gdje su s 1 označena polja ispitanika ženskog spola, a s 0 polja ispitanika muškog spola. Naravno, **dok je filter aktivan program SPSS u svim svojim analizama uvažava samo polja označena u filteru s 1**, odnosno ispitanike ženskog spola. Stupac **filter\_\$** može se vidjeti na slici 1.12.

Slika 1.12.

### "Data View" sa stupcem "filter\_\$"



	v1	v2	filter_\$
1	1	4,00	0
2	2	5,00	1
3	1	4,00	0
4	1	4,00	0
5	2	4,00	1
6	2	5,00	1

Izvor: Simulirani podaci.



#### Primjer 1.12.

Zadatak je komentirati apsolutne i relativne frekvencije varijable: Grupirana ocjena na I. godini studija (v2) samo za ženski spol.



#### Rješenje 1.12.

Nakon što je izvršena selekcija ispitanika ženskog spola (kako je objašnjeno u 1.11), potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Descriptive Statistics**. Dalje se bira **Frequencies**, gdje se u dio prozora **Variable(s)** klikom na strelicu prebaci varijabla v2 - Grupirana ocjena na I. godini studija. Klikom na **OK** u **Outputu** programa **SPSS** dobije se tablica frekvencija uzorka prema odabranom obilježju, odnosno varijabli. Rezultat je prikazan u tablici 1.10 gdje su dani rezultati samo za ženski spol.

Prva frekvencija u stupcu **Frequency** u tablici 1.10 znači da je među ispitanicima ženskog spola bilo 20 njih s ocjenom 3 na I. godini studija. Druga frekvencija istog stupca znači da je među ispitanicima bilo 59 studentica s ocjenom 4, a njih 65 s ocjenom 5. U analizi je ukupno bilo 144 ispitanika ženskog spola.

Tablica 1.10.

## Ispitanice prema obilježju:

Grupirana ocjena na I. godini studija					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3,00	20	13,9	13,9	13,9
	4,00	59	41,0	41,0	54,9
	5,00	65	45,1	45,1	100,0
	Total	144	100,0	100,0	

Izvor: Simulirani podaci.

Stupac **Percent** prikazuje relativne frekvencije u postotcima za valjane podatke. Kako se vidi iz tablice, njegove vrijednosti odgovaraju onima iz stupca **Valid Percent**. Treća proporcija ili relativna frekvencija pokazuje da su 45,1% studentica u odnosu na sve 144 studentice imala prosječnu ocjenu na I. godini studija 5.



## Primjer 1.13.

Zadatak je iz postojeće baze podataka svih ispitanika odabrati slučajni uzorak od 60% ispitanika.



## Rješenje 1.13.

U glavnom izborniku bira se **Data**, a na njegovu padajućem izborniku **Select Cases**, gdje je potrebno aktivirati opciju **Random sample of cases** kako je prikazano na slici 1.13.

Klikom na ikonu **Sample...** otvara se novi prozor **Select Cases: Random Sample**, kako je prikazano na slici 1.14. U tom prozoru potrebno je aktivirati **Approximately** i naznačiti **60% of all cases**. Klikom na **Continue** i **OK**, u dodatnom stupcu lista **Data View** javlja se nova varijabla **filter\_\$**, gdje su s 1 označena polja slučajno odabranih 60% ispitanika, a sa 0 polja 40% ispitanika koji neće biti uključeni u daljnju analizu. Naravno, **dok je filter aktivan program SPSS u svim svojim analizama uvažava samo polja označena u filteru s 1.**

Slika 1.13.

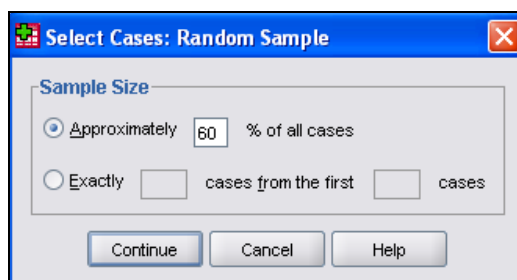
Prozor "Select Cases" s aktiviranom opcijom "Random sample of cases"



Izvor: Simulirani podaci.

Slika 1.14.

Prozor "Select Cases: Random Sample" s naznačenim slučajnim uzorkom od 60% ispitanika



Izvor: Simulirani podaci.



**Primjer 1.14.**

Zadatak je komentirati apsolutne i relativne frekvencije varijable: Grupirana ocjena na I. godini studija (v2) za 60% slučajno odabranih ispitanika.

**Rješenje 1.14.**

Nakon što je izvršena slučajna selekcija 60% ispitanika (kako je objašnjeno u 1.13), potrebno je na glavnom izborniku odabrati *Analyze*, a na njegovu padajućem izborniku *Descriptive Statistics*. Dalje se bira *Frequencies*, gdje se u dio prozora *Variable(s)* klikom na strelicu prebaci varijabla v2 - Grupirana ocjena na I. godini studija. Klikom na *OK* u *Outputu* programa SPSS dobije se tablica frekvencija uzorka prema odabranom obilježju, odnosno varijabli. Rezultat je prikazan u tablici 1.11, gdje su dani rezultati samo za 60% slučajno odabranih ispitanika.

Prva frekvencija u stupcu *Frequency* u tablici 1.11 znači da je između 60% slučajno odabranih ispitanika bio 1 s prosječnom ocjenom 2 na I. godini studija. Druga frekvencija istog stupca znači da je među tim ispitanicima bilo 17 s ocjenom 3, njih 75 s ocjenom 4, a 49 s ocjenom 5. Slučajno odabranih 60% ispitanika ima 143, a od toga 1 nije odgovorio na postavljeno pitanje.

**Tablica 1.11.****Slučajno odabranih 60% ispitanika prema obilježju:**

Grupirana ocjena na I. godini studija					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2,00	1	,7	,7	,7
	3,00	17	11,9	12,0	12,7
	4,00	75	52,4	52,8	65,5
	5,00	49	34,3	34,5	100,0
	Total	142	99,3	100,0	
Missing	System	1	,7		
Total		143	100,0		

Izvor: Simulirani podaci.

Stupac *Percent* prikazuje relativne frekvencije u postotcima za valjane podatke. Druga proporcija ili relativna frekvencija pokazuje da je od 60% ispitanika njih 12% imalo prosječnu ocjenu na I. godini studija 3.



### Primjer 1.15.

Zadatak je formirati grafikon jednostavnih stupaca za varijablu: Spol (v1).



### Rješenje 1.15.

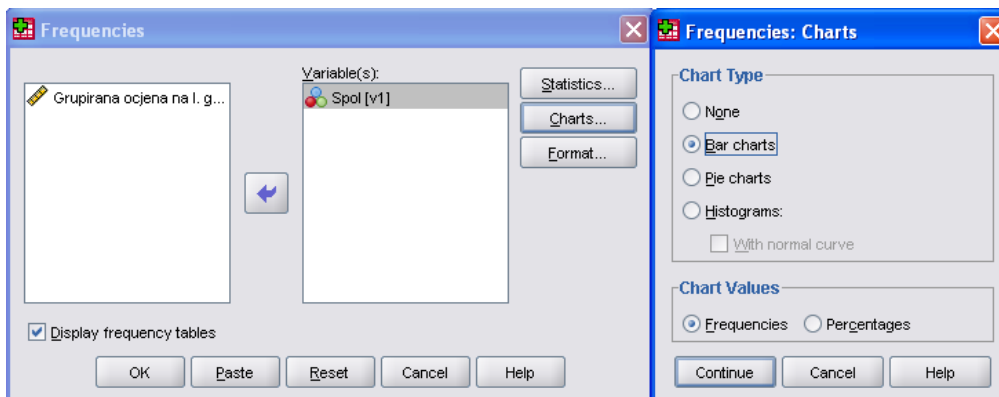
Grafikon jednostavnih stupaca za odabranu varijablu u programu **SPSS** može se konstruirati na dva načina:

#### I. način:

Potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Descriptive Statistics**. Dalje se bira **Frequencies**, gdje se u dio prozora **Variable(s)** klikom na strelicu prebaci varijabla v1 - Spol. Odabirom ikone **Charts** otvara se novi prozor **Frequencies: Charts**, gdje je potrebno aktivirati **Bar charts**. Klikom na **Continue** i **OK** u **Outputu** programa **SPSS** uz tablici svih frekvencija dobije se i traženi grafikon.

#### Slika 1.15.

Prozori "Frequencies" i "Frequencies: Charts" s odabranim "Bar Charts" i varijablom Spol (v1)



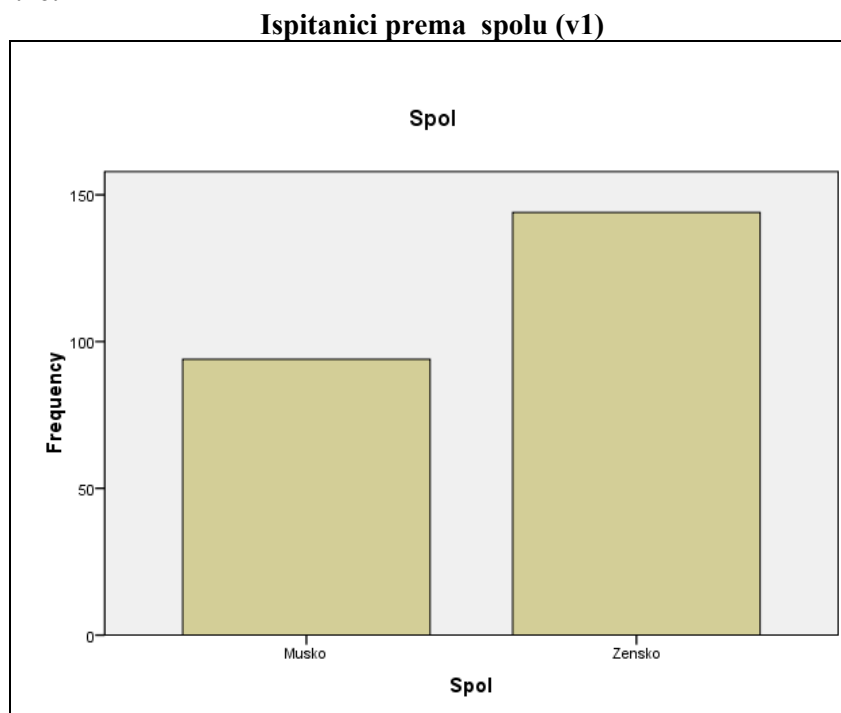
Izvor: Simulirani podaci.

Prema grafikonu prikazanom na slici 1.16 vidi se da je među ispitanicima više onih ženskog spola, preciznije njih 144, a onih muškog spola ima 94.

#### II. način:

Potrebno je na glavnom izborniku odabrati **Graphs**, a na njegovu padajućem izborniku **Legacy Dialogs: Bar**, na kojem se aktivira ikona **Simple**, što je prikazano na slici 1.17.

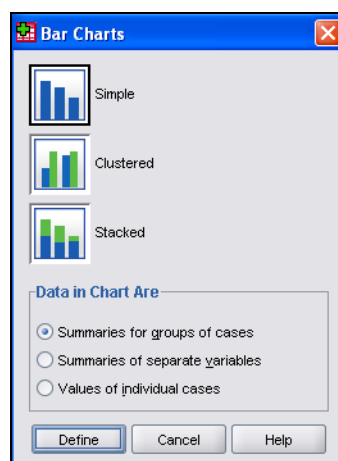
Slika 1.16.



*Izvor: Simulirani podaci.*

Slika 1.17.

**Prozori "Bar Charts" s aktiviranim "Simple" grafikonima**



*Izvor: Simulirani podaci.*

Odabirom ikone **Define** otvara se prozor **Define Simple Bar: Summaries for Groups of Cases**, gdje se u prostor **Category Axis** prebaci željena varijabla, što je u ovom slučaju Spol (v1). Grafikonu se može definirati i naslov na način da se klikne na ikonu **Titles...** Klikom na **OK** u **Outputu** programa **SPSS** dobije se traženi grafikon koji je identičan onome prikazanom na slici 1.16.



#### Primjer 1.16.

Zadatak je formirati grafikon strukturni krug za varijablu: Spol (v1).



#### Rješenje 1.16.

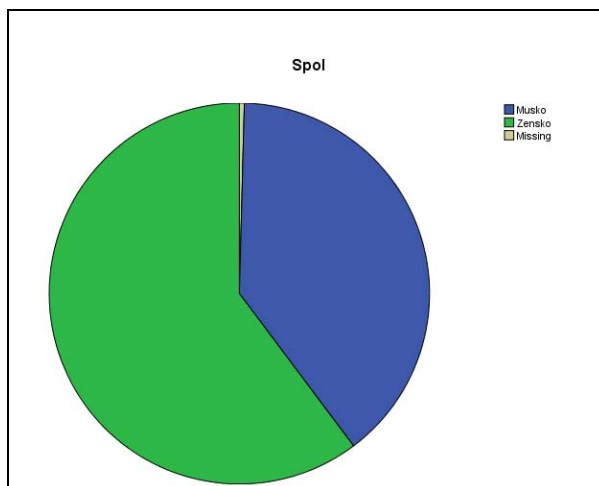
Grafikon strukturnog kruga za odabranu varijablu u programu **SPSS** može se nacrtati na dva načina:

##### I. način:

Potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Descriptive Statistics**. Dalje se bira **Frequencies**, gdje se u dio prozora **Variable(s)** klikom na strelicu prebaci varijabla v1 - Spol. Odabirom ikone **Charts** otvara se novi prozor **Frequencies: Charts**, gdje je potrebno aktivirati **Pie charts**. Klikom na **Continue** i **OK** u **Outputu** programa **SPSS** uz sve frekvencije dobije se i traženi grafikon.

#### Slika 1.18.

Ispitanici prema spolu (v1)



Izvor: Simulirani podaci.

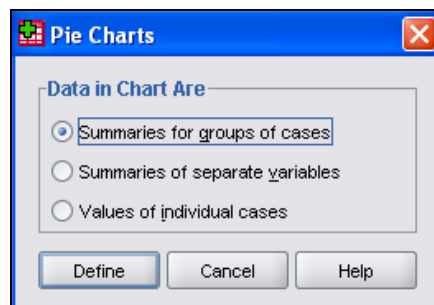
Prema slici 1.18 vidi se da je među ispitanicima više onih ženskog spola. Može se vidjeti da su na ovom grafičkom prikazu uključeni svi ispitanici, pa i oni koji nisu dali odgovor na pitanje o spolu.

## II. način:

Potrebno je na glavnom izborniku odabrati **Graphs**, a na njegovu padajućem izborniku **Legacy Dialogs:Pie**, na kojem se aktivira **Summaries of groups of case**. To je prikazano na slici 1.19.

**Slika 1.19.**

### Prozor "Pie Charts" s aktiviranim "Summaries of groups of case"



*Izvor: Simulirani podaci.*

Odabirom ikone **Define** otvara se prozor **Define Pie: Summaries for Groups of Cases**, gdje se u prostor **Define Slices by** prebaci željena varijabla, što je u ovom slučaju **Spol (v1)**. Grafikonu se može definirati i naslov, na način da se klikne na ikonu **Titles...** Klikom na **OK** u **Outputu** programa **SPSS** dobije se traženi grafikon koji je sličan onome prikazanom na slici 1.18, samo što ne sadrži ispitanike koji nisu odgovorili na pitanje o spolu (**Missing**).



### Primjer 1.17.

Za sve ispitanike zadatak je:

- Izračunati ukupan džeparac po obitelji (v4), ako sva djeca u jednoj obitelji dobivaju jednak iznos novca.
- Prikazati varijablu grupiranog džeparca (v2) grafički pomoću jednostavnih stupaca.
- Objasniti grafikon pod (b)!

**Rješenje 1.17.**

a) Za sve ispitanike potrebno je izračunati ukupan džeparac po obitelji uz pretpostavku da sva djeca u jednoj obitelji imaju jednak džeparac. Dakle, da bi se dobio ukupan džeparac po obitelji za svakog ispitanika, potrebno je džeparac ispitanika ( $v1$ ) pomnožiti s brojem djece u svakoj obitelji ( $v3$ ).

Da bi se pomoću postojećih numeričkih varijabli izračunala nova numerička varijabla, potrebno je na glavnom izborniku odabrati **Transform**, a u njegovu padajućem izborniku **Compute**. Na ekranu se pojavi prozor **Compute Variable**.

U polje **Target Variable** potrebno je unijeti ime nove varijable  $v4$ , dok se u **Numeric Expression** unosi izraz po kojem se iz postojećih varijabli kreira nova varijabla na sljedeći način:

$$v4 = v1 \times v3 . \quad (1.5)$$

Nakon što se klikne ikona **OK**, nova varijabla  $v4$  automatski se formira u dodatnom stupcu u **Data View** programa **SPSS**. Nova varijabla se može vidjeti na slici 1.20.

**Slika 1.20.****List "Data view" s novom varijablom  $v4$** 

	v1	v2	v3	v4	var
1	2000,00	3500,00	2	4000,00	
2	.	.	3	.	
3	600,00	600,00	2	1200,00	
4	1000,00	1000,00	3	3000,00	
5	.	.	2	.	
6	500,00	600,00	2	1000,00	
7	80,00	200,00	1	80,00	
8	100,00	200,00	3	300,00	
9	100,00	200,00	1	100,00	

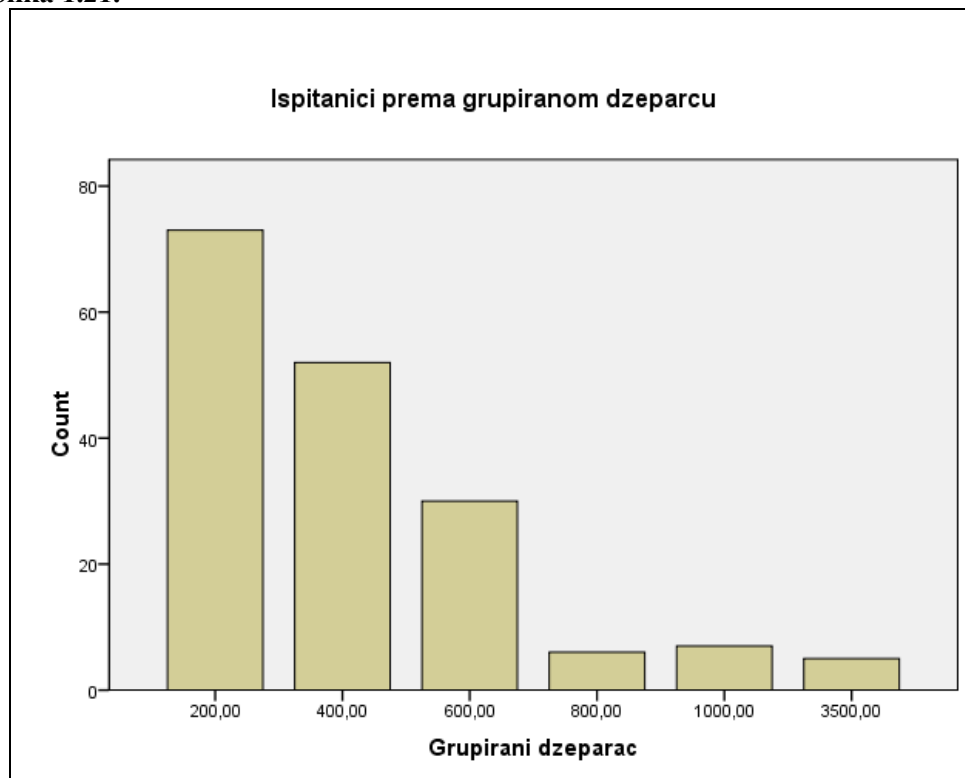
Izvor: Simulirani podaci.

U **Variable View** potrebno je na odgovarajući način definirati tu novu varijablu te u stupcu **Label** naznačiti da je to "Ukupan džeparac po obitelji". Ta nova varijabla je numerička i ima skalu mjerenja **Scale**.

b) Potrebno je na glavnom izborniku odabrati **Graphs**, a na njegovu padajućem izborniku **Bar char**, na kojem se aktivira ikona **Simple**.

Odabirom ikone **Define** otvara se prozor **Define Simple Bar: Summaries for Groups of Cases**, gdje se u prostor **Category Axis** prebaci željena varijabla, što je u ovom slučaju Grupirani džeparac (v2). Grafikonu se može definirati i naslov na način da se klikne na ikonu **Titles...** Klikom na **OK** u **Outputu** programa **SPSS** dobije se traženi grafikon koji je prikazan na slici 1.21.

**Slika 1.21.**



*Izvor: Simulirani podaci.*

c) Na grafikonu se može vidjeti da najveći broj ispitanika ima džeparac do 200 kn. Kako raste visina džeparca, smanjuje se broj ispitanika koji ima toliki džeparac.



#### **Primjer 1.18.**

Za sve ispitanike potrebno je:

a) Grupirati varijablu Visina u cm (v2) u Grupiranu visinu (v3) po principu:

- 160   ⇒ 160
- 170   ⇒ 170
- 180   ⇒ 180
- 190   ⇒ 190
- 200   ⇒ 200
- 210   ⇒ 210.

b) Prikazati varijablu grupirane visine (v3) grafički pomoću linijskog grafikona.

c) Objasniti grafikon pod (b)!



### Rješenje 1.18.

a) Da bi se pomoću varijable v2 - "Visina u cm" formirala nova varijabla v3 - "Grupirana visina" po principu koji je prikazan u zadatku pod (a), potrebno je na glavnom izborniku odabrati **Transform**, a u njegovu padajućem izborniku odabrati **Recode into Different Variables**. U polje **Numeric Variable** bira se v2 - "Visina u cm", a u **Output Variable, Name** je v3, a **Label** "Grupirana visina".

Odabirom ikone **Old and New Values** otvara se novi prozor **Recode into Different Variables**. Aktiviranjem opcije **Range: Lowest through value** upisuju se odgovarajuće grupe ocjena kako je prikazano na slici 1.22.

Nakon upisivanja svakog pojedinog **Range: Lowest through**, upisuje se i odgovarajuća **New Value**, koja se svaki put kao nova grupa dodaje klikom na ikonu **Add**. Nakon što su unesene sve grupe ikonom **Continue**, zatim **Change** i **OK**, u novom stupcu **Data View**, kreirat će se nova varijabla "Grupirana visina", odnosno v3.

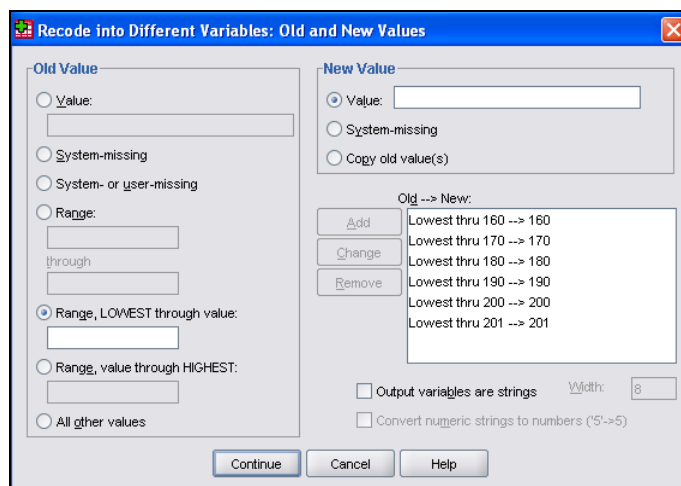
b) Potrebno je na glavnom izborniku odabrati **Graphs**, a na njegovu padajućem izborniku **Line chart** na kojem se aktivira ikona **Simple**.

Odabirom ikone **Define** otvara se prozor **Define Simple Bar: Summaries for Groups of Cases**, gdje se u prostor **Category Axis** prebaci željena varijabla, što je u ovom slučaju Grupirana visina (v3). Grafikonu se može definirati i naslov na način da se klikne na ikonu **Titles...** Klikom na **OK** u **Outputu** programa **SPSS** dobije se traženi grafikon koji je prikazan na slici 1.23.



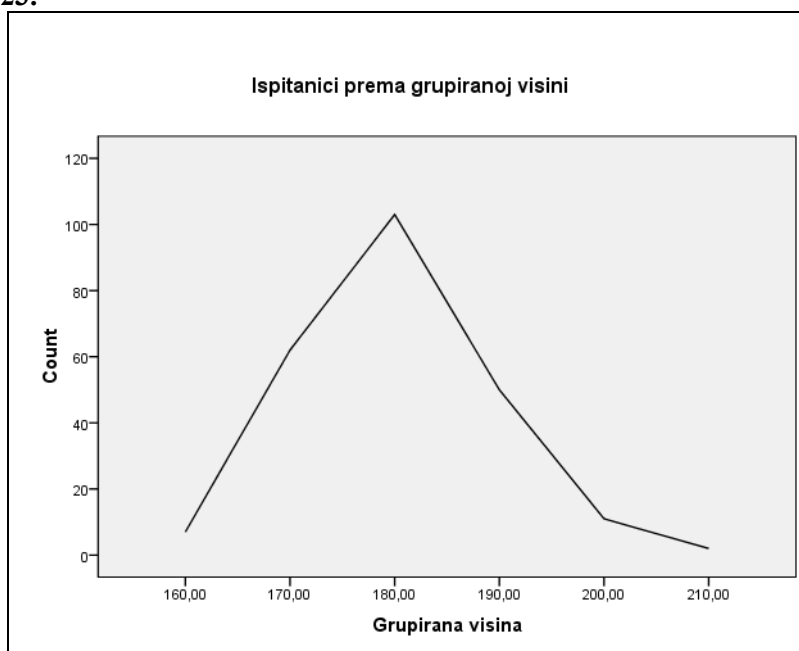
Slika 1.22.

Prozor "Recode into Different Variables: Old and New Values", pomoću kojih se kreira nova varijabla v3 - Grupirana visina



Izvor: Simulirani podaci.

Slika 1.23.



Izvor: Simulirani podaci.

c) Na grafikonu se može vidjeti da najveći broj ispitanika ima visinu 180 cm. Kako se visina smanjuje ili povećava, smanjuje se broj ispitanika koji ima toliku visinu.



#### Primjer 1.19.

Ako je prosječni džeparac 372 kn, potrebno je za sve ispitanike izračunati indekse varijable džeparac (v3) uzimajući za bazu indeksa prosječni džeparac po ispitaniku.



#### Rješenje 1.19.

Za sve ispitanike potrebno je izračunati indekse varijable džeparac (v3) uzimajući za bazu indeksa prosječni džeparac po ispitaniku koji iznosi 372. Dakle, da bi se dobili ovi indeksi, potrebno je džeparac ispitanika (v2) podijeliti s bazom koja iznosi 372 i sve to prema jednadžbi (1.6) pomnožiti sa 100.

$$I_i = \frac{f_i}{B} \cdot 100, \quad i = 1, 2, \dots, k. \quad (1.6)$$

Dakle, indeksi se računaju kao omjer frekvencije promatranog niza i odabrane baze (B).

Da bi se pomoću postojećih numeričkih varijabli izračunala nova numerička varijabla, potrebno je na glavnom izborniku odabrati **Transform**, a u njegovu padajućem izborniku **Compute Variable**. Na ekranu se pojavi prozor pod tim imenom.

U polje **Target Variable** potrebno unijeti ime nove varijable v3, dok se u **Numeric Expression** unosi izraz po kojem se iz postojećih varijabli kreira nova varijabla na sljedeći način:

$$v3 = (v2 / 372) \times 100.$$

Nakon što se klikne ikona **OK**, nova varijabla v3 automatski se formira u dodatnom stupcu lista **Data View** programa **SPSS**.

U **Variable View** potrebno je na odgovarajući način definirati tu novu varijablu te u stupcu **Label** naznačiti da su to "Indeksi džeparca po ispitaniku (372=100)". Baza indeksa se naznači tako da se izjednači sa 100. Ta nova varijabla je numerička i ima skalu mjerenja **Scale**.

Dobiveni indeks za npr. 1. ispitanika u nizu je:  $I_1 = 537,63$ , što znači da je on imao na raspolaganju džeparac 437,63% veći od baze, tj. od prosječnog džeparca po ispitaniku od 372 kn.

## 1.8 Srednje vrijednosti

Prikupljeni statistički podaci u svom izvornom obliku, često zbog svog obujma nemaju razumljivu formu. Zbog toga se i vrši njihovo grupiranje, odnosno formiranje statističkih nizova. Na taj način se dobiva detaljniji uvid u svojstva promatranog statističkog niza.

Računanjem srednjih vrijednosti dolazi se do informacija o vrijednostima statističkog obilježja oko kojih se raspoređuju elementi statističkog niza.

**Srednja vrijednost je vrijednost statističkog obilježja oko koje se grupiraju podaci statističkog niza. Još se zove i "mjera centralne tendencije".**

**Srednje vrijednosti mogu se podijeliti na:**

1. **Položajne srednje vrijednosti** određuju se položajem podataka u nizu. Najvažnije položajne srednje vrijednosti su :
  - a) **mod,**
  - b) **medijan.**
2. **Potpune srednje vrijednosti** računaju se upotrebom svih podataka u statističkom nizu. Potpune srednje vrijednost su:
  - a) **aritmetička sredina,**
  - b) **harmonijska sredina,**
  - c) **geometrijska sredina** (koja nema svoju primjenu u sociološkim istraživanjima).

### 1.8.1 Aritmetička sredina

**Aritmetička sredina spada u potpune srednje vrijednosti i računa se upotrebom svih podataka u statističkom nizu. To je najvažnija i najčešće korištena srednja vrijednost (ali ne i uvijek najbolja srednja vrijednost).**

**Aritmetička sredina je omjer zbroja svih vrijednosti numeričkog obilježja jednog niza i broja elemenata tog niza.**

Ako se radi o **negrupiranom statističkom numeričkom nizu računa se jednostavna aritmetička sredina**:

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N} \quad (1.7)$$

Ako se radi o **grupiranom statističkom numeričkom nizu računa se složena, vagana ili ponderirana aritmetička sredina**.

**Vagana ili ponderirana aritmetička sredina računa se prema izrazu:**

$$\bar{X} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} \quad (1.8)$$

Aritmetička sredina izražava se u originalnim jedinicama mjere promatranog numeričkog obilježja, obuhvaća sve elemente nekog skupa, te se pomoću nje mogu uspoređivati nizovi koji su grupirani po jednakom obilježju. Uz ove, aritmetička sredina zadovoljava i slijedeće kriterije:

- a) Aritmetička sredina nalazi su između najveće i najmanje vrijednosti promatranog numeričkog obilježja:

$$x_{\min} \leq \bar{X} \leq x_{\max}$$

- b) Zbroj odstupanja vrijednosti numeričkog obilježja od aritmetičke sredine u jednoj distribuciji je uvijek nula.

$$\sum_{i=1}^N (x_i - \bar{X}) = 0, \text{ za negrupirani niz,}$$

$$\sum_{i=1}^k f_i (x_i - \bar{X}) = 0, \text{ za grupirani niz.}$$

- c) Zbroj kvadrata odstupanja vrijednosti numeričkog obilježja od aritmetičke sredine u jednoj distribuciji je manji ili jednak zbroju kvadrata odstupanja vrijednosti numeričkog obilježja u istoj distribuciji od bilo koje druge vrijednosti (a).

$$\sum_{i=1}^N (x_i - \bar{X})^2 \leq \sum_{i=1}^N (x_i - a)^2, \text{ za negrupirani niz,}$$

$$\sum_{i=1}^k f_i (x_i - \bar{X})^2 \leq \sum_{i=1}^k f_i (x_i - a)^2, \text{ za grupirani niz.}$$

Ako neki numerički niz sadrži ekstremno male ili velike vrijednosti obilježja, aritmetička sredina kao prosječna vrijednost gubi na svojoj reprezentativnosti. Taj problem je dodatno izražen kada u distribuciji postoje razredi s otvorenim donjom, odnosno gornjom granicom razreda, i kada nije moguće te granice objektivno procijeniti.

### 1.8.2 Mod

**Mod je vrijednost statističkog obilježja koja se najčešće javlja u nekom nizu, tj. vrijednost obilježja kojoj pripada najveća frekvencija.**

Mod se može primijeniti na kvalitativne i kvantitativne statističke nizove, a spada u položajne srednje vrijednosti. U većini statističke literature oznaka mu je: *Mo*.

**Kod nominalnih obilježja mod se određuje** brojanjem, na način da se traži koja se vrijednost obilježja u nizu najčešće javlja. Ako je niz grupiran, traži se najveća apsolutna frekvencija. Vrijednost obilježja kojoj pripada ta najveća apsolutna frekvencija je mod.

Ovo je primjer tzv. **unimodalnih nizova** koji imaju samo jedan mod. Postoje statistički nizovi u kojima se dvije ili više vrijednosti obilježja mogu pojavljivati češće u odnosu na ostale modalitete obilježja. U tom slučaju kaže se su to **bimodalni ili multimodalni nizovi**. Kod bimodalne distribucije, koja ima dva vrha, postoji glavni mod i lokalni mod. U takvom slučaju kada je u nizu prisutno više od jednog moda, potrebno je statistički skup podijeliti na više podskupova, od kojih će svaki imati svoja karakteristična svojstva, te izvršiti analizu svakog podskupa posebno.

### 1.8.3 Medijan

**Medijan je vrijednost statističkog obilježja koja uređeni statistički niz dijeli na dva jednakobrojna dijela.**

**Medijan** dijeli uređeni statistički niz na dva jednaka dijela u omjeru 1:1, odnosno 50% elemenata statističkog skupa ima vrijednost obilježja manju ili jednaku

medijanu, a 50% elemenata statističkog skupa ima vrijednost obilježja veću od medijana. U većini statističke literature oznaka mu je:  $Me$ .

Medijan se može primijeniti na redoslijedne i kvantitativne statističke nizove, a spada u položajne srednje vrijednosti. Medijan se ne primjenjuje kod nominalnih nizova, jer poredak oblika ovog obilježja može biti proizvoljan.

Kod **negrupiranog, a uređenog niza (po veličini vrijednosti obilježja)**, medijan je vrijednost obilježja koja pripada elementu statističkog niza koji se nalazi u sredini niza. Ako je broj elemenata statističkog niza paran, onda se za medijan uzima jednostavan prosjek vrijednosti obilježja dvaju članova koji se nalaze na sredini statističkog niza.

U statističkim distribucijama postoji samo jedan medijan i on se nalazi između najveće i najmanje vrijednosti obilježja. Prednost medijana je da na njega ne utječu ekstremno male ili velike vrijednosti obilježja, pa je primjerena srednja vrijednost i kod izrazito asimetričnih distribucija.

### **1.8.3.1      Kvantili**

**Kvantili su vrijednosti statističkog obilježja koje uređeni statistički niz dijele na  $q$  jednakobrojnih dijelova.**

Kod analize statističkih nizova vrlo su često u upotrebi kvantili.

**Kvantili su vrijednosti statističkog obilježja koje statistički niz dijele na 4 jednakobrojna dijela.** Kvantili se mogu podijeliti na:

- a) donji kvartil ( $Q_1$ ) i
- b) gornji kvartil ( $Q_3$ )<sup>2</sup>.

**Donji kvartil** dijeli uređeni statistički niz na četiri jednaka dijela u omjeru 1:3, odnosno 25% elemenata statističkog skupa ima vrijednost obilježja manju ili jednaku donjem kvartilu, a 75% elemenata statističkog skupa ima vrijednost obilježja veću od donjeg kvartila.

**Gornji kvartil** dijeli uređeni statistički niz na četiri jednaka dijela u omjeru 3:1, odnosno 75% elemenata statističkog skupa ima vrijednost obilježja manju ili

---

<sup>2</sup> Vrijedi da je medijan:  $Me = Q_2$ .

jednaku donjem kvartilu, a 25% elemenata statističkog skupa ima vrijednost obilježja veću od gornjeg kvartila.

Kvartili se, slično kao i medijan, mogu primijeniti na redosljedne i kvantitativne statističke nizove, a određuju se položajem u nizu. Ni oni se ne primjenjuju kod nominalnih nizova, jer kako je već naglašeno poredak oblika ovog obilježja može biti proizvoljan.

## 1.9 Disperzija

Može se dogoditi da neki statistički skupovi imaju jednake npr. aritmetičke sredine, a da su njihovi elementi potpuno različiti. To znači da je raspored elemenata u tim skupovima različit. **Informaciju o rasporedu elemenata daju mjere raspršenosti ili disperzije elemenata numeričkog statističkog niza.**

Postoje apsolutne i relativne mjere raspršenosti. Apsolutni pokazatelji izraženi su u originalnim jedinicama mjere i omogućavaju usporedbu nizova prema istom obilježju. **Apsolutni pokazatelji raspršenosti su: raspon varijacije, interkvartil, varijanca i standardna devijacija.**

Usporedbu raspršenosti elemenata nizova s različitom mjernom jedinicom omogućuju relativni pokazatelji, koji su najčešće izraženi u postocima. **Relativni pokazatelji raspršenost su: koeficijent kvartilne devijacije i koeficijent varijacije.**

### 1.9.1 Apsolutne mjere disperzije

**Raspon varijacije je najjednostavnija mjera disperzije, a predstavlja razliku između najveće i najmanje vrijednosti numeričkog obilježja promatranog niza:**

$$R = x_{\max} - x_{\min} \quad (1.9)$$

Ovaj apsolutni pokazatelj raspršenosti izražen je u originalnim jedinicama mjere numeričkog obilježja. Može poprimiti vrijednost 0. To se događa u slučaju kada svi elementi niza imaju jednaku vrijednost obilježja. Najveća vrijednost ovog

pokazatelja nije ograničena, jer ona ovisi o konkretnoj raspršenosti promatranih vrijednosti obilježja.

Raspon varijacije je nepotpuna mjera disperzije jer se računa samo na temelju dvije vrijednosti obilježja, odnosno na temelju najveće i najmanje vrijednosti. Može se reći da ovo nije precizna mjera raspršenosti elemenata niza, pogotovo u slučaju postojanja ekstremno malih i/ili ekstremno velikih vrijednosti obilježja. Tada se dobije veliki raspon varijacije, a možda je većina elemenata skupa raspršena usko oko srednjih vrijednosti.

Taj problem preciznosti rješava interkvartilni raspon ili interkvartil.

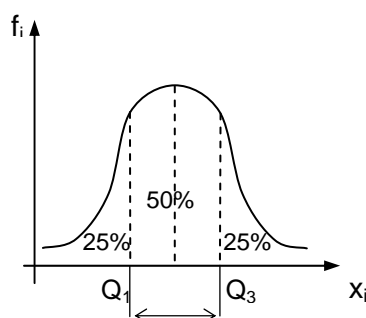
**Interkvartil** je apsolutna, nepotpuna mjera raspršenosti, koja pokazuje disperziju srednjih 50% elemenata uređenog numeričkog niza:

$$I_q = Q_3 - Q_1 \quad (1.10)$$

**Interkvartil predstavlja razliku gornjeg i donjeg kvartila.** Na taj način se eliminira 25% ekstremno malih i 25% ekstremno velikih vrijednosti obilježja u nizu.

**Slika 1.24.**

**Simetrična distribucija s označenim gornjim i donjim kvartilom**



*Izvor: Konstrukcija autora.*

Slika 1.24 prikazuje jednu simetričnu distribuciju, gdje su elementi skupa ravnomjerno raspoređeni oko srednjih vrijednosti. Poznato je da donji kvartil ( $Q_1$ ) dijeli distribuciju u omjeru 1:3, tj. da 25% elemenata skupa ima vrijednost obilježja manju od donjeg kvartila, a 75% elemenata skupa ima vrijednost obilježja veću od donjeg kvartila. Isto tako, gornji kvartil ( $Q_3$ ) dijeli distribuciju u omjeru 3:1, tj. da 75% elemenata skupa ima vrijednost obilježja manju od gornjeg kvartila, a 25% elemenata skupa ima vrijednost obilježja



veću od gornjeg kvartila. Na taj način interkvartil pokazuje disperziju srednjih 50% elemenata skupa.

**Varijanca spada u potpune mjere raspršenosti**, jer obuhvaća sve elemente odabranog numeričkog statističkog niza. Ovaj pokazatelj mjeri odstupanja, tj. raspršenost elemenata skupa od aritmetičke sredine.

S obzirom na poznato svojstvo aritmetičke sredine da je zbroj odstupanja vrijednosti obilježja od aritmetičke sredine u jednoj distribuciji uvijek nula, tj. da je:  $\sum_{i=1}^N (x_i - \bar{X}) = 0$  (za negrupirani niz) i  $\sum_{i=1}^k f_i (x_i - \bar{X}) = 0$  (za grupirani niz), varijanca se izražava preko kvadrata ovih odstupanja.

**Varijanca je prosječno kvadratno odstupanje vrijednosti numeričkog obilježja od aritmetičke sredine:**

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \bar{X})^2}{N} = \frac{\sum_{i=1}^N x_i^2}{N} - \bar{X}^2, \quad (1.11)$$

za negrupirani niz i

$$\sigma^2 = \frac{\sum_{i=1}^k f_i (x_i - \bar{X})^2}{\sum_{i=1}^k f_i} = \frac{\sum_{i=1}^k f_i x_i^2}{\sum_{i=1}^k f_i} - \bar{X}^2, \quad (1.12)$$

za grupirani niz.

Varijanca je mjera raspršenosti izražena u drugom stupnju, pa kao rezultat daje jedinice mjere numeričkog obilježja na kvadrat. Stoga je otežana njena interpretacija.

**Standardna devijacija je pozitivan korijen iz varijance i izražena je u originalnim jedinicama mjere.** Stoga je kao potpuna i apsolutna mjera disperzije vrlo često u upotrebi. Može se definirati kao **prosječno odstupanje vrijednosti numeričkog obilježja od aritmetičke sredine.**

$$\sigma = +\sqrt{\sigma^2} = +\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{X})^2}{N}} = +\sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \bar{X}^2}, \quad (1.13)$$

za negrupirani niz,

$$\sigma = +\sqrt{\sigma^2} = +\sqrt{\frac{\sum_{i=1}^k f_i (x_i - \bar{X})^2}{\sum_{i=1}^k f_i}} = +\sqrt{\frac{\sum_{i=1}^k f_i x_i^2}{\sum_{i=1}^k f_i} - \bar{X}^2}, \quad (1.14)$$

za grupirani niz. Pomoću standardne devijacije u originalnim mjernim jedinicama obilježja može se uspoređivati raspršenost oko aritmetičke sredine nizova koji su grupirani po jednakom obilježju.

### 1.9.2 Relativne mjere disperzije

**Koeficijent varijacije** spada u potpune relativne mjere raspršenosti, jer obuhvaća sve elemente odabranog numeričkog statističkog niza, a izražava se u postotcima (%).

**Koeficijent varijacije je postotak standardne devijacije od aritmetičke sredine:**

$$V = \frac{\sigma}{\bar{X}} \cdot 100. \quad (1.15)$$

Vrijednost koeficijenta varijacije se kreće u intervalu  $0 \leq V < +\infty$ . Vrijednost od 0% će poprimiti samo u slučaju kada su sve vrijednosti numeričkog obilježja u jednom nizu jednake, odnosno kada nema disperzije. Veća vrijednost ovog pokazatelja upućuje na veću disperziju elemenata promatranog niza.

Ovaj pokazatelj raspršenosti je izražen u postotcima, pa omogućuje usporedbu disperzije numeričkih nizova s različitim jedinicama mjere.

**Koeficijent kvartilne devijacije je relativna nepotpuna mjera raspršenosti. Predstavlja relativnu disperziju srednjih 50% elemenata numeričkog niza.** Računa se na temelju samo dvije vrijednosti obilježja, a to su donji i gornji kvartil:

$$V_q = \frac{Q_3 - Q_1}{Q_3 + Q_1}. \quad (1.16)$$

Koeficijent kvartilne devijacije se računa kao omjer interkvartila (razlika kvartila) i zbroja donjeg i gornjeg kvartila. Vrijedi da je:  $0 \leq V_q < 1$ .

Ovaj pokazatelj je jednak 0 samo u slučaju kada nema disperzije. Ako se njegova vrijednost približava 1, to znači da je raspršenost vrijednosti obilježja veća.

S obzirom da je ovo relativni pokazatelj, pomoću njega je moguće uspoređivati raspršenost srednjih 50% elemenata različitih numeričkih distribucija s različitim jedinicama mjere.

## 1.10 Asimetrija

Asimetrija distribucije podrazumijeva nagnutost distribucije na lijevu ili desnu stranu.

**Pearsonov koeficijent asimetrije je:**

$$\alpha_3 = \frac{\mu_3}{\sigma^3}, \quad (1.17)$$

gdje je:

$\mu_3$  - centralni moment trećeg reda,

$\sigma^3$  - standardna devijacija na treću potenciju.

Interval u kojem se uobičajeno kreće vrijednost ovog koeficijenta je:  $-2 \leq \alpha_3 \leq 2$ .

U slučaju izrazito asimetričnim distribucijama ovaj koeficijent može poprimiti i vrijednosti izvan intervala:  $[-2, 2]$ .

Ako je:  $\alpha_3 = 0 \Rightarrow$  simetrična distribucija,  
 $\alpha_3 > 0 \Rightarrow$  pozitivna ili desnostrana asimetrija,  
 $\alpha_3 < 0 \Rightarrow$  negativna ili lijevostrana asimetrija.

U programskom paketu **SPSS** mjera asimetrije ima oznaku **Skewness** i vrijedi da ako je vrijednost ovog pokazatelja:

$= 0 \Rightarrow$  simetrična distribucija,  
 $> 0 \Rightarrow$  pozitivna ili desnostrana asimetrija,  
 $< 0 \Rightarrow$  negativna ili lijevostrana asimetrija.

### 1.10.1 Odnosi među srednjim vrijednostima kod različito simetričnih distribucija

- a) **simetrična distribucija** ima jednake aritmetičku sredinu, medijan i mod:

$$\bar{X} = Me = Mo$$

- b) **pozitivno ili desnostrano asimetrična distribucija** ima aritmetičku sredinu veću od medijana, koji je veći od moda:

$$\bar{X} > Me > Mo$$

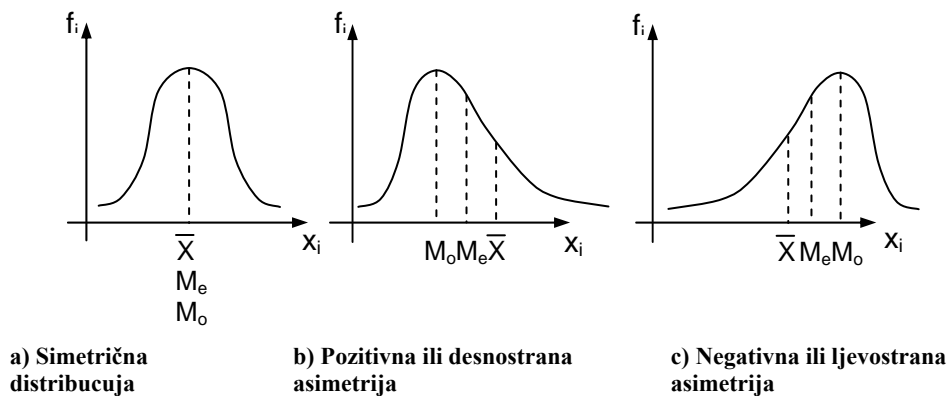
- c) **negativno ili lijevostrano asimetrična distribucija** ima aritmetičku sredinu manju od medijana, koji je manji od moda:

$$\bar{X} < Me < Mo$$

Raspored srednjih vrijednosti kod različitih distribucija prikazan je na sljedećoj slici.

**Slika 1.25.**

#### Distribucije s različitom asimetrijom



*Izvor: Konstrukcija autora.*

Prikazana su 3 slučaja raspršenosti podataka oko srednjih vrijednosti. Na dijelu slike pod (a) prikazana je jedna savršeno **simetrična distribucija u kojoj su aritmetička sredina** (prosjeak), **medijan** (vrijednost obilježja koja skup dijeli na dva

jednaka dijela) i **mod** (najčešća vrijednost obilježja, koja odgovara vrhu krivulje) u istoj točki, odnosno **jednaki**:  $\bar{X} = Me = Mo$  .

Na dijelu slike pod (b) prikazana je jedna **pozitivno ili desnostrano asimetrična distribucija** (vidi se rasipanje vrijednosti obilježja na desnu stranu) u kojoj je aritmetička sredina veća od medijana, koji je veći od moda:  $\bar{X} > Me > Mo$  .

Na dijelu slike pod (c) prikazana je jedna **negativno ili ljevostrano asimetrična distribucija** (vidi se rasipanje vrijednosti obilježja na lijevu stranu) u kojoj je aritmetička sredina manja od medijana, koji je manji od moda:  $\bar{X} < Me < Mo$  .

## 1.11 Zaobljenost

**Mjera zaobljenosti** predstavlja zaobljenost vrha krivulje distribucije frekvencija:

$$\alpha_4 = \frac{\mu_4}{\sigma^4}, \quad (1.18)$$

gdje je:

$\mu_4$  - centralni moment četvrtog reda,

$\sigma^4$  - standardna devijacija na četvrtu potenciju.

Vrijedi da ako je:

$\alpha_4 = 3 \Rightarrow$  normalno zaobljena distribucija,

$\alpha_4 > 3 \Rightarrow$  šiljatiji vrh od normalno zaobljene distribucije,

$\alpha_4 < 3 \Rightarrow$  tupi oblik distribucije.

U programskom paketu **SPSS** mjera zaobljenosti ima oznaku **Kurtosis** ( $Kurtosis = k$ ;  $k = \alpha_4 - 3$ ) i vrijedi da ako je vrijednost ovog pokazatelja:

$= 0 \Rightarrow$  normalno zaobljena distribucija,

$> 0 \Rightarrow$  šiljatiji vrh od normalno zaobljene distribucije,

$< 0 \Rightarrow$  tupi oblik distribucije.



### Primjer 1.20.

Za varijablu (obilježje) prosječna ocjena na I. godini studija (v2)<sup>3</sup> potrebno je odrediti:

a) Najmanju i najveću vrijednost obilježja, aritmetičku sredinu, mod, medijan, donji i gornji kvartil, raspon varijacije obilježja, varijancu i standardnu devijaciju, asimetriju i zaobljenost distribucije te nacrtati histogram s normalnom krivuljom.

b) Komentirati rezultate!



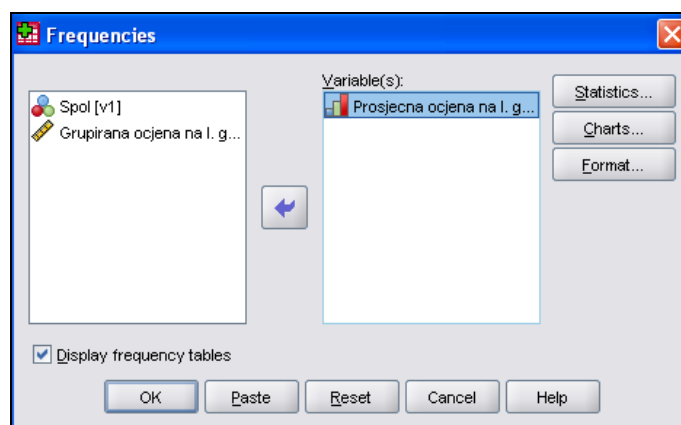
### Rješenje 1.20.

Na glavnom izborniku potrebno je izabrati ikonu *Analyze*, a na njezinu padajućem izborniku *Descriptive Statistics* i *Frequencies*.

U prozoru *Frequencies* pomoću odgovarajuće strelice varijabla v2 - prosječna ocjena na I. godini studija se prebacuje u polje *Variable(s)*, kako je prikazano na slici 1.26.

**Slika 1.26.**

**Prozor "Frequencies" sa izabranom varijablom (obilježjem) prosječna ocjena na I. godini studija (v2)**



*Izvor: Simulirani podaci.*

---

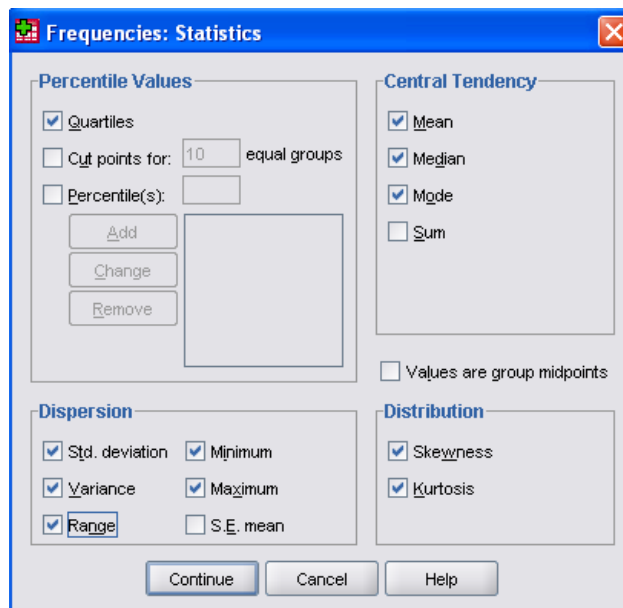
<sup>3</sup> Aritmetička sredina se po definiciji izračunava samo za numerička obilježja, ali se za cjelovitu analizu često primjenjuje i za redoslijedno obilježje.

Klikom na ikonu **Statistics** otvara se novi prozor **Frequencies Statistics**, gdje je potrebno aktivirati određene statističke veličine prema zahtjevu zadatka, a koje se odnose na odabranu varijablu prosječna ocjena na I. godini studija (v2).

Da bi se u izlaznim rezultatima dobila najmanja i najveća vrijednost niza, u spomenutom prozoru se aktivira **Minimum** i **Maximum**. Aritmetička sredina dobit će se aktiviranjem veličine **Mean**, a za mod potrebno je označiti **Mode**. Medijan je **Median**, a donji i gornji kvartil dobit će se aktiviranjem **Quartiles**. Za raspon varijacije obilježja aktivira se **Range**, za varijancu **Variance**, a standardna devijacija je **Std. deviation**. Pokazatelj asimetrije je **Skewness**, a zaobljenosti distribucije je **Kurtosis**. Prozor **Frequencies Statistics** s potrebnim aktiviranim veličinama prikazan je na slici 1.27. Kada se aktiviraju sve tražene veličine potrebno je kliknuti na **Continue**.

Slika 1.27.

Prozor "Frequencies Statistics" sa izabranim odgovarajućim statističkim veličinama

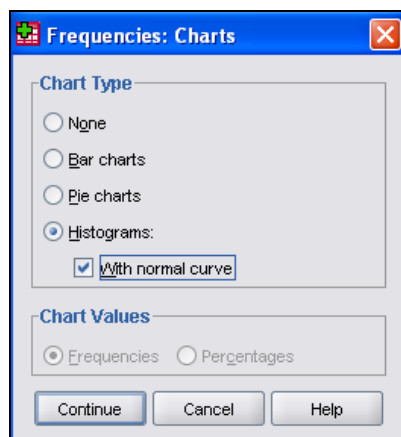


Izvor: Simulirani podaci.

Da bi se konstruirao histogram s normalnom krivuljom potrebno je na prozoru **Frequencies** kliknuti na ikonu **Charts**, čime se otvara se novi prozor **Frequencies: Charts**, gdje je potrebno aktivirati **Histograms: With normal curve**, kako je prikazano na slici 1.28.

Slika 1.28.

Prozor "Frequencies: Charts" sa izabranom opcijom histograma s ucrtanom normalnom krivuljom



Izvor: Simulirani podaci.

Tablica 1.12.

Statistički podaci o prosječnoj ocjeni na I. godini studija promatranih ispitanika

Statistics		
Prosječna ocjena na I. godini studija		
N	Valid	238,000
	Missing	1,000
Mean		4,177
Median		4,000
Mode		4,000
Std. Deviation		,662
Variance		,438
Skewness		-,457
Std. Error of Skewness		,158
Kurtosis		-,410
Std. Error of Kurtosis		,314
Range		3,000
Minimum		2,000
Maximum		5,000
Percentiles	25	4,000
	50	4,000
	75	4,800

Izvor: Simulirani podaci.



Da bi se u **Outputu** programa **SPSS** dobili željeni podaci o prosječnoj ocjeni na I. godini studija ispitanika, potrebno je kliknuti na **Continue**, zatim na **OK**. Rezultat je prikazan u tablici 1.12.

Prema podacima iz tablice 1.12. može se vidjeti da je 238 ispitanika dalo odgovor o svojim ocjenama, dok 1 nije odgovorio na to pitanje. Aritmetička sredina prosječne ocjene na I. godini studija ispitanika u odabranom uzorku (od onih koji su dali odgovor) je 4,177 cm.

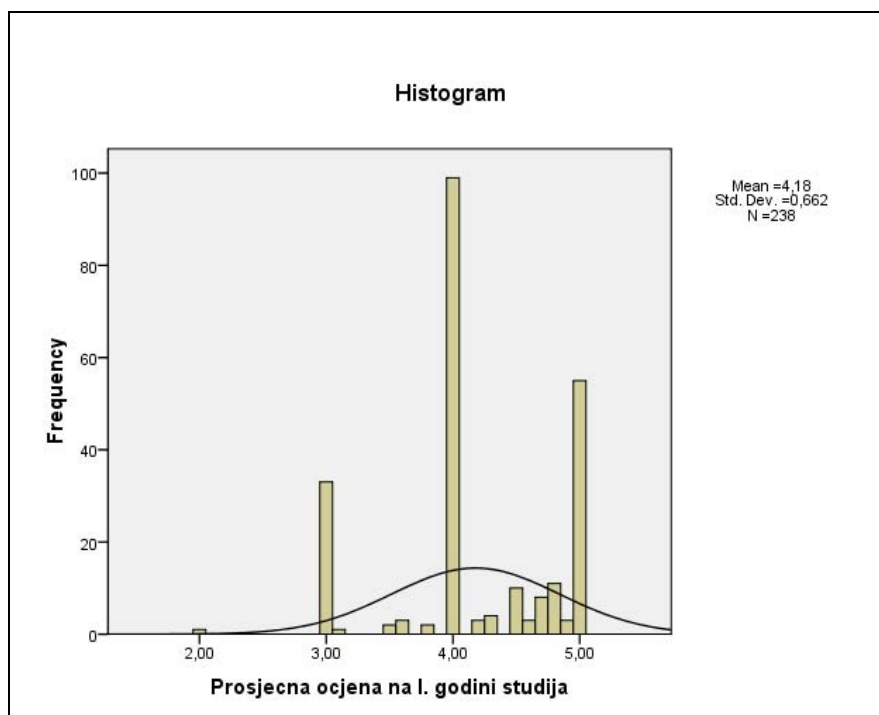
Najmanja prosječna ocjena na I. godini studija ispitanika u uzorku je 2.

Najveća prosječna ocjena na I. godini studija ispitanika u uzorku je 5.

Mod, odnosno najčešća prosječna ocjena na I. godini studija ispitanika je 4.

**Slika 1.29.**

**Histogram zadane distribucije s ucrtanom normalnom krivuljom**



*Izvor: Simulirani podaci.*

Medijan je 4, tj. može se reći da polovina ispitanika ima prosječnu ocjenu na I. godini studija 4 ili manje, dok preostala polovina ispitanika ima prosječnu ocjenu na I. godini studija veću od 4.

Donji kvartil je 4, što je ocjena koje ispitanike dijeli u omjeru 25% i 75%, a gornji kvartil je 4,8, što je ocjena koje ispitanike dijeli u omjeru 75% i 25%.

Raspon varijacije obilježja je 3, što znači da se ocjene ispitanika kreću u intervalu od 3 ocjene.

Varijanca je 0,438 što predstavlja prosječno kvadratno odstupanje ocjena ispitanika od njihove prosječne ocjene. Standardna devijacija je 0,662, što predstavlja prosječno odstupanje ocjena ispitanika od njihove aritmetičke sredine.

Asimetrija distribucije je -0,457, što upućuje na ljevostranu asimetriju. Zaobljenost distribucije je -0,410, što znači da zadana distribucija ima tupi oblik s obzirom na normalnu zaobljenost.

Histogram zadane distribucije s normalnom krivuljom prikazan je na slici 1.29, gdje se mogu vidjeti sve navedene karakteristike zadane distribucije prosječnih ocjena ispitanika na I. godini studija.



#### Primjer 1.21.

Potrebno je dokazati 3 svojstva aritmetičke sredine za anketirane ispitanike prema varijabli (obilježju) prosječna ocjena na I. godini studija (v2).

a) Provjeriti 1. svojstvo da je:  $x_{\min} \leq \bar{X} \leq x_{\max}$ .

b) Provjeriti 2. svojstvo da je:  $\sum_{i=1}^N (x_i - \bar{X}) = 0$ , (za negrupirani niz).

c) Provjeriti 3. svojstvo da je:  $\sum_{i=1}^N (x_i - \bar{X})^2 \leq \sum_{i=1}^N (x_i - a)^2$ , (za negrupirani niz) ako je npr.  $a = 3$ .



#### Rješenje 1.21.

a) Da bi se dokazalo 1. svojstvo da se aritmetička sredina nalazi između najveće i najmanje vrijednosti obilježja treba izračunati  $\bar{X}, x_{\min}$  i  $x_{\max}$ .

Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Frequencies**.

U prozoru **Frequencies** pomoću odgovarajuće strelice varijabla v2 - Prosječna ocjena na I. godini studija se prebacuje u polje **Variable(s)**.

Klikom na ikonu **Statistics** otvara se novi prozor **Frequencies Statistics**, gdje je potrebno aktivirati **Minimum**, **Maximum** i **Mean**.

Dalje je potrebno kliknuti na **Continue**, zatim na **OK**. Rezultat je prikazan u tablici 1.13.

**Tablica 1.13.**

Statistics		
Prosječna ocjena na I. godini studija		
N	Valid	238,0000
	Missing	1,0000
	Mean	4,1773
	Minimum	2,0000
	Maximum	5,0000

*Izvor: Simulirani podaci.*

Iz podataka u tablici 1.13. vidi se da je aritmetička sredina 4,1773 i vrijedi da je između najmanje (2) i najveće (5) vrijednosti obilježja:  $2 \leq 4,1773 \leq 5$ .

b) Da bi se dokazalo 2. svojstvo aritmetičke sredine da je zbroj odstupanja vrijednosti obilježja od prosjeka jednak 0, potrebno je formirati novi stupac (varijablu) gdje će se izračunati ta odstupanja. Zbroj svih podataka iz tog novog stupca mora biti 0. Na glavnom izborniku potrebno je odabrati **Transform**, a na njegovu padajućem izborniku se bira **Compute Variable**. Na taj način se otvori prozor **Compute Variable** koji je prikazan na slici 1.30.

Pod **Target Variable** upisuje se novo ime v3, a u **Numeric Expression** se upisuje:  $v2 - 4,773$ , odnosno da se od svih vrijednosti varijable v2 oduzme njihov prosjek 4,773. Potrebno je napomenuti da se **decimalni zarez** u programu **SPSS** označava **točkom**. Klikom na **OK** u prostoru **Data View** formira se nova varijabla v3.

**Tablica 1.14.**

**Zbroj odstupanja vrijednosti obilježja od prosjeka**

Statistics		
v3		
N	Valid	238
	Missing	1
	Sum	0

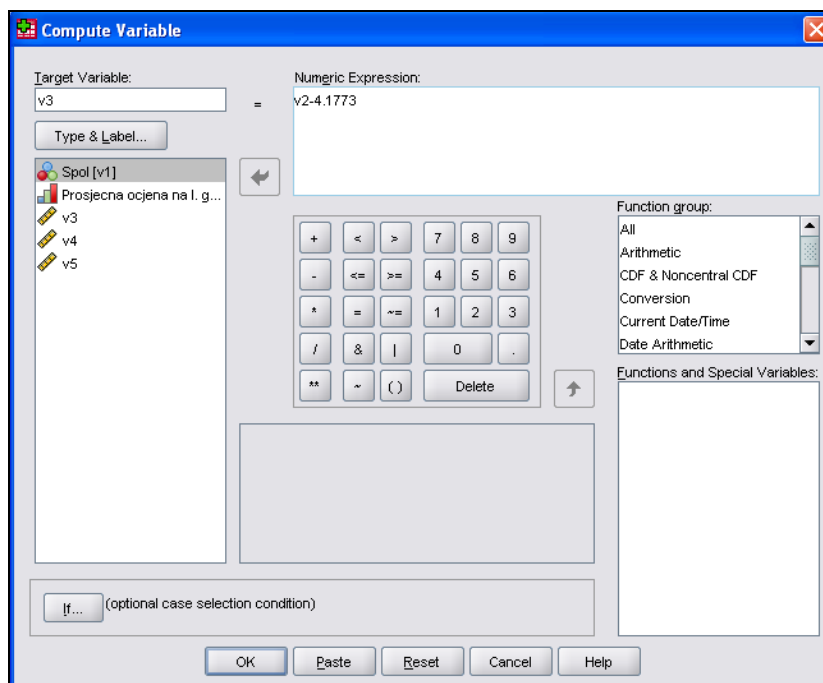
*Izvor: Simulirani podaci.*

Da bi se vidio zbroj podataka u stupcu na glavnom izborniku potrebno je izabrati ikonu *Analyze*, a na njeninu padajućem izborniku *Descriptive Statistics* i *Descriptives*. Varijabla v3 prebacuje se u *Variable(s)*. U *Options* potrebno je aktivirati *Sum* (dok se druge veličine mogu deaktivirati jer u ovom slučaju nisu potrebne).

Suma vrijednosti novoformirane varijable v3 u tablici 1.14 je 0, čime je dokazano 2. svojstvo aritmetičke sredine.

**Slika 1.30.**

**Prozor "Compute Variable" gdje se računaju odstupanja od prosjeka kao nova varijabla v3**



*Izvor: Simulirani podaci.*

c) Da bi se dokazalo 3. svojstvo aritmetičke sredine da je zbroj kvadrata odstupanja vrijednosti obilježja od prosjeka manji ili jednak zbroju kvadrata odstupanja vrijednosti obilježja od bilo kojeg drugog broja, npr. 3, potrebno je formirati nove stupce gdje će se izračunati ti kvadrati odstupanja. Usporedbom zbroja tih stupaca dokazat će se svojstvo 3. Na glavnom izborniku potrebno je odabrati *Transform*, a na njegovu padajućem izborniku bira se *Compute*. Na taj način se otvori prozor *Compute Variable*.

Pod **Target Variable** upisuje se novo ime v4, a u **Numeric Expression** se upisuje:  $(v2 - 4,773)**2$ , odnosno da se od svih vrijednosti varijable v2 oduzme njihov prosjek 4,773 i sve se stavi na kvadrat. Potrebno je napomenuti da se **potencija** u programu SPSS **označava dvjema zvjezdicama**. Klikom na **OK** na listu **Data View** formira se nova varijabla v4.

Zatim se ponovo u **Target Variable** upisuje novo ime v5, a u **Numeric Expression** se upisuje:  $(v2 - 3)**2$ , odnosno od svih se vrijednosti varijable v2 oduzme 3 i sve se stavi na kvadrat. Klikom na **OK** u prostoru **Data View** formira se nova varijabla v5.

Da bi se vidio zbroj podataka u ova dva nova stupca v4 i v5 na glavnom izborniku, potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Descriptives**. Varijable v4 i v5 prebacuju se u **Variable(s)**. U **Options** potrebno je aktivirati **Sum** (dok se druge veličine mogu deaktivirati jer u ovom slučaju nisu potrebne).

**Tablica 1.15.**

**Zbroj kvadrata odstupanja vrijednosti obilježja od prosjeka i od vrijednosti 3**

Statistics			
		v4	v5
N	Valid	238	238
	Missing	1	1
	Sum	103,84	433,72

*Izvor: Simulirani podaci.*

Sume vrijednosti novoformiranih varijabli v4 i v5 prikazane su u zadnjem stupcu tablice 1.15, čime je dokazano 3. svojstvo aritmetičke sredine, tj. da je:

$$\sum_{i=1}^N (x_i - \bar{X})^2 \leq \sum_{i=1}^N (x_i - a)^2, \text{ odnosno da je: } 103,84 \leq 433,72.$$



### **Primjer 1.22.**

Za varijablu (obilježje) džeparac ispitanika (v2) potrebno je odrediti:

a) Najmanju i najveću vrijednost obilježja, aritmetičku sredinu, mod, medijan, donji i gornji kvartil, raspon varijacije obilježja, varijancu i standardnu devijaciju, asimetriju i zaobljenost distribucije te nacrtati histogram s normalnom krivuljom.

b) Komentirati rezultate!

**Rješenje 1.22.**

Na glavnom izborniku potrebno je izabrati ikonu *Analyze*, a na njezinu padajućem izborniku *Descriptive Statistics* i *Frequencies*.

U prozoru *Frequencies* pomoću odgovarajuće strelice varijabla v2 - džeparac ispitanika se prebacuje u polje *Variable(s)*.

Klikom na ikonu *Statistics* otvara se novi prozor *Frequencies Statistics*, gdje je potrebno aktivirati određene statističke veličine prema zahtjevu zadatka, a koje se odnose na odabranu varijablu džeparac ispitanika (v2): *Minimum* i *Maximum*, *Mean*, *Mode*, *Median*, *Quartiles*, *Range*, *Variance*, *Std. deviation*, *Skewness*, *Kurtosis*. Kada se aktiviraju sve tražene veličine potrebno je kliknuti na *Continue*.

Da bi se konstruirao histogram s normalnom krivuljom potrebno je na prozoru *Frequencies* kliknuti na ikonu *Charts*, čime se otvara se novi prozor *Frequencies: Charts*, gdje je potrebno aktivirati *Histograms: With normal curve*.

**Tablica 1.16.**

**Statistički podaci o džeparcu promatranih ispitanika**

Statistics		
Koliki džeparac dobivate		
N	Valid	173,000
	Missing	66,000
Mean		371,549
Median		300,000
Mode		200,000 <sup>a</sup>
Std. Deviation		401,305
Variance		161046,098
Skewness		4,314
Std. Error of Skewness		,185
Kurtosis		26,422
Std. Error of Kurtosis		,367
Range		3500,000
Minimum		,000
Maximum		3500,000
Percentiles	25	150,000
	50	300,000
	75	500,000
a. Multiple modes exist. The smallest value is shown		

*Izvor: Simulirani podaci.*

Da bi se u **Outputu** programa **SPSS** dobili željeni podaci o džeparcu ispitanika, potrebno je kliknuti na **Continue**, zatim na **OK**. Rezultat je prikazan u tablici 1.16.

Može se vidjeti da je 173 ispitanika dalo odgovor o svojoj visini, dok ih 66 nije odgovorio na to pitanje. Aritmetička sredina džeparca ispitanika u odabranom uzorku (od onih koji su dali odgovor) je 371,55 kn.

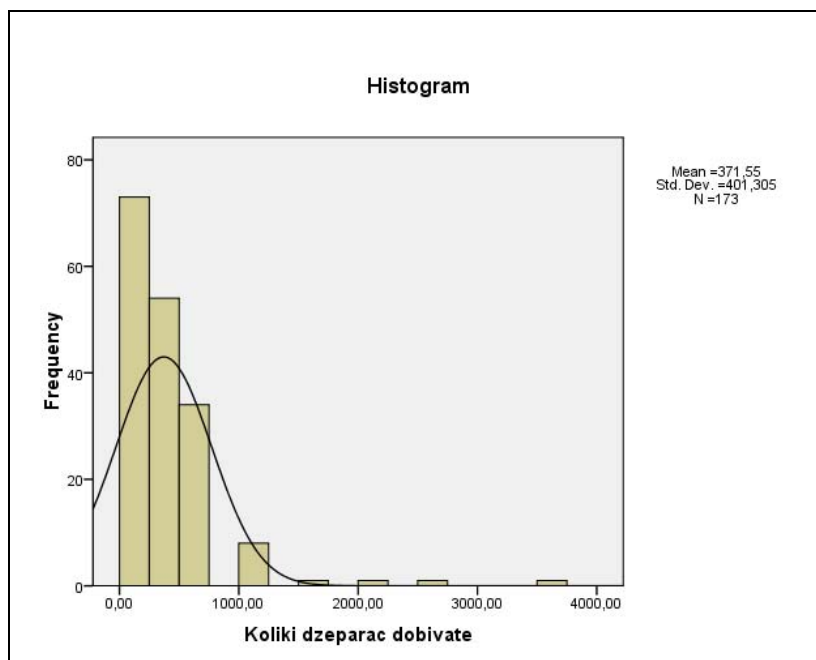
Najmanji džeparac ispitanika u uzorku je 0 kn.

Najveći džeparac ispitanika u uzorku je 3500,00 kn.

Mod, odnosno najčešći džeparac ispitanika je 200,00 kn, a ispod podataka u tablici 1.16 može se vidjeti napomena da ova distribucija ima više od jednog moda. Prema tablici frekvencija, koja se automatizmom ovom procedurom dobije u **Outputu** programa **SPSS**, postoji još jedan mod i on iznosi 300,00 kn. Naime vrijednosti džeparca od 200,00 kn i 300,00 kn odgovaraju najvećim apsolutnim frekvencijama, koje iznose 26. Dakle, ova zadana distribucija džeparca ispitanika je bimodalna distribucija.

**Slika 1.31.**

#### Histogram zadane distribucije s ucrtanom normalnom krivuljom



Izvor: Simulirani podaci.

Medijan je 300,00 kn, tj. može se reći da polovina ispitanika ima džeparac 300,00 kn ili manji, dok preostala polovina ispitanika ima džeparac veći od 300,00 kn.

Donji kvartil je 150,00 kn, što je džeparac koje ispitanike dijeli u omjeru 25% i 75%, a gornji kvartil je 500,00 kn, što je džeparac koji ispitanike dijeli u omjeru 75% i 25%.

Raspon varijacije obilježja je 3500,00 kn, što u stvari znači da se džeparac ispitanika kreće u rasponu od 3500,00 kn.

Varijanca je 161046,098 kn<sup>2</sup> što predstavlja prosječno kvadratno odstupanje džeparca ispitanika od njihovog prosječnog džeparca. Standardna devijacija je 401,31 kn, što predstavlja prosječno odstupanje džeparca ispitanika od aritmetičke sredine.

Asimetrija distribucije je 4,314, što upućuje na desnostranu asimetriju. Zaobljenost distribucije je 26,422, što znači da zadana distribucija ima šiljasti oblik s obzirom na normalnu zaobljenost.

Histogram zadane distribucije s normalnom krivuljom prikazan je na slici 1.31, gdje se može vidjeti izrazito desnostrana asimetrija, šiljasti vrh i sve ostale karakteristike zadane distribucije džeparca ispitanika.

## 1.12 Složeno prikazivanje statističkih podataka

Kada se istovremeno u istoj tablici žele prikazati dva statistička obilježja ili varijable koriste se složene statističke tablice u kojima se podaci mijenjaju po redcima i po stupcima.

Kao grafički prikaz ovakvih podataka najprikladnije je koristiti grafikon višestrukih stupaca.



### **Primjer 1.23.**

Zadatak je formirati dvostruku statističku tablicu, gdje se varijabla (obilježje) spol (v1) mijenja po redcima (row), a varijabla (obilježje) grupirani džeparac (v2) po stupcima (column), nacrtati grafikon višestrukih stupaca prema odabranim varijablama i komentirati sliku!



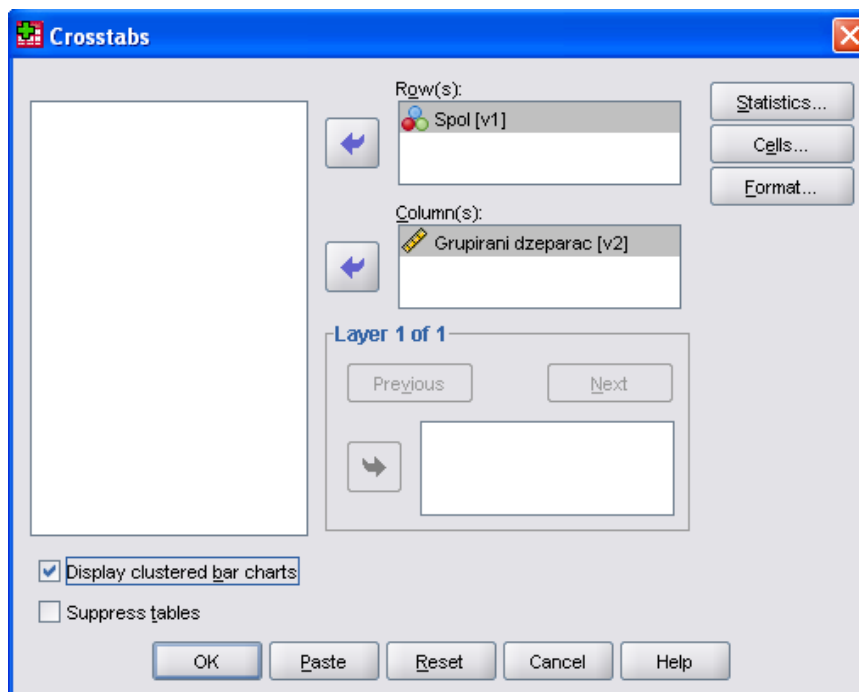


### Rješenje 1.23.

Da bi se formirala dvostruka statistička tablica sa zbirnim redcima i stupcima, potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Crosstabs**. Prema zahtjevu zadatka pomoću odgovarajućih strelica varijabla v1 - Spol se prebacuje u polje **Row(s)**, a druga varijabla v2 - Grupirani džeparac u polje **Column(s)**. To je prikazano na slici 1.33.

Slika 1.33.

Prozor "Crosstabs" s odabranim varijablama v1 u redcima i v2 u stupcima



Izvor: Simulirani podaci.

Da bi se u **Outputu** dobio i grafikon višestrukih stupaca pri kreiranju dvostruke statističke tablice u prozoru **Crosstabs** potrebno je aktivirati **Display clustered bar charts**. Klikom na **OK** u **Outputu** se za rezultat dobije tražena tablica sa zbirnim stupcima i redcima koja je prikazana u tablici 1.17.

Prema podacima u tablici može se vidjeti da najviše ispitanika ima džeparac do 200 kn. Iako je u analizi više ispitanika ženskog spola, pokazalo se da veći džeparac imaju oni muškog spola.

**Tablica 1.17.**

**Ispitanici prema spolu i grupiranom džeparcu**

Spol * Grupirani džeparac Crosstabulation								
Count		Grupirani džeparac						
		200,00	400,00	600,00	800,00	1000,00	3500,00	Total
Spol	Musko	28	18	10	4	4	4	68
	Zensko	45	34	20	2	3	1	105
	Total	73	52	30	6	7	5	173

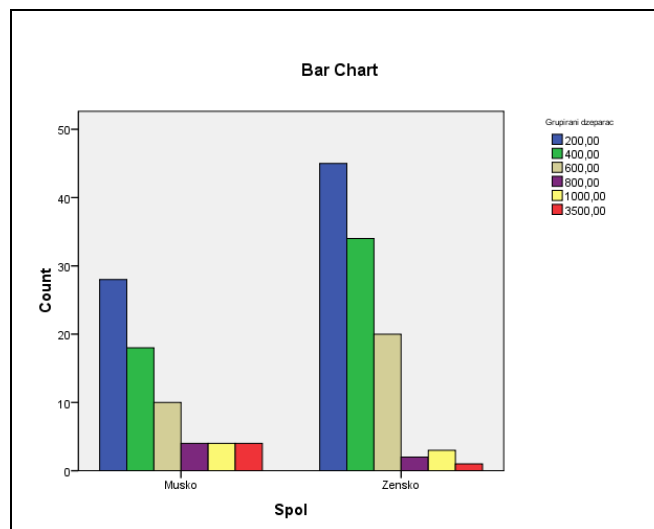
*Izvor: Simulirani podaci.*

U zbirnim redcima i stupcima nalaze se marginalne frekvencije koje kažu koliko ispitanika ima po jednoj grupi obilježja bez obzira na drugo obilježje. Na primjer, prvi podatak u zbirkom stupcu kaže da je 68 ispitanika muškog spola bez obzira na džeparac. Treći podatak u zbirkom retku kaže da je 30 ispitanika imalo džeparac između 400 i 600 kn bez obzira na spol.

Traženi grafikon višestrukih stupaca prikazan je na slici 1.34.

**Slika 1.34.**

**Višestruki stupci visine džeparca prema spolu**



*Izvor: Simulirani podaci.*

Može se vidjeti da se prvi višestruki stupac odnosi na ispitanike muškog spola, a drugi na ispitanike ženskog spola. I u jednom i u drugom dijelu grafikona najveći

stupci su oni koji prikazuju niži džeparac, dok oni manji stupci pokazuju viši iznos džeparca.



#### Primjer 1.24.

Zadatak je formirati dvostruku statističku tablicu, gdje se varijabla (obilježje) spol (v1) mijenja po stupcima (column), a varijabla (obilježje) grupirani džeparac (v2) po redcima (row), nacrtati grafikon višestrukih stupaca prema odabranim varijablama i komentirati sliku!



#### Rješenje 1.24.

Da bi se formirala dvostruka statistička tablica sa zbirnim redcima i stupcima, potrebno je na glavnom izborniku izabrati ikonu *Analyze*, a na njezinu padajućem izborniku *Descriptive Statistics* i *Crosstabs*. Prema zahtjevu zadatka pomoću odgovarajućih strelica varijabla v1 - Spol se prebacuje u polje *Column(s)*, a druga varijabla v2 - Grupirani džeparac u polje *Row(s)*.

Sličan je princip kao što je prikazano na slici 1.33, samo što odabrane varijable po stupcima i po redcima mijenjaju svoje mjesto.

Da bi se u *Outputu* dobio i grafikon višestrukih stupaca pri kreiranju dvostruke statističke tablice u prozoru *Crosstabs* potrebno je aktivirati *Display clustered bar charts*. Klikom na *OK* u *Outputu* se za rezultat dobije tražena tablica sa zbirnim stupcima i redcima koja je prikazana u tablici 1.18.

Tablica 1.18.

#### Ispitanici prema spolu i grupiranom džeparcu

Grupirani džeparac * Spol Crosstabulation				
Count		Spol		
		Musko	Zensko	Total
Grupirani džeparac	200,00	28	45	73
	400,00	18	34	52
	600,00	10	20	30
	800,00	4	2	6
	1000,00	4	3	7
	3500,00	4	1	5
	Total	68	105	173

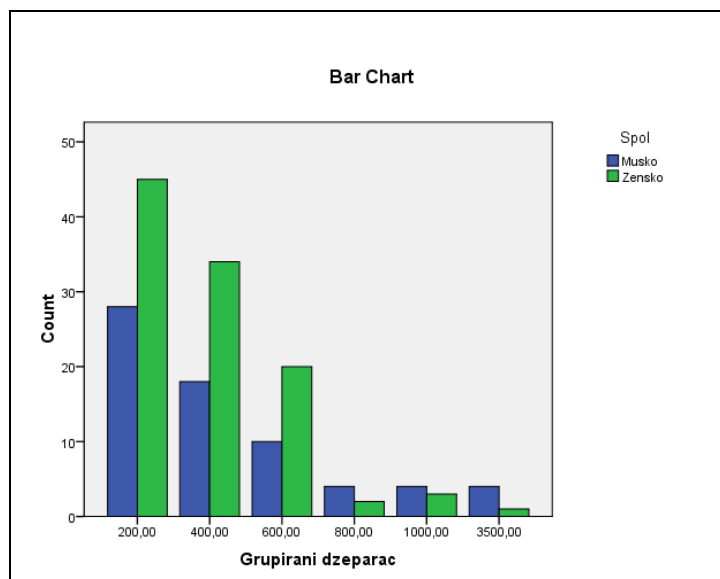
Izvor: Simulirani podaci.

Ako se napravi usporedba tablice 1.18 s tablicom 1.17, može se vidjeti da se sada varijabla spol mijenja po stupcima, a varijabla grupirani džeparac po redcima. I iz ove tablice jasno je da najveći broj ispitanika i jednog i drugog spola imaju najmanju kategoriju džeparca, tj. do 200 kn na mjesec. Na primjer, prvi podatak u zbirnom retku kaže da je 68 ispitanika muškog spola bez obzira na džeparac. Treći podatak u zbirnom stupcu kaže da je 30 ispitanika imalo džeparac između 400 i 600 kn bez obzira na spol.

Traženi grafikon višestrukih stupaca prikazan je na slici 1.35.

**Slika 1.35.**

### Višestruki stupci visine džeparca prema spolu



*Izvor: Simulirani podaci.*

Može se vidjeti da se svaki dvostruki stupac odnosi na različitu kategoriju visine džeparca. U prva tri dvostruka stupca veću zastupljenost imaju ispitanici ženskog spola, a prema posljednja tri stupca može se zaključiti da je muški spol u većini. Dakle, općenito ispitanici ženskog spola imaju niži iznos džeparca, dok ispitanici muškog spola imaju veći iznos džeparca.



### Primjer 1.25.

Zadatak je formirati dvostruku statističku tablicu, gdje se varijabla (obilježje) školska sprema oca (v1) mijenja po redcima (row), a varijabla (obilježje) priključenje EU

(v2) po stupcima (column), nacrtati grafikon višestrukih stupaca prema odabranim varijablama i komentirati sliku!



### Rješenje 1.25.

Na glavnom izborniku potrebno je izabrati ikonu *Analyze*, a na njezinu padajućem izborniku *Descriptive Statistics* i *Crosstabs*.

Tablica 1.19.

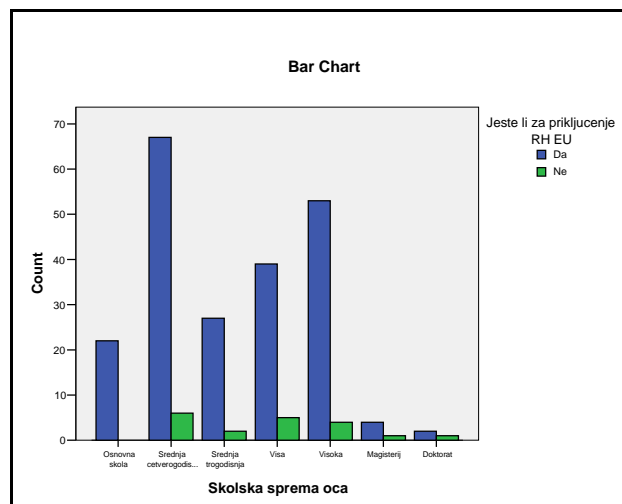
### Ispitanici prema školskoj spremi oca i mišljenju o priključenju EU

Školska sprema oca * Jeste li za priključenje RH EU Crosstabulation				
Count		Jeste li za priključenje RH EU		
		Da	Ne	Total
Školska sprema oca	Osnovna škola	22	0	22
	Srednja četverogodišnja	67	6	73
	Srednja trogodišnja	27	2	29
	Visa	39	5	44
	Visoka	53	4	57
	Magisterij	4	1	5
	Doktorat	2	1	3
	Total	214	19	233

Izvor: Simulirani podaci.

Slika 1.36.

### Ispitanici prema školskoj spremi oca i mišljenju o priključenju EU



Izvor: Simulirani podaci.

Pomoću odgovarajućih strelica varijabla v1 - Školska sprema oca se prebacuje u polje **Row(s)**, a druga varijabla v2 - Prikličenje EU u polje **Column(s)**. Da bi se u **Outputu** dobio i grafikon višestrukih stupaca pri kreiranju dvostruke statističke tablice u prozoru **Crosstabs**, potrebno je aktivirati **Display clustered bar charts**. Klikom na **OK** u **Outputu** se za rezultat dobije tražena dvostruka tablica, koja je prikazana tablicom 1.19 i grafikon, prikazan na slici 1.36.

Prema podacima u tablici 1.19 i slici 1.36 može se zaključiti da je većina ispitanika iz odabranog uzorka za priključenje Republike Hrvatske Europskoj uniji, bez obzira na školsku spremu njihova oca. Isto tako može se primijetiti da je u uzorku najveći broj ispitanika kojima otac ima srednju stručnu spremu (što je zaista i općenita situacija u Republici Hrvatskoj - prema podacima u SLJH, razne godine - pa bi se u tom smislu moglo komentirati o slučajnosti izabranog uzorka).



#### Primjer 1.26.

Zadatak je formirati dvostruku statističku tablicu, gdje se varijabla (obilježje) priključenje EU (v1) mijenja po redcima (row), a varijabla (obilježje) vjenčanje s osobom druge vjere (v2) po stupcima (column), nacrtati grafikon višestrukih stupaca prema odabranim varijablama i komentirati sliku!



#### Rješenje 1.26.

Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Crosstabs**. Pomoću odgovarajućih strelica varijabla v1 - Prikličenje EU u polje **Row(s)**, a druga varijabla v2 - Vjenčanje sa osobom druge vjere u polje **Column(s)**.

Tablica 1.20.

Ispitanici prema stavu o priključenju EU i o vjenčanju sa osobom druge vjere

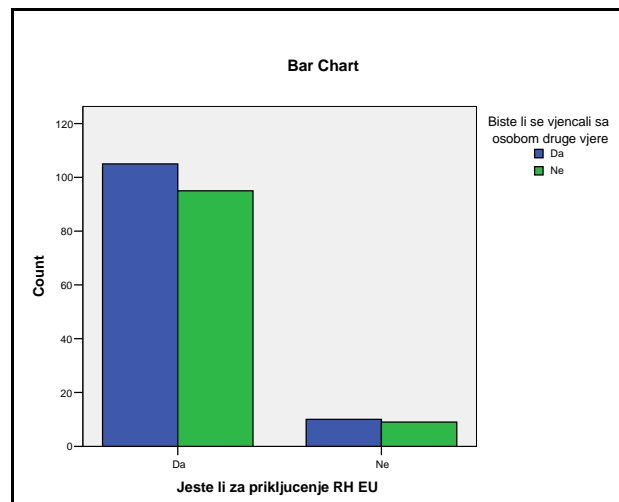
Jeste li za priključenje RH EU * Biste li se vjenicali sa osobom druge vjere Crosstabulation				
Count		Biste li se vjenicali sa osobom druge vjere		
		Da	Ne	Total
Jeste li za priključenje RH EU	Da	105	95	200
	Ne	10	9	19
	Total	115	104	219

Izvor: Simulirani podaci.

Da bi se u **Outputu** dobio i grafikon višestrukih stupaca pri kreiranju dvostruke statističke tablice u prozoru **Crosstabs**, potrebno je aktivirati **Display clustered bar charts**. Klikom na **OK** u **Outputu** se za rezultat dobije tražena dvostruka tablica koja je prikazana tablicom 1.20 i grafikon prikazan na slici 1.37.

**Slika 2.28**

**Ispitanici prema stavu o priključenju EU i o vjenčanju sa osobom druge vjere**



*Izvor: Simulirani podaci.*

Prema podacima u tablici 1.20 i slici 1.37 može se zaključiti da je većina ispitanika iz odabranog uzorka za priključenje Republike Hrvatske Europskoj uniji, što se vidi prema marginalnim frekvencijama u zbirnom stupcu tablice i prema prvom dvostrukom stupcu grafikona. Isto tako može se primijetiti da je u uzorku broj ispitanika koji bi se vjenčao sa osobom druge vjere malo veći od onih koji to ne bi napravili, te da su otprilike ravnomjerno raspoređeni bez obzira na mišljenje o tome jesu li za priključenje Republike Hrvatske Europskoj uniji.



#### **Primjer 1.27.**

Na temelju podataka sakupljenih anonimnom anketom u jednom poduzeću zadatak je formirati dvostruku statističku tablicu, gdje se varijabla (obilježje) financijske prilike u obitelji (v1) mijenja po redcima (row), a varijabla (obilježje) uživate li drogu (v2) po stupcima (column), nacrtati grafikon višestrukih stupaca prema odabranim varijablama i komentirati sliku!

**Rješenje 1.27.**

Na glavnom izborniku potrebno je izabrati ikonu *Analyze*, a na njezinu padajućem izborniku *Descriptive Statistics* i *Crosstabs*. Pomoću odgovarajućih strelica varijabla v1 - Financijske prilike u obitelji prebacuje se u polje *Row(s)*, a druga varijabla v2 - Uživete li drogu u polje *Column(s)*. Da bi se u *Outputu* dobio i grafikon višestrukih stupaca pri kreiranju dvostruke statističke tablice u prozoru *Crosstabs*, potrebno je aktivirati *Display clustered bar charts*. Klikom na *OK* u *Outputu* se za rezultat dobije tražena dvostruka tablica koja je prikazana tablicom 1.21 i grafikon prikazan na slici 1.38.

**Tablica 1.21.**

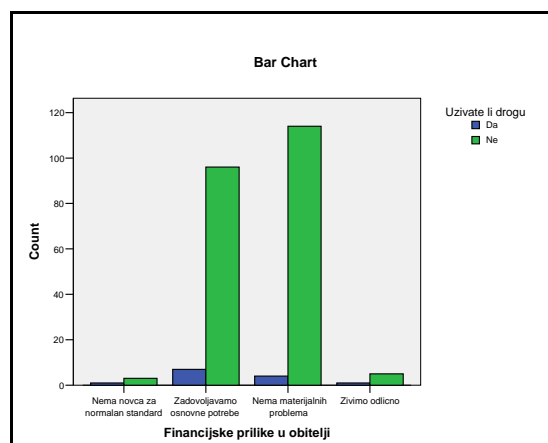
**Uzorak zaposlenih prema financijskim prilikama u obitelji i konzumiranju droge**

Financijske prilike u obitelji * Uživete li drogu Crosstabulation				
Count		Uživete li drogu		
		da	ne	Total
Financijske prilike u obitelji	nema sredstava za normalan standard	1	3	4
	zadovoljavamo osnovne potrebe	7	96	103
	nema materijalnih problema	4	114	118
	živimo odlično	1	5	6
	Total	13	218	231

*Izvor: Simulirani podaci.*

**Slika 1.38.**

**Uzorak zaposlenih prema financijskim prilikama u obitelji i konzumiranju droge**



*Izvor: Simulirani podaci.*



Prema podacima u tablici 1.21 i slici 1.38 može se zaključiti da većina ispitanika iz odabranog uzorka ipak ne konzumira drogu. Ispitanika koji su se izjasnili da nemaju novca za normalan standard te onih koji žive odlično ima znatno manje u odnosu na ostale kategorije (prvi i zadnji višestruki stupac grafikona na slici 1.38).



### Primjer 1.28.

Na temelju podataka sakupljenih anonimnom anketom u jednom poduzeću zadatak je formirati dvostruku statističku tablicu, gdje se varijabla (obilježje) uživete li drogu (v2) mijenja po redcima (row), a varijabla (obilježje) spol (v1) po stupcima (column), nacrtati grafikon višestrukih stupaca prema odabranim varijablama i komentirati sliku!



### Rješenje 1.28.

Na glavnom izborniku potrebno je izabrati ikonu *Analyze*, a na njezinu padajućem izborniku *Descriptive Statistics* i *Crosstabs*. Pomoću odgovarajućih strelica varijabla v2 - Uživete li drogu premjestiti u polje *Row(s)*, a drugu varijablu v1 - Spol u polje *Column(s)*. Da bi se u *Outputu* dobio i grafikon višestrukih stupaca pri kreiranju dvostruke statističke tablice u prozoru *Crosstabs*, potrebno je aktivirati *Display clustered bar charts*.

Tablica 1.22.

### Uzorak zaposlenih prema konzumiranju droge i spolu

Uživete li drogu * Spol Crosstabulation				
Count		Spol		
		Musko	Zensko	Total
Uživete li drogu	Da	7	6	13
	NE	84	136	220
	Total	91	142	233

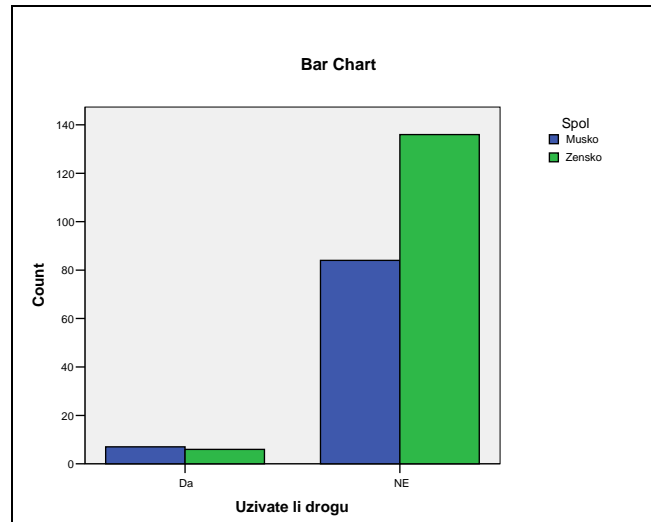
Izvor: Simulirani podaci.

Klikom na *OK* u *Outputu* se za rezultat dobije tražena dvostruka tablica koja je prikazana tablicom 1.22 i grafikom prikazan na slici 1.39.

Može se zaključiti da je većina ispitanika u odabranom uzorku ženskog spola. Prvi dvostruki stupac grafikona odnosi se na one ispitanike koji konzumiraju drogu. Može se vidjeti da je među njima više ispitanika muškog spola (i prema prvom retku tablice 1.22).

**Slika 1.39.**

**Uzorak zaposlenih prema konzumiranju droge i spolu**



*Izvor: Simulirani podaci.*

## 2 VJEROJATNOST I PROCJENA PROSJEČNE VRIJEDNOSTI

### 2.1 Vjerojatnost

Slučajni događaj je takav događaj koji se može, ali ne mora realizirati, tj. realizira se uz određenu vjerojatnost.

Vjerojatnost *realizacije slučajnog događaja "A"* jednaka je omjeru broja povoljnih ishoda i svih mogućih ishoda:

$$P(A) = \frac{m(A)}{n}, \quad 0 \leq P \leq 1, \quad (2.1)$$

gdje je:

$m(A)$  - broj svih povoljnih ishoda za događaj A,

$n$  - broj svih mogućih ishoda.

Za siguran događaj vjerojatnost je:  $\Rightarrow P = 1$  (obrat ne vrijedi).

Za nemoguć događaj vjerojatnost je:  $\Rightarrow P = 0$  (obrat ne vrijedi).

Vjerojatnost da se **slučajni događaj "A"** *ne realizira* jednaka je omjeru broja nepovoljnih ishoda i svih mogućih ishoda:

$$P(\bar{A}) = \frac{n - m(A)}{n}, \quad 0 \leq P \leq 1, \quad (2.2)$$

gdje je:

$n - m(A)$  - broj svih nepovoljnih ishoda za događaj A,

$n$  - broj svih mogućih ishoda.

Vrijedi da je zbroj vjerojatnosti da se neki događaj realizira i vjerojatnosti da se taj događaj ne realizira jednak 1:

$$P(A) + P(\bar{A}) = 1. \quad (2.3)$$

Za ovakav slučaj kada je unaprijed poznat broj svih povoljnih/nepovoljnih i ukupnih ishoda izračunata vjerojatnost se naziva **vjerojatnost "a priori"**.

Ako vjerojatnost realizacije slučajnog događaja  $A$  nije poznata unaprijed, može se izračunati tzv. **vjerojatnost "a posteriori"**:

$$P(A) = p \lim \frac{m(A)}{n}, \quad (2.4)$$

gdje je:

$p \lim$  - čita se "granična vrijednost po vjerojatnosti" (da se razlikuje od limesa u linearnoj algebri).



### Primjer 2.1.

Na temelju podataka sakupljenih anonimnom anketom u jednom poduzeću potrebno je izračunati vjerojatnost da odabrani ispitanik bude ženskog spola (varijabla: Spol (v1)).



### Rješenje 2.1.

Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Frequencies**.

U prozoru **Frequencies** pomoću odgovarajuće strelice varijabla v1 - Spol se prebacuje u polje **Variable(s)**. Konačno, da bi se u **Outputu** programa **SPSS** dobili željeni podaci o spolu ispitanika, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.1.

**Tablica 2.1.**

**Podaci o spolu ispitanika**

Spol					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Musko	94	39,3	39,5	39,5
	Zensko	144	60,3	60,5	100,0
	Total	238	99,6	100,0	
Missing	System	1	,4		
Total		239	100,0		

*Izvor: Simulirani podaci.*

Prema podacima iz tablice 5.1. u stupcu **Valid Percent** može se vidjeti da je 60,5% ispitanika ženskog spola, odnosno vjerojatnost da ispitanik bude ženskog spola je 0,605.



### Primjer 2.2.

Na temelju podataka sakupljenih anonimnom anketom u jednom poduzeću za varijablu grupirana visina (v1) na sljedeći način:

Range: Lowest through

- 160  $\Rightarrow$  160,
- 170  $\Rightarrow$  170,
- 180  $\Rightarrow$  180,
- 190  $\Rightarrow$  190,
- 200  $\Rightarrow$  200,
- 210  $\Rightarrow$  210,

potrebno je izračunati vjerojatnost da ispitanik bude visok između 171 i 180 cm.



### Rješenje 2.2.

Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Frequencies**.

U prozoru **Frequencies** pomoću odgovarajuće strelice varijabla v1 - Grupirana visina se prebacuje u polje **Variable(s)**.

Konačno, da bi se u **Outputu** programa SPSS dobili željeni podaci o visini ispitanika, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.2.

**Tablica 2.2.**

#### Podaci o visini ispitanika

Grupirana visina					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	160,00	7	2,9	3,0	3,0
	170,00	62	25,9	26,4	29,4
	180,00	103	43,1	43,8	73,2
	190,00	50	20,9	21,3	94,5
	200,00	11	4,6	4,7	99,1
	210,00	2	,8	,9	100,0
	Total	235	98,3	100,0	
Missing	System	4	1,7		
Total		239	100,0		

*Izvor: Simulirani podaci.*

Prema podacima iz tablice 2.2. u stupcu **Valid Percent** može se vidjeti da je 43,8% ispitanika s visinom između 171 i 180 cm, odnosno vjerojatnost da ispitanik bude visok između 171 i 180 cm je 0,438.



### Primjer 2.3.

Na temelju podataka sakupljenih anonimnom anketom u jednom poduzeću potrebno je izračunati vjerojatnost da se slučajno odabere osoba kojoj su roditelji u braku.



### Rješenje 2.3.

Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Frequencies**.

U prozoru **Frequencies** pomoću odgovarajuće strelice varijabla v1 - Bračno stanje roditelja se prebacuje u polje **Variable(s)**.

Konačno, da bi se u **Output** programa SPSS dobili željeni podaci o bračnom stanju roditelja ispitanika, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.3.

Tablica 2.3.

Podaci o bračnom stanju roditelja ispitanika

Bračno stanje roditelja					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	U braku	215	90,0	90,0	90,0
	Rastavljeni	10	4,2	4,2	94,1
	Umro jedan od roditelja	11	4,6	4,6	98,7
	Ostalo	3	1,3	1,3	100,0
	Total	239	100,0	100,0	

Izvor: Simulirani podaci.

Prema podacima iz tablice 2.3 u stupcu **Valid Percent** može se vidjeti da 90% ispitanika iz promatranog poduzeća ima roditelje u braku, odnosno vjerojatnost da ispitanik ima roditelje u braku je 0,9.

### Teorem o zbrajanju vjerojatnosti

**Za međusobno isključive događaje** (tj. one događaje koji ne mogu nastupiti istodobno, odnosno čiji su pripadni skupovi elementarnih događaja disjunktni), vjerojatnost realizacije jednog **ili** drugog događaja jednaka je zbroju vjerojatnosti realizacije jednog i vjerojatnosti realizacije drugog događaja.

$$P(A \text{ ili } B) = P(A) + P(B). \quad (2.5)$$

Za događaje koji se međusobno ne isključuju vjerojatnost realizacije jednog ili drugog događaja je:

$$P(A \text{ ili } B) = P(A) + P(B) - P(AB), \quad (2.6)$$

gdje je:

$P(AB)$  -vjerojatnost da slučajni događaji A i B nastupe istovremeno.

### **Teorem o množenju vjerojatnosti**

Ako se događaji A i B **međusobno ne isključuju i nezavisni su**, tj. vjerojatnost realizacije jednog ne zavisi o vjerojatnosti realizacije drugog događaja, tada je vjerojatnost istovremene realizacije događaja A i B jednaka produktu vjerojatnosti realizacije događaja A i događaja B:

$$P(A \text{ i } B) = P(A) \cdot P(B). \quad (2.7)$$



#### **Primjer 2.4.**

Na temelju anonimne ankete provedene u jednom poduzeću:

- Potrebno je izračunati vjerojatnost da ispitanik ne pije nikada alkoholna pića.
- Potrebno je izračunati vjerojatnost da ispitanik ženskog spola ne pije nikada alkoholna pića.



#### **Rješenje 2.4.**

- Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Frequencies**.

U prozoru **Frequencies** pomoću odgovarajuće strelice varijabla v1 - o Konzumiranju alkohola se prebacuje u polje **Variable(s)**.

Konačno, da bi se u **Outputu** programa **SPSS** dobili željeni podaci o konzumaciji alkohola ispitanika, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.4.

Prema podacima iz tablice 2.4. u stupcu **Valid Percent** može se vidjeti da 22,8% ispitanika ne pije nikada alkohol, odnosno vjerojatnost da ispitanik ne pije alkohol nikada je 0,228.

Tablica 2.4.

## Podaci o konzumiranju alkohola ispitanika

Konzumirate li alkohol					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nikada	54	22,6	22,8	22,8
	Samo uz obroke	9	3,8	3,8	26,6
	U posebnim prilikama	165	69,0	69,6	96,2
	Cesto i u svakoj prigodi	9	3,8	3,8	100,0
	Total	237	99,2	100,0	
Missing	System	2	,8		
Total		239	100,0		

Izvor: Simulirani podaci.

b) Da bi se izračunala vjerojatnost da ispitanik ženskog spola ne pije nikada alkoholna pića, potrebno je ostaviti u analizi samo ispitanike ženskog spola, odnosno filtrirati željeni uzorak.

Na glavnom izborniku potrebno je izabrati ikonu **Data**, a na njezinu padajućem izborniku **Select Cases**.

U prozoru **Select Cases** potrebno je aktivirati opciju **If conditions is satisfied**. Klikom na ikonu **if** otvara se prozor **Select Cases:If** gdje je potrebno odabrati varijablu **v2** i postaviti uvjet da je  $v2 = 2$ , odnosno da je spol 2, što označava ženski spol.

Konačno, da bi se filtrirali samo ispitanici ženskog spola, potrebno je kliknuti na **Continue** i **OK**. Daljnja analiza se sada odnosi samo na ženski spol.

Tablica 2.5.

## Podaci o konzumiranju alkohola ispitanika ženskog spola

Konzumirate li alkohol					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nikada	37	25,7	26,1	26,1
	Samo uz obroke	3	2,1	2,1	28,2
	U posebnim prilikama	94	65,3	66,2	94,4
	Cesto i u svakoj prigodi	8	5,6	5,6	100,0
	Total	142	98,6	100,0	
Missing	System	2	1,4		
Total		144	100,0		

Izvor: Simulirani podaci.

Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Frequencies**.



U prozoru **Frequencies** pomoću odgovarajuće strelice varijabla v1 - o Konzumiranju alkohola se prebacuje u polje **Variable(s)**.

Konačno, da bi se u **Outputu** programa SPSS dobili željeni podaci o konzumaciji alkohola ispitanika ženskog spola, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.5.

Prema podacima iz tablice 2.5 u stupcu **Valid Percent** može se vidjeti da 26,1% od ispitanika ženskog spola ne pije nikada alkohol, odnosno vjerojatnost da ispitanik ženskog spola nikada ne pije alkohol je 0,261.

## 2.2 Diskontinuirana slučajna varijabla

Diskontinuirana ili diskretna slučajna varijabla je takva varijabla koja na slučaj može poprimiti najviše prebrojivo mnogo vrijednosti i svaku od njih s određenom vjerojatnošću.

Zadovoljava uvjete:

- normativnost:  $\sum_{i=1}^{\infty} P(x_i) = 1$ ,
- nenegativnost:  $P(x_i) \geq 0, \forall i$ .

Skup svih uređenih parova vrijednosti slučajne varijable i njoj pripadajućih vjerojatnosti naziva se **distribucija vjerojatnosti slučajne varijable  $X$** :

$$\{[x_i; P(x_i)], i = 1, 2, \dots, \infty\}. \quad (2.8)$$

Zakon po kojem svakoj vrijednosti slučajne varijable  $X$  pripada vjerojatnost  $P(x_i)$  naziva se **zakon vjerojatnosti slučajne varijable  $X$** .

**Funkcija distribucije slučajne varijable  $X$**  je funkcija koja daje vjerojatnost da će slučajna varijabla  $X$  poprimiti vrijednost jednaku ili manju od nekog realnog broja  $x_k$ :

$$F(x_k) = P(X \leq x_k) = \sum_{i=1}^k P(x_i) \quad (2.9)$$

Može se reći da je to **kumulativ vjerojatnosti** od  $F(X=0)$  do  $F(X=x_k)$ . Funkcija distribucije slučajne varijable  $X$  je monotonno neopadajuća funkcija.

**Parametri distribucije:**

- **Očekivanje slučajne varijable** je zbroj umnožaka vrijednosti varijable  $X$  i njoj pripadajućih odgovarajućih vjerojatnosti  $P(x_i)$  :

$$E(X) = \sum_{i=1}^{\infty} x_i \cdot P(x_i) = \mu . \quad (2.10)$$

- **Varijanca slučajne varijable** je očekivanje kvadratnog odstupanja vrijednosti varijable  $X$  od njenog očekivanja:

$$V(X) = E(X^2) - [E(X)]^2 = E(X - \mu)^2 = \sum_{i=1}^{\infty} x_i^2 \cdot P(x_i) - \mu^2 . \quad (2.11)$$

**2.2.1 Binomna distribucija**

Ako je vjerojatnost da se dogodi neki događaj poznata, unaprijed utvrđena i konstantna tijekom cijelog istraživanja (iznosi  $p$ ) kaže se da se diskontinuirana slučajna varijabla  $X$  ravna prema tzv. **binomnoj distribuciji**.

Vjerojatnost da se neki slučajni događaj  $X$  u  $n$  pokusa realizira  $x$  puta prema **binomnom zakonu vjerojatnosti** je:

$$P(X = x) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x} = \binom{n}{x} \cdot p^x \cdot q^{n-x}, \quad x = 0, 1, 2, \dots, n \quad (2.12)$$

gdje je:

- $n$  - broj pokusa,
- $p$  - vjerojatnost realizacije slučajnog događaja,
- $q$  - vjerojatnost da se slučajni događaj ne realizira,
- $x$  - broj povoljnih ishoda u  $n$  pokusa,
- $n - x$  - broj nepovoljnih ishoda u  $n$  pokusa.

**Očekivanje** slučajne varijable  $X$  kod binomne distribucije:

$$E(X) = n \cdot p . \quad (2.13)$$

**Varijanca** slučajne varijable  $X$  kod binomne distribucije:

$$\sigma^2 = n \cdot p \cdot q . \quad (2.14)$$

**Koeficijent asimetrije** slučajne varijable  $X$  kod binomne distribucije:

$$\alpha_3 = \frac{q - p}{\sqrt{n \cdot p \cdot q}} . \quad (2.15)$$

**Koeficijent zaobljenosti** slučajne varijable  $X$  kod binomne distribucije:

$$\alpha_4 = 3 + \frac{1 - 6 \cdot p \cdot q}{n \cdot p \cdot q} . \quad (2.16)$$

Općenito se kaže da se **slučajna varijabla  $X$  ravna po binomnoj distribuciji**, koja je određena parametrima  $n$  i  $p$ :

$$X \sim B(n, p) . \quad (2.17)$$

U programskom paketu **SPSS** postoje **funkcije gustoće vjerojatnosti** i **kumulativne funkcije distribucije vjerojatnosti** *Binomne distribucije*.

**Funkcije gustoće vjerojatnosti (Probability Density Functions)**

- **Transform; Compute; Function group (izabrati: PDF & Noncentral PDF); izabrati: Pdf.Binom u Numeric Expression:**

**PDF.BINOM(quant,n,prob)**

Rezultat ove funkcije su vjerojatnosti da broj uspjeha u  $n$  pokušaja, uz vjerojatnost realizacije  $p = prob$ , budu jednake nekom broju  $x = quant$ .

**Kumulativne funkcije distribucije vjerojatnosti (Cumulative Distribution Functions)**

- **Transform; Compute; Function group (izabrati: CDF & Noncentral CDF); izabrati: Cdf.Binom u Numeric Expression:**

**CDF.BINOM(quant,n,prob)**

Rezultat ove funkcije je kumulativ vjerojatnosti da broj uspjeha u  $n$  pokušaja, uz vjerojatnost realizacije  $p = prob$ , bude manji ili jednak nekom broju  $x = quant$ .



**Primjer 2.5.**

Vjerojatnost da je zaposlenik jednog poduzeća muškog spola je:

$$p = \frac{m(M)}{n} = 0,225806451 .$$

- a) Kolika je vjerojatnost da između 7 slučajno odabranih zaposlenika budu 3 zaposlenika muškog spola?
- b) Kolika je vjerojatnost da između 7 slučajno odabranih zaposlenika budu najviše 2 zaposlenika muškog spola?
- c) Kolika je vjerojatnost da između 7 slučajno odabranih zaposlenika budu najviše 4 ili najmanje 3 zaposlenika muškog spola?
- d) Kolika je vjerojatnost da između 7 slučajno odabranih zaposlenika bude barem 1 zaposlenik muškog spola?
- e) Kolika je vjerojatnost da između 7 slučajno odabranih zaposlenika budu svi zaposlenici muškog spola?



### Rješenje 2.5.

Na glavnom izborniku potrebno je odabrati ikonu **Transform**, a na njezinu padajućem izborniku **Compute**. Nakon toga otvara se prozor **Compute Variable**, koji je prikazan na slici 2.1. Da bi se aktivirala naredba za izračun vjerojatnosti po Binomnom zakonu, potrebno je u izborniku **Function group**: odabrati **PDF & Noncentral PDF**. Nakon toga se u izborniku **Function and Special Variables**: bira **Pdf.Binom** i klikom dva puta na tu funkciju ona se pojavljuje u prozoru **Numeric Expression**: na sljedeći način: **PDF.BINOM(?,?,?)**.

U navedenoj funkciji prvi upitnik predstavlja **quant**, tj.  $x$  vrijednost koju poprima slučajna varijabla u ovom primjeru  $X1$ . Drugi upitnik funkcije predstavlja **n**, odnosno veličinu odabranog uzorka. Treći upitnik je **prob**, tj.  $p$  vjerojatnost realizacije slučajnog događaja. U ovom primjeru je zadano:

$$quant = X1 ; n = 7 ; prob = 0,225806451 ,$$

pa vrijedi da je:

$$PDF.BINOM(X1,7,0.225806451),$$

što je i prikazano u prozoru **Numeric Expression** na slici 2.1.

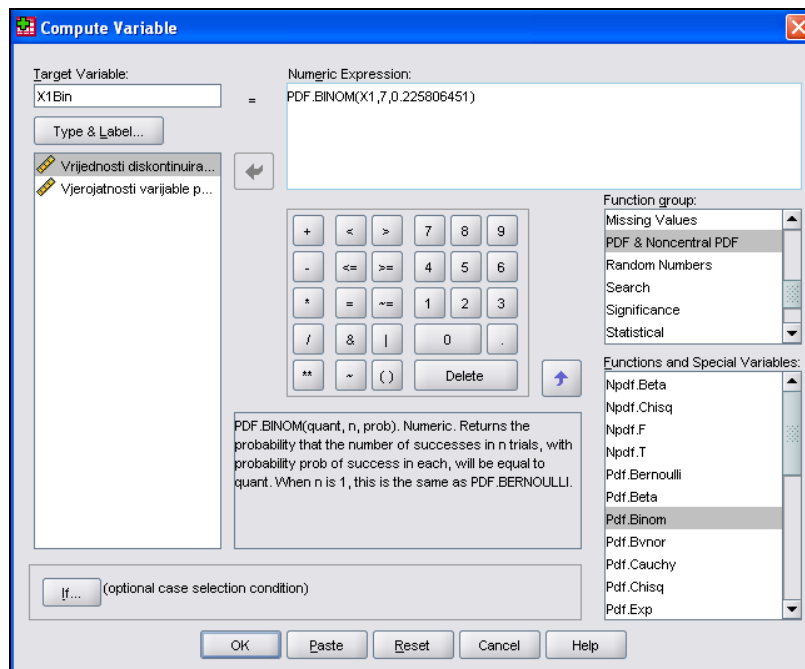
U **Target Variable**: potrebno je imenovati novu varijablu izračunatih vjerojatnosti po Binomnom zakonu i u ovom primjeru nazvana je **X1Bin**.

Konačno, da bi se u **dodatnom stupcu** programa SPSS dobili željeni podaci **vjerojatnosti slučajne varijable  $X1$  po Binomnoj distribuciji**, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.5.

Prema rezultatima iz tablice 2.5 mogu se očitati i/ili izračunati tražene vjerojatnosti.

Slika 2.1.

**Prozor u SPSS-u za izračunavanje vjerojatnosti diskontinuirane slučajne varijable po Binomnom zakonu**



*Izvor: Simulirani podaci.*

Tablica 2.5.

**Podaci u SPSS-u vrijednosti i vjerojatnosti diskontinuirane slučajne varijable po Binomnom zakonu**

X1	X1Bin
0,00	0,166704
1,00	0,340355
2,00	0,297810
3,00	0,144769
4,00	0,042224
5,00	0,007389
6,00	0,000718
7,00	0,000030

*Izvor: Simulirani podaci.*

a) Vjerojatnost da između 7 slučajno odabranih zaposlenika budu 3 zaposlenika muškog spola je:

$$p(X_1 = 3) = 0,144769.$$

b) Vjerojatnost da između 7 slučajno odabranih zaposlenika budu najviše 2 zaposlenika muškog spola je:

$$\begin{aligned} p(X_1 \leq 2) &= p(X_1 = 0) + p(X_1 = 1) + p(X_1 = 2) = \\ &= 0,166704 + 0,340355 + 0,297810 = 0,805. \end{aligned}$$

c) Vjerojatnost da između 7 slučajno odabranih zaposlenika budu najviše 4 ili najmanje 3 zaposlenika muškog spola je:

$$p(X_1 \leq 4 \text{ ili } X_1 \geq 3) = 1.$$

Ovom vjerojatnošću su obuhvaćene sve vrijednosti slučajne varijable  $X_1$ , a zbroj svih vjerojatnosti u jednoj distribuciji je 1.

d) Vjerojatnost da između 7 slučajno odabranih zaposlenika bude barem 1 zaposlenik muškog spola je:

$$\begin{aligned} p(X_1 \geq 1) &= p(X_1 = 1) + p(X_1 = 2) + \dots + p(X_1 = 7) = \\ &= 1 - p(X_1 = 0) = 1 - 0,166704 = 0,8333. \end{aligned}$$

e) Vjerojatnost da između 7 slučajno odabranih zaposlenika budu svi zaposlenici muškog spola je:

$$p(X_1 = 7) = 0,00003.$$

### 2.2.2 Poissonova distribucija

Ako je vjerojatnost da se dogodi neki događaj poznata, unaprijed utvrđena, konstantna i jako mala tijekom cijelog istraživanja (iznosi  $p$ ), te broj pokusa teži u beskonačnost ( $n \rightarrow \infty$ ) umjesto binomne distribucije, upotrebljava se **Poissonova distribucija**.

U praksi vrijedi da ako je:

$$n \geq 50 \text{ i } p \leq 0,10$$

koristi se Poissonova distribucija. Može se reći da je ova distribucija definirana za **rijetke događaje**, odnosno one događaje koji imaju veliki uzorak i malu vjerojatnost.

**Zakon vjerojatnosti** po Poissonovoj distribuciji je:

$$P(X = x) = \frac{(np)^x \cdot e^{-np}}{x!} = \frac{(\mu)^x \cdot e^{-\mu}}{x!}, \quad x = 0, 1, 2, \dots, \infty, \quad (2.18)$$

jer je:  $\mu = n \cdot p$ ;

gdje je:

$n$  - broj pokusa,

$p$  - vjerojatnost realizacije slučajnog događaja,

$x$  - broj povoljnih ishoda u  $n$  pokusa.

**Očekivanje** slučajne varijable  $X$  kod Poissonove distribucije:

$$E(X) = n \cdot p = \mu. \quad (2.19)$$

**Varijanca** slučajne varijable  $X$  kod Poissonove distribucije:

$$\sigma^2 = n \cdot p = \mu. \quad (2.20)$$

**Koeficijent asimetrije** slučajne varijable  $X$  kod Poissonove distribucije:

$$\alpha_3 = \frac{1}{\sqrt{\mu}}. \quad (2.21)$$

**Koeficijent zaobljenosti** slučajne varijable  $X$  kod Poissonove distribucije:

$$\alpha_4 = 3 + \frac{1}{\mu}. \quad (2.22)$$

Općenito se kaže da se **slučajna varijabla  $X$  ravna po Poissonovoj distribuciji**, koja je određena parametrom  $\mu$ :

$$X \sim P(\mu). \quad (2.23)$$

U programskom paketu **SPSS** postoje **funkcije gustoće vjerojatnosti** i **kumulativne funkcije distribucije vjerojatnosti Poissonove distribucije**.

**Funkcije gustoće vjerojatnosti (Probability Density Functions)**

- **Transform; Compute; Function group (izabrati: PDF & Noncentral PDF); izabrati: Pdf.Poisson u Numeric Expression:**

### PDF.POISSON(quant,mean)

Rezultat ove funkcije su vjerojatnosti da broj uspjeha u  $n$  pokušaja, uz vjerojatnost realizacije  $p = prob$ , budu jednake nekom broju  $x = quant$ . Vrijedi da je očekivanje  $mean = \mu = n \cdot p$ .

### Kumulativne funkcije distribucije vjerojatnosti (Cumulative Distribution Functions)

- **Transform; Compute; Function group (izabрати: CDF & Noncentral CDF); izabрати: Cdf.Poisson u Numeric Expression:**

### CDF.POISSON(quant,mean)

Rezultat ove funkcije je kumulativ vjerojatnosti da broj uspjeha u  $n$  pokušaja, uz vjerojatnost realizacije  $p = prob$ , bude manji ili jednak nekom broju  $x = quant$ . Vrijedi da je očekivanje  $mean = \mu = n \cdot p$ .



#### Primjer 2.6.

Vjerojatnost da se među zaposlenicima u poduzeću "P" konzumira droga je:  $p = 0,056$ .

- Kolika je vjerojatnost da između 60 slučajno odabranih zaposlenika budu 2 zaposlenika koja konzumiraju drogu?
- Kolika je vjerojatnost da između 60 slučajno odabranih zaposlenika barem 1 zaposlenik konzumira drogu?
- Kolika je vjerojatnost da između 60 slučajno odabranih zaposlenika najviše 10 zaposlenika konzumira drogu?



#### Rješenje 2.6.

Zadana vjerojatnost slučajnog događaja da se među zaposlenicima u poduzeću "P" konzumira droga je manja od 0,10, a uzorak je u svim slučajevima veći od 50, pa se u ovom primjeru pretpostavlja da se slučajna varijabla  $X_2$  ravna po Poissonovu zakonu.

Na glavnom izborniku potrebno je odabrati ikonu **Transform**, a na njezinu padajućem izborniku **Compute**. Nakon toga otvara se prozor **Compute Variable**. Da bi se aktivirala naredba za izračun vjerojatnosti po Poissonovu zakonu, potrebno je u izborniku **Function group**: odabrati **PDF & Noncentral PDF**. Nakon toga se u



izborniku **Function and Special Variables**: bira **Pdf.Poisson** i klikom dva puta na tu funkciju ona se pojavljuje u prozoru **Numeric Expression**: na sljedeći način: **PDF.POISSON(?,?)**.

U navedenoj funkciji prvi upitnik predstavlja **quant**, tj. **x** vrijednost koju poprima slučajna varijabla u ovom primjeru X2. Drugi upitnik funkcije predstavlja **mean**, odnosno očekivanje slučajne varijable koja se po Poissonovu zakonu računa na sljedeći način:

$$n = 60 ; prob = 0,056 : \Rightarrow \mu = n \cdot p = 60 \cdot 0,056 = 3,36 .$$

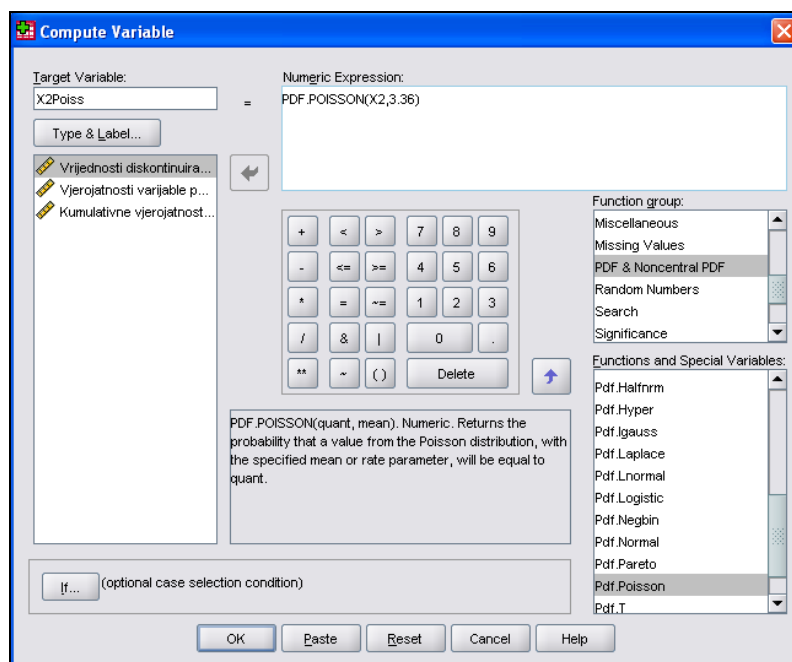
U ovom primjeru vrijedi da je:  $quant = X2 ; \mu = 3,36$ ,

pa se može upisati: **PDF.POISSON(X2,3.36)**,

što je i prikazano u prozoru **Numeric Expression** na slici 2.2.

**Slika 2.2.**

**Prozor u SPSS-u za izračunavanje vjerojatnosti diskontinuirane slučajne varijable po Poissonovu zakonu**



*Izvor: Simulirani podaci.*

U **Target Variable:** potrebno je imenovati novu varijablu izračunatih vjerojatnosti po Poissonovu zakonu i u ovom primjeru nazvana je **X2Poiss**.

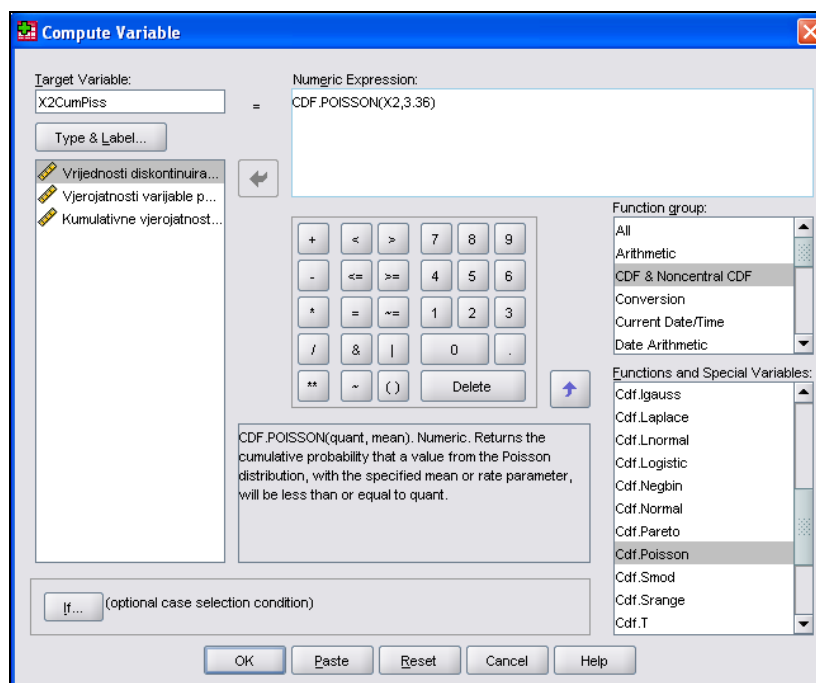
Konačno, da bi se u **dodatnom stupcu** programa SPSS dobili željeni podaci **vjerojatnosti slučajne varijable X2 po Poissonovoj distribuciji**, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.6.

Da bi se dobile kumulativne vjerojatnosti, potrebno je ponoviti sličan postupak.

Na glavnom izborniku potrebno je odabrati ikonu **Transform**, a na njezinu padajućem izborniku **Compute**. Nakon toga otvara se prozor **Compute Variable**. Da bi se aktivirala naredba za izračun kumulativnih vjerojatnosti po Poissonovu zakonu, potrebno je u izborniku **Function group:** odabrati **CDF & Noncentral CDF**. Nakon toga se u izborniku **Function and Special Variables:** bira **Cdf.Poisson** i klikom dva puta na tu funkciju ona se pojavljuje u prozoru **Numeric Expression:** na sljedeći način: **CDF.POISSON(?,?)**.

### Slika 2.3.

**Prozor u SPSS-u za izračunavanje kumulativnih vjerojatnosti diskontinuirane slučajne varijable po Poissonovu zakonu**



Izvor: Simulirani podaci.

U navedenoj funkciji prvi upitnik opet predstavlja *quant*, tj. *x* vrijednost koju poprima slučajna varijabla u ovom primjeru *X2*. Drugi upitnik funkcije predstavlja *mean*, odnosno očekivanje slučajne varijable:

$$quant = X2 ; \mu = 3,36 ,$$

pa se može upisati: **CDF.POISSON(X2,3.36)**,

što je i prikazano u dijelu prozora *Numeric Expression* na slici 2.3.

U **Target Variable**: potrebno je imenovati novu varijablu izračunatih vjerojatnosti po Poissonovu zakonu i u ovom primjeru nazvana je *X2CumPoiss*.

Konačno, da bi se u *dodatnom stupcu* programa *SPSS* dobili željeni podaci *kumulativnih vjerojatnosti slučajne varijable X2 po Poissonovoj distribuciji*, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.6.

**Tablica 2.6.**

**Podaci u SPSS-u vrijednosti, vjerojatnosti i kumulativne vjerojatnosti diskontinuirane slučajne varijable po Poissonovu zakonu**

X2	X2Poiss	X2CumPoiss
0	0,034735	0,034735
1	0,116710	0,151446
2	0,196074	0,347519
3	0,219602	0,567122
4	0,184466	0,751588
5	0,123961	0,875549
6	0,069418	0,944967
7	0,033321	0,978288
8	0,013995	0,992283
9	0,005225	0,997507
10	0,001755	0,999263
11	0,000536	0,999799
12	0,000150	0,999949
13	0,000039	0,999988

*Izvor: Simulirani podaci.*

Prema rezultatima iz tablice 2.6 (prikazano je samo 13 vrijednosti od ukupno 60 vrijednosti zadane diskontinuirane slučajne varijable *X2*) mogu se očitati i/ili izračunati tražene vjerojatnosti.

a) Vjerojatnost da između 60 slučajno odabranih zaposlenika budu 2 zaposlenika koja konzumiraju drogu je:

$$p(X1 = 2) = 0,196074.$$

b) Vjerojatnost da između 60 slučajno odabranih zaposlenika barem 1 zaposlenik konzumira drogu je:

$$\begin{aligned} p(X1 \geq 1) &= p(X1 = 1) + p(X1 = 2) + \dots + p(X1 = 60) = \\ &= 1 - p(X1 = 0) = 1 - 0,034735 = 0,9653. \end{aligned}$$

Ovdje je iskorišteno svojstvo da je zbroj svih vjerojatnosti u distribuciji je 1.

c) Kolika je vjerojatnost da između 60 slučajno odabranih zaposlenika najviše 10 zaposlenika konzumira drogu?

$$\begin{aligned} p(X1 \leq 10) &= p(X1 = 0) + p(X1 = 1) + \dots + p(X1 = 10) = \\ &= 0,999263. \end{aligned}$$

Ovdje je tražena vjerojatnost jednostavno očitana iz rezultata stupca kumulativnih vjerojatnosti Poissonove distribucije. Jednak rezultat bi se dobio postupnim zbrajanjem vjerojatnosti iz stupca vjerojatnosti Poissonove distribucije.

### 2.2.3 Dvodimenzionalna diskontinuirana distribucija

Diskontinuirana slučajna varijabla  $X$  može poprimiti vrijednosti  $x_1, x_2, x_3, \dots, x_n$ , a diskontinuirana slučajna varijabla  $Y$  može istovremeno poprimiti vrijednosti  $y_1, y_2, y_3, \dots, y_m$ .

Vjerojatnost da slučajna varijabla  $X$  poprimi vrijednost  $x_i$ , a istovremeno slučajna varijabla  $Y$  poprimi vrijednost  $y_j$  je:

$$P(X = x_i, Y = y_j) = P(x_i, y_j) \quad (2.24)$$

Za svaku distribuciju vjerojatnosti mora biti ispunjen uvjet normativnosti:

$$\sum_{i=1}^n \sum_{j=1}^m P(x_i, y_j) = 1 \quad (2.25)$$

Skup svih uređenih parova  $\{(x_i, y_j), P(x_i, y_j)\}$  je dvodimenzionalna distribucija slučajnih varijabli  $(X, Y)$ .

### 2.2.4 Marginalne distribucije diskontinuirane slučajne varijable

Marginalna distribucija slučajne varijable  $X$  je skup uređenih parova odgovarajućih vrijednosti varijable  $X$  i njima pripadajućih vjerojatnosti da slučajna varijabla  $X$  poprimi neku vrijednost  $x_i$ , bez obzira na to koju će vrijednost poprimiti slučajna varijabla  $Y$ .

Marginalna distribucija slučajne varijable  $Y$  je skup uređenih parova odgovarajućih vrijednosti varijable  $Y$  i njima pripadajućih vjerojatnosti da slučajna varijabla  $Y$  poprimi neku vrijednost  $y_j$ , bez obzira na to koju će vrijednost poprimiti slučajna varijabla  $X$ .

Marginalna vjerojatnost slučajne varijable  $X$  je:

$$P(x_i \cdot) = \sum_{j=1}^m P(x_i, y_j), \quad (2.26)$$

i obično se u dvostrukoj statističkoj tablici nalazi u zbirnom stupcu (ako se slučajna varijabla  $X$  mijenja po redcima).

Marginalna vjerojatnost slučajne varijable  $Y$  je:

$$P(\cdot y_j) = \sum_{i=1}^n P(x_i, y_j), \quad (2.27)$$

i obično se u dvostrukoj statističkoj tablici nalazi u zbirnom retku (ako se slučajna varijabla  $Y$  mijenja po stupcima).



#### Primjer 2.7.

a) Nakon prikupljenih podataka pomoću anketnog upitnika na uzorku studentske populacije zadatak je izračunati dvodimenzionalnu distribuciju vjerojatnosti: za

- varijablu (obilježje) *grupirani džeparac (v1)* koja se mijenja po redcima (row) i za

- varijablu (obilježje) *grupirana ocjena u srednjoj školi (v2)* koja se mijenja po stupcima (column).

b) Komentirati izračunate vjerojatnosti dvodimenzionalne distribucije!

c) Komentirati izračunate vjerojatnosti marginalne distribucije!

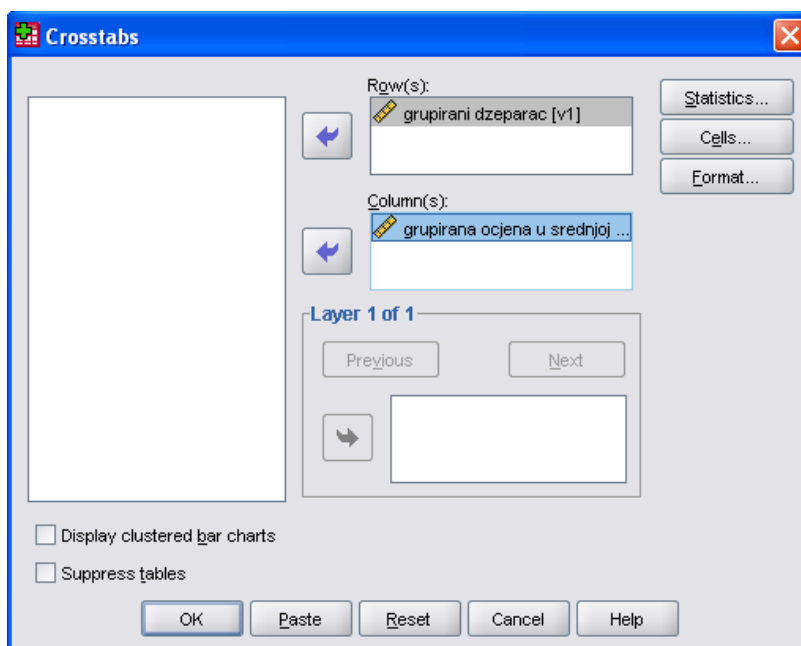


### Rješenje 2.7.

a) Na glavnom izborniku potrebno je odabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Crosstabs**. Nakon toga otvara se prozor **Crosstabs**.

### Slika 2.4.

**Prozor Crosstabs za izračunavanje dvodimenzionalnih vrijednosti zadanih slučajnih varijabli**



*Izvor: Simulirani podaci.*

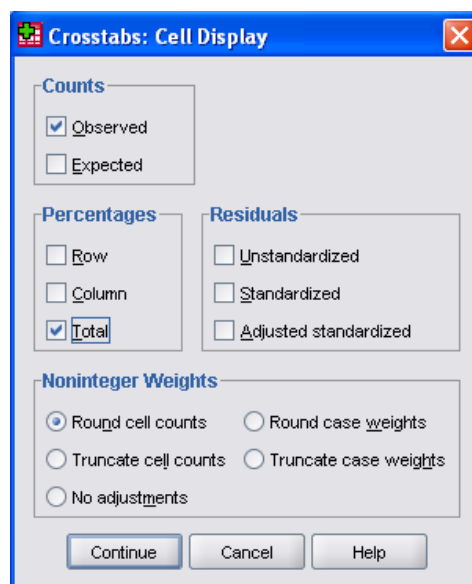
Odgovarajućim strelicama prema zahtjevu zadatka potrebno je varijablu (obilježje) *grupirani džeparac (v1)* prebaciti u prozor za retke Row(s) i varijablu (obilježje) *grupirana ocjena u srednjoj školi (v2)* prebaciti u prozor za stupce Column(s) kako je prikazano na slici 2.4.

Klikom na ikonu **Cells...** otvara se prozor **Crosstabs: Cell Display** gdje je uz postojeće originalne frekvencije (**Counts**) potrebno aktivirati **Percentages; Total** kako je prikazano na slici 2.5.

Da bi se u **Outputu** programa **SPSS** dobili traženi podaci potrebno je kliknuti na **Continue** i **OK**.

**Slika 2.5.**

**Prozor Crosstabs: Cell Display za odabir apsolutnih i/ili relativnih frekvencija dvodimenzionalne distribucije**



*Izvor: Simulirani podaci.*

Rezultati su prikazani u tablici 2.7.

b) Izračunate vjerojatnosti dvodimenzionalne distribucije u tablici 2.7 izražene su u postotcima.

Vjerojatnost u prvom retku i prvom stupcu numeričkog dijela tablice iznosi 3,5%, što znači da vjerojatnost da ispitanik ima džeparac do 200 kn i ocjenu iz srednje škole 3 iznosi 0,035.

Vjerojatnost u prvom retku i drugom stupcu numeričkog dijela tablice iznosi 22,1%, što znači da vjerojatnost da ispitanik ima džeparac do 200 kn i ocjenu iz srednje škole 4 iznosi 0,221.

**Tablica 2.7.****Dvodimenzionalna distribucija varijabli džeparca i prosječne ocjene u srednjoj školi**

grupirani džeparac * grupirana ocjena u srednjoj školi Crosstabulation						
			grupirana ocjena u srednjoj školi			
			3,00	4,00	5,00	Total
grupirani džeparac	200,00	Count	6	38	29	73
		% of Total	3,5%	22,1%	16,9%	42,4%
	400,00	Count	14	20	18	52
		% of Total	8,1%	11,6%	10,5%	30,2%
	600,00	Count	5	14	10	29
		% of Total	2,9%	8,1%	5,8%	16,9%
	800,00	Count	1	3	2	6
		% of Total	,6%	1,7%	1,2%	3,5%
	1000,00	Count	0	2	5	7
		% of Total	,0%	1,2%	2,9%	4,1%
	3500,00	Count	0	3	2	5
		% of Total	,0%	1,7%	1,2%	2,9%
	Total	Count	26	80	66	172
		% of Total	15,1%	46,5%	38,4%	100,0%

*Izvor: Simulirani podaci.*

Vjerojatnost u četvrtom retku i drugom stupcu numeričkog dijela tablice iznosi 1,7%, što znači da vjerojatnost da ispitanik ima džeparac između 600 kn i 800 kn i ocjenu iz srednje škole 4 iznosi 0,071.

Posljednja vjerojatnost numeričkog dijela tablice, tj. vjerojatnost u šestom retku i trećem stupcu iznosi 2,9%, što znači da vjerojatnost da ispitanik ima džeparac između 1000 kn i 3500 kn i ocjenu iz srednje škole 5 iznosi 0,012.

c) Izračunate vjerojatnosti marginalne distribucije nalaze se u zbirnom retku i zbirnom stupcu tablice 2.7.

Na primjer, druga marginalna vjerojatnost u zbirnom stupcu iznosi 30,2%, tj. vjerojatnost da ispitanik ima džeparac između 200 kn i 400 kn bez obzira na prosječnu ocjenu u srednjoj školi iznosi 0,302.

Isto tako treća marginalna vjerojatnost u zbirnom retku iznosi 38,4%, tj. vjerojatnost da ispitanik ima prosječnu ocjenu u srednjoj školi 5 bez obzira na visinu džeparca iznosi 0,384.

Zbroj svih vjerojatnosti u izračunatoj dvodimenzionalnoj distribuciji vjerojatnosti je 1, odnosno 100%.



**Primjer 2.8.**

Na temelju provedene anonimne ankete utvrđena je vjerojatnost da je zaposlenik u jednoj korporaciji u potpunosti zadovoljan svojim statusom i plaćom:  $p = 0,038$ .

- Kolika je vjerojatnost da između 70 slučajno odabranih zaposlenika navedene korporacije ( $X_1$ ) budu 2 zaposlena koja su u potpunosti zadovoljna svojim statusom i plaćom?
- Kolika je vjerojatnost da između 70 slučajno odabranih zaposlenika navedene korporacije ( $X_1$ ) bude najviše 15 zaposlenih koja su u potpunosti zadovoljna svojim statusom i plaćom?

**Rješenje 2.8.**

Zadana vjerojatnost slučajnog događaja da je zaposlenik u jednoj korporaciji u potpunosti zadovoljan svojim statusom i plaćom je manja od 0,10, a uzorak je veći od 50, pa se u ovom primjeru pretpostavlja da se slučajna varijabla  $X_1$  ravna po Poissonovom zakonu.

Na glavnom izborniku potrebno je odabrati ikonu **Transform**, a na njezinu padajućem izborniku **Compute**. Nakon toga otvara se prozor **Compute Variable**. Da bi se aktivirala naredba za izračun vjerojatnosti po Poissonovu zakonu, potrebno je u izborniku **Function group**: odabrati **PDF & Noncentral PDF**. Nakon toga se u izborniku **Function and Special Variables**: bira **Pdf.Poisson** i klikom dva puta na tu funkciju ona se pojavljuje u prozoru **Numeric Expression**: na sljedeći način: **PDF.POISSON(?,?)**.

U navedenoj funkciji prvi upitnik predstavlja **quant**, tj.  $x$  vrijednost koju poprima slučajna varijabla u ovom primjeru  $X_1$ . Drugi upitnik funkcije predstavlja **mean**, odnosno očekivanje slučajne varijable koja se po Poissonovu zakonu računa na sljedeći način:

$$n = 70 ; \text{prob} = 0,038 : \Rightarrow \mu = n \cdot p = 70 \cdot 0,038 = 2,66 .$$

U ovom primjeru vrijedi da je:  $\text{quant} = X_1 ; \mu = 2,66$ ,

pa se može upisati: **PDF.POISSON( $X_1, 2.66$ )**.

U **Target Variable**: potrebno je imenovati novu varijablu izračunatih vjerojatnosti po Poissonovu zakonu i u ovom primjeru nazvana je **X1Poiss**.

Konačno, da bi se u *dodatnom stupcu* programa **SPSS** dobili željeni podaci *vjerojatnosti slučajne varijable X1 po Poissonovoj distribuciji*, potrebno je kliknuti na **OK**.

Da bi se dobile kumulativne vjerojatnosti, potrebno je ponoviti sličan postupak.

Na glavnom izborniku potrebno je odabrati ikonu **Transform**, a na njezinu padajućem izborniku **Compute**. Nakon toga otvara se prozor **Compute Variable**. Da bi se aktivirala naredba za izračun kumulativnih vjerojatnosti po Poissonovu zakonu, potrebno je u izborniku **Function group**: odabrati **CDF & Noncentral CDF**. Nakon toga se u izborniku **Function and Special Variables**: bira **Cdf.Poisson** i klikom dva puta na tu funkciju ona se pojavljuje u prozoru **Numeric Expression**: na sljedeći način: **CDF.POISSON(?,?)**.

Tablica 2.8.

Podaci vrijednosti, vjerojatnosti i kumulativne vjerojatnosti diskontinuirane slučajne varijable po Poissonovu zakonu u dokumentu SPSS-a

X1	v1	X1poiss	X1cumpoiss
0,00	3	0,0699482	0,0699482
1,00	1	0,1860623	0,2560105
2,00	3	0,2474628	0,5034733
3,00	3	0,2194170	0,7228903
4,00	2	0,1459123	0,8688027
5,00	3	0,0776254	0,9464280
6,00	3	0,0344139	0,9808419
7,00	3	0,0130773	0,9939192
8,00	3	0,0043482	0,9982674
9,00	3	0,0012851	0,9995526
10,00	1	0,0003418	0,9998944
11,00	1	0,0000827	0,9999771
12,00	1	0,0000183	0,9999954
13,00	3	0,0000037	0,9999991
14,00	1	0,0000007	0,9999998

Izvor: Simulirani podaci.

U navedenoj funkciji prvi upitnik opet predstavlja *quant*, tj.  $x$  vrijednost koju poprima slučajna varijabla u ovom primjeru X2. Drugi upitnik funkcije predstavlja *mean*, odnosno očekivanje slučajne varijable:  $quant = X1 ; \mu = 2,66$ ,

pa se može upisati: **CDF.POISSON(X1,2.66)**.

U **Target Variable**: potrebno je imenovati novu varijablu izračunatih vjerojatnosti po Poissonovu zakonu i u ovom primjeru nazvana je **X1CumPoiss**.

Konačno, da bi se u *dodatnom stupcu* programa **SPSS** dobili željeni podaci *kumulativnih vjerojatnosti slučajne varijable X1 po Poissonovoj distribuciji*, potrebno je kliknuti na **OK**. Rezultat je prikazan u tablici 2.8.

Prema rezultatima iz tablice 2.8 (prikazano je samo 15 vrijednosti od ukupno 70 vrijednosti zadane diskontinuirane slučajne varijable X1) mogu se očitati i/ili izračunati tražene vjerojatnosti.

a) Vjerojatnost da između 70 slučajno odabranih zaposlenika navedene korporacije (X1) budu 2 zaposlena koja su u potpunosti zadovoljna svojim statusom i plaćom je:

$$p(X1 = 2) = 0,186.$$

b) Vjerojatnost da između 70 slučajno odabranih zaposlenika navedene korporacije (X3) bude najviše 15 zaposlenih koja su u potpunosti zadovoljna svojim statusom i plaćom je:

$$p(X1 \leq 15) = p(X1 = 0) + p(X1 = 1) + \dots + p(X1 = 15) = 0,999999 \approx 1.$$

Izračunata vjerojatnost je približno jednaka 1, jer je zbroj svih 71 vjerojatnosti 1, ali zbog manjeg broja decimalnih mjesta ovaj je rezultat izjednačen s 1.

## 2.3 Kontinuirana slučajna varijabla

Kontinuirana varijabla X je takva varijabla koja može poprimiti neprebrojivo beskonačno mnogo vrijednosti. Zato se za kontinuiranu slučajnu varijablu ne računa vjerojatnost u određenoj točki, nego nad određenim intervalom.

**Funkcija vjerojatnosti** ili funkcija gustoće vjerojatnosti kontinuirane slučajne varijable X ima svojstva:

- $f(x) \geq 0, \forall x$ ,
- $\int_{-\infty}^{+\infty} f(x)dx = 1$ , (površina ispod krivulje funkcije vjerojatnosti je 1),
- $P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)dx, [x_2 > x_1]$ , (**vjerojatnost** da slučajna varijabla X poprimi vrijednost veću od  $x_1$  i manju ili jednaku  $x_2$ , **jednaka je** određenom integralu funkcije vjerojatnosti  $f(x)$  na tom intervalu, odnosno **površini**

koju krivulja funkcije vjerojatnosti na tom intervalu zatvara s pozitivnim smjerom osi  $x$ ).

**Funkcija distribucije slučajne varijable  $X$**  je funkcija koja daje vjerojatnost da će slučajna varijabla  $X$  poprimiti vrijednost jednaku ili manju od nekog realnog broja  $x$ :

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x)dx \quad (2.27)$$

**Parametri distribucije:**

- **Očekivanje slučajne varijable** (ako integral konvergira):

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \mu. \quad (2.28)$$

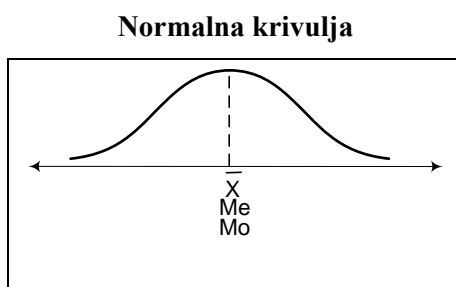
- **Varijanca slučajne varijable:**

$$V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x)dx. \quad (2.29)$$

### 2.3.1 Normalna distribucija

Normalna distribucija je simetrična ( $\alpha_3 = 0$ ), unimodalna i ima oblik zvona. Normalno je zaobljena ( $\alpha_4 = 3$ ):

**Slika 2.5.**



*Izvor: Konstrukcija autora.*

Prema slici 2.5 može se uočiti da normalna krivulja u istoj točki točno na sredini ima aritmetičku sredinu, medijan i mod.

**Funkcija vjerojatnosti** ili funkcija gustoće vjerojatnosti kod normalne distribucije je:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in (-\infty, +\infty). \quad (2.30)$$

Općenito se kaže da se **kontinuirana slučajna varijabla X ravna po normalnoj distribuciji**, koja je određena parametrima  $\mu$  (očekivanje) i  $\sigma^2$  (varijanca):

$$X \sim N(\mu, \sigma^2). \quad (2.31)$$

Za standardiziranu varijablu Z:

$$Z = \frac{x - \mu}{\sigma}, \quad (2.32)$$

funkcija vjerojatnosti je:

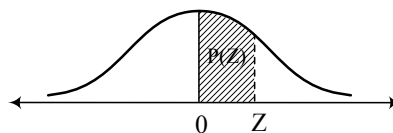
$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}. \quad (2.33)$$

Ova normalna distribucija se naziva **standardizirana ili jedinična normalna distribucija** s očekivanjem 0 i jediničnom varijancom, pa vrijedi da je:

$$X \sim N(0,1). \quad (2.34)$$

**Slika 2.6.**

#### Površina ispod normalne krivulje



*Izvor: Konstrukcija autora.*

Jedinična normalna distribucija ima uvijek iste vrijednosti parametara, pa se vrijednosti za intervale vrijednosti varijable Z, od 0 do z kako je prikazano na slici 2.6, (u statističkoj literaturi ima i drugačije koncipiranih tablica), mogu prikazati u jedinstvenoj **tablici površina (vjerojatnosti) ispod normalne krivulje**.

Na primjer:

$$P(x_1 < X \leq x_2) = P_2\left(\frac{x_2 - \mu}{\sigma}\right) - P_1\left(\frac{x_1 - \mu}{\sigma}\right) = P_2(Z_2) - P_1(Z_1). \quad (2.35)$$

### 2.3.2 Studentova distribucija

Studentova distribucija se još naziva t-distribucija. Ima svoju primjenu u statistici kod procjene parametara osnovnog skupa i kod testiranja hipoteza na osnovu uzorka.

Varijabla "t" je definirana na području:  $-\infty, +\infty$ . Studentova distribucija je simetrična s obzirom na  $t = 0$ .

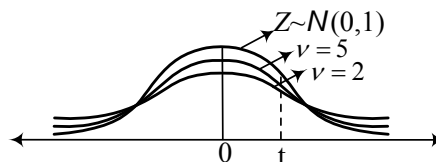
Ako je  $Z \sim N(0,1)$  i  $\chi^2$  varijabla tzv. hi-kvadrat ili gama-distribucije s  $\nu$  (ili *df* - degree of freedom) stupnjeva slobode i ako su one nezavisne slučajne varijable, tada je:

$$t_\nu = \frac{z}{\sqrt{\frac{\chi^2}{\nu}}}, \quad (2.36)$$

t - varijabla Studentove distribucije sa stupnjevima slobode  $\nu$  (ili *df*).

**Slika 2.7.**

**Studentova distribucija za različite stupnjeve slobode  $\nu$**



*Izvor: Konstrukcija autora.*

Na slici se vidi da se s povećanjem stupnjeva slobode  $\nu$ , studentova distribucija približava normalnoj distribuciji. Već za  $\nu = 30$  u praksi se umjesto t - distribucije upotrebljava normalna distribucija (pogreška aproksimacije u tom je slučaju manja od 0,03).

### 2.3.3 Hi-kvadrat distribucija

Hi-kvadrat distribucija još se naziva gama-distribucija. Ako su  $X_1, X_2, \dots, X_n$  nezavisne normalne varijable koje imaju jednaka očekivanja,  $E(X_1) = E(X_2) = \dots = E(X_n) = \mu$  i jednake varijance  $V(X_1) = V(X_2) = \dots = V(X_n) = \sigma^2$ , tada je:

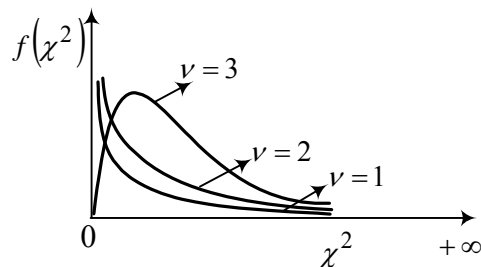
$$\chi^2 = \sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^2, \quad (2.37)$$

$\chi^2$  - gama varijabla sa stupnjem slobode  $\nu = n$ .

Varijabla " $\chi^2$ " je definirana na području:  $<0, +\infty>$ , i za različite stupnjeve slobode distribucija ima različit oblik, što se jasno vidi na slici.

**Slika 2.8.**

**Hi-kvadrat distribucija za različite stupnjeve slobode  $\nu$**



*Izvor: Konstrukcija autora.*

Slika prikazuje da za broj stupnjeva slobode  $\nu \geq 3$  krivulja gama-distribucije ima pozitivnu ili desnostranu asimetriju. Kako broj stupnjeva slobode  $\nu$  raste, to se gama-distribucija približava obliku normalne distribucije.

Varijabla  $\chi^2$ :

$$\chi^2 = \sum_{i=1}^r \left( \frac{f_i - f_{ii}}{f_{ii}} \right)^2, \quad (2.38)$$

pripada  $\chi^2$  distribuciji sa stupnjevima slobode:  $\nu = r - 1$  (gdje je  $r$  - broj nepoznatih parametara u pretpostavljenoj distribuciji), pri čemu je  $f_i$  - empirijska frekvencija, a  $f_{ii}$  - teorijska frekvencija.

Kada je broj stupnjeva slobode  $\nu > 30$ , upotrebljava se aproksimacija normalnom distribucijom:

$$\chi^2 = \frac{1}{2} \left( z + \sqrt{2 \cdot \nu - 1} \right)^2, \quad (2.39)$$

gdje je  $z$  - varijabla iz jedinične normalne distribucije.

### 2.3.4 F - distribucija

F-distribucija je određena s dva parametra  $\nu_1$  i  $\nu_2$  koji predstavljaju stupnjeve slobode. Ako su  $\chi_1^2$  i  $\chi_2^2$  dvije nezavisne Hi-kvadrat (gama) distribucije sa stupnjevima slobode  $\nu_1$  i  $\nu_2$  tada varijabla F:

$$F = \frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2}, \quad (2.40)$$

pripada F - distribuciji sa stupnjevima slobode  $\nu_1$  i  $\nu_2$ .

## 2.4 Procjena prosječne vrijednosti

Pomoću uzorka procjenjuju se određeni parametri osnovnog skupa i testiraju se hipoteze o nepoznatim parametrima osnovnog skupa.

Ako se iz osnovnog skupa veličine  $N$  izaberu svi mogući uzorci veličine  $n$ , te se za svaki uzorak izračuna neki odgovarajući parametar, distribucija tih parametara naziva se **sampling distribucija**.



**Priistranost** je **razlika** između očekivane vrijednosti nekog parametra iz sampling distribucije i toga istog parametra iz osnovnog skupa:

$$E(\hat{\Theta}) \neq \Theta, \quad (2.41)$$

pri čemu je  $E(\hat{\Theta})$  očekivana vrijednost parametra iz sampling distribucije, a  $\Theta$  je parametar iz osnovnog skupa.

Ako između očekivane vrijednosti nekog parametra iz sampling distribucije i toga istog parametra iz osnovnog skupa **ne postoji razlika**, onda se to svojstvo naziva **nepriistranost**:

$$E(\hat{\Theta}) = \Theta. \quad (2.42)$$

Standardna devijacija sampling distribucije parametra naziva se **standardna greška**.

Na temelju reprezentativnog uzorka vrši se procjena prosječne vrijednosti, tj. aritmetičke sredine osnovnog skupa.

Ako veličina uzorka teži prema beskonačno, **sampling distribucija aritmetičkih sredina** teži **normalnom obliku**. Kod malih uzoraka sampling distribucija aritmetičkih sredina ima oblik Studentove ili t-distribucije. Stoga vrijedi da ako je:

$$n > 30 \quad \Rightarrow \quad \text{koristi se normalna distribucija i}$$

$$n \leq 30 \quad \Rightarrow \quad \text{koristi se Studentova ili t-distribucija.}$$

**Standardna greška aritmetičke sredine** (standardna devijacija sampling distribucije aritmetičke sredine) je:

$$Se(\hat{X}) = \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad \text{ako je } f \leq 0,05 \text{ i} \quad (2.43)$$

$$Se(\hat{X}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \Rightarrow \quad \text{ako je } f > 0,05. \quad (2.44)$$

Ako nije poznata standardna devijacija osnovnog skupa  $\sigma$ , računa se **nepriistrana (točkasta) ocjena varijance osnovnog skupa na osnovu uzorka  $S^2$** :

$$S^2 = \hat{\sigma}^2 \cdot \left( \frac{n}{n-1} \right) \quad \Rightarrow \quad \text{ako je } n \leq 30 \text{ i} \quad (2.45)$$

$$S^2 = \hat{\sigma}^2 \quad \Rightarrow \quad \text{ako je } n > 30. \quad (2.46)$$

**Točkasta procjena aritmetičke sredine** osnovnog skupa na osnovu uzorka je:

$$\hat{\bar{X}}.$$

**Intervalna procjena aritmetičke sredine** osnovnog skupa na osnovu uzorka je:

$$\Pr\left\{\hat{\bar{X}} - Z \cdot Se(\hat{\bar{X}}) < \bar{X} < \hat{\bar{X}} + Z \cdot Se(\hat{\bar{X}})\right\} = 1 - \alpha, \quad (2.47)$$

gdje je:

$\hat{\bar{X}}$  - aritmetička sredina izračunata na osnovu uzorka,

$Z$  - ako je ( $n > 30$ ) računa se vrijednost iz tablice površina normalne distribucije (pomoću  $Z_{\frac{1-\alpha}{2}}$ ),

$t$  - ako je ( $n \leq 30$ ) umjesto ( $Z$ ) računa se vrijednost iz tablice površina studentove ili t-distribucije (pomoću  $t_{\frac{\alpha}{2}, df=n-1}$ ),

$Se(\hat{\bar{X}})$  - standardna greška aritmetičke sredine,

$1 - \alpha$  - nivo pouzdanosti procjene (ako nije određeno drugačije, najčešće se uzima da je  $1 - \alpha = 95\%$ ).

**Veličina uzorka** uz zadani nivo pouzdanosti i maksimalnu grešku određuje se na sljedeći način:

$$n' = \left[ \frac{Z \cdot \sigma}{greška} \right]^2 \Rightarrow \text{ako je } f \leq 0,05 \text{ ili} \quad (2.48)$$

$$n = \frac{n'}{1 + \frac{n'}{N}} \Rightarrow \text{ako je } f > 0,05, \quad (2.49)$$

gdje je:

$Z$  - vrijednost iz tablice površina normalne distribucije (pomoću  $Z_{\frac{1-\alpha}{2}}$ ),

$\sigma$  - standardna devijacija iz osnovnog skupa,

$greška$  - maksimalna greška.

Ako je **maksimalna greška** određena **u relativnom izrazu** umjesto  $\sigma$  (standardne devijacije) veličina uzorka se računa pomoću  $V$  (koeficijenta varijacije):

$$V = \frac{\sigma}{\bar{X}} \cdot 100. \quad (2.50)$$



### Primjer 2.9.

Na temelju provedene ankete za zadani uzorak ispitanika potrebno je:

- Izračunati prosječnu visinu ispitanika u uzorku (varijabla  $v1$ )!
- Izračunati interval procjene za prosječnu visinu ispitanika osnovnog skupa uz pouzdanost procjene od 95%!
- Izračunati interval procjene za prosječnu visinu ispitanika osnovnog skupa uz pouzdanost procjene od 99%!



### Rješenje 2.9.

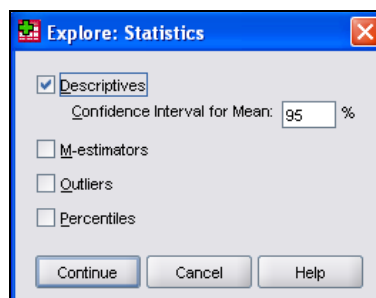
a) Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Explore**.

U otvorenom prozoru **Explore** potrebno je zadanu numeričku varijablu  $v1$  - visina u cm pomoću odgovarajuće strelice prebaciti u **Dependent List**. U prostoru **Display** potrebno je aktivirati **Statistics**.

Klikom na ikonu **Statistics** otvara se novi prozor **Explore: Statistics** koji je prikazan na slici 2.9.

Slika 2.9.

**Odabrani nivo pouzdanosti procjene za aritmetičku sredinu od 95%**



*Izvor: Simulirani podaci.*

Prema zahtjevu zadatka od (b) ovdje je izabran nivo pouzdanosti procjene od 95%. Konačno, da bi se u **Outputu** programa SPSS dobili željeni podaci o visini ispitanika, potrebno je kliknuti na **Continue** i **OK**. Rezultat je prikazan u tablici 2.9.

**Tablica 2.9.**

**Podaci o visini ispitanika**

Descriptives			Statistic	Std. Error
Visina u cm	Mean		175,77	,571
	95% Confidence Interval for Mean	Lower Bound	174,64	
		Upper Bound	176,89	
	5% Trimmed Mean		175,54	
	Median		174,00	
	Variance		76,539	
	Std. Deviation		8,749	
	Minimum		154	
	Maximum		206	
	Range		52	
	Interquartile Range		12	
	Skewness		,485	,159
	Kurtosis		,184	,316

*Izvor: Simulirani podaci.*

Prema podacima iz tablice 2.9. može se vidjeti da je prosječna visina ispitanika u uzorku 175,77cm.

b) Intervalna procjena aritmetičke sredine (**Confidence Interval for Mean**) osnovnog skupa pomoću zadanog uzorka uz nivo pouzdanosti procjene od 95% je (**Lower Bound** i **Upper Bound**):

$$\Pr\{174,64 < \bar{X} < 176,89\} = 95\% .$$

Uz traženu intervalnu procjenu u izlaznoj tablici **Descriptives** u **Outputu** programa **SPSS** dobiju se i sljedeći podaci uzorka ispitanika:

Median (medijan), Variance (varijanca), Std. Deviation (standardna devijacija), Minimum (najmanja vrijednost varijable), Maximum (najveća vrijednost varijable), Range (raspon varijacije obilježja), Interquartile Range (interkvartil), Skewness (mjera asimetrije), Kurtosis (mjera zaobljenosti), **Std. Error Mean (standardna greška aritmetičke sredine)**, Std. Error Skewness (standardna greška mjere asimetrije) i Std. Error Kurtosis (standardna greška mjere zaobljenosti).

c) Da bi se dobila intervalna procjena aritmetičke sredine visine osnovnog skupa pomoću zadanog uzorka uz nivo pouzdanosti procjene od 99%, potrebno je ponoviti sličan postupak, samo se u prozor **Explore: Statistics** upiše novi nivo pouzdanosti procjene 99%.

Rezultat je prikazan u tablici 2.10.

**Tablica 2.10.**

**Podaci o visini ispitanika**

Descriptives			
		Statistic	Std. Error
Visina u cm	Mean	175,77	,571
	99% Confidence Interval for Mean		
	Lower Bound	174,28	
	Upper Bound	177,25	
	5% Trimmed Mean	175,54	
	Median	174,00	
	Variance	76,539	
	Std. Deviation	8,749	
	Minimum	154	
	Maximum	206	
	Range	52	
	Interquartile Range	12	
	Skewness	,485	,159
	Kurtosis	,184	,316

*Izvor: Simulirani podaci.*

Intervalna procjena aritmetičke sredine osnovnog skupa pomoću zadanog uzorka uz nivo pouzdanosti procjene od 99% je:

$$\Pr\{174,28 < \bar{X} < 177,25\} = 99\%.$$

Može se zaključiti da s povećanjem nivoa pouzdanosti procjene interval procjene postaje širi, odnosno procjena postaje manje precizna.



**Primjer 2.10.**

Na temelju provedene ankete za zadani uzorak ispitanika potrebno je:

- Izračunati prosječnu težinu ispitanika u uzorku (varijabla v1)!
- Izračunati interval procjene za prosječnu težinu ispitanika osnovnog skupa uz pouzdanost procjene od 95%!
- Izračunati interval procjene za prosječnu težinu ispitanika osnovnog skupa uz pouzdanost procjene od 99%!



**Rješenje 2.10.**

- Na glavnom izborniku potrebno je izabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Descriptive Statistics** i **Explore**.

U otvorenom prozoru **Explore** potrebno je zadatu numeričku varijablu **v1** - težina u kg pomoću odgovarajuće strelice prebaciti u **Dependent List**. U prostoru **Display** potrebno je aktivirati **Statistics**.

Klikom na ikonu **Statistics** otvara se novi prozor **Explore: Statistics**.

Prema zahtjevu zadatka od (b) izabran je nivo pouzdanosti procjene od 95%. Konačno, da bi se u **Outputu** programa SPSS dobili željeni podaci o težini ispitanika, potrebno je kliknuti na **Continue** i **OK**. Rezultat je prikazan u tablici 2.11.

Prema podacima iz tablice 2.11 može se vidjeti da je prosječna težina ispitanika u uzorku 61,18 kg.

**Tablica 2.11.**

**Podaci o težini ispitanika**

Descriptives				Statistic	Std. Error
Težina u kg	Mean			67,1795	,88431
	95% Confidence Interval for Mean	Lower Bound		65,4372	
		Upper Bound		68,9218	
	5% Trimmed Mean			66,3257	
	Median			64,0000	
	Variance			182,989	
	Std. Deviation			13,52735	
	Minimum			46,00	
	Maximum			110,00	
	Range			64,00	
	Interquartile Range			19,00	
	Skewness			,840	,159
	Kurtosis			,214	,317

*Izvor: Simulirani podaci.*

b) Intervalna procjena aritmetičke sredine težine u osnovnom skupu pomoću zadanog uzorka uz nivo pouzdanosti procjene od 95% je:

$$\Pr\{65,44 < \bar{X} < 68,92\} = 95\% .$$

c) Da bi se dobila intervalna procjena aritmetičke sredine težine populacije osnovnog skupa pomoću zadanog uzorka uz nivo pouzdanosti procjene od 99%, potrebno je ponoviti sličan postupak, samo se u prozor **Explore: Statistics** upiše novi nivo pouzdanosti procjene 99%.

Rezultat je prikazan u tablici 2.12. Intervalna procjena aritmetičke sredine težine stanovnika u osnovnom skupu pomoću zadanog uzorka uz nivo pouzdanosti procjene od 99% je:

$$\Pr\{64,88 < \bar{X} < 69,48\} = 99\% .$$

Može se opet zaključiti da s povećanjem nivoa pouzdanosti procjene interval procjene postaje širi, odnosno procjena postaje manje precizna.

**Tablica 2.12.**

**Podaci o težini ispitanika**

Descriptives				Statistic	Std. Error
Tezina u kg	Mean			67,1795	,88431
	99% Confidence Interval for Mean	Lower Bound		64,8829	
		Upper Bound		69,4761	
	5% Trimmed Mean			66,3257	
	Median			64,0000	
	Variance			182,989	
	Std. Deviation			13,52735	
	Minimum			46,00	
	Maximum			110,00	
	Range			64,00	
	Interquartile Range			19,00	
	Skewness			,840	,159
	Kurtosis			,214	,317

*Izvor: Simulirani podaci.*





### 3 TESTIRANJE HIPOTEZA SA ZAVISNIM I NEZAVISNIM UZORCIMA

#### 3.1 Znanstvene i statističke hipoteze

Istraživanja koja se provode na cijelom osnovnom skupu često su skupa i zahtijevaju mnogo vremena, pa se u praksi često, koristeći metode i tehnike inferencijalne statistike, na temelju podataka iz uzorka donose zaključci vezani za osnovni skup. Pri takvom zaključivanju postavljaju se različite pretpostavke ili hipoteze. Potrebno je razlikovati **znanstvene hipoteze** i **statističke hipoteze**.

**Znanstvene hipoteze** predstavljaju nagađanje, naslućivanje i pretpostavke koje motiviraju istraživača. Iz znanstvene hipoteze, tj. hipoteze istraživača (koja je najčešće afirmativna) izvodi se statistička hipoteza.

**Statističke hipoteze** postavljaju se na način da mogu biti vrednovane statističko-analitičkim postupcima. One su u stvari matematički izraz koji predstavlja polaznu osnovu na kojoj se temelji kalkulacija statističkog testa.

**Testiranje hipoteza** je statistički postupak kojim se određuje je li i koliko pouzdano raspoloživi podaci iz reprezentativnog uzorka podupiru pretpostavljenu pretpostavku.

Pri testiranju hipoteza **potrebno je**:

- postaviti *nultu ili početnu hipotezu* ( $H_0$ ) i *alternativnu hipotezu* ( $H_1$ )
- izabrati *razinu značajnosti ili signifikantnosti* ( $\alpha$ )
- prikupiti primjerene podatke na *reprezentativnom uzorku*
- *izračunati vrijednost rezultata statističkog testa* (*empirijska vrijednost testa* iz uzorka) specifičnog za nultu hipotezu ( $H_0$ )
- *usporediti empirijsku vrijednost testa* s vrijednosti iz poznate distribucije vjerojatnosti (*s tabličnom vrijednosti testa*) specifičnog za nultu hipotezu ( $H_0$ )
- *interpretirati rezultat statističkog testa u terminima vjerojatnosti (signifikantnosti).*

**Nulta hipoteza**,  $H_0$  (eng. null hypothesis) pretpostavka je o izostanku efekta, tj. da ne postoji razlika među uzorcima u promatranoj populaciji. Ta početna hipoteza je u stvari ona pretpostavka koja se testira, tj. hipoteza da nema razlike (eng. hypothesis of no difference). Postavlja se najčešće u svrhu odbacivanja.

**Alternativna hipoteza**,  $H_1$  (eng. alternative hypothesis) vrijedi ako nulta hipoteza nije istinita. Ona se najčešće direktno odnosi na teorijsku pretpostavku koja se želi istražiti, tj. može reći da je alternativna hipoteza ustvari hipoteza istraživača.

Kada se ne može unaprijed sa sigurnošću odrediti smjer neke razlike, a ona postoji, primjenjuje se **dvosmjerni test** (eng. two-tailed test).

**Jednosmjerni test** (eng. one-tailed test) primjenjuje se kada je smjer razlike specificiran u alternativnoj hipotezi ( $H_1$ ).

S obzirom da se zaključivanje provodi na temelju informacija o uzorku, moguće je pogriješiti i donijeti krivi zaključak.

Veličina signifikantnosti ili značajnosti testa  $\alpha$  je u literaturi poznata kao **Greška tipa I**, odnosno kao vjerojatnost da se odbaci nulta hipoteza premda je ona istinita. U praktičnim istraživanjima se najčešće uzima da je  $\alpha = 5\%$ .

**Greška tipa II** ( $\beta$ ) predstavlja vjerojatnost da se prihvati nulta hipoteza premda ona nije istinita. Ako je nulta hipoteza istinita Greška tipa II postaje  $1 - \alpha$ . **Snaga testa** ( $1 - \beta$ ) je vjerojatnost da se ne prihvati lažna nulta hipoteza. Navedene **vjerojatnosti** prikazane su u tablici 3.1.

**Tablica 3.1.**

**Vjerojatnosti prihvatanja/odbacivanja lažne/istinite  $H_0$  hipoteze**

	$H_0$ <b>prihvaćena</b>	$H_0$ <b>odbaćena</b>
$H_0$ <b>istinita</b>	$1 - \alpha$	$\alpha$
$H_0$ <b>lažna</b>	$\beta$	$1 - \beta$

*Izvor: Konstrukcija autora prema teorijskim postavkama.*

### 3.2 Testiranje hipoteze o prosječnoj vrijednosti jednog osnovnog skupa

Postavljaju se **hipoteze za dvosmjerno testiranje** da je prosječna vrijednost tj. aritmetička sredina jednog osnovnog skupa  $\bar{X}$  jednaka nekoj pretpostavljenoj vrijednosti  $\bar{X}_0$ :

$$H_0 : \dots\dots\dots \bar{X} = \bar{X}_0$$

$$H_1 : \dots\dots\dots \bar{X} \neq \bar{X}_0$$

**Interval prihvatanja hipoteze**  $H_0$  glasi:

$$\bar{X}_0 \pm Z \cdot Se(\bar{X}), \quad (3.1)$$

gdje je:

$\bar{X}_0$  - neka pretpostavljena aritmetička sredina

$Z$  - vrijednost za  $Z$  normalne distribucije, ako je ( $n > 30$ ), računa se vrijednost iz tablice površina normalne distribucije (pomoću  $Z_{\frac{1-\alpha}{2}}$ )

$t$  - ako je ( $n \leq 30$ ) umjesto ( $Z$ ) računa se vrijednost iz tablice površina studentove ili t-distribucije (pomoću  $t_{[\frac{\alpha}{2}, df=n-1]}$ )

$Se(\bar{X})$  - standardna greška aritmetičke sredine (standardna devijacija sampling distribucije aritmetičke sredine).

Ako je poznata standardna devijacija osnovnoga skupa:

$$Se(\hat{X}) = \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad \text{ako je } f \leq 0,05 \text{ i} \quad (3.2)$$

$$Se(\hat{X}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \Rightarrow \quad \text{ako je } f > 0,05, \quad (3.3)$$

pri čemu je  $\sigma$  standardna devijacija osnovnog skupa.

Ako nije poznata standardna devijacija osnovnog skupa  $\sigma$ , računa se **nepristrana (točkasta) ocjena varijance osnovnog skupa na osnovi uzorka**  $S^2$ :

$$S^2 = \hat{\sigma}^2 \cdot \left( \frac{n}{n-1} \right) \quad \Rightarrow \quad \text{ako je } n \leq 30 \text{ i} \quad (3.4)$$

$$S^2 = \hat{\sigma}^2 \quad \Rightarrow \quad \text{ako je } n > 30, \quad (3.5)$$

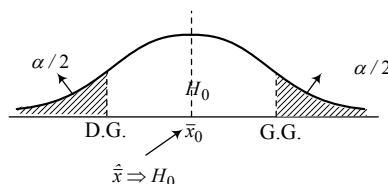
gdje je:  $\hat{\sigma}^2$  varijanca uzorka.

$\alpha$  - nivo signifikantnosti ili značajnosti testa (ako nije određeno drukčije, najčešće se uzima da je  $\alpha = 5\%$ ).

Zaključak o prihvatanju ili odbacivanju nulte  $H_0$  hipoteze donosi se na osnovi  $\hat{\bar{X}}$  aritmetičke sredine iz uzorka (prema slici).

**Slika 3.1.**

**Odluka o prihvatanju hipoteza kod dvosmjernog testiranja o pretpostavljenoj prosječnoj vrijednosti osnovnog skupa**



*Izvor: Konstrukcija autora.*

Prema slici 3.1 ako se aritmetička sredina iz uzorka  $\hat{\bar{X}}$  nalazi između donje (D.G.) i gornje granice (G.G.) intervala prihvatanja hipoteze  $H_0$ , ta se hipoteza prihvata kao istinita uz odgovarajući nivo signifikantnosti testa ( $\alpha$ ). U suprotnom se ta hipoteza odbacuje.

Testirati se može Z-testom (ako je  $n > 30$ ) ili t-testom (ako je  $n \leq 30$ ):

$$Z^* = \frac{\hat{\bar{X}} - \bar{X}_0}{Se(\bar{X})}; \quad Z_{tab\left[\frac{1-\alpha}{2}\right]} \quad \text{ili} \quad t^* = \frac{\hat{\bar{X}} - \bar{X}_0}{Se(\bar{X})}, \quad (3.6)$$

gdje je:

$Z^*$  - Z empirijska vrijednost izračunata na osnovi uzorka

$Z_{tab}$  - vrijednost iz tablice površina normalne distribucije (pomoću  $Z_{\frac{1-\alpha}{2}}$ )

$t^*$  - t empirijska vrijednost izračunata na osnovu uzorka (ako je  $n \leq 30$ )

$t_{tab}$  - tablice površina studentove ili t-distribucije (pomoću  $t_{[df=n-1]}^{\alpha/2}$ ).

Zaključak se donosi ako je:

$|Z^*| < Z_{tab} \Rightarrow H_0$ ; odnosno  $|t^*| < t_{tab} \Rightarrow H_0$ , dok se u suprotnom slučaju ta početna hipoteza odbacuje.

Testirati se može i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $Z^*$  ili  $t^*$  (Tablica A ili B): ako je  $\alpha^* > 5\% \Rightarrow H_0$ , dok se u suprotnom slučaju hipoteza  $H_0$  odbacuje.

**Hipoteze** se na odgovarajući način mogu postaviti i za **jednosmjerno testiranje**.

### 3.3 Testiranje hipoteze o razlici prosječnih vrijednosti dvaju nezavisnih osnovnih skupova

Postavlja se početna ili nulta hipoteza da su aritmetičke sredine dvaju nezavisnih osnovnih skupova  $\bar{X}_1$  i  $\bar{X}_2$  jednake tj. da je njihova razlika nula. Suprotna ili alternativna hipoteza pretpostavlja da razlika između aritmetičkih sredina dvaju osnovnih skupova postoji:

$$H_0 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 = 0$$

$$H_1 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 \neq 0$$

**Interval prihvatanja hipoteze**  $H_0$  glasi:

$$0 \pm Z \cdot Se(\bar{X}_1 - \bar{X}_2), \quad (3.7)$$

gdje je:

$\bar{X}_1$ ;  $\bar{X}_2$  - aritmetičke sredine dvaju nezavisnih osnovnih skupova

$Z$  - ako su veličine uzoraka ( $n_1 + n_2 - 2 > 30$ ), računa se vrijednost iz tablice površina normalne distribucije (pomoću  $Z_{\frac{1-\alpha}{2}}$ )

$t$  - ako su veličine uzoraka ( $n_1 + n_2 - 2 \leq 30$ ) umjesto ( $Z$ ), računa se vrijednost iz tablice površina studentove ili t-distribucije (pomoću  $t_{[df=n_1+n_2-2]}^{\alpha/2}$ )

$\alpha$  - nivo signifikantnosti ili značajnosti testa (ako nije određeno drukčije, najčešće se uzima da je  $\alpha = 5\%$ ).

$Se(\bar{X}_1 - \bar{X}_2)$  - **standardna greška razlike aritmetičkih sredina** koja se računa:

- Ako su standardne devijacije osnovnih skupova poznate i jednake za oba skupa  $\sigma_1 = \sigma_2 = \sigma$ :

$$Se(\bar{X}_1 - \bar{X}_2) = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \quad (3.8)$$

- Ako standardne devijacije osnovnih skupova nisu poznate, već se procjenjuju, i ako je uzorak mali, tj.  $n_1 + n_2 - 2 \leq 30$ :

$$Se(\bar{X}_1 - \bar{X}_2) = \sqrt{\left( \frac{n_1 \cdot \hat{\sigma}_1^2 + n_2 \cdot \hat{\sigma}_2^2}{n_1 + n_2 - 2} \right) \cdot \left( \frac{n_1 + n_2}{n_1 \cdot n_2} \right)}. \quad (3.9)$$

- Ako standardne devijacije osnovnih skupova nisu poznate, već se procjenjuju, i ako je uzorak velik, tj.  $n_1 + n_2 - 2 > 30$ :

$$Se(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad (3.10)$$

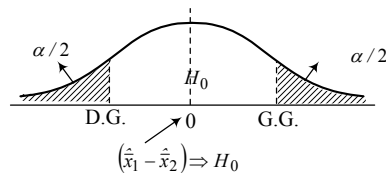
gdje su:

- $\sigma$  - standardna devijacija osnovnog skupa (jednaka za oba skupa)
- $n_1$  - veličina jednog uzorka
- $n_2$  - veličina drugog uzorka
- $\hat{\sigma}_1^2$  - varijanca jednog uzorka
- $\hat{\sigma}_2^2$  - varijanca drugog uzorka
- $S_1^2$  - nepristrana ocjena varijance jednoga osnovnog skupa
- $S_2^2$  - nepristrana ocjena varijance drugoga osnovnog skupa.

Zaključak o prihvatanju ili odbacivanju nulte hipoteze  $H_0$  donosi se na osnovi  $(\hat{\bar{X}}_1 - \hat{\bar{X}}_2)$  razlike aritmetičkih sredina iz promatranih uzoraka. Prema slici ako se razlika aritmetičkih sredina iz uzorka  $(\hat{\bar{X}}_1 - \hat{\bar{X}}_2)$  nalazi između donje (D.G.) i gornje granice (G.G.) intervala prihvatanja hipoteze  $H_0$ , ta se hipoteza prihvata kao istinita uz odgovarajući nivo signifikantnosti testa ( $\alpha$ ). U suprotnom se navedena hipoteza odbacuje.

### Slika 3.2.

#### Odluka o prihvatanju hipoteza kod testiranja o razlici prosječnih vrijednosti dvaju nezavisnih osnovnih skupova



Izvor: Konstrukcija autora.

Testirati se može Z-testom (ako je:  $n_1 + n_2 - 2 > 30$ ) ili t-testom (ako je:  $n_1 + n_2 - 2 \leq 30$ ):

$$Z^* = \frac{\hat{\bar{X}}_1 - \hat{\bar{X}}_2}{Se(\bar{X}_1 - \bar{X}_2)}; \quad Z_{tab\left[\frac{1-\alpha}{2}\right]} \quad \text{ili} \quad t^* = \frac{\hat{\bar{X}}_1 - \hat{\bar{X}}_2}{Se(\bar{X}_1 - \bar{X}_2)}, \quad (3.11)$$

gdje je:

$Z^*$  - Z empirijska vrijednost izračunata na osnovi uzoraka

$Z_{tab}$  - vrijednost iz tablice površina normalne distribucije (pomoću  $Z_{\frac{1-\alpha}{2}}$ )

$t^*$  - t empirijska vrijednost izračunata na osnovi uzorka (ako je  $n_1 + n_2 - 2 \leq 30$ )

$t_{tab}$  - tablice površina studentove ili t-distribucije (pomoću  $t_{\frac{\alpha}{2}, [df=n_1+n_2-2]}$ ).

Zaključak se donosi ako je:

$|Z^*| < Z_{tab} \Rightarrow H_0$ ; odnosno  $|t^*| < t_{tab} \Rightarrow H_0$ , dok se u suprotnom odbacuje navedena hipoteza.

Testirati se može i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $Z^*$  ili  $t^*$  (Tablica A ili B): ako je  $\alpha^* > 5\% \Rightarrow H_0$ , dok se u suprotnom ta hipoteza odbacuje.



### Primjer 3.1.

Na temelju podataka prikupljenih anketnim upitnikom za zadani uzorak ispitanika potrebno je riješiti sljedeće postavke:

- Izračunati prosječnu visinu uzorka ispitanika (varijabla v1) posebno za muški i za ženski spol!
- Može li se na osnovu zadanog uzorka prihvatiti pretpostavka da je prosječna visina osoba ženskog spola 168 cm na promatranom području uz graničnu signifikantnost od 5%!
- Može li se na osnovu zadanog uzorka prihvatiti pretpostavka da ne postoji značajna razlika u visini između osoba muškog i ženskog spola na promatranom području uz graničnu signifikantnost od 5%!

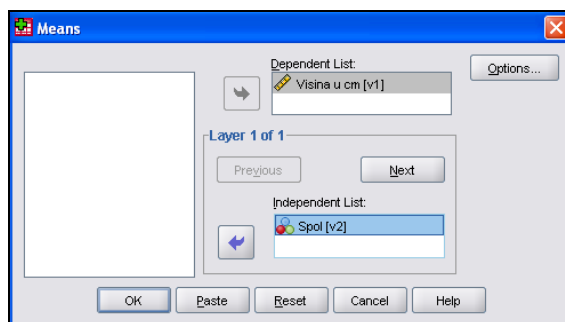


### Rješenje 3.1.

- U programskom paketu **SPSS** potrebno je izračunati prosječnu visinu ispitanika muškog i ženskog spola. Zbog zahtjeva stavki zadatka b) i c) izračunate su standardne devijacije uzoraka i standardne greške odgovarajućih aritmetičkih sredina.

### Slika 3.3.

Prozor "Means" iz izbornika "Compare Means" s odabranim varijablama v1 i v2



*Izvor: Simulirani podaci.*

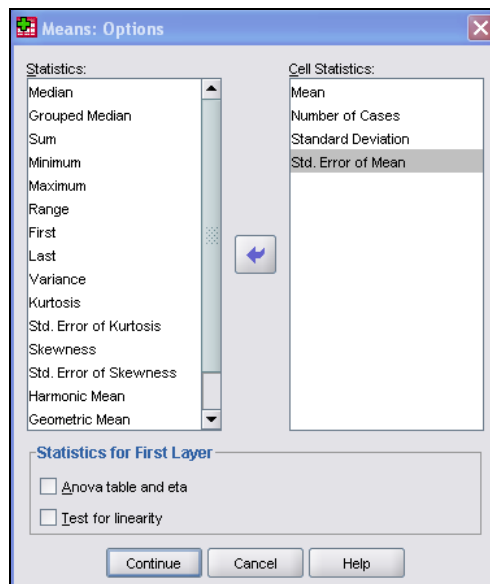


Potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku odabrati **Compare Means**. Dalje se bira **Means**. Najprije se varijabla v1 - Visina u cm (v1) prebaci u polje **Dependent List**, a varijabla v2 - Spol u polje **Independent List**, kako je prikazano na slici 3.3.

Da bi se u izlaznim rezultatima dobile tražene veličine klikom na **Options** na novom otvorenom prozoru iz izbornika **Statistics** u **Cell Statistics** izabere se: **Mean**; **Number of Cases**; **Standard Deviation**; **Std. Error of Mean**, kako je prikazano na slici 3.4.

Slika 3.4.

Prozor "Means:Options" s odabranim veličinama u "Cell Statistics"



Izvor: Simulirani podaci.

Tablica 3.2.

Podaci o prosječnoj težini ispitanika u uzorku prema spolu

Report				
Visina u cm				
Spol	Mean	N	Std. Deviation	Std. Error of Mean
Musko	183,51	91	7,151	,750
Zensko	170,87	144	5,554	,463
Total	175,77	235	8,749	,571

Izvor: Simulirani podaci.

Klikom na ikone **Continue** i **OK** u **Outputu** programa SPSS dobiju se tražene veličine, kako je prikazano u tablici 3.2.

Prema dobivenim podacima u tablici **Report** može se vidjeti da je prosječna težina osoba muškog spola u uzorku 183,51 cm, a prosječna težina osoba ženskog spola u uzorku 170,88 cm.

b) Da bi se donio zaključak uz graničnu signifikantnost od 5% o prihvatanju hipoteze da je prosječna visina osoba ženskog spola na promatranom području 168 cm potrebno je postaviti **hipoteze za dvosmjerno testiranje**:

$$H_0 : \dots\dots\dots \bar{X} = 168$$

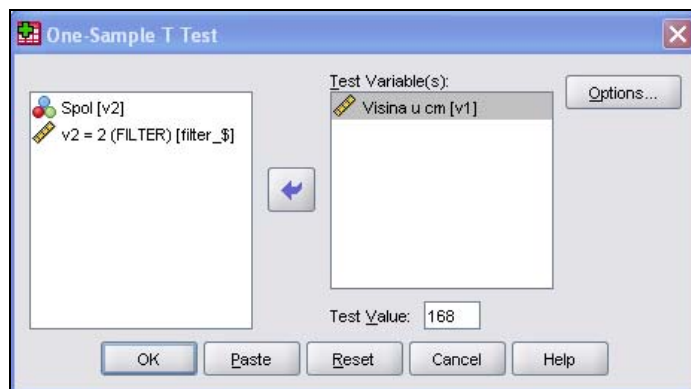
$$H_1 : \dots\dots\dots \bar{X} \neq 168$$

Nakon odabira (filtriranja) ispitanika ženskog spola (**Data; Select cases; If condition is satisfied: v1 = 2**) koji je u obilježju spol kodiran s 2, potrebno je na glavnom izborniku odabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Compare Means** i **One-Sample T-Test**.

U prozoru **One-Sample T-Test** bira se numerička varijabla **Visina u cm (v1)** u **Test Variable(s)**. S obzirom da početna hipoteza pretpostavlja da je prosjek osnovnog skupa jednak 168 cm, u **Test Value**: potrebno je upisati 168. Sve je prikazano na slici 3.5.

**Slika 3.5.**

**Prozor "One-Sample T-Test" s definiranim "Test Variable(s)" i "Test Value"**



*Izvor: Simulirani podaci.*

Klikom na ikonu **OK** u **Output-u** programskog paketa SPSS dobije se rezultat testiranja.

Tablica 3.3.

## Podaci o prosječnoj težini ispitanika u uzorku prema spolu

One-Sample Test						
	Test Value = 168					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Visina u cm	6,211	143	,000	2,875	1,96	3,79

Izvor: Simulirani podaci.

Prema podacima u tablicama 3.2 i 3.3 vrijedi da je veličina uzorka, tj. broj ispitanika ženskog spola 144, što je veće od 30, pa je empirijska vrijednost Z testa:

$$(Z^*) = t^* = \frac{\hat{\bar{X}} - \bar{X}_0}{Se(\bar{X})} = \frac{170,88 - 168}{0,463} = 6,22.$$

Uz zadanu graničnu signifikantnost od 5%, tablična vrijednost Z testa je:

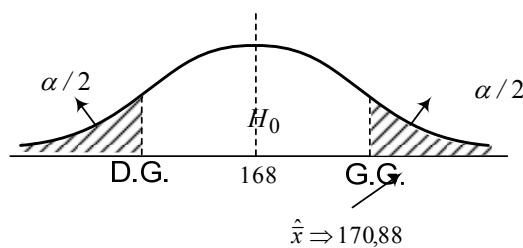
$$Z_{\frac{1-\alpha}{2}} = \frac{1-0,05}{2} = 0,475 \stackrel{tab}{=} 1,96.$$

Dakle, može se zaključiti da je  $|Z^*| > Z_{tab}$  tj. nulta hipoteza da je prosječna visina osoba ženskog spola na promatranom području 168 cm nije istinita.

Do jednakog zaključka može se doći i usporedbom empirijske s graničnom signifikantnošću. Prema rezultatima iz tablice 2.3 vrijedi da je signifikantnost iz uzorka za ovo dvosmjerno testiranje (**Sig.(2-tailed)**) približno 0, pa vrijedi da je:  $\alpha^* \approx 0\%$ , odnosno  $\alpha^* < 5\%$ , tj. odbacuje se nulta hipoteza da je prosječna visina osoba ženskog spola na promatranom području 168 cm.

Slika 3.6.

## Odluka o prihvatanju hipoteza o nepoznatoj prosječnoj vrijednosti osnovnog skupa



Izvor: Simulirani podaci.

Zaključiti se može i na temelju intervala prihvatanja hipoteze  $H_0$  :

$$\bar{X}_0 \pm Z \cdot Se(\bar{X}) \Rightarrow 168 \pm 1,96 \cdot 0,463 \Rightarrow 168 \pm 0,90748 \Rightarrow 167,093 \dots 168,907$$

Prema tablici 3.2 aritmetička sredina (prosječna visina za ženski spol) iz uzorka je  $\hat{\bar{X}} = 170,88$  cm.

Na slici 3.6 može se vidjeti da se aritmetička sredina iz uzorka  $\hat{\bar{X}} = 170,88$  ne nalazi između donje ( D.G. = 167,093 ) i gornje granice ( G.G. = 168,907 ) intervala prihvatanja hipoteze  $H_0$ , pa se ta hipoteza odbacuje, odnosno uz nivo signifikantnosti testa od 5% može se zaključiti da prosječna visina osoba ženskog spola u osnovnom skupu ne iznosi 168 cm.

c) Da bi se donio zaključak uz graničnu signifikantnost od 5% o prihvatanju hipoteze da ne postoji značajna razlika u visini između osoba muškog i ženskog spola na promatranom području potrebno je postaviti **hipoteze o razlici aritmetičkih sredina dvaju nezavisnih osnovnih skupova**:

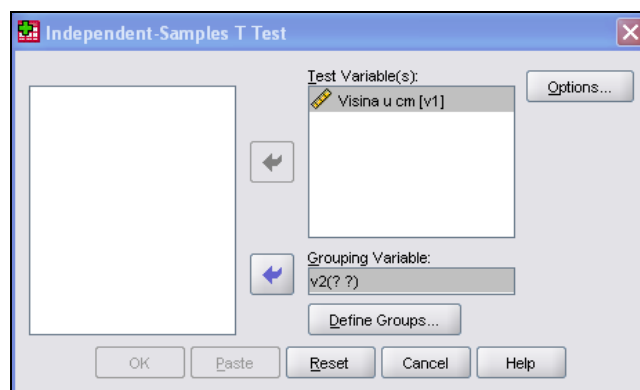
$$H_0 : \dots \bar{X}_1 - \bar{X}_2 = 0$$

$$H_1 : \dots \bar{X}_1 - \bar{X}_2 \neq 0$$

U programskom paketu SPSS potrebno je na glavnom izborniku odabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Compare Means** i **Independent-Samples T Test**.

**Slika 3.7.**

**Prozor " Independent-Samples T Test " s definiranim "Test Variable(s)" i "Grouping Variable"**



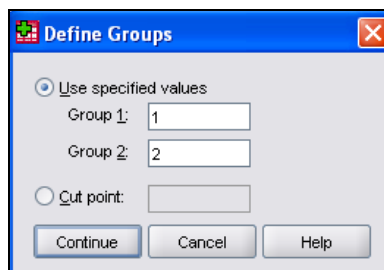
Izvor: Simulirani podaci.

U prozoru **Independent-Samples T Test**, kako je prikazano na slici 3.7, za **Test Variable(s)** bira se Visina u cm (v1) a za **Grouping Variable** bira se Spol (v2).

Zatim je potrebno definirati grupe za **Grouping Variable**. Klikom na ikonu **Define Groups** otvara se novi prozor gdje je u ovom slučaju **Group 1:** 1, što znači Muški spol, a **Group 2:** 2, što označava ženski spol. Definiranje grupa prikazano je na slici 3.8.

Slika 3.8.

#### Prozor " Define Groups" s definiranim grupama



Izvor: Simulirani podaci.

Klikom na **Continue** i **OK** u **Outputu** se dobije rješenje analize **Independent Samples Test** koje je prikazano u tablici 3.4.

Tablica 3.4.

#### Rezultati testiranja nezavisnih uzoraka

Independent Samples Test								
		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Visina u cm	Equal variances assumed	6,444	,012	15,163	233	,000	12,630	,833
	Equal variances not assumed			14,336	157,311	,000	12,630	,881

Izvor: Simulirani podaci.

Zaključak se može donijeti na temelju izračunatog intervala prihvatanja hipoteze  $H_0$ :

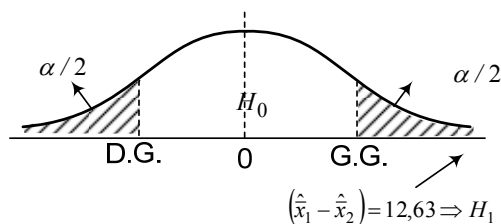
$$0 \pm Z \cdot Se(\bar{X}_1 - \bar{X}_2) \Rightarrow 0 \pm 1,96 \cdot 0,881 \Rightarrow 0 \pm 1,7268,$$

gdje je, na temelju podataka iz tablice 2.4, standardna greška razlike aritmetičkih sredina  $Se(\bar{X}_1 - \bar{X}_2) = 0,881$  (uz pretpostavku da varijance nisu jednake, tj. **Equal variances not assumed**).

Zaključak o prihvatanju ili odbacivanju nulte  $H_0$  hipoteze donosi se na osnovu razlike aritmetičkih sredina iz uzorka:  $(\hat{X}_1 - \hat{X}_2) = 12,63$ .

**Slika 3.9.**

**Odluka o prihvatanju hipoteza o razlici prosječnih vrijednosti dvaju nezavisnih osnovnih skupova**



*Izvor: Simulirani podaci.*

Prema slici 3.9 razlika aritmetičkih sredina iz uzorka  $(\hat{X}_1 - \hat{X}_2)$  ne nalazi se između donje ( $D.G. = -1,7268$ ) i gornje granice ( $G.G. = +1,7268$ ) intervala prihvatanja hipoteze  $H_0$ , pa se nulta hipoteza odbacuje. Može se zaključiti da postoji statistički značajna razlika u prosječnoj visini između ispitanika muškog i ženskog spola.

Testiranje se može izvesti Z-testom ( $n_1 + n_2 - 2 > 30$ , a u SPSS **Output-u** naveden je t-test) čija je empirijska vrijednost također dana u izlaznoj tablici 3.4:

$$Z^*(t^*) = 14,336 ; \quad Z_{tab} \left[ \frac{1-\alpha}{2} \right] = 1,96 .$$

Može se zaključiti da je  $|Z^*| > Z_{tab}$ , što opet vodi k zaključku o odbacivanju nulte hipoteze.

Isto testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  koja je isto prikazana u izlaznoj tablici 3.4:  $\alpha^* \approx 0\%$ , tj.  $\alpha^* < 5\%$ , čime se potvrđuje jednak zaključak.

Dakle, na sva tri načina testiranja dolazi se do jednakog zaključka o odbacivanju  $H_0$  hipoteze da je prosječna visina stanovnika muškog i ženskog spola na promatranom području jednaka.

### 3.4 Testiranje hipoteze o razlici prosječnih vrijednosti dvaju zavisnih osnovnih skupova

Kod ovog testiranja postavljaju se hipoteze i donose zaključci o njihovu prihvatanju na jednak način **kao kod testiranja hipoteze o razlici aritmetičkih sredina dvaju nezavisnih osnovnih skupova**. Postavlja se početna ili nulta hipoteza da su aritmetičke sredine dvaju zavisnih osnovnih skupova  $\bar{X}_1$  i  $\bar{X}_2$  jednake, tj. da je njihova razlika nula. Suprotna ili alternativna hipoteza pretpostavlja da razlika između aritmetičkih sredina dvaju osnovnih skupova postoji:

$$H_0 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 = 0$$

$$H_1 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 \neq 0$$

**Interval prihvatanja hipoteze  $H_0$  glasi:**

$$0 \pm Z \cdot Se(\bar{X}_1 - \bar{X}_2), \quad (3.12)$$

gdje je:

$\bar{X}_1; \bar{X}_2$  - aritmetičke sredine dvaju nezavisnih osnovnih skupova

$Z$  - ako su veličine uzoraka ( $n_1 + n_2 - 2 > 30$ ), računa se vrijednost iz tablice površina normalne distribucije (pomoću  $Z_{\frac{1-\alpha}{2}}$ )

$t$  - ako su veličine uzoraka ( $n_1 + n_2 - 2 \leq 30$ ) umjesto ( $Z$ ), računa se vrijednost iz tablice površina studentove ili t-distribucije (pomoću  $t_{[df=n_1+n_2-2]}^{\alpha/2}$ )

$\alpha$  - nivo signifikantnosti ili značajnosti testa (ako nije određeno drukčije, najčešće se uzima da je  $\alpha = 5\%$ ).

$Se(\bar{X}_1 - \bar{X}_2)$  - **standardna greška razlike aritmetičkih sredina koja se računa:**

$$Se(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2 \cdot r_{1,2} \cdot \sqrt{\frac{S_1^2}{n_1}} \cdot \sqrt{\frac{S_2^2}{n_2}}}, \quad (3.13)$$

gdje je:

$r_{1,2}$  - Pearsonov koeficijent linearne korelacije između dvaju mjerenja iste slučajne varijable na istom uzorku (zavisni skupovi)

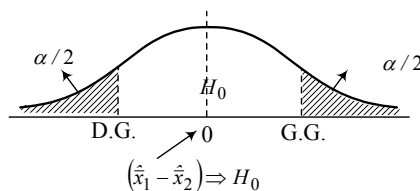
$\sqrt{\frac{S_1^2}{n_1}}$  - standardna greška aritmetičke sredine jednog uzorka ( $Se(\bar{X}_1)$ )

$\sqrt{\frac{S_2^2}{n_2}}$  - standardna greška aritmetičke sredine drugog uzorka ( $Se(\bar{X}_2)$ ).

Zaključak o prihvatanju ili odbacivanju nulte hipoteze  $H_0$  donosi se na osnovi ( $\hat{\bar{X}}_1 - \hat{\bar{X}}_2$ ) razlike aritmetičkih sredina iz promatranih zavisnih uzoraka (prema slici 3.10).

**Slika 3.10.**

**Odluka o prihvatanju hipoteza kod testiranja o razlici prosječnih vrijednosti dvaju zavisnih osnovnih skupova**



Izvor: Konstrukcija autora.

Prema slici 3.10 ako se razlika aritmetičkih sredina iz uzorka ( $\hat{\bar{X}}_1 - \hat{\bar{X}}_2$ ) nalazi između donje (D.G.) i gornje granice (G.G.) intervala prihvatanja hipoteze  $H_0$ , ta se hipoteza prihvaća kao istinita uz odgovarajući nivo signifikantnosti testa ( $\alpha$ ). U suprotnom se navedena hipoteza odbacuje.

Testirati se može Z-testom (ako je:  $n_1 + n_2 - 2 > 30$ ) ili t-testom (ako je:  $n_1 + n_2 - 2 \leq 30$ ):

$$Z^* = \frac{\hat{\bar{X}}_1 - \hat{\bar{X}}_2}{Se(\bar{X}_1 - \bar{X}_2)}; \quad Z_{tab} \left[ \frac{1-\alpha}{2} \right] \quad \text{ili} \quad t^* = \frac{\hat{\bar{X}}_1 - \hat{\bar{X}}_2}{Se(\bar{X}_1 - \bar{X}_2)}, \quad (3.14)$$

gdje je:

$Z^*$  - Z empirijska vrijednost izračunata na osnovi uzoraka

$Z_{tab}$  - vrijednost iz tablice površina normalne distribucije (pomoću  $Z_{\frac{1-\alpha}{2}}$ )



$t^*$  -  $t$  empirijska vrijednost izračunata na osnovi uzorka (ako je  $n_1 + n_2 - 2 \leq 30$ )

$t_{tab}$  - tablice površina studentove ili t-distribucije (pomoću  $t_{[df=n_1+n_2-2]}^{\alpha/2}$ ).

Zaključak se donosi ako je:

$|Z^*| < Z_{tab} \Rightarrow H_0$ ; odnosno  $|t^*| < t_{tab} \Rightarrow H_0$ , dok se u suprotnom odbacuje navedena hipoteza.

Testirati se može i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $Z^*$  ili  $t^*$  (Tablica A ili B): ako je  $\alpha^* > 5\% \Rightarrow H_0$ , dok se u suprotnom ta hipoteza odbacuje.



### Primjer 3.2.

Na temelju podataka prikupljenih anketnim upitnikom za zadani uzorak ispitanika potrebno je riješiti sljedeće postavke:

- Izračunati prosječnu težinu uzorka ispitanika od prije 5 godina (varijabla v1) i prosječnu težinu uzorka ispitanika sada (varijabla v2)!
- Može li se na osnovu zadanog uzorka prihvatiti pretpostavka da se prosječna težina ispitanika na promatranom području ne razlikuje od prije 5 godina i danas uz graničnu signifikantnost od 5%! **Ovdje se radi o zavisnim uzorcima!**



### Rješenje 3.2.

a) U programskom paketu **SPSS** potrebno je izračunati prosječnu težinu uzorka ispitanika od prije 5 godina i prosječnu težinu uzorka ispitanika sada. Osim na uobičajen način (**Analyze; Descriptive statistics; Frequencies**; gdje se onda odabire odgovarajuća statistika) tražene odgovarajuće aritmetičke sredine mogu se izračunati u opciji **Compare Means**, što je detaljno opisano u zadatku pod b). Rezultat je prikazan u tablici 3.5. Prosječna težina ispitanika prije 5 godina je 67,54 kg, a prosječna težina ispitanika danas je 68,68 kg.

b) Da bi se donio zaključak o prihvatanju pretpostavke da se prosječna težina ispitanika na promatranom području ne razlikuje od prije 5 godina i danas, tj. da bi se izvršilo testiranje hipoteza o razlici aritmetičkih sredina dvaju zavisnih osnovnih skupova postavljaju se hipoteze:

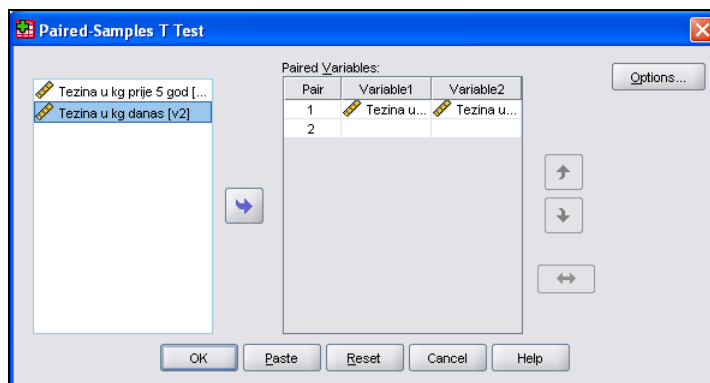
$$H_0 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 = 0$$

$$H_1 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku odabrati **Compare Means**. Dalje se bira **Paired - Samples T Test**.

**Slika 3.11.**

**Prozor " Paired - Samples T Test " iz izbornika "Compare Means" s odabranim varijablama v1 i v2**



*Izvor: Simulirani podaci.*

Zatim se numeričke varijable v1 - Težina u kg prije 5 god. i v2 - Težina u kg danas prebace u polje **Paired Variables**, kako je prikazano na slici 3.11. Klikom na ikonu **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine, kako je prikazano u tablicama 3.5 i 3.6.

**Tablica 3.5.**

**Rezultati o aritmetičkim sredinama dvaju zavisnih uzoraka**

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Tezina u kg prije 5 god	67,5400	50	14,56612	2,05996
	Tezina u kg danas	68,6800	50	14,16872	2,00376

*Izvor: Simulirani podaci.*

Na temelju rezultata iz tablice 3.6 vrijedi da je:

$$Z^{*}(t^{*}) = \frac{\hat{\bar{X}}_1 - \hat{\bar{X}}_2}{Se(\bar{X}_1 - \bar{X}_2)} = -2,773 ; \quad Z_{tab} = 1,96$$

t	df	Sig. (2-tailed)
-2,773	49	,008

b) Može li se li se na osnovu zadanog uzorka prihvatiti pretpostavka da se prosječni džeparac ispitanika na promatranom području ne razlikuje od prije 3 godine i danas uz graničnu signifikantnost od 5%! **Ovdje se radi o zavisnim uzorcima!**



### Rješenje 3.3.

a) U programskom paketu **SPSS** potrebno je izračunati prosječni džeparac uzorka od 60 ispitanika od prije 3 godine i prosječni džeparac uzorka ispitanika danas. Osim na uobičajen način (**Analyze; Descriptive statistics; Frequencies**; gdje se onda odabire odgovarajuća statistika); tražene odgovarajuće aritmetičke sredine mogu se izračunati u opciji **Compare Means**, što je detaljno opisano u zadatku pod b). Prosječni džeparac ispitanika prije 3 godine je 447,96 kn, a prosječni džeparac ispitanika danas je 313,27 kn.

b) Da bi se donio zaključak o prihvatanju pretpostavke da se prosječni džeparac ispitanika na promatranom području ne razlikuje od prije 3 godine i danas, tj. da bi se izvršilo testiranje hipoteza o razlici aritmetičkih sredina **dvaju zavisnih osnovnih skupova** postavljaju se hipoteze:

$$H_0 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 = 0$$

$$H_1 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku odabrati **Compare Means**. Dalje se bira **Paired - Samples T Test**.

Zatim se numeričke varijable v1 - Visina džeparca i v2 - Visina džeparca prije 3 godine prebace u polje **Paired Variables**. Klikom na ikonu **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine, kako je prikazano u tablicama 3.7 i 3.8.

**Tablica 3.7.**

#### Rezultati o aritmetičkim sredinama dvaju zavisnih uzoraka

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Visina džeparca	313,2653	49	317,79440	45,39920
	Visina džeparca prije 3 godine	447,9592	49	589,31097	84,18728

Izvor: Simulirani podaci.

Na temelju rezultata iz tablice 3.8 vrijedi da je:

$$Z^*(t^*) = \frac{\hat{X}_1 - \hat{X}_2}{Se(\hat{X}_1 - \hat{X}_2)} = -1,730; \quad Z_{tab} = 1,96.$$

Očigledno je  $|Z^*| < Z_{tab}$ , pa se prihvata početna hipoteza, tj. donosi se zaključak da prosječni džeparac populacije na promatranom području prije 3 godine i prosječni džeparac populacije na promatranom području danas nemaju statistički značajnu razliku uz signifikantnost testa od 5%.

**Tablica 3.8.**

**Rezultati testiranja o aritmetičkim sredinama dvaju zavisnih uzoraka**

Paired Samples Test					
		Paired Differences			
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference
					Lower Upper
Pair 1	Visina džeparca - Visina džeparca prije 3 godine	-134,69388	544,93540	77,84791	-291,21760 21,82985

t	df	Sig. (2-tailed)
-1,730	48	,090

*Izvor: Simulirani podaci.*

Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  koja je isto prikazana u izlaznoj tablici 3.8:  $\alpha^* = 0,090 = 9\%$  tj.  $\alpha^* > 5\%$ , čime se potvrđuje jednak zaključak.

Dakle, i na ovaj način testiranja dolazi se do jednakog zaključka o prihvatanju  $H_0$  hipoteze, da ne postoji statistički značajna razlika u prosječnom džeparcu stanovnika na promatranom području od prije 3 godine i danas.



**Primjer 3.4.**

Na temelju podataka prikupljenih anketnim upitnikom za zadani uzorak ispitanika potrebno je izračunati prosječnu težinu uzorka ispitanika (varijabla v7) koji su se izjasnili da zadovoljavaju osnovne potrebe i onih ispitanika koji za sebe kažu da nemaju materijalnih problema.

Može li se na osnovu zadanog uzorka prihvatiti pretpostavka da ne postoji značajna razlika u težini između stanovnika koji zadovoljavaju osnovne potrebe i onih ispitanika koji su se izjasnili da nemaju materijalnih problema na promatranom području uz graničnu signifikantnost od 5%? (Zaključak je potrebno donijeti na osnovu odgovarajućeg intervala prihvatanja  $H_0$  hipoteze i pomoću empirijske signifikantnosti!)



#### Rješenje 3.4.

U programskom paketu **SPSS** potrebno je izračunati prosječnu težinu ispitanika koji zadovoljavaju osnovne potrebe i onih ispitanika koji su se izjasnili da nemaju materijalnih problema.

Potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku odabrati **Compare Means**. Dalje se bira **Means**. Najprije se varijabla v1 - Težina u kg prebaci u polje **Dependent List**, a varijabla v2 - Financijske prilike u polje **Independent List**.

Da bi se u izlaznim rezultatima dobile tražene veličine klikom na **Options** na novom otvorenom prozoru iz izbornika **Statistics** u **Cell Statistics** izabere se: **Mean**; **Number of Cases**; **Standard Deviation**; **Std. Error of Mean**.

Klikom na ikone **Continue** i **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine, kako je prikazano u tablici 3.9.

Prema dobivenim podacima u tablici **Report** može se vidjeti da je prosječna težina osoba koje zadovoljavaju osnovne potrebe u uzorku 65,02 kg, a prosječna težina osoba koje nemaju materijalnih problema u uzorku 68,10 kg.

**Tablica 3.9.**

**Podaci o prosječnoj težini ispitanika u uzorku prema financijskoj situaciji**

Report			
Težina u kg			
Financijske prilike	Mean	N	Std. Deviation
Nema sredstava za normalan standard	70,7500	4	22,29163
Zadovoljava osnovne potrebe	65,0194	103	12,50724
Nema materijalnih problema	68,1008	119	13,43496
Zivi odlicno	84,3333	6	16,46410
Total	67,1983	232	13,56384

Izvor: Simulirani podaci.

Da bi se donio zaključak uz graničnu signifikantnost od 5% o prihvatanju hipoteze da ne postoji značajna razlika u težini između stanovnika koji zadovoljavaju osnovne potrebe i onih ispitanika koji su se izjasnili da nemaju materijalnih problema na promatranom području potrebno je postaviti **hipoteze o razlici aritmetičkih sredina dvaju nezavisnih osnovnih skupova**:

$$H_0 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 = 0$$

$$H_1 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Compare Means** i **Independent-Samples T Test**.

U prozoru **Independent-Samples T Test**, za **Test Variable(s)** bira se Težina u kg (v1) a za **Grouping Variable** bira se Financijske prilike (v2).

Zatim je potrebno definirati grupe za **Grouping Variable**. Klikom na ikonu **Define Groups** otvara se novi prozor gdje je u ovom slučaju **Group 1**: 2, što znači da ispitanik zadovoljava osnovne potrebe, a **Group 2**: 3, što označava da ispitanik nema materijalnih problema.

Klikom na **Continue** i **OK** u **Outputu** se dobije rješenje analize **Independent Samples Test** koje je prikazano u tablici 3.10.

**Tablica 3.10.**

### Rezultati testiranja nezavisnih uzoraka

Independent Samples Test								
		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Težina u kg	Equal variances assumed	1,137	,288	-1,759	220	,080	-3,08142	1,75131
	Equal variances not assumed			-1,769	218,816	,078	-3,08142	1,74228

*Izvor: Simulirani podaci.*

Zaključak se može donijeti na temelju izračunatog intervala prihvatanja hipoteze  $H_0$ :

$$0 \pm Z \cdot Se(\bar{X}_1 - \bar{X}_2) \Rightarrow 0 \pm 1,96 \cdot 1,74228 \Rightarrow 0 \pm 3,4149,$$

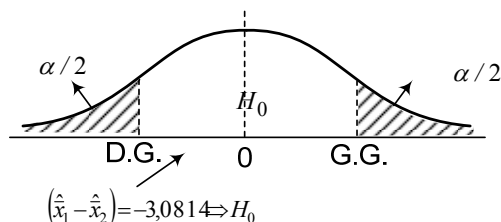
gdje je, na temelju podataka iz tablice 3.10, standardna greška razlike aritmetičkih sredina  $Se(\bar{X}_1 - \bar{X}_2) = 1,74228$  (uz pretpostavku da varijance nisu jednake, tj. **Equal variances not assumed**).

Zaključak o prihvatanju ili odbacivanju nulte  $H_0$  hipoteze donosi se na osnovu razlike aritmetičkih sredina iz uzorka:  $(\hat{X}_1 - \hat{X}_2) = -3,0814$ .

Prema slici 3.12 razlika aritmetičkih sredina iz uzorka  $(\hat{X}_1 - \hat{X}_2)$  nalazi se između donje ( $D.G. = -3,4149$ ) i gornje granice ( $G.G. = +3,4149$ ) intervala prihvatanja hipoteze  $H_0$ , pa se nulta hipoteza prihvata. Može se zaključiti da ne postoji statistički značajna razlika u prosječnoj težini stanovnika koji zadovoljavaju osnovne potrebe i onih koji nemaju materijalnih problema uz signifikantnost od 5%.

**Slika 3.12.**

**Odluka o prihvatanju hipoteza o razlici prosječnih vrijednosti dvaju nezavisnih osnovnih skupova**



Izvor: Simulirani podaci.

Testiranje se može izvesti Z-testom (iako je  $n_1 + n_2 - 2 > 30$ , u SPSS **Output-u** naveden je t-test) čija je empirijska vrijednost također dana u izlaznoj tablici 3.10:

$$Z^*(t^*) = 1,769; \quad Z_{tab}\left[\frac{1-\alpha}{2}\right] = 1,96.$$

Može se zaključiti da je  $|Z^*| < Z_{tab}$ , što opet vodi ka zaključku o prihvatanju nulte hipoteze.

Isto testiranje i zaključak može se donijeti i izračunavanjem granične signifikantnosti  $\alpha^*$  koja je isto prikazana u izlaznoj tablici 3.10:  $\alpha^* = 0,078 = 7,8\%$ , tj.  $\alpha^* > 5\%$ , čime se potvrđuje jednak zaključak.



### **Primjer 3.5.**

Potrebno je izračunati prosječni džeparac uzorka ispitanika koji su se izjasnili da zadovoljavaju osnovne potrebe i onih ispitanika koji za sebe kažu da nemaju materijalnih problema.



Može li se na osnovu zadanog uzorka prihvatiti pretpostavka da ne postoji značajna razlika u prosječnom džeparcu između stanovnika koji zadovoljavaju osnovne potrebe i onih ispitanika koji su se izjasnili da nemaju materijalnih problema na promatranom području uz graničnu signifikantnost od 1%? (Zaključak je potrebno donijeti na osnovu odgovarajućeg intervala prihvatanja  $H_0$  hipoteze i pomoću empirijske signifikantnosti!)



### Rješenje 3.5.

U programskom paketu **SPSS** potrebno je izračunati prosječni džeparac ispitanika koji zadovoljavaju osnovne potrebe i onih ispitanika koji su se izjasnili da nemaju materijalnih problema.

Potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku odabrati **Compare Means**. Dalje se bira **Means**. Najprije se varijabla v2 - Džeparac prebaci u polje **Dependent List**, a varijabla v1 - Financijske prilike u polje **Independent List**.

Tablica 3.11.

Podaci o prosječnom džeparcu ispitanika u uzorku prema financijskoj situaciji

Report			
Džeparac			
Financijske prilike	Mean	N	Std. Deviation
Nema sredstava za normalan standard	125,0000	2	106,06602
Zadovoljava osnovne potrebe	311,6667	78	276,48732
Nema materijalnih problema	405,4419	86	437,20391
Zivi odlično	900,0000	5	946,04440
Total	373,8480	171	403,08014

Izvor: Simulirani podaci.

Da bi se u izlaznim rezultatima dobile tražene veličine klikom na **Options** na novom otvorenom prozoru iz izbornika **Statistics** u **Cell Statistics** izabere se: **Mean**; **Number of Cases**; **Standard Deviation**; **Std. Error of Mean**.

Klikom na ikone **Continue** i **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine, kako je prikazano u tablici 3.11.

Prema dobivenim podacima u tablici **Report** može se vidjeti da je prosječni džeparac osoba koji zadovoljavaju osnovne potrebe **311,67 kn** i onih ispitanika koji su se izjasnili da nemaju materijalnih problema u uzorku **405,44 kn**.

Da bi se donio zaključak uz graničnu signifikantnost od 1% o prihvatanju hipoteze da ne postoji značajna razlika u prosječnom džeparcu između stanovnika koji zadovoljavaju osnovne potrebe i onih ispitanika koji su se izjasnili da nemaju materijalnih problema na promatranom području potrebno je postaviti **hipoteze o razlici aritmetičkih sredina dvaju nezavisnih osnovnih skupova**:

$$H_0 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 = 0$$

$$H_1 : \dots\dots\dots \bar{X}_1 - \bar{X}_2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati ikonu **Analyze**, a na njezinu padajućem izborniku **Compare Means** i **Independent-Samples T Test**.

U prozoru **Independent-Samples T Test**, za **Test Variable(s)** bira se Džeparac (v2) a za **Grouping Variable** bira se Financijske prilike (v1).

Zatim je potrebno definirati grupe za **Grouping Variable**. Klikom na ikonu **Define Groups** otvara se novi prozor gdje je u ovom slučaju **Group 1**: 2, što znači da ispitanik zadovoljava osnovne potrebe, a **Group 2**: 3, što označava da ispitanik nema materijalnih problema.

Klikom na **Continue** i **OK** u **Outputu** se dobije rješenje analize **Independent Samples Test** koje je prikazano u tablici 3.12.

**Tablica 3.12.**

### Rezultati testiranja nezavisnih uzoraka

Independent Samples Test								
		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Džeparac	Equal variances assumed	2,635	,106	-1,623	162	,107	-93,77519	57,79575
	Equal variances not assumed			-1,657	145,301	,100	-93,77519	56,59249

*Izvor: Simulirani podaci.*

Zaključak se može donijeti na temelju izračunatog intervala prihvatanja hipoteze  $H_0$ :

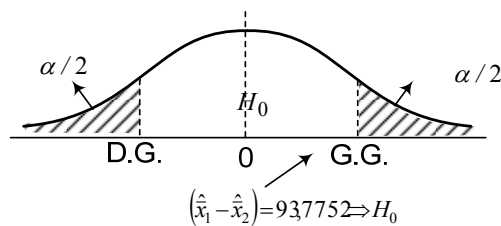
$$0 \pm Z \cdot Se(\bar{X}_1 - \bar{X}_2) \Rightarrow 0 \pm 2,58 \cdot 56,59249 \Rightarrow 0 \pm 146,009,$$

gdje je, na temelju podataka iz tablice 3.12, standardna greška razlike aritmetičkih sredina  $Se(\bar{X}_1 - \bar{X}_2) = 56,59249$  (uz pretpostavku da varijance nisu jednake, tj. **Equal variances not assumed**).

Zaključak o prihvatanju ili odbacivanju nulte  $H_0$  hipoteze donosi se na osnovu razlike aritmetičkih sredina iz uzorka:  $(\hat{\bar{X}}_1 - \hat{\bar{X}}_2) = -93,77519$ .

**Slika 3.13.**

**Odluka o prihvatanju hipoteza o razlici prosječnih vrijednosti dvaju nezavisnih osnovnih skupova**



*Izvor: Simulirani podaci.*

Prema slici 3.13 razlika aritmetičkih sredina iz uzorka  $(\hat{\bar{X}}_1 - \hat{\bar{X}}_2)$  nalazi se između donje ( $D.G. = -146,009$ ) i gornje granice ( $G.G. = +146,009$ ) intervala prihvatanja hipoteze  $H_0$ , pa se nulta hipoteza prihvata. Može se zaključiti da ne postoji statistički značajna razlika u prosječnom džeparcu stanovnika koji zadovoljavaju osnovne potrebe i onih koji nemaju materijalnih problema uz signifikantnost od 1%.

Testiranje se može izvesti Z-testom ( $n_1 + n_2 - 2 > 30$ , a u SPSS **Output-u** naveden je t-test) čija je empirijska vrijednost također dana u izlaznoj tablici 3.12:

$$Z^*(t^*) = -1,657 ; \quad Z_{tab\left[\frac{1-\alpha}{2}\right]} = 2,58 .$$

Može se zaključiti da je  $|Z^*| < Z_{tab}$ , što opet vodi ka zaključku o prihvatanju nulte hipoteze uz signifikantnost testa od 1%.

Isto testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  koja je isto prikazana u izlaznoj tablici:  $\alpha^* = 0,100 = 10,0\%$ , tj.  $\alpha^* > 1\%$ , čime se potvrđuje jednak zaključak.

### 3.5 Testiranje hipoteze o nezavisnosti dvaju kvalitativnih obilježja elemenata osnovnog skupa

Ovo testiranje vrši se pomoću **Hi-kvadrat testa**.

Hi-kvadrat test ne pretpostavlja oblik distribucije i svrstava se u neparametrijske testove. Temelji se na rasporedu frekvencija unutar tablice kontigence (tablica po kojoj se izračunava empirijska vrijednost Hi-kvadrata). To znači da je zbroj originalnih apsolutnih frekvencija i očekivanih teorijskih frekvencija (koje se izračunavaju uz pretpostavku nulte hipoteze  $H_0$ ) uvijek jednak, a pri donošenju zaključka bitan je njihov raspored u distribuciji. Ako je razlika originalnih i teorijskih frekvencija velika početna hipoteza  $H_0$  se odbacuje, a ako njihova razlika statistički nije značajna ta hipoteza se prihvaća kao istinita. Stoga se Hi-kvadrat test u literaturi naziva "frequency based statistic".

Prema časopisu *Science* ovaj test nalazi se između *dvadeset najvažnijih znanstvenih otkrića dvadesetog stoljeća*.

Da bi se testirala hipoteza o nezavisnosti dvaju kvalitativnih obilježja elemenata osnovnog skupa postavljaju se hipoteze:

$$\begin{aligned} H_0 : & \dots\dots\dots P_{ij} = P_{i\bullet} \cdot P_{\bullet j}, \quad \forall i, \forall j \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c \\ H_1 : & \dots\dots\dots \exists P_{ij} \neq P_{i\bullet} \cdot P_{\bullet j} \end{aligned}$$

gdje nulta hipoteza  $H_0$  pretpostavlja da nema ovisnosti između dvaju obilježja.

Empirijska vrijednost  $\chi^2$  testa je:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(m_{ij} - e_{ij})^2}{e_{ij}}, \quad (3.15)$$

gdje su:

- $m_{ij}$  - originalne ili empirijske frekvencije
- $e_{ij}$  - očekivane ili teorijske frekvencije koje se izračunavaju pod pretpostavkom nulte hipoteze  $H_0$ , tj. da nema ovisnosti između dvaju obilježja osnovnog skupa
- $r$  - broj redaka u tablici kontigence
- $c$  - broj stupaca u tablici kontigence.

Očekivane frekvencije  $e_{ij}$  se izračunavaju na sljedeći način:

$$e_{ij} = \frac{m_{i\bullet} \cdot m_{\bullet j}}{n}, \quad (3.16)$$

gdje je:

$m_{i\bullet}$  - marginalna frekvencija i-tog retka (u tablici kontingence nalazi se u zbirnom stupcu)

$m_{\bullet j}$  - marginalna frekvencija j-tog stupca (u tablici kontingence nalazi se u zbirnom retku)

$n$  - veličina uzorka.

Tablična vrijednost  $\chi^2$  - testa traži se iz tablica Hi-kvadrat distribucije (tablice C1 i C2):

$$\chi^2_{tab}[\alpha, df = (r-1) \cdot (c-1)] \quad (3.17)$$

uz odgovarajući nivo signifikantnosti  $\alpha$  i stupnjeve slobode  $df$ .

Zaključak se donosi na način da se uspoređi Hi-kvadrat empirijska i tablična vrijednost:

$\chi^2_{emp} < \chi^2_{tab} \Rightarrow H_0$ ; što znači da ne postoji ovisnost obilježja elemenata osnovnog skupa, dok se u suprotnom ta početna hipoteza odbacuje.

Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $\chi^2$  (Tablica C1 i C2): ako je  $\alpha^* > 5\% \Rightarrow H_0$ , dok se u suprotnom ta hipoteza odbacuje.

Potrebno je napomenuti da kod ovakvog testiranja frekvencije u poljima tablice kontingence ne smiju biti po volji malene. Opći princip je da najmanja očekivana frekvencija ne smije biti manja od 5, iako se u praksi odstupa od toga pravila, ali u manjem broju slučajeva. Npr. u tablicama kontingence koje su veće od 2x2 (dva retka i dva stupca) može se dozvoliti da najmanja očekivana frekvencija bude 1, ako nema više od 20% svih frekvencija koje su manje od 5.

Za najmanje tablice veličine 2x2 (dva retka i dva stupca) da bi se upotrijebio Hi-kvadrat test potrebno je da veličina uzorka bude veća od 40, te se još koriste neki dodatni testovi.

Pomoću hi-kvadrat testa utvrđuje se postoji li povezanost između dvije promatrane varijable, ali ne utvrđuje se visina povezanosti. Aproksimativna visina

povezanosti ili ovisnosti dviju varijabli može se odrediti pomoću **Pearsonovog koeficijenta kontigence**:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \quad (3.18)$$

čija je najmanja vrijednost nula. Nedostatak ovog koeficijenta je u tome što ovisi o samim podacima, te što ne može poprimiti vrijednost 1. Stoga se ne mogu međusobno uspoređivati pojedinačne vrijednosti koeficijenta  $C$ , ako su dobivene iz različitih uzoraka.

Još je važno navesti neke **karakteristike Hi - kvadrat testa**:

- Hi-kvadrat test računa se samo s frekvencijama (u polja hi-kvadrat testa ne unose se aritmetičke sredine ili postotci);
- zbroj očekivanih frekvencija jednak je zbroju opaženih frekvencija, tj. frekvencijama iz uzorka;
- frekvencije u pojedinim poljima moraju biti nezavisne, tako da svaka frekvencija u pojedinom polju mora pripadati drugom individuumu (npr. u tablicu se ne smije unositi nekoliko odgovora jednog ispitanika);
- očekivane frekvencije ne smiju biti po volji male:
  - ako je broj stupnjeva slobode veći od 1 (tablice veće od  $2 \times 2$ ), hi-kvadrat se može računati ako manje od 20% polja ima očekivanu frekvenciju manju od 5, a ni jedno polje manje od 1. Ako to nije postignuto treba spajati polja u kojima su očekivane frekvencije previše male.
  - za tablice ( $2 \times 2$ ), hi-kvadrat se smije upotrijebiti, ako je  $N > 40$ . Ako je  $20 < N < 40$  ne smije ni jedna očekivana frekvencija biti manja od 5.
  - u novije vrijeme pojavile su se studije koje dokazuju da nije posebno važno pridržavati se ovog pravila.
- kada postoji samo 1 stupanj slobode (za tablice  $2 \times 2$ ), potrebno je provesti tzv. Yatesovu korekciju za kontinuitet. Ova korekcija se radi na način da za 0,5 smanji svaka opažena frekvencija, koja je veća od očekivane, te da se za 0,5 poveća svaka opažena frekvencija koja je manja od očekivane. Na taj se način svaka razlika između očekivane i opažene frekvencije smanji za 0,5. Ova korekcija nema smisla ako su razlike između opaženih i očekivanih

frekvencija toliko male, da se Yatesovom korekcijom dobije razlika koja je numerički veća.

Umjesto Pearsonovog koeficijenta kontigence može se računati  $Fi(\phi)$  **koeficijent** za tablice  $(2 \times 2)$ :

$$Fi = \sqrt{\frac{\chi^2}{N}}, \quad (3.19)$$

koji je statistički značajan, ako je i pripadajući  $\chi^2$  statistički značajan i **Cramerov**  $Fi(\phi)$  za tablice reda većeg od  $(2 \times 2)$ :

$$\text{Cramerov } Fi(\phi) = \sqrt{\frac{\chi^2}{N(s-1)}}, \quad (3.20)$$

gdje je:

$s$  - manji broj stupaca ili redova u tablici kontigence.

**Cramerov**  $Fi(\phi)$  **koeficijent** je statistički značajan ako je i pripadajući  $\chi^2$  statistički značajan.



### Primjer 3.6.

Putem anketnog upitnika u jednom poduzeću izvršeno je ispitivanje zadovoljstva zaposlenih po različitim kategorijama. Na temelju odabranog uzorka potrebno je riješiti sljedeće postavke:

- Ispitati postoji li ovisnost između spola i vrste glazbe koju slušaju zaposleni u promatranom poduzeću uz graničnu signifikantnost od 1%!
- U slučaju prihvatanja hipoteze da postoji ovisnost između promatranih obilježja potrebno je izračunati Pearsonov koeficijent kontigence!



### Rješenje 3.6.

- Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li ovisnost između spola i vrste glazbe koju slušaju zaposleni u promatranom poduzeću potrebno je postaviti hipoteze:

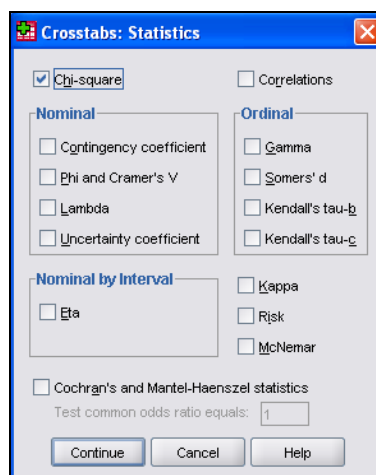
$$H_0 : \dots P_{ij} = P_{i\bullet} \cdot P_{\bullet j}, \quad \forall i, \forall j \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c$$

$$H_1 : \dots \exists P_{ij} \neq P_{i\bullet} \cdot P_{\bullet j}$$

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Descriptive statistics** i **Crosstabs**. U otvorenom prozoru bira se varijabla Spol u: **Rows** i druga varijabla Vrsta glazbe koja se sluša u: **Columns**. Aktiviranjem ikone **Statistics** otvara se prozor **Crosstabs: Statistics**, u kojem je potrebno aktivirati **Chi-square**. Navedeni prozor prikazan je na slici 3.14. Klikom na ikone **Continue** i **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.13. prikazani su podaci dvostruke statističke tablice prema nominalnim obilježjima spol i vrsta glazbe koja se sluša.

Slika 3.14.

### Prozor "Crosstabs:Statistics" s odabranim hi-kvadrat testom



Izvor: Simulirani podaci.

Tablica 3.13.

### Rezultati o spolu i vrsti glazbe koja se sluša za zadani uzorak ispitanika

Spol * Vrsta glazbe koja se slusa Crosstabulation						
Count		Vrsta glazbe koja se slusa				Total
		Klasicnu	Zabavnu	Narodnu	Ostalo	
Spol	Musko	7	36	8	42	93
	Zensko	9	85	2	48	144
Total		16	121	10	90	237

Izvor: Simulirani podaci.

U tablici 3.14. prikazani su podaci testiranja nezavisnosti dvaju nominalnih obilježja hi-kvadrat testom.



Tablica 3.14.

## Rezultati testiranja nezavisnosti dvaju nominalnih obilježja hi-kvadrat testom

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13,755 <sup>a</sup>	3	,003
Likelihood Ratio	13,869	3	,003
Linear-by-Linear Association	5,743	1	,017
N of Valid Cases	237		

a. 1 cells (12,5%) have expected count less than 5. The minimum expected count is 3,92.

Izvor: Simulirani podaci.

Prema rezultatima iz tablice 3.14 empirijska vrijednost  $\chi^2$  testa je:

$$\chi^2_* = \sum_{i=1}^r \sum_{j=1}^c \frac{(m_{ij} - e_{ij})^2}{e_{ij}} = 13,755.$$

Tablična vrijednost  $\chi^2$  - testa uz signifikantnost od 1% je:

$$\chi^2_{tab}[\alpha, df = (r-1) \cdot (c-1)] \Rightarrow [\alpha = 1\%; df = 3] = 11,34.$$

Vrijedi da je:  $\chi^2_* > \chi^2_{tab}$ ; što znači da se uz značajnost od 1% ne može prihvatiti početna pretpostavka da ne postoji ovisnost između spola i vrste glazbe koja se sluša u osnovnom skupu. Dakle, postoji statistički značajna ovisnost između spola i vrste glazbe koju slušaju zaposleni u promatranom poduzeću uz signifikantnost testa od 1%.

Prema tablici 3.14 može se vidjeti da je empirijska signifikantnost  $\alpha^* = 0,003 = 0,3\% \Rightarrow \alpha^* < 1\%$ , pa se donosi jednak zaključak o postojanju ovisnosti promatranih obilježja.

b) S obzirom da je donesen zaključak o odbacivanju hipoteze  $H_0$ , tj. donesen je zaključak o prihvatanju hipoteze da postoji ovisnost između spola i vrste glazbe koju slušaju zaposleni u promatranom poduzeću, izračunat je Pearsonov koeficijent kontigence. U programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Descriptive statistics** i **Crosstabs**. U otvorenom prozoru bira se varijabla Spol u: **Rows** i druga varijabla Vrsta glazbe koja se sluša u: **Columns**. Aktiviranjem ikone **Statistics** otvara se prozor **Crosstabs: Statistics**, u kojem je potrebno aktivirati **Chi-square** i **Contingency Coefficient**. Klikom na ikone **Continue** i **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. Rezultat je prikazan u tablici 3.15.

Tablica 3.15.

**Pearsonov koeficijent kontigence pri testiranju nezavisnosti dvaju nominalnih obilježja hi-kvadrat testom**

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	,234	,003
	N of Valid Cases	237	

Izvor: Simulirani podaci.

Pearsonov koeficijent kontigence je:  $C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{13,755}{13,755 + 237}} = 0,234$ .

Pri tom je empirijska signifikantnost  $\alpha^* = 0,003 = 0,3\% \Rightarrow \alpha^* < 1\%$ , što potvrđuje značajnost izračunatog koeficijenta.



### Primjer 3.7.

Nakon najave i uvođenja novog zakona o zabrani pušenja u Republici Hrvatskoj putem anketnog upitnika među zaposlenima u jednom poduzeću izvršeno je ispitivanje navike pušenja cigareta. Na temelju odabranog uzorka potrebno je ispitati postoji li ovisnost između spola i navike pušenja cigareta kod zaposlenih u promatranom poduzeću uz graničnu signifikantnost od 5%!



### Rješenje 3.7.

a) Da bi se donio zaključak o prihvaćanju hipoteze o tome postoji li ovisnost između spola i navike pušenja cigareta kod zaposlenih u promatranom poduzeću potrebno je postaviti hipoteze:

$$H_0 : \dots\dots\dots P_{ij} = P_{i\bullet} \cdot P_{\bullet j}, \quad \forall i, \forall j \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c$$

$$H_1 : \dots\dots\dots \exists P_{ij} \neq P_{i\bullet} \cdot P_{\bullet j}$$

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Descriptive statistics** i **Crosstabs**. U otvorenom prozoru bira se varijabla Spol u: **Rows** i druga varijabla Navika pušenja cigareta u: **Columns**. Aktiviranjem ikone **Statistics** otvara se prozor **Crosstabs: Statistics**, u kojem je potrebno aktivirati **Chi-square**. Klikom na ikone **Continue** i **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.16. prikazani su podaci

dvostruke statističke tablice prema nominalnim obilježjima Spol i Navika pušenja cigareta.

**Tablica 3.16.**

**Rezultati o spolu i navici pušenja cigareta zadanog uzorka ispitanika**

Spol * Navika pusenja cigareta Crosstabulation						
Count		Navika pusenja cigareta				
		Ne	Do 10 cigareta dnevno	Do 1 kutije dnevno	Vise od 1 kutije dnevno	Total
Spol	Musko	75	9	5	4	93
	Zensko	97	16	24	7	144
	Total	172	25	29	11	237

*Izvor: Simulirani podaci.*

U tablici 3.17. prikazani su podaci testiranja nezavisnosti dvaju nominalnih obilježja hi-kvadrat testom. Empirijska vrijednost  $\chi^2$  testa je:  $\chi^2* = \sum_{i=1}^r \sum_{j=1}^c \frac{(m_{ij} - e_{ij})^2}{e_{ij}} = 7,704$ .

**Tablica 3.17.**

**Rezultati testiranja nezavisnosti dvaju nominalnih obilježja hi-kvadrat testom**

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,409 <sup>a</sup>	3	,060
Likelihood Ratio	8,116	3	,044
Linear-by-Linear Association	4,805	1	,028
N of Valid Cases	237		

a. 1 cells (12,5%) have expected count less than 5. The minimum expected count is 4,32.

*Izvor: Simulirani podaci.*

Tablična vrijednost  $\chi^2$  - testa uz signifikantnost od 5% je:

$$\chi_{tab}^2[\alpha, df = (r-1) \cdot (c-1)] \Rightarrow [\alpha = 5\%; df = 3] = 7,815.$$

Vrijedi da je:  $\chi^2* < \chi_{tab}^2 \Rightarrow H_0$ ; što znači da se uz značajnost od 5% može prihvatiti početna pretpostavka da ne postoji ovisnost između spola i navike pušenja cigareta kod zaposlenih u promatranom poduzeću.

Prema tablici 3.17 može se vidjeti da je empirijska signifikantnost ovog testa  $\alpha^* = 0,06 = 6\% \Rightarrow \alpha^* > 5\% \Rightarrow H_0$ , pa se i na ovaj način donosi jednak zaključak o nezavisnosti promatranih obilježja.

**Primjer 3.8.**

Putem anketnog upitnika u jednom poduzeću izvršeno je ispitivanje zadovoljstva zaposlenih po različitim kategorijama. Na temelju odabranog uzorka potrebno je riješiti sljedeće postavke:

- Ispitati postoji li ovisnost između navike pušenja cigareta i konzumiranja alkohola zaposlenih u promatranom poduzeću uz graničnu signifikantnost od 5%!
- U slučaju prihvatanja hipoteze da postoji ovisnost između promatranih obilježja potrebno je izračunati Pearsonov koeficijent kontigence!
- U slučaju prihvatanja hipoteze da postoji ovisnost između promatranih obilježja potrebno je izračunati  $F_i$  koeficijent i Cramerov  $F_i$  koeficijent!

**Rješenje 3.8.**

- Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li ovisnost između navike pušenja cigareta i konzumiranja alkohola zaposlenih u promatranom poduzeću, potrebno je postaviti hipoteze:

$$H_0 : \dots\dots\dots P_{ij} = P_{i\bullet} \cdot P_{\bullet j}, \quad \forall i, \forall j \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c$$

$$H_1 : \dots\dots\dots \exists P_{ij} \neq P_{i\bullet} \cdot P_{\bullet j}$$

**Tablica 3.18.**

**Rezultati o navici pušenja cigareta i konzumiranju alkohola za zadani uzorak ispitanika**

Navika pušenja cigareta 1 * Konzumirate li alkoholna pica Crosstabulation						
Count		Konzumirate li alkoholna pica				
		Nikada	Samo uz obroke	U posebnim prilikama	Cesto i u svakoj prigodi	Total
Navika pušenja cigareta 1	Ne	51	6	109	5	171
	Do 10 cigareta dnevno	1	1	22	1	25
	Do 1 kutije dnevno	2	1	25	1	29
	Više od 1 kutije dnevno	0	0	9	2	11
	Total	54	8	165	9	236

*Izvor: Simulirani podaci.*

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost Hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinu padajućem izborniku **Descriptive statistics** i **Crosstabs**. U otvorenom prozoru bira se varijabla Navika pušenja cigareta u **Rows** i druga varijabla Konzumiranja alkohola u **Columns**. Aktiviranjem ikone **Statistics** otvara se prozor

**Crosstabs: Statistics** u kojem je potrebno aktivirati **Chi-square**. Klikom na ikone **Continue** i **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine. U tablici 3.18. prikazani su podaci dvostruke statističke tablice prema nominalnim obilježjima navika pušenja cigareta i konzumiranje alkohola.

U tablici 3.19. prikazani su podaci testiranja nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom.

**Tablica 3.19.**

**Rezultati testiranja nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom**

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	23,408 <sup>a</sup>	9	,005
Likelihood Ratio	26,286	9	,002
Linear-by-Linear Association	16,438	1	,000
N of Valid Cases	236		

a. 7 cells (43,8%) have expected count less than 5. The minimum expected count is ,37.

*Izvor: Simulirani podaci.*

Prema rezultatima iz tablice 3.19 empirijska vrijednost  $\chi^2$ -testa je:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(m_{ij} - e_{ij})^2}{e_{ij}} = 23,408.$$

Međutim ispod tablice 3.19 nalazi se napomena da je u 7 polja tablice čak 43,8% vrijednosti očekivanih frekvencija manje od 5. Stoga izračunata vrijednost hi-kvadrat testa nije pouzdana zbog opasnosti da se vrijednost  $\chi^2$ \* ne precijeni, pa je dopušteno "spajanje" nekih polja u tablici kontingence. To se radi "spajanjem" polja s najmanjim frekvencijama.

**Tablica 3.20.**

**Rezultati o navici pušenja cigareta 2 i konzumiranju alkohola za zadani uzorak ispitanika**

Navika pušenja cigareta 2 * Konzumirate li alkoholna pica Crosstabulation						
Count		Konzumirate li alkoholna pica				
		Nikada	Samo uz obroke	U posebnim prilikama	Cesto i u svakoj prigodi	Total
Navika pušenja cigareta 2	Ne	51	6	109	5	171
	Do 10 cigareta dnevno	1	1	22	1	25
	Do 1 kutije dnevno i više	2	1	34	3	40
	Total	54	8	165	9	236

*Izvor: Simulirani podaci.*

U ovom primjeru za obilježje navika pušenja cigareta 1 svi ispitanici koji konzumiraju "više od 1 kutije dnevno" priključeni su u skupinu "do jedne kutije na dan". Na taj način formirana je nova varijabla Navika pušenja cigareta 2. Sada je potrebno napraviti jednako testiranje s tom novom varijablom. Rezultati su prikazani u tablicama 3.20 i 3.21.

**Tablica 3.21.**

**Rezultati testiranja nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom**

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18,401 <sup>a</sup>	6	,005
Likelihood Ratio	22,350	6	,001
Linear-by-Linear Association	15,946	1	,000
N of Valid Cases	236		

a. 4 cells (33,3%) have expected count less than 5. The minimum expected count is ,85.

*Izvor: Simulirani podaci.*

Prema rezultatima iz tablice 3.21 empirijska vrijednost  $\chi^2$ -testa je:

$$\chi^{2*} = \sum_{i=1}^r \sum_{j=1}^c \frac{(m_{ij} - e_{ij})^2}{e_{ij}} = 18,401.$$

Tablična vrijednost  $\chi^2$ -testa uz signifikantnost od 5% je:

$$\chi_{tab}^2[\alpha, df = (r-1) \cdot (c-1)] \Rightarrow [\alpha = 5\%; df = 6] = 12,59.$$

Vrijedi da je:  $\chi^{2*} > \chi_{tab}^2$ ; što znači da se uz značajnost od 5% ne može prihvatiti početna pretpostavka  $H_0$  da ne postoji ovisnost navike pušenja cigareta i konzumiranja alkohola u osnovnom skupu. Dakle, postoji statistički značajna ovisnost između navike pušenja cigareta i konzumiranja alkohola zaposlenih u promatranom poduzeću. Da bi se potvrdio doneseni zaključak u ovom primjeru, može se napraviti dodatno spajanje ćelija (jer je najmanja očekivana frekvencija u tablici kontigence opet manja od 1 i iznosi 0,85).

Prema tablici 3.21 može se vidjeti da je empirijska signifikantnost  $\alpha^* = 0,005 = 0,5\% \Rightarrow \alpha^* < 5\%$ , pa se donosi jednak zaključak o postojanju ovisnosti promatranih obilježja.

b) S obzirom na to da je donesen zaključak o odbacivanju hipoteze  $H_0$ , tj. donesen je zaključak o prihvatanju hipoteze da postoji ovisnost između navike pušenja cigareta i konzumiranja alkohola zaposlenih u promatranom poduzeću, izračunat je Pearsonov koeficijent kontigence.

U programskom paketu **SPSS** napravljen je jednak postupak kao za izračun hi-kvadrat testa, samo je još u **Statistics** uz **Chi-square** aktiviran i **Contingency Coefficient**. Rezultat je prikazan u tablici 3.22. Pearsonov koeficijent kontigence je:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{18,401}{18,401 + 236}} = 0,269.$$

**Tablica 3.22.**

**Pearsonov koeficijent kontigence pri testiranju nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom**

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	,269	,005
	N of Valid Cases	236	

*Izvor: Simulirani podaci.*

Pri tom je empirijska signifikantnost  $\alpha^* = 0,005 = 0,5\% \Rightarrow \alpha^* < 5\%$ , što potvrđuje značajnost izračunatog koeficijenta.

c) S obzirom na to da je donesen zaključak o odbacivanju hipoteze  $H_0$ , tj. donesen je zaključak o prihvatanju hipoteze da postoji ovisnost između navike pušenja cigareta i konzumiranja alkohola zaposlenih u promatranom poduzeću, izračunati su  $F_i$  koeficijent i Cramerov  $F_i$  koeficijent.

U programskom paketu **SPSS** napravljen je jednak postupak kao za izračun hi-kvadrat testa, samo su još u **Statistics** uz **Chi-square** aktivirani i **Phi** i **Cramer's V**. Rezultat je prikazan u tablici 3.23.

**Tablica 3.23.**

**$F_i$  koeficijent i Cramerov  $F_i$  koeficijent pri testiranju nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom**

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,279	,005
	Cramer's V	,197	,005
	N of Valid Cases	236	

*Izvor: Simulirani podaci.*

Prema rezultatima u tablici 3.23  $F_i$  koeficijent je:  $\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{18,401}{236}} = 0,279$ .

Pri tom je empirijska signifikantnost  $\alpha^* = 0,005 = 0,5\% \Rightarrow \alpha^* < 5\%$ , što potvrđuje njegovu značajnost.

$$\text{Cramerov } \phi = \sqrt{\frac{\chi^2}{N(s-1)}} = \sqrt{\frac{18,401}{236(3-1)}} = 0,197.$$

Empirijska signifikantnost je  $\alpha^* = 0,005 = 0,5\% \Rightarrow \alpha^* < 5\%$ , što potvrđuje i njegovu statističku značajnost.



### Primjer 3.9.

U jednoj javnoj instituciji putem anonimne ankete među zaposlenima želi se ispitati postoji li ovisnost između spola i razmišljanja o davanju mita uz graničnu signifikantnost od 5%!



### Rješenje 3.9.

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li ovisnost između spola i razmišljanja o davanju mita kod zaposlenih u odabranoj javnoj instituciji potrebno je postaviti hipoteze:

$$H_0 : \dots P_{ij} = P_{i\cdot} \cdot P_{\cdot j}, \quad \forall i, \forall j \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c$$

$$H_1 : \dots \exists P_{ij} \neq P_{i\cdot} \cdot P_{\cdot j}$$

Tablica 3.24.

Rezultati o spolu i stajalištu o mitu za odabrani uzorak ispitanika

Spol * Mito biste dali Crosstabulation						
Count		Mito biste dali				
						Total
		Nikada	U izuzetnim okolnostima kada bi mi zdravije bilo ugroženo	Kada bi mi to pomoglo u rješavanju životnog problema	Uvijek kada je to najlakši i najsigurniji put do cilja	
Spol	Žensko	12	12	3	1	28
	Muško	8	6	4	4	22
	Total	20	18	7	5	50

Izvor: Simulirani podaci.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost Hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinu padajućem izborniku **Descriptive statistics** i **Crosstabs**. U otvorenom prozoru bira se varijabla Spol u **Rows** i druga varijabla Mito biste dali u **Columns**.



Aktiviranjem ikone **Statistics** otvara se prozor **Crosstabs: Statistics** u kojem je potrebno aktivirati **Chi-square**. Klikom na ikone **Continue** i **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine. U tablici 3.24 prikazani su podaci dvostruke statističke tablice prema nominalnim obilježjima spol i razmišljanje o mitu.

U tablici 3.25 prikazani su podaci testiranja nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom.

**Tablica 3.25.**

**Rezultati testiranja nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom**

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,082 <sup>a</sup>	3	,253
Likelihood Ratio	4,193	3	,241
Linear-by-Linear Association	2,404	1	,121
N of Valid Cases	50		

a. 4 cells (50,0%) have expected count less than 5. The minimum expected count is 2,20.

*Izvor: Simulirani podaci.*

Prema rezultatima iz tablice 3.25 empirijska vrijednost  $\chi^2$ -testa je:

$$\chi^2_{*} = \sum_{i=1}^r \sum_{j=1}^c \frac{(m_{ij} - e_{ij})^2}{e_{ij}} = 4,082.$$

Tablična vrijednost  $\chi^2$  - testa uz signifikantnost od 5% je:

$$\chi^2_{tab}[\alpha, df = (r-1) \cdot (c-1)] \Rightarrow [\alpha = 5\%; df = 3] = 7,815.$$

Vrijedi da je:  $\chi^2_{*} < \chi^2_{tab} \Rightarrow H_0$ ; što znači da se uz značajnost od 5% može prihvatiti početna pretpostavka da ne postoji statistički značajna ovisnost između spola i stava o mitu zaposlenih u promatranoj javnoj instituciji.

Prema podacima u tablici 3.25 može se vidjeti da je empirijska signifikantnost  $\alpha^{*} = 0,253 = 25,3\% \Rightarrow \alpha^{*} > 5\% \Rightarrow H_0$ , pa se donosi jednak zaključak o nezavisnosti promatranih obilježja.

Međutim, ispod tablice dana je napomena da u 4 polja tablice kontigence (50%) postoje očekivane frekvencije manje od 5 i da je najmanja očekivana frekvencija 2,2. U ovom slučaju, rezultate ovog testa treba uzeti s rezervom. Da bi se pokušao riješiti problem mogu se spojiti neka polja. To za obilježje spol nije moguće, jer je to alternativno obilježje (ima samo 2 modaliteta), pa spajanje ne bi imalo smisla. Mogu se spojiti neki modaliteti drugog obilježja u ovoj analizi, tj. stava ispitanika o mitu.

**Primjer 3.10.**

U jednom poduzeću putem anonimne ankete među zaposlenima želi se ispitati postoji li ovisnost između spola i religioznosti uz graničnu signifikantnost od 5%! Za zadani uzorak potrebno je izračunati  $F(\phi)$  koeficijent za promatrane varijable i testirati njegovu značajnost (Zaključak je potrebno donijeti pomoću empirijske signifikantnosti!)

**Rješenje 3.10.**

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li ovisnost između spola i religioznosti zaposlenika u promatranom poduzeću potrebno je postaviti hipoteze:

$$H_0 : \dots\dots\dots P_{ij} = P_{i\bullet} \cdot P_{\bullet j}, \quad \forall i, \forall j \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c$$

$$H_1 : \dots\dots\dots \exists P_{ij} \neq P_{i\bullet} \cdot P_{\bullet j}$$

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost Hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinu padajućem izborniku **Descriptive statistics** i **Crosstabs**. U otvorenom prozoru bira se varijabla Spol u **Rows** i druga varijabla Mito biste dali u **Columns**. Aktiviranjem ikone **Statistics** otvara se prozor **Crosstabs: Statistics** u kojem je potrebno aktivirati **Chi-square**, **Phi** and **Cramer's V**. Klikom na ikone **Continue** i **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine. U tablici 3.24 prikazani su podaci dvostruke statističke tablice prema nominalnim obilježjima spol i razmišljanje o mitu.

**Tablica 3.26.****Rezultati o spolu i religioznosti za odabrani uzorak ispitanika**

Spol * Jeste li religiozni Crosstabulation				
Count		Jeste li religiozni		
		da	ne	Total
Spol	Musko	77	14	91
	Zensko	118	22	140
Total		195	36	231

*Izvor: Simulirani podaci.*

U tablici 3.27 prikazani su podaci testiranja nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom.

**Tablica 3.27.****Rezultati testiranja nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom**

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,005 <sup>a</sup>	1	,946		
Continuity Correction <sup>b</sup>	,000	1	1,000		
Likelihood Ratio	,005	1	,946		
Fisher's Exact Test				1,000	,551
Linear-by-Linear Association	,005	1	,946		
N of Valid Cases	231				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14,18.  
b. Computed only for a 2x2 table

*Izvor: Simulirani podaci.*

Prema rezultatima iz tablice 3.27 empirijska vrijednost  $\chi^2$ -testa je:

$$\chi^2_{*} = \sum_{i=1}^r \sum_{j=1}^c \frac{(m_{ij} - e_{ij})^2}{e_{ij}} = 0,005.$$

Tablična vrijednost  $\chi^2$  - testa uz signifikantnost od 5% je:

$$\chi^2_{tab}[\alpha, df = (r-1) \cdot (c-1)] \Rightarrow [\alpha = 5\%; df = 1] = 3,841.$$

Vrijedi da je:  $\chi^2_{*} < \chi^2_{tab} \Rightarrow H_0$ ; što znači da se uz značajnost od 5% može prihvatiti početna pretpostavka da ne postoji ovisnost između spola i religioznosti zaposlenih u promatranom poduzeću.

Prema tablici 3.27 može se vidjeti da je empirijska signifikantnost  $\alpha^{*} = 0,946 = 94,6\% \Rightarrow \alpha^{*} > 5\% \Rightarrow H_0$ , pa se donosi jednak zaključak o nezavisnosti promatranih obilježja.

**Tablica 3.28.****Fi koeficijent i Cramerov Fi koeficijent pri testiranju nezavisnosti dvaju nominalnih obilježja Hi-kvadrat testom**

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,004	,946
	Cramer's V	,004	,946
	N of Valid Cases	231	

*Izvor: Simulirani podaci.*

Prema podacima u tablici 3.28  $Fi(\phi)$  koeficijent je:  $Fi = \sqrt{\frac{\chi^2}{N}} = 0,004$ .

Pri tom je njegova empirijska signifikantnost  $\alpha^* = 0,946 = 94,6\% \Rightarrow \alpha^* > 5\%$ , što vodi ka zaključku da koeficijent nije statistički značajan, što opet potvrđuje zaključak o nezavisnosti promatranih obilježja.

### 3.6 Testiranje hipoteze da distribucija ima određeni oblik

Pomoću Hi-kvadrat testa mogu se testirati hipoteze je li neka promatrana distribucija ima oblik po nekom teorijskom zakonu. U tom smislu mogu se vršiti testiranja ima li distribucija oblik jednolike, binomne, Poissonove ili normalne distribucije (mnogi statistički testovi za testiranje hipoteza pretpostavljaju da je neka konkretna varijabla distribuirana po normalnom zakonu).

Postavljaju se hipoteze:

- za testiranje **Jednolike distribucije**:  
 $H_0 : \dots\dots\dots P_1 = P_2 = \dots = P_k = P$   
 $H_1 : \dots\dots\dots \exists P_i \neq P$
- za testiranje **Binomne distribucije**:  
 $H_0 : \dots\dots\dots X \sim B(n, p)$   
 $H_1 : \dots\dots\dots X \not\sim B(n, p)$
- za testiranje **Poissonove distribucije**:  
 $H_0 : \dots\dots\dots X \sim P(\mu)$   
 $H_1 : \dots\dots\dots X \not\sim P(\mu)$
- za testiranje **Normalne distribucije**:  
 $H_0 : \dots\dots\dots X \sim N(\mu, \sigma)$   
 $H_1 : \dots\dots\dots X \not\sim N(\mu, \sigma)$

Empirijska vrijednost Hi-kvadrat testa za svako od ovih testiranja je:

$$\chi^2_* = \sum_{i=1}^k \frac{(f_i - f_{ii})^2}{f_{ii}}, \quad (3.21)$$

gdje su:

$f_i$  - originalne frekvencije (iz distribucije uzorka)

$f_{ti}$  - teorijske frekvencije koje se izračunavaju pod pretpostavkom početne ili nulte hipoteze, tj. da distribucija ima oblik neke teorijske distribucije.

Teorijske frekvencije su:

$$f_{ti} = p(x_i) \cdot \sum_{i=1}^k f_i, \quad (3.22)$$

gdje je:

$\sum_{i=1}^k f_i$  - suma originalnih frekvencija (iz distribucije uzorka),

$p(x_i)$  - vjerojatnost da slučajna varijabla  $X$  poprimi vrijednost  $x_i$  prema teorijskom zakonu, koji se pretpostavlja u nultoj  $H_0$  hipotezi.

Vjerojatnost da slučajna varijabla  $X$  poprimi vrijednost  $x_i$  prema **Jednolikoj distribuciji**:

$$p(x_i) = \frac{1}{n}. \quad (3.23)$$

Vjerojatnost da slučajna varijabla  $X$  poprimi vrijednost  $x_i$  prema **Binomnoj distribuciji**:

$$p(X = x) = \binom{n}{x} \cdot p^x \cdot q^{n-x}. \quad (3.24)$$

Vjerojatnost da slučajna varijabla  $X$  poprimi vrijednost  $x_i$  prema **Poissonovoj distribuciji**:

$$p(X = x) = \frac{\mu^x \cdot e^{-\mu}}{x!}. \quad (3.25)$$

Vjerojatnost da slučajna varijabla  $X$  poprimi vrijednost na intervalu između  $x_1$  i  $x_2$  prema **Normalnoj distribuciji** u nekim slučajevima prema položaju ispod normalne krivulje može biti:

$$p(x_1 < X \leq x_2) = p_2\left(\frac{x_2 - \mu}{\sigma}\right) - p_1\left(\frac{x_1 - \mu}{\sigma}\right). \quad (3.26)$$

Tablična vrijednost se traži iz tablica Hi-kvadrat distribucije (Tablice C1 i C2):  $\chi_{tab}^2[\alpha, df]$ , uz odgovarajući nivo signifikantnosti  $\alpha$  i stupnjeve slobode  $df$ , koji se računaju u ovisnosti o testiranoj distribuciji:

- za **jednoliku distribuciju**:  $df = k - 1$ ,
- za **binomnu distribuciju**:  $df = k - 2$ ,
- za **Poissonovu distribuciju**:  $df = k - 2$ ,
- za **normalnu distribuciju**:  $df = k - 3$ ,

gdje "k" predstavlja broj frekvencija.

Zaključak se donosi na način da se uspoređi Hi-kvadrat empirijska i tablična vrijednost:

$\chi^2 * < \chi^2_{tab} \Rightarrow H_0$ ; što znači da konkretna distribucija ima oblik teorijske distribucije (jednolike, binomne, Poissonove ili normalne) koja se testira, a u suprotnom se ta početna hipoteza odbacuje.

Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $\chi^*$  (Tablice C1 i C2): ako je  $\alpha^* > 5\% \Rightarrow H_0$ , dok se u suprotnom ta hipoteza odbacuje.

### 3.6.1 Kolmogorov - Smirnov test

Testiranje hipoteze ima li neka promatrana distribucija oblik po nekom teorijskom zakonu može se izvršiti i pomoću Kolmogorov-Smirnovog testa.

Postavljaju se hipoteze:

- za testiranje **Jednolike distribucije**:  
 $H_0 : \dots\dots\dots P_1 = P_2 = \dots = P_k = P$   
 $H_1 : \dots\dots\dots \exists P_i \neq P$
- za testiranje **Binomne distribucije**:  
 $H_0 : \dots\dots\dots X \sim B(n, p)$   
 $H_1 : \dots\dots\dots X \not\sim B(n, p)$
- za testiranje **Poissonove distribucije**:  
 $H_0 : \dots\dots\dots X \sim P(\mu)$   
 $H_1 : \dots\dots\dots X \not\sim P(\mu)$
- - za testiranje **Normalne distribucije**:  
 $H_0 : \dots\dots\dots X \sim N(\mu, \sigma)$   
 $H_1 : \dots\dots\dots X \not\sim N(\mu, \sigma)$

Ovaj test se temelji na najvećoj apsolutnoj razlici (ili diferenciji)  $D$  između empirijskih (iz uzorka) kumulativnih frekvencija i očekivanih kumulativnih frekvencija (koje se računaju uz pretpostavku  $H_0$  hipoteze):

$$D = \max_x |Ft_i(x) - F_i(x)|, \quad (3.26)$$

gdje su:

$Ft_i$  - kumulativne teorijske frekvencije,

$F_i$  - kumulativne empirijske frekvencije.

U programskom paketu **SPSS** kao rezultat ovog testiranja dobije se empirijska signifikantnost. Ako je  $\alpha^* > 5\% \Rightarrow H_0$ , tj. prihvaća se pretpostavka da zadana empirijska distribucija ima teorijski oblik koji se testira (jednoliki, binomni, Poissonov ili normalni).



### Primjer 3.11.

Na fakultetu "E" izvršeno je anonimno istraživanje anketnim upitnikom među studentskom populacijom. Potrebno je riješiti sljedeće postavke:

- Pomoću  $\chi^2$ - testa potrebno je ispitati može li se prihvatiti pretpostavka da džeparac studentske populacije na promatranom fakultetu "E" ima oblik jednolike distribucije uz graničnu signifikantnost od 5%!
- Testiranje pod (a) potrebno je napraviti i pomoću Kolmogorov - Smirnov testa!



### Rješenje 3.11.

- Da bi se donio zaključak o prihvatanju hipoteze o tome da džeparac studentske populacije na promatranom fakultetu "E" ima oblik jednolike distribucije (tj. da svi studenti imaju jednoliko raspoređen džeparac) uz signifikantnost od 5%, potrebno je postaviti hipoteze:

$$H_0 : \dots\dots\dots P_1 = P_2 = \dots P_i = \dots = P_k = P$$

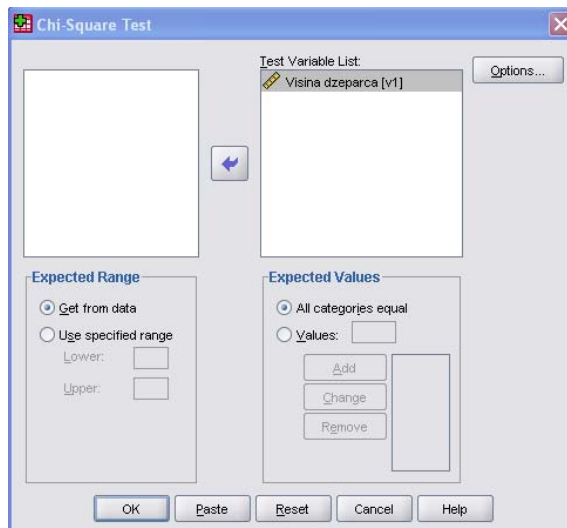
$$H_1 : \dots\dots\dots \exists P_i \neq P \quad i = 1, 2, \dots, k$$

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost Hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** a u pomoćnom izborniku **Chi-Square**. U otvorenom prozoru bira se varijabla Visina džeparca (v1) u **Test Variable**

**List.** U *Expected Values* već je aktivirano *all categories equal* kako je prikazano na slici 3.15.

**Slika 3.15.**

**Prozor "Chi-Square Test" s odabranom varijablom**



*Izvor: Simulirani podaci.*

Klikom na ikonu **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine. U tablici 3.29 prikazani su rezultati Hi-kvadrat testa.

**Tablica 3.29.**

**Hi-kvadrat test hipoteze da zadana distribucija ima jednoliki oblik**

Test Statistics <sup>a</sup>	
	Visina dzeparca
Chi-Square	266,023 <sup>a</sup>
df	26
Asymp. Sig.	,000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6,4.

*Izvor: Simulirani podaci.*

Prema rezultatima iz tablice 3.29 empirijska vrijednost  $\chi^2$  testa je:

$$\chi^2_{*} = \sum_{i=1}^k \frac{(f_i - f_{ti})^2}{f_{ti}} = 266,023.$$

Tablična vrijednost  $\chi^2$  - testa uz signifikantnost od 5% je:



$$\chi^2_{tab}[\alpha = 5\%; df = 26] = 38,89.$$

Vrijedi da je:  $\chi^2_* > \chi^2_{tab}$ ; što znači da se uz značajnost od 5% ne može prihvatiti početna pretpostavka  $H_0$  da džeparac studentske populacije na promatranom fakultetu "E" ima oblik jednolike distribucije, tj. da je jednoliko raspoređen. Dakle, postoji značajna razlika u raspodjeli džeparca studenata na promatranom fakultetu.

Također se prema tablici 3.29 može vidjeti da je empirijska signifikantnost  $\alpha^* \approx 0\% \Rightarrow \alpha^* < 5\%$ , pa se donosi isti zaključak o nejednako raspoređenom džeparcu studenata na promatranom području.

b) Da bi se donio zaključak o prihvatanju hipoteze o tome da džeparac studentske populacije na fakultetu "E" ima oblik jednolike distribucije (tj. da studenti imaju jednoliko raspoređen džeparac) uz signifikantnost od 5% pomoću Kolmogorov-Smirnov testa, potrebno je postaviti hipoteze:

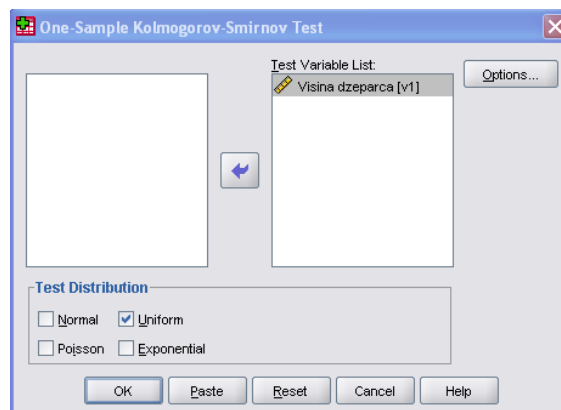
$$H_0 : \dots\dots\dots P_1 = P_2 = \dots P_i = \dots = P_k = P$$

$$H_1 : \dots\dots\dots \exists P_i \neq P \quad i = 1, 2, \dots, k$$

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju čega će se donijeti odgovarajući zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinu padajućem izborniku **1-Sample K-S**.

**Slika 3.16.**

**Prozor One-Sample Kolmogorov-Smirnov Test s odabranom varijablom i Uniform distribucijom**



*Izvor: Simulirani podaci.*

U otvorenom prozoru bira se varijabla Visina džeparca (v1) u **Test Variable List**. U **1-Sample K-S** potrebno je aktivirati **Uniform** kako je prikazano na slici 3.16.

Tablica 3.30.

## Kolmogorov-Smirnov test za jednoliku distribuciju

One-Sample Kolmogorov-Smirnov Test		
		Visina džeparca
N		173
Uniform Parameters <sup>a</sup>	Minimum	,00
	Maximum	3500,00
Most Extreme Differences	Absolute	,731
	Positive	,731
	Negative	-,006
Kolmogorov-Smirnov Z		9,610
Asymp. Sig. (2-tailed)		,000
a. Test distribution is Uniform.		

Izvor: Simulirani podaci.

Klikom na ikonu **OK** u **Outputu** programa **SPSS** dobije se empirijska signifikantnost. U tablici 3.30 prikazan je rezultat. Na temelju dobivenih rezultata prema tablici 3.30 može se zaključiti da je:  $\alpha^* \approx 0 \Rightarrow \alpha^* < 5\%$ , tj. odbacuje se početna hipoteza da distribucija džeparca ima jednoliki oblik. Dakle, studenti na fakultetu "E" nemaju jednoliko raspoređen džeparac.



## Primjer 3.12.

Na fakultetu "E" napravljeno je istraživanje na studentskoj populaciji pomoću anketnog upitnika. Potrebno je pomoću Kolmogorov-Smirnov testa ispitati može li se prihvatiti pretpostavka da visina studenata na promatranom fakultetu ima oblik normalne distribucije uz graničnu signifikantnost od 5%!



## Rješenje 3.12.

a) Da bi se donio zaključak o prihvatanju hipoteze o tome da visina studenata na fakultetu "E" ima oblik normalne distribucije uz graničnu signifikantnost od 5% pomoću Kolmogorov-Smirnov testa, potrebno je postaviti hipoteze:

$$H_0 : \dots\dots\dots X \sim N(\mu, \sigma)$$

$$H_1 : \dots\dots\dots X \not\sim N(\mu, \sigma)$$

U programskom paketu **SPSS-a** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju čega će se donijeti odgovarajući zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **1-Sample K-S**. U otvorenom prozoru bira se varijabla Visina u cm (v1) u **Test Variable List**. U **1-**

**Sample K-S** potrebno je aktivirati **Normal**. Klikom na ikonu **OK** u **Outputu** programa **SPSS** dobije se empirijska signifikantnost. U tablici 3.31 prikazan je rezultat.

**Tablica 3.31.**

**Kolmogorov-Smirnov test za normalnu distribuciju**

One-Sample Kolmogorov-Smirnov Test		
		Visina u cm
N		235
Normal Parameters <sup>a</sup>	Mean	175,77
	Std. Deviation	8,749
Most Extreme Differences	Absolute	,096
	Positive	,096
	Negative	-,051
Kolmogorov-Smirnov Z		1,478
Asymp. Sig. (2-tailed)		,025
a. Test distribution is Normal.		

*Izvor: Simulirani podaci.*

Na temelju dobivenih rezultata iz tablice 3.31 može se zaključiti da je empirijske signifikantnost:  $\alpha^* = 0,025 = 2,5\% \Rightarrow \alpha^* < 5\%$ , tj. odbacuje se početna hipoteza da distribucija visine studenata na fakultetu "E" ima normalan oblik. Dakle, uz signifikantnost testa od 5% zaključuje se da visina studenata na promatranom fakultetu nema oblik normalne distribucije.



**Primjer 3.13.**

Na fakultetu "E" napravljeno je istraživanje na studentskoj populaciji pomoću anketnog upitnika. Pomoću  $\chi^2$ -testa potrebno je ispitati može li se na osnovu zadanog uzorka prihvatiti pretpostavka da je težina studenata na promatranom fakultetu jednolike (uniformne) distribucije uz graničnu signifikantnost testa od 5%!

Testiranje je potrebno izvršiti i pomoću Kolmogorov-Smirnov testa!



**Rješenje 3.13.**

Da bi se donio zaključak o prihvatanju hipoteze o tome da težina studenata na fakultetu "E" ima oblik jednolike distribucije (tj. da svi studenti imaju ujednačeno raspoređenu težinu) uz signifikantnost od 5% potrebno je postaviti hipoteze:

$$H_0 : \dots\dots\dots P_1 = P_2 = \dots = P_k = P$$

$$H_1 : \dots\dots\dots \exists P_i \neq P$$

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost Hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** a u pomoćnom izvorniku **Chi-Square**. U otvorenom prozoru bira se varijabla Težina u kg (v1) u **Test Variable List**. U **Expected Values** već je aktivirano *all categories equal*. Klikom na ikonu **OK** u **Outputu** programa **SPSS** dobiju se tražene veličine. U tablici 3.32 prikazani su rezultati Hi-kvadrat testa.

**Tablica 3.32.**

**Hi-kvadrat test hipoteze da zadana distribucija ima jednoliki oblik**

Test Statistics <sup>a</sup>	
	Težina u kg
Chi-Square	153,761 <sup>a</sup>
df	52
Asymp. Sig.	,000

a. 53 cells (100,0%) have expected frequencies less than 5. The minimum expected cell frequency is 4,4.

*Izvor: Simulirani podaci.*

Prema dobivenim rezultatima empirijska vrijednost  $\chi^2$  testa je:

$$\chi^2* = \sum_{i=1}^k \frac{(f_i - f_{ii})^2}{f_{ii}} = 153,761 .$$

Tablična vrijednost  $\chi^2$  - testa uz signifikantnost od 5% je:

$$\chi_{tab}^2[\alpha = 5\%; df = 52] = 67,5 .$$

Vrijedi da je:  $\chi^2* > \chi_{tab}^2$  ; što znači da se uz značajnost od 5% ne može prihvatiti početna pretpostavka  $H_0$  da težina studenata na fakultetu "E" ima oblik jednolike distribucije, tj. da je jednoliko raspoređena. Dakle, postoji značajna razlika u rasporedu težine studenata na promatranom fakultetu.

Također prema tablici u outputu može se vidjeti da je empirijska signifikantnost  $\alpha^* \approx 0\% \Rightarrow \alpha^* < 5\%$  , pa se donosi isti zaključak o nejednako raspoređenoj težini studenata na fakultetu "E".

Da bi se donio zaključak o prihvaćanju navedenih hipoteza pomoću Kolmogorov-Smirnov testa potrebno je u programskom paketu **SPSS** odabrati **Analyze**, a na njezinu padajućem izborniku **Nonparametric Tests** na pomoćnom izborniku **1-Sample K-S**, gdje je potrebno je aktivirati **Uniform**. Klikom na ikonu **OK** u **Outputu**

programa **SPSS** dobije se empirijska signifikantnost. U tablici 3.33 prikazan je rezultat.

**Tablica 3.33.**

**Kolmogorov-Smirnov test za jednoliku distribuciju**

One-Sample Kolmogorov-Smirnov Test		
		Tezina u kg
N		234
Uniform Parameters <sup>a</sup>	Minimum	46,00
	Maximum	110,00
Most Extreme Differences	Absolute	,323
	Positive	,323
	Negative	-,054
Kolmogorov-Smirnov Z		4,948
Asymp. Sig. (2-tailed)		,000
a. Test distribution is Uniform.		

*Izvor: Simulirani podaci.*

Na temelju dobivenih rezultata prema tablici 3.33 može se zaključiti da je:  $\alpha^* \approx 0 \Rightarrow \alpha^* < 5\%$ , tj. odbacuje se početna hipoteza da distribucija težine studenata ima jednoliki oblik. Dakle uz signifikantnost testa od 5% studenti na fakultetu "E" nemaju jednoliku raspoređenu težinu.

## 3.7 Testiranje hipoteza sa zavisnim uzorcima

### 3.7.1 McNemarov test za dva zavisna uzorka

Ako se uspoređuju rezultati iste grupe ispitanika "prije" i "poslije", ili se ista grupa ispitanika uspoređuje u 2 različite aktivnosti, ispituje se korelacija između 1. i 2. rezultata pomoću hi-kvadrat McNemarovog testa.

Kod ovog testiranja hipoteze su:

$H_0$  - ne postoji razlika u rezultatima ispitanika u aktivnosti 1. i aktivnosti 2.

$H_1$  - postoji razlika u rezultatima ispitanika u aktivnosti 1. i aktivnosti 2.

**Tablica 3.34.**

**Rezultati mjerenja aktivnosti ispitanika "prije" i "poslije" kod McNemarovog testa**

Aktivnost 1	Aktivnost 2		
		Zadovoljili	Nisu zadovoljili
	Zadovoljili	B	A
	Nisu zadovoljili	D	C

*Izvor: Konstrukcija autora.*

Postupak testiranja je prikazan u tablici. Razlika između 1. i 2. testa je u poljima A i D. U poljima B i C su oni koji su ili uspjeli ili nisu uspjeli u oba testa. Dakle, A+D je ukupan broj onih ispitanika kod kojih se ne slaže uspjeh kod 1. i 2. mjerenja, odnosno onih koji su promijenili svoj uspjeh.

Empirijska vrijednost McNemarovog hi-kvadrat testa je:

$$\chi^2_* = \frac{(A - D)^2}{A + D}. \quad (3.27)$$

Ako je  $(A+D) < 20$ , uz Yatesovu korekciju, vrijedi da je:

$$\chi^2_* = \frac{(|A - D| - 1)^2}{A + D}. \quad (3.28)$$

Tablična vrijednost  $\chi^2$ - testa traži se iz tablica Hi-kvadrat distribucije (tablice C1 i C2):

$$\chi^2_{tab}[\alpha, df = (r - 1) \cdot (c - 1)] \quad (3.29)$$

uz odgovarajući nivo signifikantnosti  $\alpha$  i stupnjeve slobode  $df$ .

Zaključak se donosi na način da se uspoređi Hi-kvadrat empirijska i tablična vrijednost:

$\chi^2_* < \chi^2_{tab} \Rightarrow H_0$ ; što znači da ne postoji razlika u rezultatima ispitanika u aktivnosti 1. i aktivnosti 2., dok se u suprotnom ta početna hipoteza odbacuje.

Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $\chi^*$  (Tablica C1 i C2): ako je  $\alpha^* > 5\% \Rightarrow H_0$ , dok se u suprotnom ta hipoteza odbacuje.

**Primjer 3.14.**

Nakon održana 2 edukativna management seminara u jednoj korporaciji na uzorku od 100 ispitanika izvršeno je testiranje usvojenog znanja za svaki seminar s 2 različita testa. Zadatak je utvrditi postoji li razlika u prolaznosti ispitanika između ovih testova.

**Rješenje 3.14.**

a) Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u prolaznosti između navedenih testova za zaposlenike u promatranoj korporaciji potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rezultatima ispitanika u testu 1. i testu 2.

$H_1$  - postoji razlika u rezultatima ispitanika u testu 1. i testu 2.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Descriptive statistics** i **Crosstabs**. U otvorenom prozoru bira se varijabla Test 1 u: **Rows** i druga varijabla Test 2 u: **Columns**. Aktiviranjem ikone **Statistics** otvara se prozor **Crosstabs: Statistics**, u kojem je potrebno aktivirati **McNemar**.

Klikom na ikone **Continue** i **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.35. prikazani su podaci dvostruke statističke tablice prema rezultatima kandidata na testu 1 i 2.

**Tablica 3.35.****Rezultati Testa 1 i Testa 2 za zadani uzorak ispitanika**

Test 1 * Test 2 Crosstabulation				
Count		Test 2		
		zadovoljili	nisu zadovoljili	Total
Test 1	zadovoljili	55	5	60
	nisu zadovoljili	15	25	40
	Total	70	30	100

*Izvor: Simulirani podaci.*

Rezultati McNemarovog hi-kvadrat testa su prikazani u tablici 3.36.

Tablica 3.36.

## McNemarov hi-kvadrat test za zavisne uzorke

Chi-Square Tests		
	Value	Exact Sig. (2-sided)
McNemar Test		,041 <sup>a</sup>
N of Valid Cases	100	

a. Binomial distribution used.

Izvor: Simulirani podaci.

Prema rezultatima iz tablice 3.36 vrijedi da je empirijska signifikantnost  $\alpha^* = 0,041 = 4,1\% \Rightarrow \alpha^* < 5\%$ , pa se donosi zaključak o odbacivanju početne hipoteze, tj. postoji statistički značajna razlika u rezultatima ispitanika na testu 1. i testu 2.

Na temelju izraza (3.28) uvažavajući Yatesovu korekciju, jer je  $(A+D)=20$ , vrijedi da je empirijska vrijednost McNemarovog hi-kvadrat testa:

$$\chi^2_* = \frac{(|A-D|-1)^2}{A+D} \Rightarrow \chi^2_* = \frac{(10-1)^2}{20} = 4,05.$$

Tablična vrijednost hi-kvadrat testa (Tablice C1 i C2) uz signifikantnost od 5% je:

$$\chi^2_{tab}[\alpha; df = (r-1) \cdot (c-1)] \Rightarrow \chi^2_{tab}[\alpha = 5\%; df = 1 \cdot 1 = 1] = 3,841.$$

Vrijedi da je:  $\chi^2_* > \chi^2_{tab}$  što opet znači da postoji razlika u prolaznosti kod testa 1. i testa 2., tj. prema rezultatima iz tablice 3.35 može se zaključiti da je test 2. lakši, odnosno da je pripadni edukacijski seminar bio učinkovitiji.

**Primjer 3.15.**

U poduzeću "P" izvršeno je ispitivanje na uzorku od 70 zaposlenika o njihovom zadovoljstvu na poslu 10 dana prije i 5 dana poslije dijeljenja božićnice. Zadatak je utvrditi postoji li razlika u zadovoljstvu zaposlenika na poslu prije i poslije božićnice.

**Rješenje 3.15.**

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u zadovoljstvu zaposlenika na poslu prije i poslije dijeljenja božićnice u promatranom poduzeću potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rezultatima ispitanika prije i poslije.



$H_1$  - postoji razlika u rezultatima ispitanika prije i poslije.

U programskom paketu **SPSS** potrebno odabrati **Analyze; Descriptive statistics** i **Crosstabs**.

- varijabla Zadovoljstvo na poslu prije božićnice u: **Rows** i

- varijabla Zadovoljstvo na poslu nakon božićnice u: **Columns**.

Aktiviranjem ikone **Statistics**; u prozoru **Crosstabs: Statistics**; aktivirati **McNemar**.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku vrijednost hi-kvadrat testa. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Descriptive statistics** i **Crosstabs**. U otvorenom prozoru bira se varijabla Zadovoljstvo na poslu prije božićnice u: **Rows** i druga varijabla Zadovoljstvo na poslu nakon božićnice u: **Columns**. Aktiviranjem ikone **Statistics** otvara se prozor **Crosstabs: Statistics**, u kojem je potrebno aktivirati **McNemar**.

Klikom na ikone **Continue** i **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.37 prikazani su podaci dvostruke statističke tablice prema rezultatima o zadovoljstvu zaposlenih na poslu prije i poslije božićnice za zadani uzorak ispitanika.

**Tablica 3.37.**

**Rezultati o zadovoljstvu na poslu prije i poslije božićnice za zadani uzorak ispitanika**

Zadovoljstvo na poslu prije božićnice * Zadovoljstvo na poslu nakon božićnice Crosstabulation				
Count		Zadovoljstvo na poslu nakon božićnice		
		zadovoljan/na	nije zadovoljan/na	Total
Zadovoljstvo na poslu prije božićnice	zadovoljan/na	32	8	40
	nije zadovoljan/na	20	10	30
Total		52	18	70

Izvor: Simulirani podaci.

Rezultati McNemarovog hi-kvadrat testa su prikazani u tablici 3.38.

Uvažavajući Yatesovu korekciju (sadržanu u izračunu SPSS-a) vrijedi da je empirijska vrijednost McNemarovog hi-kvadrat testa:

$$\chi^2* = \frac{(|A - D| - 1)^2}{A + D} \Rightarrow \chi^2* = \frac{(12 - 1)^2}{28} = 4,321.$$

Tablična vrijednost hi-kvadrat testa (Tablice C1 i C2) uz signifikantnost od 5% je:

$$\chi^2_{tab}[\alpha; df = (r-1) \cdot (c-1)] \Rightarrow \chi^2_{tab}[\alpha = 5\%; df = 1 \cdot 1 = 1] = 3,841.$$

**Tablica 3.38.**

**McNemarov hi-kvadrat test za zavisne uzorke**

Chi-Square Tests		
	Value	Exact Sig. (2-sided)
McNemar Test		,036 <sup>a</sup>
N of Valid Cases	70	
a. Binomial distribution used.		

Izvor: Simulirani podaci.

Vrijedi da je:  $\chi^2_* > \chi^2_{tab}$  što opet znači da postoji razlika u zadovoljstvu na poslu kod zaposlenika prije i poslije dijeljenja božićnice, tj. prema podacima u tablici 3.37 može se zaključiti da je zadovoljstvo nakon božićnice poraslo.

Prema rezultatima iz tablice 3.38 vrijedi da je empirijska signifikantnost  $\alpha^* = 0,038 = 3,8\% \Rightarrow \alpha^* < 5\%$ , pa se donosi zaključak o odbacivanju početne hipoteze, tj. uz signifikantnost testa od 5% zaključuje se da postoji statistički značajna razlika u zadovoljstvu na poslu kod zaposlenika prije i poslije dijeljenja božićnice.

### 3.7.2 Friedman test za više od dva zavisna uzorka

Ovaj test *primjenjuje se za više od dva zavisna uzorka varijabli koje se mjere pomoću redoslijedne skale*. Naime, ako se na istoj grupi ispitanika vrši mjerenje u različitim uvjetima kaže se da se radi o zavisnim uzorcima. Može se reći da se ovo testiranje temelji na testu analize varijance, gdje se umjesto brojčanih mjernih podataka koriste rangovi. Postavljaju se hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

Test se provodi na sljedeći način:

1. Potrebno je sve podatke razvrstati u  $N$  redova (što odgovara broju ispitanika) i  $k$  stupaca (svaki stupac predstavlja jedan zavisni uzorak, tj. jedno ponovljeno mjerenje). Svaki redak odnosi se na istog ispitanika.

Potrebno je rangirati rezultate za svaki redak, tj. za svakog ispitanika. (Ako postoji više podataka s istom vrijednosti u retku potrebno im je dodijeliti jednak rang na način da se računa njihov prosječni rang!)

Potrebno je izračunati zbroj rangova u svakom uzorku (stupcu):  $T_i$ .

2. Potrebno je izračunati empirijsku vrijednost Friedmanovog testa:  $\chi_r^{2*}$ :

$$\chi_r^{2*} = \frac{12}{N k(k+1)} \cdot \left( \sum_{i=1}^k T_i^2 \right) - 3N(k+1), \quad (3.30)$$

gdje je:

- $N$  - ukupan broj ispitanika (podataka),
- $T_i$  - zbroj rangova u svakom uzorku,
- $k$  - broj uzoraka, tj. broj ponovljenih mjerenja.

3. Ako su broj ispitanika (podataka) i broj uzoraka dovoljno veliki vrijednost Friedmanovog testa  $\chi_r^{2*}$  ima jednaku distribuciju kao i hi-kvadrat, pa se zaključak donosi na način da se empirijska vrijednost testa usporedi s tabličnom vrijednosti testa:

$$\chi_{tab}^{2[\alpha]}[df=k-1], \quad (3.31)$$

gdje je:

- $\alpha$  - granična razina signifikantnosti,
- $df$  - stupnjevi slobode.

Zaključuje se usporedbom:  $\chi_r^{2*} < \chi_{tab} \Rightarrow H_0$ , odnosno ako je  $\alpha^* > 5\% \Rightarrow H_0$ , tj. prihvaća se pretpostavka da ne postoji razlika u rangovima odabranih uzoraka, tj. ponovljenih mjerenja, odnosno da zavisni uzorci nisu različiti.

4. Ako su broj ispitanika (podataka)  $N$  i broj uzoraka  $k$  mali, vrijednost Friedmanovog testa  $\chi_r^{2*}$  nema jednaku distribuciju kao i hi-kvadrat, pa se zaključak donosi pomoću posebnih tablica.



### Primjer 3.16.

Na temelju provedenog anketnog upitnika na uzorku od 20 studenata fakulteta "E" potrebno je pomoću Friedmanovog testa ispitati može li se prihvatiti pretpostavka da

ne postoji razlika u prosječnoj ocjeni studenata na I, II, III i IV godini studija uz graničnu signifikantnost testa od 5%!



### Rješenje 3.16.

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u prosječnoj ocjeni na I, II, III i IV godini studija između 20 odabranih studenata uz graničnu signifikantnost testa od 5% potrebno je postaviti hipoteze:

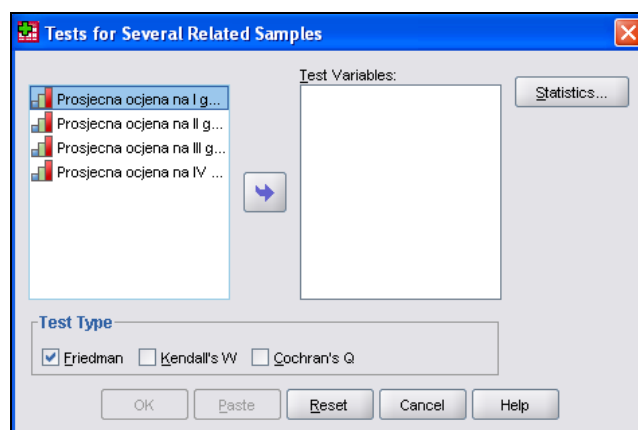
$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **K Related Samples**. U otvorenom prozoru biraju se varijable Prosječna ocjena na I godini studija (v1), Prosječna ocjena na II godini studija (v2), Prosječna ocjena na III godini studija (v3), Prosječna ocjena na IV godini studija (v4) u: **Test Variable List**. Na slici 3.17 prikazan je prozor **Tests for Several Related Samples** s odabranim veličinama za Friedmanov test.

### Slika 3.17.

Prozor "Tests for Several Related Samples" s odabranim veličinama za Friedmanov test



Izvor: Simulirani podaci.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine.

**Tablica 3.39.****Prosječni rangovi ocjena ispitanika na različitoj godini studija**

Ranks	
	Mean Rank
Prosječna ocjena na I godini studija	1,92
Prosječna ocjena na II godini studija	2,25
Prosječna ocjena na III godini studija	2,98
Prosječna ocjena na IV godini studija	2,85

*Izvor: Simulirani podaci.*

U tablici 3.39 prikazani su odgovarajući prosječni rangovi uzorka studenata na različitim godinama studija.

U tablici 3.40 prikazani su odgovarajući rezultati Friedmanovog testa za zadani uzorak ispitanika na različitim godinama studija.

**Tablica 3.40.****Rezultati Friedmanovog testa za zadani uzorak ispitanika**

Test Statistics <sup>a</sup>	
N	20,000
Chi-Square	9,123
df	3,000
Asymp. Sig.	,028
a. Friedman Test	

*Izvor: Simulirani podaci.*

Prema dobivenim rezultatima iz tablice 3.40 može se vidjeti da je empirijska vrijednost Friedmanovog testa za zadani uzorak ispitanika:

$$\chi_r^{2*} = \frac{12}{Nk(k+1)} \cdot \left( \sum_{i=1}^k T_i^2 \right) - 3N(k+1) = 9,123,$$

a tablična vrijednost odgovarajuće hi-kvadrat distribucije je:  $\chi_{tab}^{2[\alpha=5\%]}[df=k-1=3] = 7,81$ .

Stoga se može zaključiti da je  $H^* > \chi_{tab}$ , pa se odbacuje početna hipoteza, tj. zaključuje se uz signifikantnost testa od 5% da postoji statistički značajna razlika u rangovima u prosječnoj ocjeni na I, II, III i IV godini studija između studenata na fakultetu "E".

Prema istoj tablici vidi se da je empirijska signifikantnost  $\alpha^* = 0,28 = 2,8\% \Rightarrow \alpha^* < 5\%$ , čime se opet potvrđuje zaključak o odbacivanju početne hipoteze.

Na temelju podataka u tablici 3.39 može se vidjeti da prosječna ocjena na III godini studija ima najbolji prosječni rang 2,98, a prosječni rang za I godinu studija je najlošiji i iznosi 1,92.



### Primjer 3.17.

Na temelju provedenog anketnog upitnika na uzorku od 20 studenata diplomskog studija fakulteta "E" potrebno je pomoću Friedmanovog testa ispitati može li se prihvatiti pretpostavka da ne postoji razlika u njihovoj prosječnoj ocjeni na ispitima A, B, i C uz graničnu signifikantnost testa od 5%!



### Rješenje 3.17.

Da bi se donio zaključak o prihvaćanju hipoteze o tome postoji li razlika u prosječnoj ocjeni studenata na ispitima A, B, i C uz graničnu signifikantnost testa od 5% potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **K Related Samples**. U otvorenom prozoru biraju se varijable Prosječna ocjena na ispitu A, Prosječna ocjena na ispitu B, Prosječna ocjena na ispitu C u: **Test Variable List**. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine.

Tablica 3.41.

### Prosječni rangovi ocjena ispitanika na različitim ispitima

Ranks	
	Mean Rank
Prosječna ocjena na ispitu A	1,68
Prosječna ocjena na ispitu B	1,90
Prosječna ocjena na ispitu C	2,42

Izvor: Simulirani podaci.

U tablici 3.41 prikazani su odgovarajući prosječni rangovi uzorka studenata na različitim kolegijima.

U tablici 3.42 prikazani su odgovarajući rezultati Friedmanovog testa za zadani uzorak studenata na različitim kolegijima.

**Tablica 3.42.**

**Rezultati Friedmanovog testa za zadani uzorak ispitanika**

Test Statistics <sup>a</sup>	
N	20,000
Chi-Square	6,156
df	2,000
Asymp. Sig.	,046
a. Friedman Test	

Izvor: Simulirani podaci.

Prema dobivenim rezultatima iz tablice 3.42 može se vidjeti da je empirijska vrijednost Friedmanovog testa za zadani uzorak ispitanika:

$$\chi_r^2 = \frac{12}{Nk(k+1)} \cdot \left( \sum_{i=1}^k T_i^2 \right) - 3N(k+1) = 6,156,$$

a tablična vrijednost odgovarajuće hi-kvadrat distribucije je:  $\chi_{tab}^{2[\alpha=5\%]} = 5,991$ .

Stoga se može zaključiti da je  $H^* > \chi_{tab}$ , pa se odbacuje početna hipoteza, tj. zaključuje se uz signifikantnost testa od 5% da postoji statistički značajna razlika u rangovima u prosječnoj ocjeni na kolegijima A, B i C između studenata diplomskog studija na fakultetu "E".

Prema istoj tablici vidi se da je empirijska signifikantnost  $\alpha^* = 0,046 = 4,6\% \Rightarrow \alpha^* < 5\%$ , čime se opet potvrđuje zaključak o odbacivanju početne hipoteze.

Na temelju podataka u tablici 3.41 može se vidjeti da prosječna ocjena na kolegiju C ima najbolji prosječni rang 2,42, a prosječni rang za kolegij A je najlošiji i iznosi 1,68.



**Primjer 3.18.**

Prilikom promocije novog proizvoda napravljena je reklamna kampanja u 2 ciklusa. Odabran je slučajni uzorak od 30 potrošača koji su kupovali navedeni proizvod. Potrebno je, pomoću Friedmanovog testa ispitati može li se prihvatiti pretpostavka da ne postoji razlika u kupovini navedenog proizvoda prije reklamnih kampanja, nakon

I. ciklusa reklama te nakon II. ciklusa reklama uz graničnu signifikantnost testa od 5%!



### Rješenje 3.18.

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u kupovini navedenog proizvoda prije reklamnih kampanja, nakon I. ciklusa reklama te nakon II. ciklusa reklama uz graničnu signifikantnost testa od 5% potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

U programskom paketu **SPSS** potrebno je odabrati **Analyze; Nonparametric Tests; K Related Samples**.

- varijable: Mjesečna potrošnja prije reklame;  
Mjesečna potrošnja nakon I. reklame;  
Mjesečna potrošnja nakon II. reklame; u: **Test Variable List**.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **K Related Samples**. U otvorenom prozoru biraju se varijable Mjesečna potrošnja prije reklame, Mjesečna potrošnja nakon I. reklame, Mjesečna potrošnja nakon II. reklame u: **Test Variable List**. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine.

Tablica 3.43.

Prosječni rangovi mjesečne potrošnje prije i nakon reklamnih kampanja

Ranks	
	Mean Rank
Mjesečna potrošnja prije reklame	1,42
Mjesečna potrošnja nakon I. reklame	2,10
Mjesečna potrošnja nakon II. reklame	2,48

Izvor: Simulirani podaci.

U tablici 3.43 prikazani su odgovarajući prosječni rangovi mjesečne potrošnje uzorka potrošača prije i nakon reklamnih kampanja.



U tablici 3.44 prikazani su odgovarajući rezultati Friedmanovog testa za zadani uzorak potrošača.

**Tablica 3.44.**

**Rezultati Friedmanovog testa za zadani uzorak potrošača**

Test Statistics <sup>a</sup>	
N	30,000
Chi-Square	27,658
df	2,000
Asymp. Sig.	,000
a. Friedman Test	

*Izvor: Simulirani podaci.*

Prema dobivenim rezultatima iz tablice 3.44 može se vidjeti da je empirijska vrijednost Friedmanovog testa za zadani uzorak ispitanika:

$$\chi_r^{2*} = \frac{12}{Nk(k+1)} \cdot \left( \sum_{i=1}^k T_i^2 \right) - 3N(k+1) = 27,658,$$

a tablična vrijednost odgovarajuće hi-kvadrat distribucije je:  $\chi_{tab}^{2[\alpha=5\%]}[df=k-1=2] = 5,991$ .

Stoga se može zaključiti da je  $H^* > \chi_{tab}$ , pa se odbacuje početna hipoteza, tj. zaključuje se uz signifikantnost testa od 5% da postoji statistički značajna razlika u rangovima između kupaca u kupovini navedenog proizvoda prije reklamnih kampanja, nakon I. ciklusa reklama te nakon II. ciklusa reklama.

Prema istoj tablici vidi se da je empirijska signifikantnost  $\alpha^* \approx 0\% \Rightarrow \alpha^* < 5\%$ , čime se opet potvrđuje zaključak o odbacivanju početne hipoteze.

Na temelju podataka u tablici 3.43 može se vidjeti da mjesečna potrošnja nakon II ciklusa reklame ima najbolji prosječni rang 2,48, a prosječni rang mjesečne potrošnje prije reklama je najlošiji i iznosi 1,42. Dakle, može se zaključiti da su provedene reklame bile učinkovite i da su statistički značajno povećale prodaju.

### 3.8 Testiranje hipoteza s nezavisnim uzorcima

#### 3.8.1 Mann-Whitney U-test za dva nezavisna uzorka (Wilcoxon T-test ili test zbroja rangova)

Ovaj test *primjenjuje se za dva nezavisna uzorka koja se mjere pomoću redoslijedne skale*. Postavljaju se hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

Testom zbroja rangova testira se je li dva nezavisna uzorka pripadaju populaciji s istim medijanom. Za uzorke koji su veći od 8 ( $n_i \geq 8$ ) u svakoj od dvije promatrane grupe može se upotrijebiti  $z$ -test, čija je empirijska vrijednost:

$$z^* = \frac{|2 \cdot T_i - n_i \cdot (n+1)| - 2}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n+1)}{3}}}, \quad (3.31)$$

gdje je:

$T_i$  - suma rangova jednog uzorka,

$n_i$  - broj ispitanika u grupi u kojoj je uzeta suma rangova,

$n = n_1 + n_2$  - ukupan broj ispitanika u oba uzorka.

Isti rezultat za vrijednost  $z$  dobije se za obje grupe. S obzirom da varijabla  $z$  pripada normalnoj distribuciji, zaključak se donosi na uobičajeni način, tj. ako je  $z^* < z_{tab} \Rightarrow H_0$  ili ako je empirijska signifikantnost  $\alpha^* > 5\% \Rightarrow H_0$ , odnosno ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

Ako su uzorci veličine  $n_i < 8$  za ovo testiranje koriste se posebne tablice.



#### Primjer 3.19.

Na Splitskom sveučilištu provedena je anonimna anketa na odabranom uzorku studenata. Pomoću Mann-Whitney U-testa potrebno je na temelju zadanog uzorka ispitati postoji li razlika u rangovima u prosječnoj ocjeni na I godini studija između

studenata muškog i ženskog spola na Splitskom sveučilištu uz signifikantnost testa od 5%!



### Rješenje 3.19.

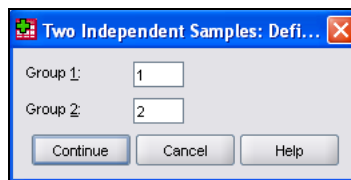
Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u rangovima u prosječnoj ocjeni na I godini studija između studenata muškog i ženskog spola na Splitskom sveučilištu uz signifikantnost testa od 5% potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

Slika 3.18.

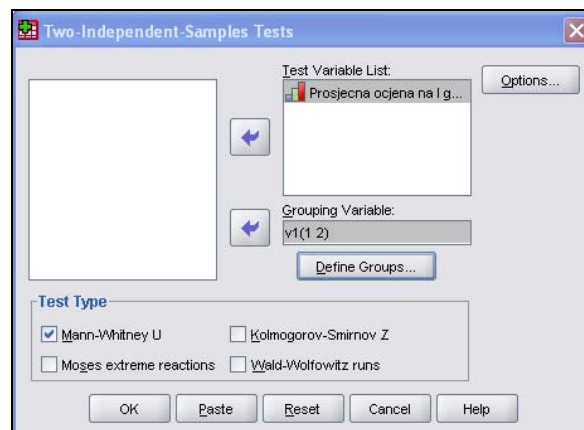
Prozor "Two Independent Samples: Define Groups" s odabranim grupama



Izvor: Simulirani podaci.

Slika 3.19.

Prozor "Two Independent Samples Tests" s odabranim veličinama za Mann-Whitney U-test



Izvor: Simulirani podaci.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **2 Independent Samples**. U otvorenom prozoru bira se varijabla Prosječna ocjena na I god. studija (v2) u: **Test Variable List**. U **Grouping Variable** bira se varijabla Spol (v1). Izborom ikone **Define Groups**, u otvorenom prozoru bira se za **Group1**:1, a za **Group 2**:2, što odgovara kategorijama muškog spola (1) i ženskog spola (2) prema kojima i treba napraviti navedeno testiranje u ovom primjeru. Izbor grupa prikazan je na slici 3.18.

Na slici 3.19. prikazan je prozor "**Two Independent Samples Tests**" s odabranim veličinama za Mann-Whitney U-test.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.45. prikazani su odgovarajući rangovi zadanog uzorka studenata prema spolu.

**Tablica 3.45.**

**Rangovi prosječne ocjene na I godini studija uzorka ispitanika prema spolu**

Ranks				
	Spol	N	Mean Rank	Sum of Ranks
Prosječna ocjena na I god studija	Musko	89	115,13	10246,50
	Zensko	135	110,77	14953,50
	Total	224		

*Izvor: Simulirani podaci.*

**Tablica 3.46.**

**Rezultati Mann-Whitney U-testa za zadani uzorak ispitanika**

Test Statistics <sup>a</sup>	
	Prosječna ocjena na I god studija
Mann-Whitney U	5773,500
Wilcoxon W	14953,500
Z	-,512
Asymp. Sig. (2-tailed)	,609

a. Grouping Variable: Spol

*Izvor: Simulirani podaci.*

U tablici 3.46. prikazani su odgovarajući rezultati Mann-Whitney U-testa za zadani uzorak ispitanika prema spolu. Može se vidjeti da je empirijska signifikantnost  $\alpha^* = 0,609 = 60,9\% \Rightarrow \alpha^* > 5\% \Rightarrow H_0$ , pa se donosi zaključak o prihvatanju početne hipoteze, tj. da ne postoji statistički značajna razlika u rangovima u prosječnoj ocjeni na I godini studija na Splitskom sveučilištu između studenata muškog i ženskog spola uz signifikantnost testa od 5%.

**Primjer 3.20.**

Na Splitskom sveučilištu provedena je anonimna anketa na odabranom uzorku studenata. Pomoću Mann-Whitney U-testa potrebno je na temelju zadanog uzorka ispitati postoji li razlika u rangovima u prosječnoj ocjeni na I godini studija između studenata koji su se izjasnili da su religiozni i onih koji to nisu na Splitskom sveučilištu uz signifikantnost testa od 5%!

**Rješenje 3.20.**

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u rangovima u prosječnoj ocjeni na I godini studija između studenata koji su se izjasnili da su religiozni i onih koji to nisu na Splitskom sveučilištu uz signifikantnost testa od 5% potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **2 Independent Samples**. U otvorenom prozoru bira se varijabla Prosječna ocjena na I god. studija (v1) u: **Test Variable List**. U **Grouping Variable** bira se varijabla Jeste li religiozni (v2). Izborom ikone **Define Groups**, u otvorenom prozoru bira se za **Group1**:1, a za **Group 2**:2, što odgovara kategorijama da - religioznost (1) i ne - religioznost (2) prema kojima i treba napraviti navedeno testiranje u ovom primjeru.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.45. prikazani su odgovarajući rangovi zadanog uzorka studenata prema religioznosti.

**Tablica 3.45.**

**Rangovi prosječne ocjene na I godini studija uzorka studenata prema religioznosti**

Ranks				
	Je...	N	Mean Rank	Sum of Ranks
Prosječna ocjena na I	da	186	110,77	20604,00
god studija	ne	33	105,64	3486,00
	Total	219		

Izvor: Simulirani podaci.

Tablica 3.46.

## Rezultati Mann-Whitney U-testa za zadani uzorak ispitanika

Test Statistics <sup>a</sup>	
	Prosječna ocjena na I god studija
Mann-Whitney U	2925,000
Wilcoxon W	3486,000
Z	-,445
Asymp. Sig. (2-tailed)	,657
a. Grouping Variable: Jeste li religiozni	

Izvor: Simulirani podaci.

U tablici 3.46 prikazani su odgovarajući rezultati Mann-Whitney U-testa za zadani uzorka ispitanika prema njihovom izjašnjavanju o religioznosti. Može se vidjeti da je empirijska signifikantnost  $\alpha^* = 0,657 = 65,7\% \Rightarrow \alpha^* > 5\% \Rightarrow H_0$ , pa se donosi zaključak o prihvatanju početne hipoteze, tj. da ne postoji statistički značajna razlika u rangovima u prosječnoj ocjeni na I godini studija na Splitskom sveučilištu između studenata koji su se izjasnili da su religiozni i onih koji to nisu uz signifikantnost testa od 5%.



## Primjer 3.21.

Na fakultetu "E" provedena je anonimna anketa na odabranom uzorku studenata. Pomoću Mann-Whitney U-testa potrebno je na temelju zadanog uzorka ispitati postoji li razlika u rangovima u prosječnoj ocjeni iz kolegija "Statističke metode u ekonomiji" između studenata koji zadovoljavaju osnovne potrebe i onih koji su se izjasnili da nemaju materijalnih problema uz signifikantnost testa od 5%!



## Rješenje 3.21.

a) Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u rangovima uspjeha na kolegiju "Statističke metode u ekonomiji" za studentsku populaciju na promatranom fakultetu "E" između studenata koji zadovoljavaju osnovne potrebe i onih koji su se izjasnili da nemaju materijalnih problema uz signifikantnost testa od 5% potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira

se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **2 Independent Samples**. U otvorenom prozoru bira se varijabla Ocjena na kolegiju Statističke metode u ekonomiji (v1) u: **Test Variable List**. U **Grouping Variable** bira se varijabla Financijske prilike u obitelji (v2). Izborom ikone **Define Groups**, u otvorenom prozoru bira se za **Group1:2**, a za **Group 2:3**, što odgovara kategorijama zadovoljava osnovne potrebe (2) i nema materijalnih problema (3) prema kojima i treba napraviti navedeno testiranje u ovom primjeru.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.47 prikazani su odgovarajući rangovi zadanog uzorka studenata prema financijskoj situaciji.

**Tablica 3.47.**

**Rangovi prosječne ocjene na kolegiju Statističke metode u ekonomiji uzorka studenata prema financijskoj situaciji**

Ranks				
	Financijske prilike u ...	N	Mean Rank	Sum of Ranks
Ocjena na kolegiju "Statističke metode u ekonomiji"	Zadovoljava osnovne potrebe	97	94,18	9135,00
	Nema materijalnih problema	111	113,52	12601,00
	Total	208		

*Izvor: Simulirani podaci.*

**Tablica 3.48.**

**Rezultati Mann-Whitney U-testa za zadani uzorak ispitanika**

Test Statistics <sup>a</sup>	
	Ocjena na kolegiju "Statističke metode u ekonomiji"
Mann-Whitney U	4382,000
Wilcoxon W	9135,000
Z	-2,452
Asymp. Sig. (2-tailed)	,014
a. Grouping Variable: Financijske prilike u obitelji	

*Izvor: Simulirani podaci.*

Prema dobivenim rezultatima može se vidjeti da je empirijska signifikantnost  $\alpha^* = 0,014 = 1,4\% \Rightarrow \alpha^* < 5\%$ , pa se odbacuje početna hipoteza, tj. zaključuje se da postoji statistički značajna razlika u rangovima u ocjeni na kolegiju Statističke metode u ekonomiji na fakultetu "E" između studenata koji zadovoljavaju osnovne potrebe i onih koji su se izjasnili da nemaju materijalnih problema uz signifikantnost testa od 5%.

Nakon odbacivanja početne hipoteze, na temelju podataka u tablici 3.47 može se vidjeti da oni studenti kojima je financijska situacija takva da zadovoljavaju osnovne potrebe imaju lošiji prosječni rang ocjene na kolegiju Statističke metode u ekonomiji 94,18, a prosječni rang onih koji nemaju materijalnih problema je bolji i iznosi 113,52. Dakle, može se zaključiti da oni studenti koji nemaju materijalnih problema imaju veći uspjeh u svladavanju nastavnih obveza iz navedenog kolegija.

### 3.8.2 Kruskal-Wallis test za više od dva nezavisna uzorka

Ovaj test *primjenjuje se za više od dva nezavisna uzorka koja se mjere pomoću redoslijedne skale*. Može se reći da se ovo testiranje temelji na testu analize varijance, gdje se umjesto brojčanih mjernih podataka koriste rangovi. Postavljaju se hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

Test se provodi na sljedeći način:

1. Potrebno je rangirati sve podatke iz svih uzoraka na način da se najnižoj vrijednosti dodijeli rang 1. (Ako postoji više podataka s istom vrijednosti potrebno im je dodijeliti jednak rang na način da se računa njihov prosječni rang!)
2. Potrebno je izračunati zbroj rangova u svakom uzorku:  $T_i$ . Vrijedi da je zbroj svih rangova za sve uzorke zajedno:

$$\sum_{i=1}^k T_i = \frac{n(n+1)}{2}, \quad (3.32)$$

gdje je:

$n$  - ukupan broj vrijednosti (podataka).

3. Potrebno je izračunati empirijsku vrijednost Kruskal-Wallis testa:  $H^*$ :

$$H^* = \frac{12}{n(n+1)} \cdot \left( \sum_{i=1}^k \frac{T_i^2}{n_i} \right) - 3(n+1), \quad (3.33)$$

gdje je:



- $n$  - ukupan broj vrijednosti (podataka),
- $T_i$  - zbroj rangova u svakom uzorku,
- $n_i$  - broj vrijednosti (podataka) u svakom uzorku,
- $k$  - broj uzoraka.

4. Ako je broj vrijednosti (podataka) u svakom uzorku dovoljno velik (smatra se da su uzorci dovoljno veliki ako je  $\forall n_i > 5$ ) vrijednost Kruskal-Wallis testa  $H^*$  ima jednaku distribuciju kao i hi-kvadrat, pa se zaključak donosi na način da se empirijska vrijednost testa uspoređi s:

$$\chi_{tab[df=k-1]}^{2[\alpha]}, \quad (3.34)$$

gdje je:

- $\alpha$  - granična razina signifikantnosti,
- $df$  - stupnjevi slobode.

Zaključak se donosi usporedbom empirijske i tablične vrijednosti testa:  $H^* < \chi_{tab} \Rightarrow H_0$ , odnosno ako je empirijska signifikantnost  $\alpha^* > 5\% \Rightarrow H_0$ , tj. prihvaća se početna pretpostavka da ne postoji razlika u rangovima odabranih uzoraka, odnosno da uzorci nisu različiti.

5. Ako je vrijednost Kruskal-Wallis testa  $H^*$  nešto manja od  $\chi_{tab}^2$  vrijednosti, a u uzorcima postoji veći broj vezanih rangova upotrebljava se korekcija:

$$H^{*'} = \frac{H^*}{1 - \frac{\sum_{j=1}^r T_j^2}{n(n^2 - 1)}}, \quad (3.35)$$

gdje je:

$$T_j = n_j(n_j^2 - 1), \quad (3.36)$$

- $n_j$  - broj podataka koji dijele jednaki rang,
- $r$  - broj svih zajedničkih rangova.

Potrebno je  $T_j$  izračunati za svaki zajednički rang. Pomoću ove korekcije za rezultat se dobije nešto veća vrijednost Kruskal-Wallis testa  $H^*$ , tj.:

$H^* > H^*$ , pa se može dogoditi da premaši  $\chi^2_{tab}$  vrijednost što vodi ka odbacivanju pretpostavke  $H_0$  odnosno zaključku da su uzorci različiti.

6. Ako broj vrijednosti (podataka) u svakom uzorku nije dovoljno velik (tj. ako su  $n_i < 5$ ) vrijednost Kruskal-Wallis testa  $H^*$  nema jednaku distribuciju kao i hi-kvadrat, pa se zaključak donosi pomoću posebnih tablica.



### Primjer 3.22.

Na fakultetu "E" provedena je anonimna anketa na odabranom uzorku studenata. Potrebno je, pomoću Kruskal-Wallis testa, ispitati može li se prihvatiti pretpostavka da ne postoji razlika u prosječnoj ocjeni u srednjoj školi između studenata koji su završili različite srednje škole uz graničnu signifikantnost testa od 5%!



### Rješenje 3.22.

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u prosječnoj ocjeni u srednjoj školi između studenata koji su završili različite srednje škole uz graničnu signifikantnost testa od 5% potrebno je postaviti hipoteze:

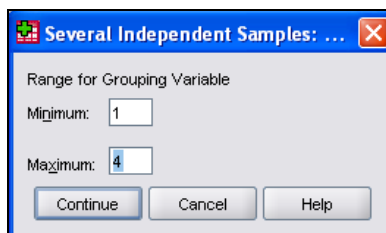
$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **K Independent Samples**. U otvorenom prozoru bira se varijabla Prosječna ocjena u srednjoj školi (v1) u: **Test Variable List**.

### Slika 3.20.

Prozor "Several Independent Samples: Define Groups" s odabranim grupama



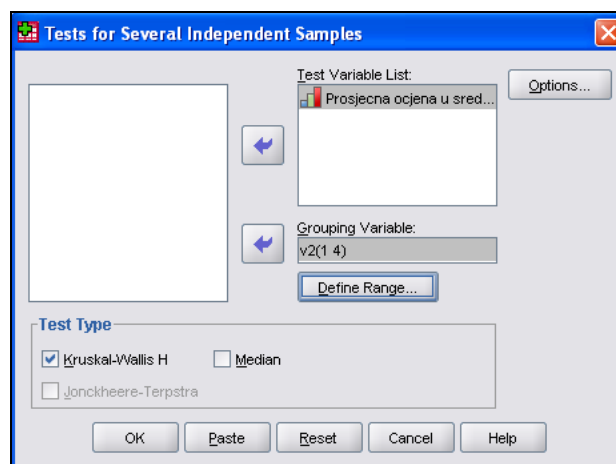
Izvor: Simulirani podaci.

U **Grouping Variable** bira se varijabla Srednja škola (v2). Izborom ikone **Define Groups**, u otvorenom prozoru bira se za **Minimum**:1, a za **Maximum**:4, što predstavlja 4 odgovarajuće završene srednje škole ispitanika prema kojima i treba napraviti navedeno testiranje u ovom primjeru. Izbor grupa prikazan je na slici 3.20.

Na slici 3.21. prikazan je prozor "Tests for Several Independent Samples" s odabranim veličinama za Kruskal-Wallis test.

Slika 3.21.

**Prozor "Tests for Several Independent Samples" s odabranim veličinama za Kruskal-Wallis test**



Izvor: Simulirani podaci.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.49 prikazani su odgovarajući rangovi zadanog uzorka ispitanika prema spolu.

Tablica 3.49.

**Rangovi prosječne ocjene u srednjoj školi uzorka ispitanika prema srednjoj školi**

Ranks			
	Srednja škola	N	Mean Rank
Prosječna ocjena u srednjoj školi	Gimnazija	134	126,54
	Zdravstvena škola	74	113,78
	Neka trogodišnja škola	8	125,75
	Ostalo	20	79,22
	Total	236	

Izvor: Simulirani podaci.

U tablici 3.50 prikazani su odgovarajući rezultati Kruskal-Wallis testa za zadani uzorak studenata prema srednjoj školi.

**Tablica 3.50.**

**Rezultati Kruskal-Wallis testa za zadani uzorak ispitanika**

Test Statistics <sup>a,b</sup>	
	Prosječna ocjena u srednjoj školi
Chi-Square	9,798
df	3
Asymp. Sig.	,020
a. Kruskal Wallis Test	
b. Grouping Variable: Srednja škola	

*Izvor: Simulirani podaci.*

Prema dobivenim rezultatima iz tablice 3.50 može se vidjeti da je empirijska vrijednost Kruskal-Wallis testa za zadani uzorak ispitanika:

$$H^* = \frac{12}{n(n+1)} \cdot \left( \sum_{i=1}^k \frac{T_i^2}{n_i} \right) - 3(n+1) = 9,798,$$

a tablična vrijednost odgovarajuće hi-kvadrat distribucije je:  $\chi_{tab}^{2[\alpha=5\%]}[df=k-1=3] = 7,815$ .

Stoga se može zaključiti da je  $H^* > \chi_{tab}$ , pa se odbacuje početna hipoteza, tj. zaključuje se da postoji statistički značajna razlika u rangovima u prosječnoj ocjeni u srednjoj školi između studenata koji potječu iz različitih srednjih škola uz graničnu signifikantnost testa od 5%.

Iz iste tablice vidi se da je empirijska signifikantnost  $\alpha^* = 0,02 = 2\% \Rightarrow \alpha^* < 5\%$ , čime se opet potvrđuje zaključak o odbacivanju početne hipoteze.

Nakon odbacivanja početne hipoteze, na temelju podataka u tablici 3.49 može se vidjeti da oni studenti koji su završili gimnaziju imaju najbolji prosječni rang ocjene iz srednje škole i da on iznosi 126,54, a prosječni rang onih koji su završili neku drugu srednju školu u kategoriji ostalih imaju najlošiji prosječni rang i on iznosi 79,22. Dakle, može se zaključiti da oni studenti koji su završili gimnaziju imaju i najbolji prosjek u srednjoj školi.



**Primjer 3.23.**

Na Splitskom sveučilištu provedena je anonimna anketa na odabranom uzorku studenata. Za zadani uzorak ispitanika potrebno je, pomoću Kruskal-Wallis testa

ispitati može li se prihvatiti pretpostavka da ne postoji razlika u prosječnoj ocjeni na I godini studija između studenata s različitim finansijskim prilikama u obitelji na Splitskom sveučilištu uz graničnu signifikantnost testa od 5%!



### Rješenje 3.23.

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u prosječnoj ocjeni na I godini studija između studenata s različitim finansijskim prilikama u obitelji uz graničnu signifikantnost testa od 5% potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **K Independent Samples**. U otvorenom prozoru bira se varijabla Prosječna ocjena na I godini studija (v1) u: **Test Variable List**. U **Grouping Variable** bira se varijabla Finansijske prilike u obitelji (v2). Izborom ikone **Define Groups**, u otvorenom prozoru bira se za **Minimum**:1, a za **Maximum**:4, što predstavlja 4 odgovarajuće kategorije finansijskih prilika u obitelji studenata prema kojima i treba napraviti navedeno testiranje u ovom primjeru. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.51 prikazani su odgovarajući rangovi zadanog uzorka ispitanika prema spolu.

Tablica 3.51.

#### Rangovi prosječne ocjene na I. godini studija uzorka ispitanika prema finansijskim prilikama

Ranks			
	Financijske prilike u ...	N	Mean Rank
Prosječna ocjena na I god studija	Nema sredstava za normalan standard	6	125,42
	Zadovoljava osnovne potrebe	97	99,20
	Nema materijalnih problema	111	122,27
	Zivi odlično	9	114,39
	Total	223	

Izvor: Simulirani podaci.

U tablici 3.52 prikazani su odgovarajući rezultati Kruskal-Wallis testa za zadani uzorak studenata prema srednjoj školi.

Prema dobivenim rezultatima iz tablice 3.52 može se vidjeti da je empirijska vrijednost Kruskal-Wallis testa za zadani uzorak ispitanika:

$$H^* = \frac{12}{n(n+1)} \cdot \left( \sum_{i=1}^k \frac{T_i^2}{n_i} \right) - 3(n+1) = 7,423$$
, a tablična vrijednost odgovarajuće hi-kvadrat distribucije je:  $\chi_{tab}^{2[\alpha=5\%]}[df=k-1=3] = 7,815$ .

**Tablica 3.52.**

**Rezultati Kruskal-Wallis testa za zadani uzorak ispitanika**

Test Statistics <sup>a,b</sup>	
	Prosječna ocjena na I god studija
Chi-Square	7,423
df	3
Asymp. Sig.	,060
a. Kruskal Wallis Test	
b. Grouping Variable: Financijske prilike u obitelji	

*Izvor: Simulirani podaci.*

Stoga se može zaključiti da je  $H^* < \chi_{tab}$ , pa se prihvća početna hipoteza, tj. zaključuje se da ne postoji statistički značajna razlika u rangovima u prosječnoj ocjeni na I godini studija između studenata koji imaju različitu financijsku situaciju u obitelji uz graničnu signifikantnost testa od 5%.

Iz iste tablice vidi se da je empirijska signifikantnost  $\alpha^* = 0,06 = 6\% \Rightarrow \alpha^* > 5\%$ , čime se opet potvrđuje zaključak o prihvatanju početne hipoteze.



**Primjer 3.24.**

Na Splitskom sveučilištu provedena je anonimna anketa na odabranom uzorku studenata. Za zadani uzorak ispitanika potrebno je, pomoću Kruskal-Wallis testa ispitati može li se prihvatiti pretpostavka da ne postoji razlika u visini džeparca između studenata s različitim financijskim prilikama u obitelji uz graničnu signifikantnost testa od 5%!



**Rješenje 3.24.**

Da bi se donio zaključak o prihvatanju hipoteze o tome postoji li razlika u visini džeparca između studenata s različitim financijskim prilikama u obitelji uz graničnu signifikantnost testa od 5% potrebno je postaviti hipoteze:

$H_0$  - ne postoji razlika u rangovima ispitanika u odabranim uzorcima.

$H_1$  - postoji razlika u rangovima ispitanika u odabranim uzorcima.

U programskom paketu **SPSS** potrebno je izračunati odgovarajuću empirijsku signifikantnost na temelju koje će se donijeti zaključak. Na glavnom izborniku bira se ikona **Analyze**, a na njezinom padajućem izborniku **Nonparametric Tests** i **K Independent Samples**. U otvorenom prozoru bira se varijabla Mjesečni džeparac u kn (v2) u: **Test Variable List**. U **Grouping Variable** bira se varijabla Financijske prilike u obitelji (v1). Izborom ikone **Define Groups**, u otvorenom prozoru bira se za **Minimum**:1, a za **Maximum**:4, što predstavlja 4 odgovarajuće kategorije finansijskih prilika u obitelji studenata prema kojima i treba napraviti navedeno testiranje u ovom primjeru. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.53 prikazani su odgovarajući rangovi zadanog uzorka ispitanika prema spolu.

**Tablica 3.53.**

**Rangovi mjesečnog džeparca u kn uzorka studenata prema financijskim prilikama u obitelji**

Ranks			
	Financijske prilike u ...	N	Mean Rank
Mjesečni džeparac u kn	Nema sredstava za normalan standard	5	66,40
	Zadovoljava osnovne potrebe	77	79,36
	Nema materijalnih problema	82	89,61
	Zivi odlično	7	130,79
	Total	171	

*Izvor: Simulirani podaci.*

U tablici 3.54 prikazani su odgovarajući rezultati Kruskal-Wallis testa za zadani uzorak studenata prema srednjoj školi.

**Tablica 3.54.**

**Rezultati Kruskal-Wallis testa za zadani uzorak ispitanika**

Test Statistics <sup>a,b</sup>	
	Mjesečni džeparac u kn
Chi-Square	8,431
df	3
Asymp. Sig.	,038
a. Kruskal Wallis Test	
b. Grouping Variable: Financijske prilike u obitelji	

*Izvor: Simulirani podaci.*

Prema dobivenim rezultatima iz tablice 3.54 može se vidjeti da je empirijska vrijednost Kruskal-Wallis testa za zadani uzorak ispitanika:

$$H^* = \frac{12}{n(n+1)} \cdot \left( \sum_{i=1}^k \frac{T_i^2}{n_i} \right) - 3(n+1) = 8,431$$
, a tablična vrijednost odgovarajuće hi-kvadrat distribucije je:  $\chi_{tab[df=k-1=3]}^{2[\alpha=5\%]} = 7,815$ .

Stoga se može zaključiti da je  $H^* > \chi_{tab}$ , pa se odbacuje početna hipoteza, tj. zaključuje se da postoji statistički značajna razlika u visini džeparca između studenata na Splitskom sveučilištu koji imaju različitu financijsku situaciju u obitelji uz graničnu signifikantnost testa od 5%.

Iz iste tablice vidi se da je empirijska signifikantnost  $\alpha^* = 0,038 = 3,8\% \Rightarrow \alpha^* < 5\%$ , čime se opet potvrđuje zaključak o odbacivanju početne hipoteze.

Nakon odbacivanja početne hipoteze, na temelju podataka u tablici 3.53 može se vidjeti da oni studenti koji su se izjasnili da nemaju sredstava za normalan standard imaju i najmanji prosječni rang mjesečnog džeparca i da on iznosi 66,4, a prosječni rang džeparca onih koji su se izjasnili da žive odlično je najveći i iznosi 130,79. Dakle, može se zaključiti da oni studenti koji imaju bolju financijsku situaciju u obitelji imaju i veći mjesečni džeparac.

### 3.9 Analiza utjecaja promjenjivog/ih faktora na kretanje slučajne varijable

Djelovanje promjenjivih faktora na numeričku vrijednost slučajne varijable  $X$  ispituje se analizom varijance.

Pomoću analize varijance ukupna varijanca se dijeli na dio koji se pripisuje djelovanju jednog ili više promjenjivih faktora i na dio koji se pripisuje utjecaju svih ostalih faktora, tj. nerazjašnjen ili rezidualni dio.



### 3.9.1 Djelovanje jednog promjenjivog faktora na kretanje slučajne varijable (nezavisni uzorci)

Kod analize varijance s jednim promjenjivim faktorom ispituje se djelovanje jednog promjenjivog faktora  $A$  na numeričku vrijednost slučajne varijable  $X$ .

Uvjet za ovo testiranje je da uzorci potječu iz normalnih populacija i da imaju jednake varijance. Toleriraju se i uzorci slične veličine, ako populacije slično odstupaju od normalne distribucije.

Uzima se onoliko uzoraka ( $j$ ) koliko ima varijacija faktora  $A$ :  $A_1, A_2, \dots, A_k$ .

Pri tom veličina ( $i$ ) svakog uzorka ne mora biti jednaka.

Postavljaju se hipoteze:

$$H_0 : \dots \sigma_A^2 = 0$$

$$H_1 : \dots \sigma_A^2 \neq 0$$

gdje nulta hipoteza  $H_0$  pretpostavlja da je varijanca promjenjivog faktora  $A$  jednaka nuli, odnosno da djelovanje tog faktora na slučajnu varijablu  $X$  nije statistički značajno. Alternativna hipoteza  $H_1$  tvrdi suprotno.

Testiranje se vrši  $F$ -testom pomoću analize varijance usporedbom empirijske i tablične vrijednosti  $F$ -testa.

Empirijska vrijednost  $F$ -testa računa se pomoću podataka iz tablice analize varijance, tzv. **tablice ANOVA**:

**Tablica 3.54.**

**Tablica analize varijance s jednim promjenjivim faktorom (ANOVA)**

Izvor varijacije	Zbroj kvadrata odstupanja	Stupnjevi slobode	Ocjena varijance
između uzoraka	$\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2$	k-1	$S_A^2$
unutar uzoraka	$\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{\bullet j})^2$	n-k	$S_u^2$
ukupno	$\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{\bullet\bullet})^2$	n-1	

*Izvor: Konstrukcija autora prema teorijskim postavkama.*

gdje je:  $k$  - broj uzoraka,

$n_j$  - veličina pojedinog  $j$  uzorka,

$\bar{X}_{\bullet j}$  - aritmetička sredina pojedinog  $j$  uzorka,

$\bar{X}_{\bullet\bullet}$  - aritmetička sredina svih podataka iz svih uzorka zajedno,

$X_{ij}$  -  $i$ -ti podatak iz  $j$ -tog uzorka.

Empirijska vrijednost F-testa je:

$$F^* = \frac{\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 / (k-1)}{\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 / (n-k)} = \frac{S_A^2}{S_u^2}. \quad (3.36)$$

Tablična vrijednost F-testa je:

$$F_{tab}^{[\alpha]}[df_1=(k-1); df_2=(n-k)], \quad (3.37)$$

sa stupnjevima slobode  $df_1$  i  $df_2$ .

Zaključak se donosi usporedbom empirijske i tablične F vrijednosti, i ako je:

$$F^* = \frac{S_A^2}{S_u^2} < F_{tab} \Rightarrow H_0, \quad (3.38)$$

**prihvaća se početna hipoteza**  $H_0$  tj. zaključuje se da varijanca promjenjivog faktora  $A$  nije statistički značajna, odnosno da taj promjenjivi faktor ne djeluje značajno na slučajnu varijablu  $X$ .

Testiranje se može izvršiti i izračunavanjem granične empirijske signifikantnosti  $\alpha^*$  pomoću  $F^*$  (Tablica D1 i D2):

ako je  $\alpha^* > 5\% \Rightarrow H_0$ , tj. prihvaća se početna pretpostavka, dok se u suprotnom ta hipoteza odbacuje.

Potrebno je naglasiti da se analiza varijance s jednim promjenjivim faktorom provodi uz **pretpostavku** da je ispunjen uvjet **homogenosti varijance promatranih uzoraka**, tj. da vrijedi da je:

$$H_0: \dots \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

Jedan od testova homogenosti je **Levene-ov test homogenosti varijanci** (temelji se na mjerenju apsolutnih odstupanja rezultata od aritmetičke sredine za svaku grupu promjenjivog faktora) koji je uključen i u **SPSS**. Ako se kod ovog testa pokaže da je empirijska signifikantnost  $\alpha^* > 5\% \Rightarrow H_0$ . U suprotnom se početna pretpostavka odbacuje, tj. ne vrijedi nulta hipoteza o homogenosti varijanci uzoraka. U tom slučaju praksa je pokazala da treba:

1. ... odustati od analize varijance,
2. ... upotrijebiti neko drugo neparametrijsko testiranje,
3. ... povećati graničnu signifikantnost testa na npr.  $\alpha = 10\%$ .

Ako je prihvaćena  $H_0$  hipoteza o homogenosti može se izvršiti test analize varijance.



#### Primjer 3.25.

Na jednom sveučilištu na slučajno odabranom uzorku studenata provedena je anonimna anketa. Potrebno je utvrditi je li financijske prilike u obitelji značajno utječu na broj djece u obitelji studenata na promatranom sveučilištu uz signifikantnost testa od 5%!



#### Rješenje 3.25.

Da bi se donio zaključak o prihvatanju hipoteze o tome je li financijske prilike u obitelji značajno utječu na broj djece u obitelji studenata na promatranom sveučilištu uz signifikantnost testa od 5% potrebno je postaviti hipoteze o analizi varijance s jednom promjenjivim faktorom što u ovom slučaju predstavljaju financijske prilike u obitelji stanovnika (faktor A):

$$H_0 : \dots\dots\dots \sigma_A^2 = 0$$

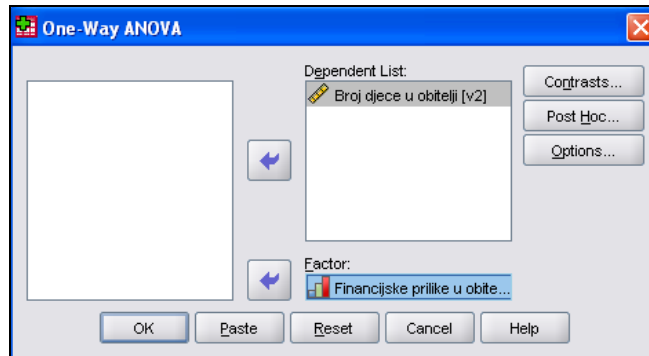
$$H_1 : \dots\dots\dots \sigma_A^2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Compare Means** i **One-Way ANOVA**. U otvorenom prozoru bira se varijabla Broj djece u obitelji (v2) u: **Dependent List**, a varijabla Financijske prilike u obitelji (v1) u: **Factor**. Izborom ikone **Options** potrebno je aktivirati **Homogeneity of variance test**.

Na slici 3.22 prikazan je prozor "One-Way ANOVA" s odabranim veličinama za zadano testiranje.

Slika 3.1.

Prozor "One-Way ANOVA" s odabranim veličinama za zadano testiranje



Izvor: Simulirani podaci.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.55 prikazani su rezultati **Levene-ovog testa** homogenosti varijanci.

Tablica 3.55.

Rezultati Levene-ovog testa homogenosti varijanci

Test of Homogeneity of Variances			
Broj djece u obitelji			
Levene Statistic	df1	df2	Sig.
1,381	3	232	,249

Izvor: Simulirani podaci.

Početna hipoteza za Levene-ov test homogenosti varijance iz uzoraka u ovom primjeru je:  $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ .

Iz rezultata u tablici vidi se da je empirijska signifikantnost  $\alpha^* = 0,249 = 24,9\%$  što znači da je:  $\alpha^* > 5\% \Rightarrow H_0$ , tj. zadovoljen je uvjet da vrijedi nulta hipoteza o homogenosti varijanci uzoraka. To omogućuje nastavak testiranja analize varijance s jednim promjenjivim faktorom.

U tablici 3.56 prikazani su rezultati u tablici ANOVA za zadani uzorak ispitanika.

Prema dobivenim rezultatima iz tablice 3.56 može se vidjeti da je empirijska vrijednost F-testa:

$$F^* = \frac{\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 / (k-1)}{\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{\bullet j})^2 / (n-k)} = \frac{S_A^2}{S_u^2} = 0,403,$$

a empirijska signifikantnost je  $\alpha^* = 0,751 = 75,1\% \Rightarrow \alpha^* > 5\% \Rightarrow H_0$ .

**Tablica 3.56.**

**Rezultati analize varijance (ANOVA) za zadani uzorak ispitanika**

ANOVA					
Broj djece u obitelji					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,818	3	,273	,403	,751
Within Groups	157,059	232	,677		
Total	157,877	235			

Izvor: Simulirani podaci.

Dakle, može se zaključiti da varijanca promjenjivog faktora A nije statistički značajna (jednaka je od 0), tj. da finansijske prilike u obitelji (faktor A) ne djeluju značajno na broj djece u obitelji.



**Primjer 3.26.**

Na jednom sveučilištu na slučajno odabranom uzorku studenata provedena je anonimna anketa. Potrebno je utvrditi je li odnosi u obitelji značajno utječu na broj djece u obitelji studenata na promatranom sveučilištu uz signifikantnost testa od 5%!



**Rješenje 3.26.**

Da bi se donio zaključak o prihvatanju hipoteze o tome je li odnosi u obitelji značajno utječu na broj djece u obitelji studenata na promatranom sveučilištu uz signifikantnost testa od 5% potrebno je postaviti hipoteze o analizi varijance s jednom promjenjivim faktorom što u ovom slučaju predstavljaju odnosi u obitelji studenata (faktor A):

$$H_0 : \dots \sigma_A^2 = 0$$

$$H_1 : \dots \sigma_A^2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Compare Means** i **One-Way ANOVA**. U otvorenom prozoru bira se varijabla Broj djece u obitelji (v1) u: **Dependent List**, a

varijabla Odnosi u obitelji (v2) u: **Factor**. Izborom ikone **Options** potrebno je aktivirati **Homogeneity of variance test**.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.57 prikazani su rezultati **Levene-ovog testa** homogenosti varijanci.

**Tablica 3.57.**

**Rezultati Levene-ovog testa homogenosti varijanci**

Test of Homogeneity of Variances			
Broj djece u obitelji			
Levene Statistic	df1	df2	Sig.
4,177	1	232	,042

*Izvor: Simulirani podaci.*

Početna hipoteza za Levene-ov test homogenosti varijance iz uzoraka u ovom primjeru je:  $H_0: \dots \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ .

Iz rezultata u tablici 3.57 vidi se da je empirijska signifikantnost  $\alpha^* = 0,042 = 4,2\%$  što znači da je:  $\alpha^* < 5\%$ , tj. nije zadovoljen uvjet da vrijedi nulta hipoteza o homogenosti varijanci uzoraka. Stoga je za ovo testiranje analize varijance granična signifikantnost povećana na 10%.

U tablici 3.58 prikazani su rezultati u tablici ANOVA za zadani uzorak ispitanika.

**Tablica 3.58.**

**Rezultati analize varijance (ANOVA) za zadani uzorak ispitanika**

ANOVA					
Broj djece u obitelji					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3,361	2	1,680	2,552	,080
Within Groups	152,750	232	,658		
Total	156,111	234			

*Izvor: Simulirani podaci.*

Prema dobivenim rezultatima iz tablice 3.58 može se vidjeti da je empirijska vrijednost F-testa:

$$F^* = \frac{\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 / (k-1)}{\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{\bullet j})^2 / (n-k)} = \frac{S_A^2}{S_u^2} = 2,552,$$

a empirijska signifikantnost je  $\alpha^* = 0,080 = 8\% \Rightarrow \alpha^* < 10\%$ .

Dakle, odbacuje se početna hipoteza  $H_0$  i može se zaključiti da je varijanca promjenjivog faktora A statistički značajna (različita je od 0), tj. da odnosi u obitelji (faktor A) djeluju značajno na broj djece u obitelji.



### Primjer 3.27.

Na jednom sveučilištu na slučajno odabranom uzorku studenata provedena je anonimna anketa. Potrebno je utvrditi je li školska sprema oca u jednoj obitelji značajno utječe na broj djece u toj obitelji uz signifikantnost testa od 5%!



### Rješenje 3.27.

Da bi se donio zaključak o prihvatanju hipoteze o tome je li školska sprema oca u jednoj obitelji značajno utječe na broj djece u toj obitelji na promatranom sveučilištu uz signifikantnost testa od 5% potrebno je postaviti hipoteze o analizi varijance s jednom promjenjivim faktorom što u ovom slučaju predstavlja školska sprema oca u obitelji studenata (faktor A):

$$H_0 : \dots\dots\dots \sigma_A^2 = 0$$

$$H_1 : \dots\dots\dots \sigma_A^2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Compare Means** i **One-Way ANOVA**. U otvorenom prozoru bira se varijabla Broj djece u obitelji (v1) u: **Dependent List**, a varijabla Školska sprema oca (v2) u: **Factor**. Izborom ikone **Options** potrebno je aktivirati **Homogeneity of variance test**.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.59 prikazani su rezultati **Levene-ovog testa** homogenosti varijanci.

**Tablica 3.59.**

### Rezultati Levene-ovog testa homogenosti varijanci

Test of Homogeneity of Variances			
Broj djece u obitelji			
Levene Statistic	df1	df2	Sig.
4,764	5	230	,000

Izvor: Simulirani podaci.

Početna hipoteza za Levene-ov test homogenosti varijance iz uzoraka u ovom primjeru je:  $H_0: \dots \sigma_1^2 = \dots = \sigma_6^2$ .

Iz rezultata u tablici 3.59 vidi se da je empirijska signifikantnost  $\alpha^* \approx 0\%$  što znači da je:  $\alpha^* < 5\%$ , tj. nije zadovoljen je uvjet da vrijedi nulta hipoteza o homogenosti varijanci uzoraka. To ne omogućuje nastavak testiranja analize varijance s jednim promjenjivim faktorom.

U tablici 3.60 prikazani su rezultati u tablici ANOVA za zadani uzorak ispitanika.

**Tablica 3.60.**

**Rezultati analize varijance (ANOVA) za zadani uzorak ispitanika**

ANOVA					
Broj djece u obitelji					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11,572	5	2,314	3,638	,003
Within Groups	146,305	230	,636		
Total	157,877	235			

*Izvor: Simulirani podaci.*

Stoga se rezultati ANOVA za zadani uzorak ispitanika prikazani u tablici 3.60 trebaju uzeti s rezervom.

Empirijska vrijednost F-testa:

$$F^* = \frac{\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 / (k-1)}{\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{\bullet j})^2 / (n-k)} = \frac{S_A^2}{S_u^2} = 3,638,$$

a empirijska signifikantnost je  $\alpha^* = 0,003 = 0,3\% \Rightarrow \alpha^* < 5\%$ . Navedeni rezultat bi vodio zaključku je da varijanca promjenjivog faktora A statistički značajna (različita je od 0), tj. da školska sprema oca u obitelji (faktor A) djeluje značajno na broj djece u obitelji (ovdje nije zadovoljen uvjet homogenosti varijanci uzoraka!).



**Primjer 3.28.**

Na jednom sveučilištu na slučajno odabranom uzorku studenata provedena je anonimna anketa. Potrebno je ispitati može li se prihvatiti pretpostavka da različiti odnosi u obitelji značajno djeluju na visinu džeparca između studenata uz graničnu signifikantnost testa od 5%!



**Rješenje 3.28.**

Da bi se donio zaključak o prihvatanju hipoteze o tome je li različiti odnosi u obitelji značajno djeluju na visinu džeparca između studenata na promatranom sveučilištu uz signifikantnost testa od 5% potrebno je postaviti hipoteze o analizi varijance s jednom promjenjivim faktorom što u ovom slučaju predstavljaju odnosi u obitelji studenata (faktor A):

$$H_0 : \dots\dots\dots \sigma_A^2 = 0$$

$$H_1 : \dots\dots\dots \sigma_A^2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Compare Means** i **One-Way ANOVA**. U otvorenom prozoru bira se varijabla Visina džeparca (v2) u: **Dependent List**, a varijabla Odnosi u obitelji (v1) u: **Factor**. Izborom ikone **Options** potrebno je aktivirati **Homogeneity of variance test**.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se tražene veličine. U tablici 3.61 prikazani su rezultati **Levene-ovog testa** homogenosti varijanci.

**Tablica 3.61.****Rezultati Levene-ovog testa homogenosti varijanci**

Test of Homogeneity of Variances			
Mjesečni džeparac u kn			
Levene Statistic	df1	df2	Sig.
2,793	1	169	,097

Izvor: Simulirani podaci.

Početna hipoteza za Levene-ov test homogenosti varijance iz uzoraka u ovom primjeru je:  $H_0 : \dots\dots \sigma_1^2 = \sigma_2^2$ .

Iz rezultata u tablici 3.61 vidi se da je empirijska signifikantnost  $\alpha^* = 0,097\% = 9,7\%$  što znači da je:  $\alpha^* > 5\%$ , tj. zadovoljen je uvjet da vrijedi nulta hipoteza o homogenosti varijanci uzoraka. To omogućuje nastavak testiranja analize varijance s jednim promjenjivim faktorom.

U tablici 3.62 prikazani su rezultati u tablici ANOVA za zadani uzorak ispitanika.

Empirijska vrijednost F-testa:

$$F^* = \frac{\sum_{j=1}^k n_j \cdot (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2 / (k-1)}{\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{\cdot j})^2 / (n-k)} = \frac{S_A^2}{S_u^2} = 1,763,$$

a empirijska signifikantnost je  $\alpha^* = 0,186 = 18,6\% \Rightarrow \alpha^* > 5\%$ .

**Tablica 3.62.**

**Rezultati analize varijance (ANOVA) za zadani uzorak ispitanika**

ANOVA					
Mjesečni džeparac u kn					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	284658,531	1	284658,531	1,763	,186
Within Groups	2,729E7	169	161475,330		
Total	2,757E7	170			

*Izvor: Simulirani podaci.*

Prema tome uz signifikantnost testa od 5% može se prihvatiti početna hipoteza  $H_0$ , može se zaključiti da varijanca promjenjivog faktora A nije statistički značajna (jednaka je 0), tj. da odnosi u obitelji (faktor A) ne djeluju značajno na visinu džeparca studenata na promatranom sveučilištu.

### 3.9.2 Djelovanje dva promjenjiva faktora na kretanje slučajne varijable

Kod analize varijance s dva promjenjiva faktora ispituje se djelovanje promjenjivih faktora A i B na numeričku vrijednost slučajne varijable X.

Uvjet za ovo testiranje je da uzorci potječu iz normalnih populacija i da imaju jednake varijance. Toleriraju se i uzorci slične veličine, ako populacije slično odstupaju od normalne distribucije.

Postavljaju se hipoteze o djelovanju jednog promjenjivog faktora A:

$$H_0 : \dots\dots\dots \sigma_A^2 = 0$$

$$H_1 : \dots\dots\dots \sigma_A^2 \neq 0$$

gdje nulta hipoteza  $H_0$  pretpostavlja da je varijanca promjenjivog faktora  $A$  jednaka nuli, odnosno da djelovanje tog faktora na slučajnu varijablu  $X$  nije statistički značajno. Alternativna hipoteza  $H_1$  tvrdi suprotno.

Postavljaju se hipoteze o djelovanju drugog promjenjivog faktora  $B$  :

$$H_0 : \dots\dots\dots \sigma_B^2 = 0$$

$$H_1 : \dots\dots\dots \sigma_B^2 \neq 0$$

gdje nulta hipoteza  $H_0$  pretpostavlja da je varijanca promjenjivog faktora  $B$  jednaka nuli, odnosno da djelovanje tog faktora na slučajnu varijablu  $X$  nije statistički značajno. Alternativna hipoteza  $H_1$  tvrdi suprotno.

Testiranje se vrši  $F$ -testom pomoću analize varijance usporedbom **empirijske i tablične vrijednosti  $F$  - testa**. Empirijska vrijednost  $F$ -testa računa se pomoću podataka iz tablice analize varijance, tzv. **tablice ANOVA**:

**Tablica 3.63.**

**Tablica analize varijance s dva promjenjiva faktora (ANOVA)**

Izvor varijacije	Zbroj kvadrata odstupanja	Stupnjevi slobode	Ocjena varijance
između redaka	$\sum_{i=1}^c n_i \cdot (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2$	$c-1$	$S_A^2$
između stupaca	$\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2$	$k-1$	$S_B^2$
ostatak	$\sum_{j=1}^k \sum_{i=1}^c (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})^2$	$n-k-c+1$	$S_R^2$

*Izvor: Konstrukcija autora prema teorijskim postavkama.*

gdje je:

$c$  - broj redaka,

$k$  - broj stupaca,

$n_i$  - veličina pojedinog  $i$  retka,

$n_j$  - veličina pojedinog  $j$  stupca,

$n$  - broj podataka,

$\bar{X}_{i\bullet}$  - aritmetička sredina pojedinog  $i$  retka,

$\bar{X}_{\bullet j}$  - aritmetička sredina pojedinog  $j$  stupca,

$\bar{X}_{\bullet\bullet}$  - aritmetička sredina svih podataka zajedno,

$X_{ij}$  - podatak iz  $i$ -tog retka i iz  $j$ -tog stupca.

**Empirijska vrijednost F-testa za promjenjivi faktor  $A$  je (ako se faktor  $A$  mijenja po redcima):**

$$F_A^* = \frac{\sum_{i=1}^c n_i \cdot (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 / (c-1)}{\sum_{j=1}^k \sum_{i=1}^c (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})^2 / (n-k-c+1)} = \frac{S_A^2}{S_R^2}. \quad (3.37)$$

**Tablična vrijednost F-testa je za promjenjivi faktor  $A$  je:**

$$F_{tabA}^{[\alpha]}[df_1=(c-1); df_2=(n-k-c+1)], \quad (3.38)$$

sa stupnjevima slobode  $df_1$  i  $df_2$ .

Zaključak se donosi usporedbom empirijske i tablične F vrijednosti, i ako je:

$$F^* = \frac{S_A^2}{S_R^2} < F_{tabA} \Rightarrow H_0, \quad (3.39)$$

**prihvaća se početna hipoteza  $H_0$**  tj. zaključuje se da varijanca promjenjivog faktora  $A$  nije statistički značajna, odnosno da taj promjenjivi faktor ne djeluje značajno na slučajnu varijablu  $X$ .

Testiranje se može izvršiti i izračunavanjem granične empirijske signifikantnosti  $\alpha^*$  pomoću  $F^*$  (Tablica D1 i D2):

ako je  $\alpha^* > 5\% \Rightarrow H_0$ , tj. prihvaća se početna pretpostavka, dok se u suprotnom ta hipoteza odbacuje.

**Empirijska vrijednost F-testa za promjenjivi faktor  $B$  je (ako se faktor  $B$  mijenja po stupcima):**

$$F_B^* = \frac{\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 / (k-1)}{\sum_{i=1}^c \sum_{j=1}^k (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})^2 / (n-k-c+1)} = \frac{S_B^2}{S_R^2}. \quad (3.40)$$

**Tablična vrijednost F-testa za promjenjivi faktor  $B$  je:**

$$F_{tabB}^{[\alpha]}[df_1=(k-1); df_2=(n-k-c+1)], \quad (3.41)$$

sa stupnjevima slobode  $df_1$  i  $df_2$ .

Zaključak se donosi usporedbom empirijske i tablične F vrijednosti, i ako je:

$$F^* = \frac{S_B^2}{S_R^2} < F_{tabB} \Rightarrow H_0, \quad (3.42)$$

**prihvaća se početna hipoteza  $H_0$**  tj. zaključuje se da varijanca promjenjivog faktora  $B$  nije statistički značajna, odnosno da taj promjenjivi faktor ne djeluje značajno na slučajnu varijablu  $X$ .

Naravno da zaključci o djelovanju promjenjivih faktora  $A$  i  $B$  na slučajnu varijablu  $X$  ne moraju biti jednaki.

Testiranje se može izvršiti i izračunavanjem granične empirijske signifikantnosti  $\alpha^*$  pomoću  $F^*$  (Tablica D1 i D2):

ako je  $\alpha^* > 5\% \Rightarrow H_0$ , tj. prihvaća se početna pretpostavka, dok se u suprotnom ta hipoteza odbacuje.



### Primjer 3.29.

Na jednom sveučilištu na slučajno odabranom uzorku studenata provedena je anonimna anketa. Potrebno je utvrditi je li odnosi u obitelji i financijske prilike u obitelji značajno djeluju na težinu studenata na promatranom sveučilištu uz signifikantnost testa od 5%!



### Rješenje 3.29.

Da bi se donio zaključak o prihvaćanju hipoteze o tome je li odnosi u obitelji i financijske prilike u obitelji značajno djeluju na težinu studenata na promatranom sveučilištu uz signifikantnost testa od 5% potrebno je postaviti hipoteze o analizi varijance s dva promjenjiva faktora. Odnosi u obitelji su jedan promjenjivi faktor (faktor A):

$$H_0 : \dots \sigma_A^2 = 0$$

$$H_1 : \dots \sigma_A^2 \neq 0$$

Financijske prilike u obitelji su drugi promjenjivi faktor (faktor B):

$$H_0 : \dots \sigma_B^2 = 0$$

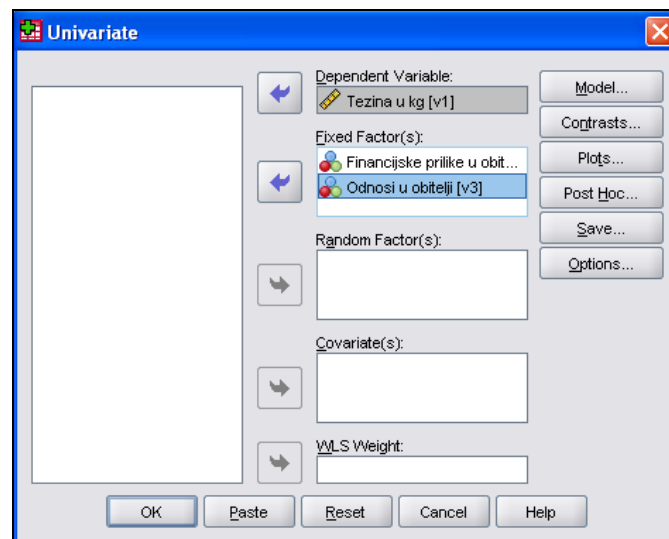
$$H_1 : \dots \sigma_B^2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **General Linear Model** i **Univariate**. U otvorenom prozoru bira se varijabla Težina u kg (v1) u: **Dependent List**. Varijable Odnosi u obitelji (v3) i Financijske prilike u obitelji (v2) u: **Fixed Factor(s)**.

Na slici 3.23 prikazan je prozor "Univariate" s odabranim varijablama za zadano testiranje analize varijance s dva promjenjiva faktora.

**Slika 3.23.**

**Prozor "Univariate" s odabranim veličinama za analizu varijance s dva promjenjiva faktora**



*Izvor: Simulirani podaci.*

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se traženi rezultati, koji su prikazani u tablici 3.64 može se vidjeti da je empirijska vrijednost F-testa za **faktor A** (Odnosi u obitelji - v3):

$$F_A^* = \frac{\sum_{i=1}^c n_i \cdot (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 / (c-1)}{\sum_{j=1}^k \sum_{i=1}^c (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})^2 / (n-k-c+1)} = \frac{S_A^2}{S_R^2} = 2,167,$$

a empirijska signifikantnost je  $\alpha^* = 0,117 = 11,7\% \Rightarrow \alpha^* > 5\% \Rightarrow H_0$ . Dakle, može se zaključiti da varijanca promjenjivog faktora A nije statistički značajna (jednaka je 0), tj. da odnosi u obitelji (faktor A) ne djeluju statistički značajno na težinu studenata na promatranom sveučilištu uz signifikantnost od 5%.

**Tablica 3.64.**

**Rezultati analize varijance (ANOVA) a dva promjenjiva faktora za zadani uzorak ispitanika**

Tests of Between-Subjects Effects					
Dependent Variable: Tezina u kg					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3866,128 <sup>a</sup>	8	483,266	2,798	,006
Intercept	71845,827	1	71845,827	415,969	,000
v2	1706,068	3	568,689	3,293	,021
v3	748,484	2	374,242	2,167	,117
v2 * v3	1191,039	3	397,013	2,299	,078
Error	37998,213	220	172,719		
Total	1074675,000	229			
Corrected Total	41864,341	228			

a. R Squared = ,092 (Adjusted R Squared = ,059)

Izvor: Simulirani podaci.

Na temelju rezultata iz tablice 3.64 može se vidjeti da je empirijska vrijednost F-testa **za faktor B** (Financijske prilike u obitelji - v2):

$$F_B^* = \frac{\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 / (k-1)}{\sum_{j=1}^k \sum_{i=1}^c (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})^2 / (n-k-c+1)} = \frac{S_B^2}{S_R^2} = 3,293,$$

a empirijska signifikantnost je  $\alpha^* = 0,021 = 2,1\% \Rightarrow \alpha^* < 5\%$ . Dakle, može se zaključiti da se odbacuje početna hipoteza faktora B, odnosno da je varijanca promjenjivog faktora B statistički značajna (različita je od 0), tj. financijske prilike u obitelji (faktor B) statistički značajno djeluju na težinu studenata na promatranom sveučilištu uz signifikantnost od 5%.



**Primjer 3.30.**

Na jednom sveučilištu na slučajno odabranom uzorku studenata provedena je anonimna anketa. Potrebno je utvrditi je li školska sprema oca i/ili školska sprema

majke u jednoj obitelji značajno utječe na broj djece u toj obitelji na promatranom sveučilištu uz signifikantnost testa od 5%!



### Rješenje 3.30.

Da bi se donio zaključak o prihvatanju hipoteze o tome je li školska sprema oca i/ili školska sprema majke u jednoj obitelji značajno utječe na broj djece u toj obitelji na promatranom sveučilištu uz signifikantnost testa od 5% potrebno je postaviti hipoteze o analizi varijance s dva promjenjiva faktora.

Školska sprema oca je jedan promjenjivi faktor (faktor A):

$$H_0 : \dots\dots\dots \sigma_A^2 = 0$$

$$H_1 : \dots\dots\dots \sigma_A^2 \neq 0$$

Školska sprema majke je drugi promjenjivi faktor (faktor B):

$$H_0 : \dots\dots\dots \sigma_B^2 = 0$$

$$H_1 : \dots\dots\dots \sigma_B^2 \neq 0$$

U programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **General Linear Model** i **Univariate**. U otvorenom prozoru bira se varijabla Broj djece u obitelji (v1) u: **Dependent List**. Varijable Školska sprema oca (v2) i Školska sprema majke (v3) u: **Fixed Factor(s)**.

**Tablica 3.65.**

**Rezultati analize varijance (ANOVA) a dva promjenjiva faktora za zadani uzorak ispitanika**

Tests of Between-Subjects Effects					
Dependent Variable: Broj djece u obitelji					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	36,799 <sup>a</sup>	32	1,150	1,928	,004
Intercept	426,721	1	426,721	715,443	,000
v2	13,635	5	2,727	4,572	,001
v3	12,338	6	2,056	3,448	,003
v2 * v3	16,710	21	,796	1,334	,157
Error	121,078	203	,596		
Total	1435,000	236			
Corrected Total	157,877	235			

a. R Squared = ,233 (Adjusted R Squared = ,112)

Izvor: Simulirani podaci.



Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se traženi rezultati, koji su prikazani u tablici 3.65 može se vidjeti da je empirijska vrijednost F-testa **za faktor A** (Školska sprema oca - v2):

$$F_A^* = \frac{\sum_{i=1}^c n_i \cdot (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 / (c-1)}{\sum_{j=1}^k \sum_{i=1}^c (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})^2 / (n-k-c+1)} = \frac{S_A^2}{S_R^2} = 4,571,$$

a empirijska signifikantnost je  $\alpha^* = 0,001 = 0,1\% \Rightarrow \alpha^* < 5\%$ . Dakle, može se zaključiti da je varijanca promjenjivog faktora A statistički značajna (različita je od 0), tj. da školska sprema oca (faktor A) statistički značajno djeluje na broj djece u obitelji studenata na promatranom sveučilištu uz signifikantnost od 5%.

Na temelju rezultata iz tablice 3.65 može se vidjeti da je empirijska vrijednost F-testa **za faktor B** (Školska sprema majke - v3):

$$F_B^* = \frac{\sum_{j=1}^k n_j \cdot (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 / (k-1)}{\sum_{j=1}^k \sum_{i=1}^c (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})^2 / (n-k-c+1)} = \frac{S_B^2}{S_R^2} = 3,448,$$

a empirijska signifikantnost je  $\alpha^* = 0,003 = 0,3\% \Rightarrow \alpha^* < 5\%$ . Dakle, može se zaključiti da se odbacuje početna hipoteza faktora B, odnosno da je varijanca promjenjivog faktora B statistički značajna (različita je od 0), tj. školska sprema majke (faktor B) statistički značajno djeluje na broj djece u obitelji studenata na promatranom sveučilištu uz signifikantnost od 5%.

### 3.10 Multivariantna Cluster analiza

Multivariantna analiza (MVA) temelji se na principima multivariantne statistike, koja uključuje promatranja i analize dvije ili više statističkih varijabli istovremeno. Ove tehnike u praksi se koriste u smislu više dimenzijskih analiza u kojima se uvažavaju utjecaji i efekti svih relevantnih varijabli.

Multivariantna analiza ocjenjuje međuovisnosti varijabli i vrši njihovo grupiranje u skladu s njihovom sličnošću (**faktorska analiza**) i/ili vrši grupiranje slučajeva (podataka) opet u skladu s njihovom sličnosti tj. povezanosti (**cluster analiza**).

Multivariantna analiza može se provoditi u smislu istraživanja i potvrđivanja. Istraživačka analiza (exploratory analysis) provodi se da bi se otkrile glavne zajedničke karakteristike većeg broja varijabli. Pri istraživanju postavlja se à priori pretpostavka o potencijalnoj povezanosti promatranih faktora. Kod analize gdje se vrši potvrđivanje, utvrđuje se jesu li odgovarajući faktori i varijable koje utječu na njih zaista u skladu s načelima unaprijed utvrđene ekonomske teorije.

**Cluster analiza** je vrsta multivariantne statističke analize koja spada u metode klasificiranja. Temelji se na matematički formuliranim mjerama sličnosti i obuhvaća različite postupke, algoritme i metode grupiranja podataka. Osnovni problem s kojim se istraživači susreću u praksi je na koji način najprije organizirati sakupljene podatke, a zatim koji je algoritam najbolje upotrijebiti.

Cluster analiza spada u istraživačke analize čiji je osnovni cilj sortirati različite podatke u grupe na način da se maksimizira stupanj sličnosti unutar grupe uz uvjet da je sličnost s drugim grupama minimalna.

Metode grupiranja ili klasifikacije u osnovi se mogu podijeliti na:

- hijerarhijske i
- nehijerarhijske.

**Hijerarhijske metode** su:

- **metoda udruživanja ili aglomerativna metoda** je metoda gdje je u početku svaka jedinica poseban klaster, a zatim se jedinice grupiraju u sve manji broj grupa, dok se sve ne svrstaju u jednu veliku grupu i
- **metoda dijeljenja** po kojoj se velika grupa, koja u početku sadrži sve jedinice, dijeli na sve veći broj grupa, sve dok svaka jedinica ne postane zasebna grupa.

Ove metode u svakom koraku daju različito rješenje s obzirom na broj i sastav klastera. U praktičnim istraživanjima se postavlja pitanje o odabiru optimalnog, tj. "pravog" rješenja. Sve to ovisi o vrsti istraživanja uz uvažavanje ekonomske teorije i tzv. efekta "parasimonije" i mogućnosti interpretacije.

Pri provođenju hijerarhijskih grupiranja u stvari se traži najmanja od udaljenosti:

- **udaljenost između dva najsličnija klastera,**
- **udaljenost između postojećih klastera i negrupiranih jedinica i**
- **udaljenost između dvije najsličnije negrupirane jedinice.**

Sljedeći korak pri provođenju metode ovisi o tome koja je od ovih udaljenosti najmanja. Različite hijerarhijske metode na različit način određuju udaljenosti između klastera tj. između klastera i negrupiranih jedinica, npr.:

- **metoda međusobnog povezivanja** (Between-groups linkage), koja se temelji na maksimiziranju udaljenosti između svakog para jedinica iz dva različita klastera. Udaljenost između dva klastera ovdje se računa kao prosjek udaljenosti svih kombinacija parova jedinica iz ta dva klastera.
- **metoda povezivanja unutar grupa** (Within-groups linkage), koja se temelji na minimalnoj udaljenosti svih jedinica unutar klastera. I ovdje je udaljenost jedinica prosjek udaljenosti svih kombinacija parova jedinica iz tog novonastalog klastera.
- **metoda najbližeg susjeda** (Nearest neighbor) ili jednostrukog povezivanja (Single linkage) pretpostavlja da je udaljenost među dva klastera jednaka udaljenosti između dvije najbliže jedinice iz ta dva klastera.
- **metoda najdaljeg susjeda** (Furthest neighbor) ili potpunog povezivanja (Complete linkage) pretpostavlja da je udaljenost među dva klastera jednaka udaljenosti između dvije najdalje jedinice iz ta dva klastera.
- tzv. **centroidna metoda** (Centroid clustering) ili ponderirana centroidna metoda (Weighted pair-group centroid) pretpostavlja da je udaljenost među dva klastera jednaka udaljenosti između aritmetičkih sredina svih jedinica iz ta dva klastera.

- **metoda medijana** (Median clustering) ili neponderirana centroidna metoda (Unweighted pair-group centroid) je slična prethodnoj metodi, samo bez ponderiranja.
- **Ward-ova metoda** (Ward's method) temelji se na najmanjoj Euklidskoj udaljenosti svake jedinice od aritmetičke sredine (što u stvari predstavlja analizu varijance) za cijeli klaster kojem ta jedinica pripada.

Grafički prikaz grupiranja je tzv. **grafikon matričnog stabla ili dendrogram** koji vizuelno u dvodimenzionalnom prostoru pokazuje hijerarhiju unutar konačnog klastera, gdje je napravljena podjela kroz binarno stablo. Dendrogram prikazuje i stvarne udaljenosti na skali od 0 do 25.

**Cluster postupak** (aglomerativna) se može opisati na sljedeći način:

- svaka jedinica na početku je zaseban klaster
- ocjenjuju se sve odgovarajuće udaljenosti između jedinica i/ili klastera
- formira se matrica udaljenosti koristeći dobivene udaljenosti
- traži se najmanja udaljenost
- par jedinica ili klastera s najmanjom udaljenosti se odvaja iz matrice
- nakon toga ocjenjuju se sve udaljenosti između tog „novog“ klastera sa svim preostalim jedinicama i/ili klasterima i formira se nova matrica udaljenosti
- cijeli postupak se ponavlja dok u matrici udaljenosti ne ostane samo jedan element.

Najčešće korištena mjera udaljenosti između numeričkih podataka je Euklidska udaljenost i Euklidska kvadratna udaljenost.

Euklidska udaljenost između dvije točke  $X(x_1, x_2, \dots, x_n)$  i  $Y(y_1, y_2, \dots, y_n)$  je:

$$d = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} . \quad (3.43)$$

Euklidska kvadratna udaljenost između točaka  $X(x_1, x_2, \dots, x_n)$  i  $Y(y_1, y_2, \dots, y_n)$  je:

$$d' = \sum_{i=1}^n (x_i - y_i)^2 . \quad (3.44)$$

Od **nehijerarhijskih metoda** u praksi je najviše u upotrebi **k-means metoda** tj. metoda k-prosjeka. Prednost i/ili nedostatak ove metode, što naravno ovisi i o vrsti istraživanja, je što se unaprijed treba odabrati broj klastera. Jedinice se tada pridružuju klasteru kojem su najbliži, tj. s kojim imaju najmanju Euklidsku udaljenost. Ako je potrebno, neke jedinice se premještaju iz klastera u klaster sve dok se ne postigne stabilnost sustava. Naravno i ovdje vrijedi kriterij minimalnih udaljenosti jedinica unutar klastera i maksimalnih udaljenosti između klastera.

Pri određivanju broja klastera može se početi od teorijskih postavki (ekonomskih) vezanih za istraživanje koje se provodi. **ANOVA testiranje** odnosi se na svaku promatranu varijablu i upućuje na zaključak je li se sredine između predloženih klastera signifikantno razlikuju. Ako empirijske p-vrijednosti ne premašuju graničnu signifikantnost od 5% može se zaključiti da se sredine između predloženih klastera značajno razlikuju. U suprotnom je potrebno promijeniti predloženi broj klastera i/ili promatrane varijable.

Ako se neka jedinica nikako ne može klasterirati ni u višim fazama klasteriranja, ona se smatra netipičnom vrijednošću (outlier). Takva jedinica se zove Runt ili Entropy grupa.

Cluster analiza **ne pretpostavlja nikakvu statističku značajnost**, pa je u praksi uobičajeno koristiti i dodatne odgovarajuće statističke testove da bi se potvrdili zaključci na znanstvenoj razini.

Cluster analizom uvijek se postiže neka klasifikacija. Rješenja nisu uvijek jedinstvena jer zavise o varijablama koje su uključene u analizu, ali i o načinu kako je kluster definiran. Međutim potrebno je napomenuti da je cluster metoda grupiranja nepristrana i transparentna. Uvažava konkretne matematičke izračune i za rezultat ima nepristrano grupiranje promatranih jedinica, naravno ako su podaci koji se koriste u analizi također nepristrani.



### **Primjer 3.31.**

Zadatak je izvršiti grupiranje odabranih 16 tranzicijskih zemalja prema 8 odgovarajućih socio-političkih pokazatelja za 2007. godinu. Neke od odabranih tranzicijskih zemalja su članice EU, a neke nisu.

Grupiranje je potrebno izvršiti:

- a) Hijerarhijskom metodom udruživanja ili aglomerativnom metodom!
- b) Nehijerarhijskom metodom k-prosjeka!

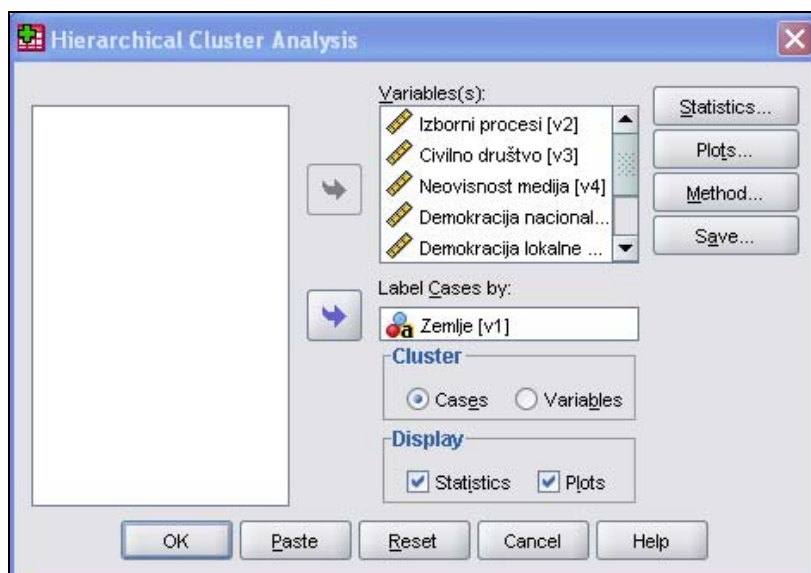


### Rješenje 3.31.

a) U programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Classify**; **Hierarchical Cluster**. U prostor **Variable(s)** potrebno je odabrati sve promatrane pokazatelje zemalja, a u prostor **Label Cases by** potrebno je odabrati varijablu **Zemlje**, kako je prikazano na slici 3.24. Odabirom opcije **Plots** treba aktivirati prikaz **Dendrogram** kako je prikazano na slici 3.25.

### Slika 3.24.

Prozor "Hierarchical Cluster Analysis" s odabranim varijablama i zemljama za klasifikaciju



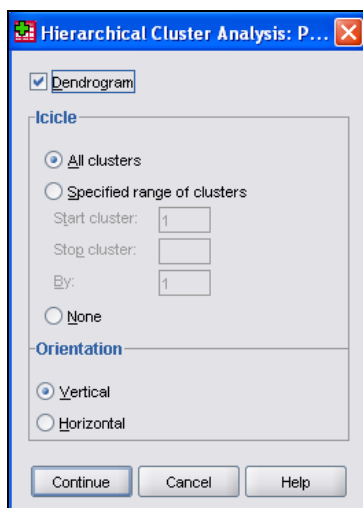
Izvor: Simulirani podaci.

Klikom na ikonu **Continue**, odabirom opcije **Method** bira se **Between-groups linkage**, a za mjeru udaljenosti izabrana je **Measure: Squared Euclidean distance** što je prikazano na slici 3.26.

Klikom na ikone **Continue** i **OK** u **Outputu** programa SPSS dobije se dendrogram, kako je prikazano na slici 3.27.

Slika 3.25.

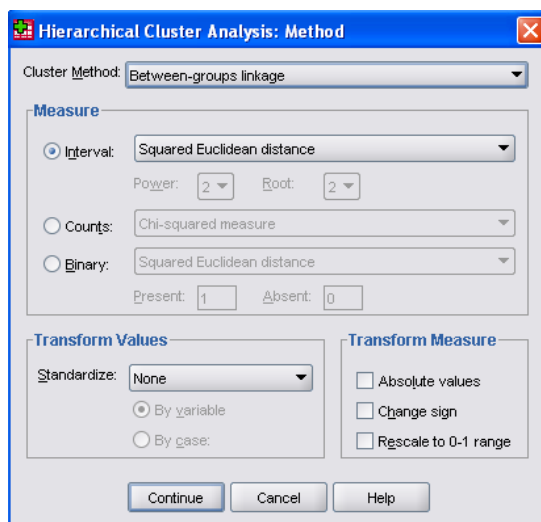
Prozor "Hierarchical Cluster Analysis: Plots" s aktiviranim prikazom dendrograma



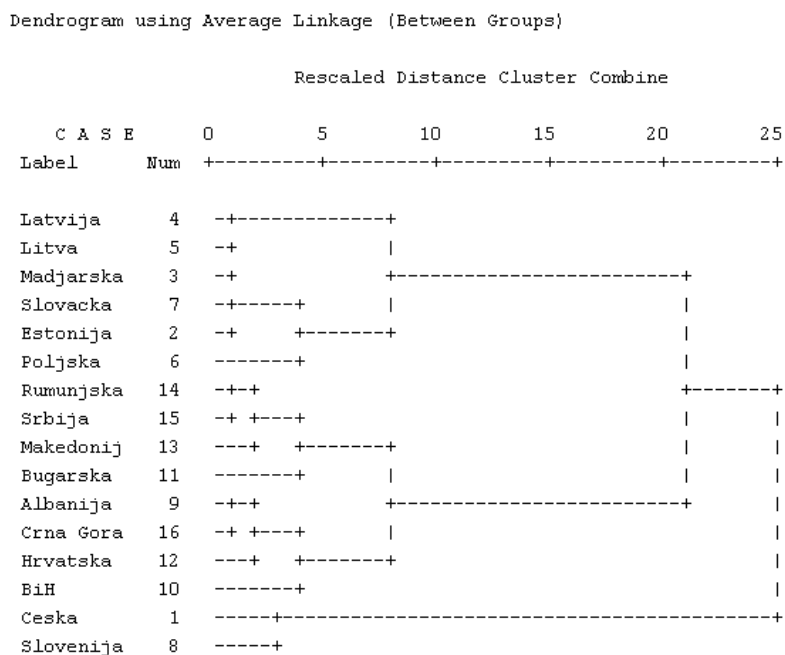
Izvor: Simulirani podaci.

Slika 3.26.

Prozor "Hierarchical Cluster Analysis: Method" s odabranom metodom *Between-groups linkage* i mjerom *Squared Euclidean distance*



Izvor: Simulirani podaci.

**Slika 3.27.****Klasifikacija odabranih zemalja na osnovu socio-političkih pokazatelja pomoću dendrograma**

Izvor: Prema podacima Svjetske banke i Freedom House (2009).

Prema dendrogramu može se vidjeti da su Latvija, Litva, Mađarska, Slovačka i Estonija klasificirane kao slične zemlje prema navedenim pokazateljima. S druge strane posebnu grupu čine Češka i Slovenija, dok su preostale zemlje u kojima prevladavaju one koje nisu članice EU klasificirane kao slične.

b) Da bi se promatrane zemlje klasificirale nehijerarhijskom metodom k-prosjeka u programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Classify; K-Means Cluster**. U prostor **Variable(s)** potrebno je odabrati sve promatrane pokazatelje zemalja, a u prostor **Label Cases by** potrebno je odabrati varijablu **Zemlje**. Željeni i unaprijed određeni **Number of Clusters** je 3 što je prikazano na slici 3.28.

Odabirom opcije **Save** treba aktivirati prikaz **Cluster membership** kako je prikazano na slici 3.29.

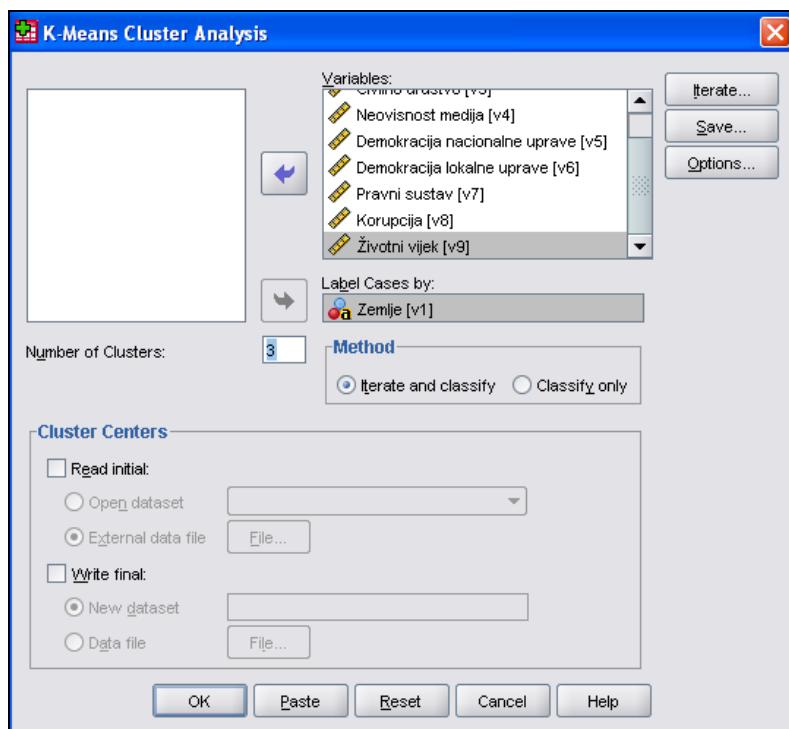
Odabirom opcije **Options** treba aktivirati prikaz **Initial cluster centres**, **ANOVA table** i **Cluster information for each case** kako je prikazano na slici 3.30. Klikom na



ikone **Continue** i **OK** u **Outputu** programa SPSS dobije se klasifikacija zemalja, kako je prikazano u tablici 3.67.

Slika 3.28.

**Prozor "K-Means Cluster Analysis: Method" s odabranim varijablama za klasifikaciju zemalja u 3 klastera**



*Izvor: Simulirani podaci.*

Slika 3.29.

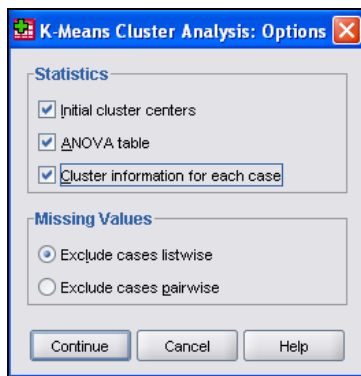
**Prozor "K-Means Cluster Analysis: Save..." s aktiviranim Cluster membership**



*Izvor: Simulirani podaci.*

Slika 3.30.

Prozor "K-Means Cluster Analysis: Options..." s aktiviranim Initial cluster centres; ANOVA table; Cluster information for each case



Izvor: Simulirani podaci.

**ANOVA testiranje** odnosi se na svaku promatranu varijablu i upućuje na zaključak je li se sredine između predloženih klastera signifikantno razlikuju. Empirijske p-vrijednosti ne premašuju graničnu signifikantnost od 5% i može se zaključiti da se sredine između predloženih klastera značajno razlikuju. Rezultati su prikazani u tablici 3.66.

Tablica 3.66.

#### ANOVA testiranje za klaster metodu k-prosjeka

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Izborni procesi	4,814	2	,095	13	50,856	,000
Civilno društvo	2,750	2	,139	13	19,775	,000
Neovisnost medija	5,391	2	,266	13	20,230	,000
Demokracija nacionalne uprave	4,570	2	,228	13	20,013	,000
Demokracija lokalne uprave	3,828	2	,264	13	14,497	,000
Pravni sustav	8,612	2	,127	13	67,962	,000
Korupcija	4,778	2	,289	13	16,538	,000
Životni vijek	18,278	2	1,683	13	10,860	,002

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Izvor: Prema podacima Svjetske banke i Freedom House (2009).

**Tablica 3.67.****Klasifikacija tranzicijskih zemalja pomoću klaster metode k-prosjeka**

Zemlje	Cluster
Ceska	1
Estonija	2
Mađarska	2
Latvija	2
Litva	2
Poljska	1
Slovačka	2
Slovenija	1
Albanija	3
BiH	3
Bugarska	2
Hrvatska	3
Makedonij	3
Rumunjska	3
Srbija	3
Crna Gora	3

*Izvor: Prema podacima Svjetske banke i Freedom House (2009).*

Prema dobivenim podacima u tablici 3.67 može se vidjeti da su Latvija, Litva, Mađarska, Slovačka, Estonija i Bugarska klasificirane u jednu grupu. U grupu Češke i Slovenije pridružila se Poljska, dok su preostale zemlje u kojima prevladavaju one koje nisu članice EU klasificirane kao slične. Rezultati su u slični onima dobivenim prema hijerarhijskoj metodi.

**Primjer 3.32.**

Zadatak je izvršiti grupiranje odabranih 10 tranzicijskih zemalja prema 5 odgovarajućih pokazatelja vanjske zaduženosti za 2008. godinu. Neki od odabranih pokazatelja su pokazatelji stanja vanjskog duga, a neki su pokazatelji tijeka vanjskog duga.

Grupiranje je potrebno izvršiti:

- Hijerarhijskom metodom udruživanja ili aglomerativnom metodom!
- Nehijerarhijskom metodom k-prosjeka!

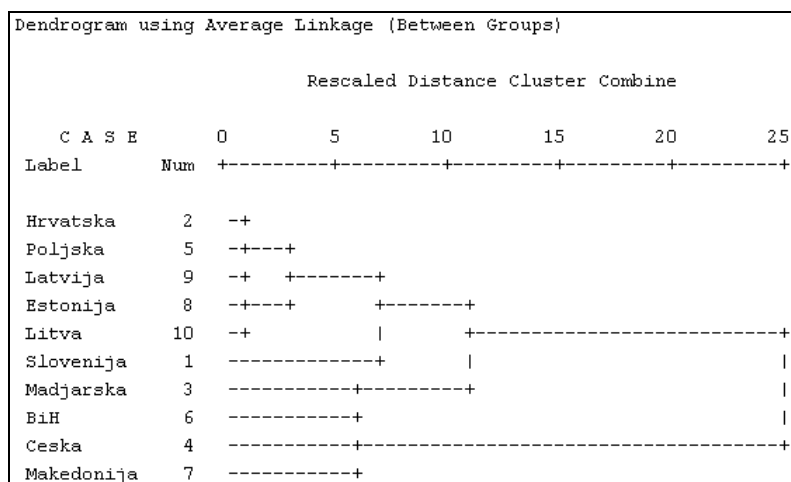
**Rješenje 3.32.**

a) U programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Classify**; **Hierarchical Cluster**. U prostor **Variable(s)** potrebno je odabrati sve promatrane pokazatelje zemalja, a u prostor **Label Cases by** potrebno je odabrati varijablu **Zemlje**. Odabirom opcije **Plots** treba aktivirati prikaz **Dendrogram**.

Klikom na ikonu **Continue**, odabirom opcije **Method** bira se **Between-groups linkage**, a za mjeru udaljenosti izabrana je **Measure: Euclidean distance**. Klikom na ikone **Continue** i **OK** u **Outputu** programa SPSS dobije se dendrogram, kako je prikazano na slici 3.31.

**Slika 3.31.**

**Klasifikacija odabranih zemalja na temelju odabranih pokazatelja vanjskog  
duga pomoću dendrograma**



*Izvor: Prema podacima Centralnih banaka promatranih zemalja i MMF-a (2010).*

Prema dendrogramu može se vidjeti da su Češka i Makedonija klasificirane u zasebnu grupu zajedno s BiH. Hrvatska je prema pokazateljima vanjske zaduženosti klasificirana kao slična s Poljskom, Latvijom i Estonijom i Litvom, a pridružuju im se i Mađarska i Slovenija. Može se zaključiti da se Hrvatska prema stanju i tijeku vanjske zaduženosti ne razlikuje značajno od nekih zemalja EU.

b) Da bi se promatrane zemlje klasificirale nehijerarhijskom metodom k-prosjeka u programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Classify**; **K-Means Cluster**. U prostor **Variable(s)**

potrebno je odabrati sve promatrane pokazatelje zemalja, a u prostor *Label Cases by* potrebno je odabrati varijablu *Zemlje*. Željeni i unaprijed određeni *Number of Clusters* je 2. Odabirom opcije *Options* treba aktivirati prikaz *Initial cluster centres* i *Cluster information for each case*. Klikom na ikone *Continue* i *OK* u *Outputu* programa SPSS dobije se klasifikacija zemalja, kako je prikazano u tablici 3.69.

**ANOVA testiranje** u tablici 3.68 odnosi se na svaku promatranu varijablu i upućuje na zaključak je li se sredine između predloženih klastera signifikantno razlikuju. Empirijske p-vrijednosti kreću se oko granične signifikantnosti od 5% i može se zaključiti da se sredine između predloženih klastera značajno razlikuju.

**Tablica 3.68.**

**ANOVA testiranje za klaster metodu k-prosjeka**

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
odnos ukupnog duga prema BDP-u (S)	,533	1	,067	8	7,919	,023
odnos ukupnog ino.duga prema izvozu roba i usluga (S)	1,904	1	,363	8	5,170	,053
odnos oplate duga prema izvozu roba i usluga (T)	,083	1	,005	8	15,781	,004
međunarodne pričuve kao poslotak vanjskog duga (S)	,232	1	,008	8	29,193	,001
odnos međunarodnih pričuva i oplate vanjskog duga (u mjesecima) (T)	2530,595	1	79,659	8	31,768	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Izvor: Prema podacima Centralnih banaka promatranih zemalja i MMF-a (2010).

**Tablica 3.69.**

**Klasifikacija tranzicijskih zemalja pomoću klaster metode k-prosjeka**

zemlje	Cluster
Slovenija	1
Hrvatska	1
Madjarska	1
Ceska	2
Poljska	1
BiH	2
Makedonija	2
Estonija	1
Latvija	1
Litva	1

Izvor: Prema podacima Centralnih banaka promatranih zemalja i MMF-a (2010).

Prema dobivenim podacima klasifikacije u tablici 3.69 može se vidjeti da su Češka i Makedonija opet klasificirane s BiH, kao što je prikazano i na dendrogramu pomoću hijerarhijske metode. Preostale zemlje su klasificirane u zasebnu grupu. Rezultati su u cijelosti slični onima dobivenim prema hijerarhijskoj metodi.



### Primjer 3.33.

Zadatak je izvršiti grupiranje odabranih 5 vodećih hrvatskih banaka prema 4 odabranih financijskih pokazatelja (od kojih su 3 pokazatelja profitabilnosti i 1 pokazatelj povrata na investirano) za 2009. godinu.

Grupiranje je potrebno izvršiti:

- Hijerarhijskom metodom udruživanja ili aglomerativnom metodom!
- Nehijerarhijskom metodom k-prosjeka!

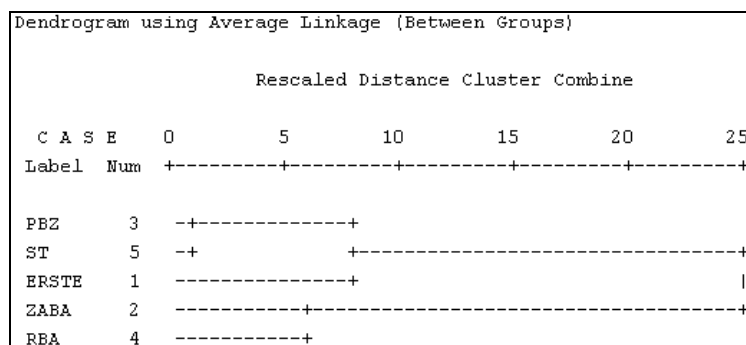


### Rješenje 3.33.

a) U programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Classify**; **Hierarchical Cluster**. U prostor **Variable(s)** potrebno je odabrati sve odabrane financijske pokazatelje banaka, a u prostor **Label Cases by** potrebno je odabrati varijablu **Banke**. Odabirom opcije **Plots** treba aktivirati prikaz **Dendrogram**.

### Slika 3.32.

#### Klasifikacija 5 vodećih banaka u RH na osnovu odabranih financijskih pokazatelja pomoću dendrograma



Izvor: Prema podacima [www.zse.hr](http://www.zse.hr) (2010).

Klikom na ikonu **Continue**, odabirom opcije **Method** bira se **Between-groups linkage**, a za mjeru udaljenosti izabrana je **Measure: Squared Euclidean distance**. Klikom na ikone **Continue** i **OK** u **Outputu** programa SPSS dobije se dendrogram, kako je prikazano na slici 3.32.

Prema dendrogramu može se vidjeti da su banke ZABA i RBA klasificirane kao slične prema odabranim pokazateljima. U drugoj grupi su PBZ, ST i ERSTE banka.

b) Da bi se promatrane banke klasificirale nehijerarhijskom metodom k-prosjeka u programskom paketu **SPSS** potrebno je na glavnom izborniku odabrati **Analyze**, a na njegovu padajućem izborniku **Classify; K-Means Cluster**. U prostor **Variable(s)** potrebno je odabrati sve promatrane financijske pokazatelje, a u prostor **Label Cases by** potrebno je odabrati varijablu **Banke**. Željeni i unaprijed određeni **Number of Clusters** je 2.

Odabirom opcije **Options** treba aktivirati prikaz **Initial cluster centres** i **Cluster information for each case**. Klikom na ikone **Continue** i **OK** u **Outputu** programa SPSS dobije se klasifikacija banaka, kako je prikazano u tablici 3.71.

**ANOVA testiranje** u tablici 3.70 odnosi se na svaku promatranu varijablu i upućuje na zaključak je li se sredine između predloženih klastera signifikantno razlikuju. Empirijske p-vrijednosti kreću se oko granične signifikantnosti od 5% i može se zaključiti da se sredine između predloženih klastera značajno razlikuju.

**Tablica 3.70.**

**ANOVA testiranje za klaster metodu k-prosjeka**

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Povrat na imovinu ROA (pok. prof.) 2009	,110	1	,011	3	9,974	,051
Neto kamatna marža (pok. prof.) 2009	,065	1	,013	3	3,700	,150
Neto nekamatna marža (pok. prof.) 2009	,054	1	,006	3	9,416	,055
Povrat na vlasnički kapital ROE (pok. pov. na inv.) 2009	4,888	1	,609	3	8,019	,066

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Izvor: Prema podacima [www.zse.hr](http://www.zse.hr) (2010).

Prema dobivenim podacima klasifikacije u tablici 3.71 može se vidjeti da su rezultati slični onima dobivenim prema hijerarhijskoj metodi. ZABA i RBA klasificirane su u jednu grupu, a u drugoj grupi su PBZ, ST i ERSTE banka.

**Tablica 3.71.**

**Klasifikacija 5 vodećih banaka u RH na osnovu odabranih financijskih pokazatelja pomoću klaster metode K-prosjeka**

banke	Cluster
ERSTE	1
ZABA	2
PBZ	1
RBA	2
ST	1

*Izvor: Prema podacima [www.zse.hr](http://www.zse.hr) (2010).*



## 4 ISPITIVANJE OVISNOSTI IZMEĐU NUMERIČKIH I NEKIH NENUMERIČKIH (DUMMY) VARIJABLI

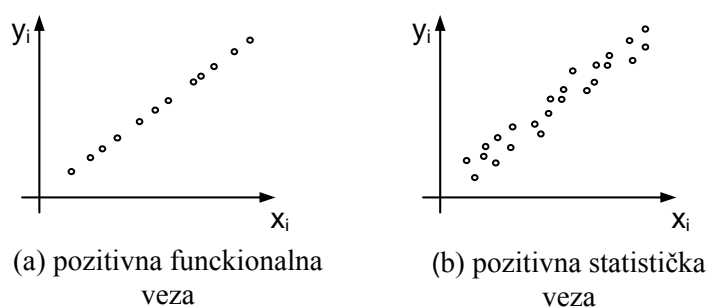
### 4.1 Dijagram rasipanja

U statističkim analizama često, uz analizu kretanja jedne pojave, postoji potreba istražiti ovisnosti dviju ili više pojava, odnosno numeričkih nizova, zajedno.

Prvi korak u istraživanju ovisnosti varijabli je crtanje grafičkog prikaza koji se zove dijagram rasipanja. **Dijagram rasipanja u pravokutnom koordinatnom sustavu točkama  $(x_i, y_i)$  prikazuje parove vrijednosti dviju promatranih numeričkih varijabli.** Na osnovu takve slike mogu se odmah uočiti osnovne veze među promatranim varijablama.

**Slika 4.1.**

**Dijagram rasipanja s pozitivnom vezom između promatranih varijabli**



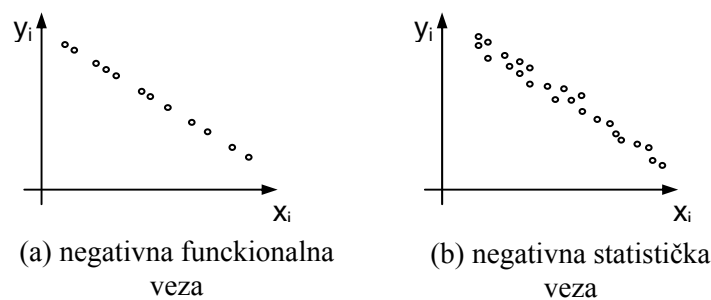
*Izvor: Konstrukcija autora.*

Na slici 4.1 su prikazana 2 dijagrama rasipanja. Slika (a) prikazuje funkcionalnu vezu između 2 varijable  $X$  i  $Y$ . Zamišljena linija koja povezuje sve točke na slici je pravac. Matematički oblik veze ove dvije promatrane varijable je jednačba pravca, od čije linije nema nikakvog odstupanja. Stoga se kaže da je ova veza strogo funkcionalna. Zamišljeni pravac je rastući, odnosno porast vrijednosti jedne varijable prati porast vrijednosti druge promatrane varijable pa je ova veza pozitivna. U praksi je čest slučaj prikazan na slici (b). Ako se između točaka ovog dijagrama zamisli krivulja, to bi opet bio pravac. Međutim ovdje su prisutna

pozitivna i negativna odstupanja od linije pravca, što se tumači raznim utjecajima drugih varijabli iz okoline. Stoga ova veza više nije strogo funkcionalna, već se kaže da je to statistička (stohastička ili slučajna) veza. I ovdje porast vrijednosti jedne varijable u prosjeku prati porast druge varijable, pa je i ova veza pozitivna.

#### Slika 4.2.

##### Dijagram rasipanja s negativnom vezom između promatranih varijabli



Izvor: Konstrukcija autora.

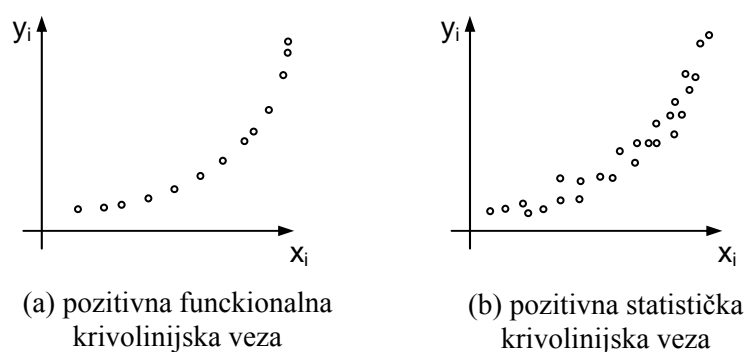
Na slici 4.2 su opet prikazana 2 dijagrama rasipanja. Slika (a) prikazuje funkcionalnu vezu između 2 varijable  $X$  i  $Y$ , a zamišljena linija koja povezuje sve točke na slici je opet pravac. Matematički oblik veze ove dvije promatrane varijable je jednačba pravca, od čije linije nema nikakvog odstupanja, pa je i ova veza strogo funkcionalna. Zamišljeni pravac je padajući, odnosno porast vrijednosti jedne varijable prati pad vrijednosti druge promatrane varijable pa je ova veza negativna. U praksi je čest slučaj prikazan na slici (b). Ako se između točaka ovog dijagrama zamisli krivulja, to bi opet bio pravac. Međutim ovdje su prisutna pozitivna i negativna odstupanja od linije pravca, što se, kako je već rečeno, tumači raznim utjecajima drugih varijabli iz okoline. Stoga ova veza više nije strogo funkcionalna, već se kaže da je to statistička (stohastička ili slučajna) veza. Porast vrijednosti jedne varijable u prosjeku prati pad druge varijable, pa je i ova veza negativna.

Veza između promatranih varijabli ne mora uvijek odgovarati jednačbi pravca. Na slici 4.3 su prikazana dva dijagrama rasipanja. Slika (a) prikazuje funkcionalnu krivolinijsku pozitivnu vezu između 2 varijable  $X$  i  $Y$ . Zamišljena linija koja povezuje sve točke na slici je krivulja. Matematički oblik veze ove dvije promatrane varijable je neka eksponencijalna jednačba, od čije linije nema nikakvog odstupanja, pa je ova veza strogo funkcionalna. I ovdje vrijedi da porast vrijednosti jedne varijable prati porast vrijednosti druge promatrane varijable pa je ova veza pozitivna. U praksi se češće događa slučaj prikazan na slici (b). Ako se između točaka ovog dijagrama zamisli linija, to bi opet bila krivulja. Međutim ovdje

su prisutna pozitivna i negativna odstupanja zbog utjecajima drugih varijabli iz prakse. Ova veza je statistička (stohastička ili slučajna). I ovdje porast vrijednosti jedne varijable u prosjeku prati porast druge varijable, pa je i ova veza pozitivna.

**Slika 4.3.**

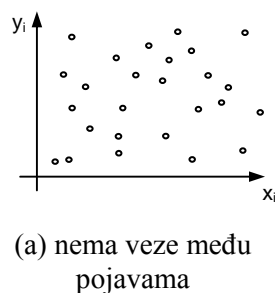
**Dijagram rasipanja s pozitivnom nelinearnom vezom između promatranih varijabli**



*Izvor: Konstrukcija autora.*

**Slika 4.3.**

**Dijagram rasipanja s varijablama koje ne pokazuju ovisnost**

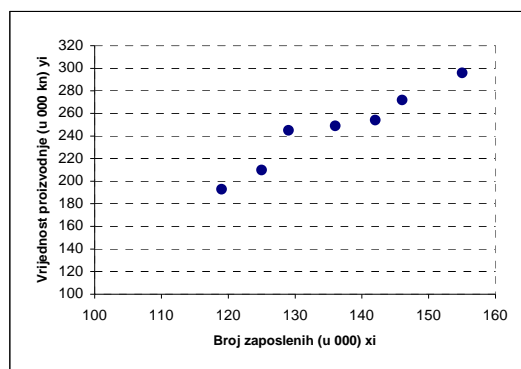


*Izvor: Konstrukcija autora.*

Na slici 4.3 je prikazan dijagram rasipanja koji upućuje na zaključak da nema povezanosti među promatranim pojavama. Naime, zamišljena krivulja koja prolazi između točaka na ovom grafikonu ne postoji, i ne može se definirati je li porast jedne pojave prati rast ili pad druge promatrane pojave, jer pri jednoj vrijednosti varijable  $x_i$  može se dogoditi više različitih vrijednosti druge varijable  $y_i$ .

**Slika 4.4.**

**Dijagram rasipanja vrijednosti proizvodnje u tekućim cijenama (u 000 kn) i broj zaposlenih (u tis.)**



*Izvor: Simulirani podaci.*

Raspored točaka dijagrama rasipanja na slici 4.4 upućuje na **pozitivnu statističku vezu** između vrijednosti proizvodnje i broja zaposlenih. Između točaka može se zamisliti linija pravca, ali sve točke ne leže na toj liniji već postoje pozitivna i/ili negativna odstupanja. Dakle, u ovom slučaju, porast broja zaposlenih na promatranom području prati porast vrijednosti proizvodnje.



#### **Primjer 4.1.**

Za uzorak od 233 ispitanika studentske populacije na jednom području evidentirani su podaci o njihovoj visini u cm i težini u kg. Potrebno je zadane varijable prikazati pomoću dijagrama rasipanja!



#### **Rješenje 4.1.**

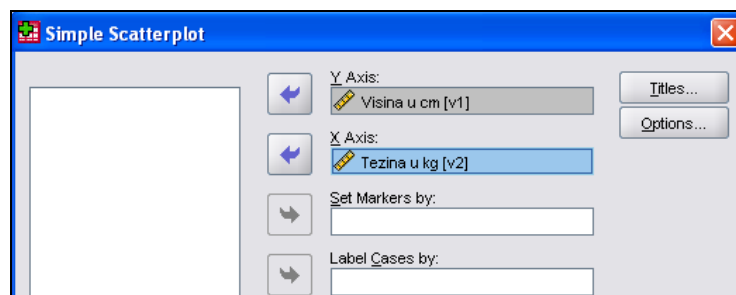
U programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Graphs**, a na njezinom padajućem izborniku **Legacy Dialogs** i na pomoćnom izborniku **Scatter/Dot** gdje se bira grafikon **Simple Scatter**. U otvorenom prozoru definira se varijabla Visina u cm (v1) u: **Y Axis**, a Težina u kg (v2) u: **X Axis**.

Na slici 4.5 prikazan je dio prozora "Simple Scatterplot" s odabranim varijablama za dijagram rasipanja.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi dijagram rasipanja, koji je prikazan na slici 4.6.

Slika 4.5.

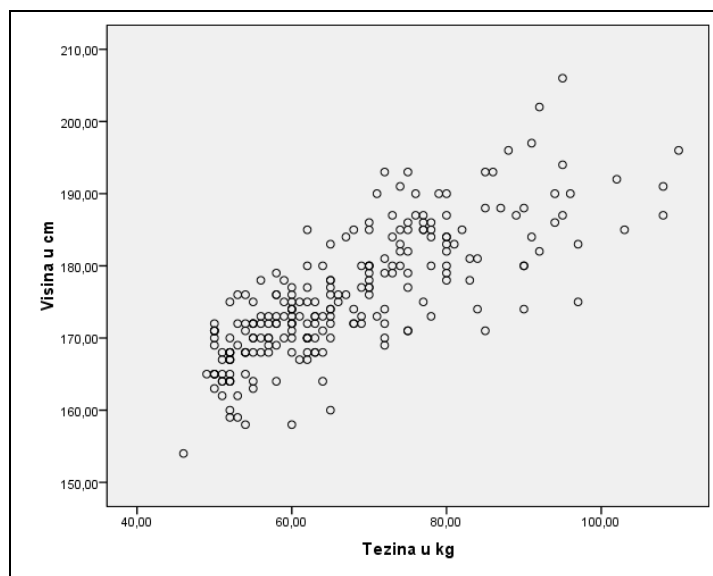
Dio prozora "Simple Scatterplot" s odabranim varijablama za dijagram rasipanja



Izvor: Simulirani podaci.

Slika 4.6.

Dijagram rasipanja visine i težine uzorka studenata



Izvor: Simulirani podaci.

Može se zaključiti da postoji pozitivna statistička veza između visine i težine studenata na promatranom području, odnosno prema rasporedu točaka dijagrama rasipanja, ako neki student ima veću težinu, može se očekivati i njegova veća visina.

## 4.2 Koeficijent linearne korelacije

Pod pojmom **korelacija** podrazumijeva se **međuzavisnost ili povezanost slučajnih numeričkih varijabli**. Po smjeru korelacija može biti pozitivna i negativna.

**Pozitivna korelacija** je prisutna kada rast jedne varijable prati rast druge promatrane varijable, odnosno kada pad jedne prati pad druge varijable. **Negativna korelacija** prisutna je kada rast jedne varijable prati pad druge varijable i obratno.

Najpoznatija mjera linearne korelacije između slučajnih varijabli je **Pearsonov koeficijent linearne korelacije (r)**:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}}, \quad \text{ili} \quad (4.1)$$

$$r = \frac{\sum_{i=1}^n X_i Y_i - n \cdot \bar{X} \cdot \bar{Y}}{n \cdot \sigma_x \cdot \sigma_y}, \quad (4.2)$$

gdje su  $\sigma_x$  i  $\sigma_y$  jednostavne standardne devijacije promatranih varijabli:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \bar{X}^2} \quad \text{i} \quad (4.3)$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \bar{Y}^2}. \quad (4.4)$$

Vrijednost koeficijenta linearne korelacije kreće se u intervalu:  $-1 \leq r \leq 1$ .

U skladu s veličinom ovog koeficijenta može se zaključiti **smjer i intenzitet linearne korelacije** među promatranim varijablama:

$r = -1$ ; $r = 1$	$\Rightarrow$ funkcionalna negativna/pozitivna korelacija
$-1 < r \leq -0,8$ ; $0,8 \leq r < 1$	$\Rightarrow$ jaka negativna/pozitivna korelacija
$-0,8 < r \leq -0,5$ ; $0,5 \leq r < 0,8$	$\Rightarrow$ srednje jaka negativna/pozitivna korelacija
$-0,5 < r < 0$ ; $0 < r < 0,5$	$\Rightarrow$ slaba negativna/pozitivna korelacija

$r = 0 \quad \Rightarrow \text{nema korelacije.}$

#### 4.2.1 Testiranje hipoteze da je koeficijent korelacije jednak nuli

Da bi se izvršilo testiranje pretpostavke da je koeficijent korelacije jednak nuli postavlja se nulta hipoteza koja pretpostavlja da je vrijednost koeficijenta korelacije osnovnog skupa jednaka 0, tj. da ne postoji korelacija između slučajnih varijabli. **Sampling distribucija** koeficijenta korelacije za dovoljno veliki uzorak i za  $r = 0$  ima oblik normalne distribucije, pa u ovom testiranju nisu potrebne nikakve transformacije.

$$H_0: \dots r = 0$$

$$H_1: \dots r \neq 0$$

**Interval prihvatanja nulte hipoteze je:**

$$0 \pm Z \cdot Se(r),$$

gdje standardna greška koeficijenta korelacije o ovisnosti o veličini uzorka  $n$ :

$$Se(r) = \sqrt{\frac{1}{n-1}}, \quad n > 30, \quad (4.5)$$

$$Se(r) = \sqrt{\frac{1-\hat{r}^2}{n-2}}, \quad n \leq 30, \quad (4.6)$$

$Z_{\frac{1-\alpha}{2}}$  - odgovarajuća vrijednost  $Z$  testa iz tablica površina ispod normalne krivulje

$\alpha$  - nivo signifikantnosti ili značajnosti testa (ako nije određeno drukčije, najčešće se uzima da je  $\alpha = 5\%$ ).

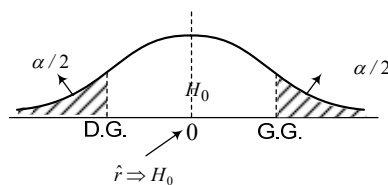
Kod malog uzorka, umjesto  $Z$  koristi se vrijednost  $t$  - iz Studentove distribucije:  $t_{\frac{\alpha}{2}, [df=n-2]}$ .

Zaključak o prihvatanju ili odbacivanju nulte  $H_0$  hipoteze donosi se pomoću  $\hat{r}$  koeficijenta korelacije iz uzorka.

Prema slici 4.8, ako se koeficijent korelacije iz uzorka  $\hat{r}$  nalazi između donje (D.G.) i gornje granice (G.G.) intervala prihvatanja hipoteze  $H_0$ , ta se hipoteza prihvata kao istinita uz odgovarajući nivo signifikantnosti testa ( $\alpha$ ), tj. zaključuje se da ne postoji korelacija između slučajnih varijabli. U suprotnom se ta hipoteza odbacuje.

**Slika 4.8.**

**Odluka o prihvatanju hipoteza kod testiranja značajnosti koeficijenta korelacije**



Izvor: Konstrukcija autora.

Testiranje se može napraviti usporedbom empirijske i tablične vrijednosti  $Z$ -testa (ili  $t$ -testa za manji uzorak):

$$\left| Z^* = \frac{\hat{r}}{Se(r)} \right| < Z_{tab} \Rightarrow H_0,$$

što znači da korelacija između promatranih varijabli nije statistički značajna.

Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $Z^*$  ili  $t^*$ : ako je  $\alpha^* > 5\% \Rightarrow H_0$ , dok se u suprotnom slučaju hipoteza  $H_0$  odbacuje.



**Primjer 4.2.**

Za uzorak od 233 ispitanika studentske populacije na jednom području evidentirani su podaci o njihovoj visini u cm i težini u kg. Potrebno je izračunati linearnu korelaciju između odabranih varijabli te izvršiti testiranje značajnosti izračunatog koeficijenta uz signifikantnost od 5%.



**Rješenje 4.3.**

Da bi se u programskom paketu **SPSS** dobio Pearsonov koeficijent linearne korelacije između odabranih varijabli potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Correlate** gdje se bira **Bivariate**.

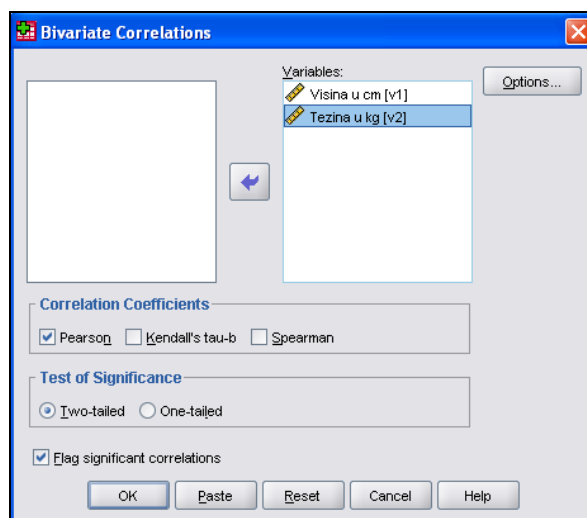


U otvorenom prozoru biraju se varijable Visina u cm (v1) i Težina u kg (v2) u: **Variables**. U **Correlation Coefficients**: već je aktiviran **Pearson** koeficijent.

Na slici 4.9 prikazan je prozor "Bivariate Correlations" s odabranim varijablama za izračunavanje Pearsonovog koeficijenta korelacije.

Slika 4.9.

**Prozor "Bivariate Correlations" s odabranim varijablama za izračunavanje Pearsonovog koeficijenta korelacije**



*Izvor: Simulirani podaci.*

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi rezultat, Pearsonovog koeficijenta korelacije između varijabli Visina u cm i Težina u kg, koji je prikazan u tablici 4.1.

Tablica 4.1.

**Tablica s koeficijentom korelacije između visine i težine uzorka studenata**

Correlations			
		Visina u cm	Težina u kg
Visina u cm	Pearson Correlation	1,000	,781**
	Sig. (2-tailed)		,000
	N	235,000	233
Težina u kg	Pearson Correlation	,781**	1,000
	Sig. (2-tailed)	,000	
	N	233	234,000

\*\* . Correlation is significant at the 0.01 level (2-tailed).

*Izvor: Simulirani podaci.*

Može se vidjeti da je korelacija između visine i težine uzorka od 233 stanovnika na promatranom području:  $\hat{r} = 0,781$ . To znači pozitivnu vezu između promatranih varijabli, odnosno ako neki stanovnik ima veću težinu, može se očekivati da je i njegova visina veća.

Da bi se testirala značajnost izračunatog Pearsonovog koeficijenta korelacije postavljaju se hipoteze:

$$H_0: \dots r = 0$$

$$H_1: \dots r \neq 0$$

Prema rezultatima iz tablice 4.1 može se vidjeti da je empirijska signifikantnost koeficijenta korelacije  $\alpha^* \approx 0\%$ , pa se može zaključiti da je  $\alpha^* < 5\%$  i da se početna hipoteza može odbaciti. Dakle, koeficijent korelacije između visine i težine stanovnika na promatranom području je statistički značajan uz signifikantnost testa od 5% (ispod tablice 4.1 navedeno je da se do istog zaključka o značajnosti koeficijenta može doći i uz signifikantnost od 1%).

### 4.3 Koeficijent korelacije ranga

Ako se želi istražiti međuovisnost pojava koje su izražene modalitetima redosljednog obilježja, odnosno ako su im modaliteti pridruženi na temelju ordinarne skale računa se korelacija ranga.

Najpoznatija **mjera korelacije ranga između dvije varijable je Spearmanov koeficijent korelacije ranga ( $r_s$ )**:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^N d_i^2}{N^3 - N}, \quad (4.7)$$

gdje je:

$N$  - broj parova vrijednosti varijabli X i Y,

$d_i = r(x_i) - r(y_i)$  - razlika rangova vrijednosti varijabli X i Y.

Za izračunavanje ovog koeficijenta potrebno je svakoj vrijednosti varijabli X i Y dodijeliti rang iskazan prvim N prirodnim brojevima. Pri tome se rangiranje može započeti rangom 1, počevši od najmanje vrijednosti obilježja ili počevši od najveće vrijednosti obilježja. Rangiranje se mora provesti na jednak način za obje promatrane

varijable. Ako se javi **više jednakih vrijednosti jedne varijable** mora im se dodijeliti jednak rang na način da se **izračuna aritmetička sredina njihovih rangova**.

Spearmanov koeficijent korelacije ranga može poprimiti vrijednosti u intervalu:  $-1 \leq r_s \leq 1$ . Kada ovaj koeficijent poprimi vrijednosti -1 i 1, riječ je o potpunoj korelaciji ranga među varijablama. Vrijednost ovog koeficijenta 0 znači da nema nikakve korelacije ranga među pojavama. Najčešće se vrijednost Spearmanovog koeficijenta kreće u rasponu  $-1 < r_s < 1$ . Koeficijent bliži rubovima ovog intervala, tj. -1 i 1 upućuje na veću korelaciju ranga promatranih dviju varijabli.

### **Primjer 4.3.**

U tablici 4.2 su zadani podaci za 12 studenata fakulteta "E" o bodovima na testu iz predmeta: Mikroekonomija i Statistika.

Zadatak je izračunati Spearmanov koeficijent korelacije ranga uspjeha studenata na testovima.

### **Rješenje 4.3.**

Prema podacima iz zadane tablice Spearmanov koeficijent korelacije ranga je:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^N d_i^2}{N^3 - N} = 1 - \frac{6 \cdot 19,5}{12^3 - 12} = 0,931818.$$

**Tablica 4.2.**

**Bodovi na testu iz Mikroekonomije i Statistike studenata fakulteta "E"**

Studenti	Bodovi na testu iz Mikroekonomije ( $x_i$ )	Bodovi na testu iz Statistike ( $y_i$ )	Rang $r(x_i)$	Rang $r(y_i)$	Razlika rangova $d_i$	$d_i^2$
A	50*	45*	6,5	5	1,5	2,25
B	64	70	9	11	-2	4
C	43	45*	5	5	0	0
D	80	75	12	12	0	0
E	21	38	2	2	0	0
F	57	60	8	8	0	0
G	50*	45*	6,5	5	1,5	2,25
H	37	50	4	7	-3	9
I	10	25	1	1	0	0
J	75	68	11	10	1	1
K	65	63	10	9	1	1
L	35	40	3	3	0	0
<b>Ukupno:</b>	-	-	-	-	0	19,5

*Izvor: Simulirani podaci.*

Spearmanov koeficijent korelacije ranga je pozitivan i iznosi 0.93. Može se zaključiti da u ovom primjeru između uspjeha na testu iz Mikroekonomije i Statistike postoji visoka korelacija ranga. Odnosno, ako je student postigao dobar rezultat na testu iz Mikroekonomije, može se očekivati da je postigao dobar rezultat i na testu iz Statistike i obrnuto.



#### Primjer 4.4.

Na fakultetu "E" izvršeno je anonimno istraživanje anketnim upitnikom među studentskom populacijom. Na taj način došlo se do podataka o prosječnoj ocjeni u srednjoj školi i prosječnoj ocjeni na I godini studija za uzorak od 224 studenta. Potrebno je izračunati korelaciju ranga između odabranih varijabli te izvršiti testiranje značajnosti izračunatog koeficijenta uz signifikantnost od 5%.

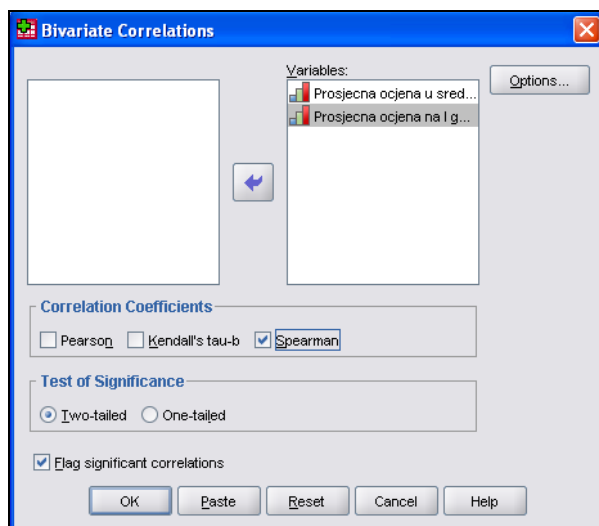


#### Rješenje 4.4.

Da bi se u programskom paketu **SPSS** dobio Spearmanov koeficijent korelacije ranga između odabranih varijabli potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Correlate** gdje se bira **Bivariate**.

Slika 4.10.

#### Prozor "Bivariate Correlations" s odabranim varijablama za izračunavanje Spearmanovog koeficijenta korelacije



Izvor: Simulirani podaci.

U otvorenom prozoru biraju se varijable Prosječna ocjena u srednjoj školi (v1) i Prosječna ocjena na I. godini studija (v2) u: **Variables**. U **Correlation Coefficients**: treba aktivirati **Spearman** koeficijent. Na slici 4.10 prikazan je prozor "Bivariate Correlations" s odabranim varijablama za izračunavanje Spearmanovog koeficijenta korelacije.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi rezultat, Spearmanovog koeficijenta korelacije između varijabli Prosječna ocjena u srednjoj školi i Prosječna ocjena na I. godini, koji je prikazan u tablici 4.3.

**Tablica 4.3.**

**Tablica s koeficijentom korelacije ranga između prosječne ocjene u srednjoj školi i prosječne ocjene na I. godini**

Correlations				
			Prosječna ocjena u srednjoj školi	Prosječna ocjena na I. god
Spearman's rho	Prosječna ocjena u srednjoj školi	Correlation Coefficient	1,000	,388**
		Sig. (2-tailed)		,000
		N	238	224
	Prosječna ocjena na I. god	Correlation Coefficient	,388**	1,000
		Sig. (2-tailed)	,000	
		N	224	225

\*\* . Correlation is significant at the 0.01 level (2-tailed).

*Izvor: Simulirani podaci.*

Prema tablici 4.3 može se zaključiti da je korelacija između prosječne ocjene u srednjoj školi i prosječne ocjene na I. godini uzorka od 224 studenta:  $\hat{r} = 0,338$ . To znači ne izrazito jaku, ali pozitivnu vezu između promatranih varijabli, odnosno ako je neki student imao veću prosječnu ocjenu u srednjoj školi, može se očekivati i njegova veća prosječna ocjena na I. godini. Da bi se testirala značajnost izračunatog Spearmanovog koeficijenta korelacije postavljaju se hipoteze:

$$H_0, \dots, r = 0$$

$$H_1, \dots, r \neq 0$$

Prema rezultatima iz tablice 4.3 može se vidjeti da je empirijska signifikantnost koeficijenta korelacije  $\alpha^* \approx 0\%$ , pa se može zaključiti da je  $\alpha^* < 5\%$  i da se početna hipoteza može odbaciti. Dakle, koeficijent korelacije ranga između prosječne ocjene u srednjoj školi i prosječne ocjene na I. godini studenata na promatranom području je statistički značajan uz signifikantnost testa od 5% (ispod tablice 4.3 navedeno je da se do istog zaključka o značajnosti koeficijenta može doći i uz signifikantnost od 1%).

## 4.4 Koeficijent parcijalne korelacije

**Koeficijent parcijalne korelacije** je pokazatelj korelacije između dvije varijable uz istodobno isključenje utjecaja drugih varijabli.

Ako se računa **parcijalna korelacija između tri varijable** u kombinaciji vrijedi da je:

- korelacija između 1. i 2. varijable uz isključenje utjecaja 3. varijable:

$$\rho_{12.3} = \frac{r_{12} - (r_{13} \cdot r_{23})}{\sqrt{(1 - r_{13}^2) \cdot (1 - r_{23}^2)}}, \quad (4.8)$$

- korelacija između 1. i 3. varijable uz isključenje utjecaja 2. varijable:

$$\rho_{13.2} = \frac{r_{13} - (r_{12} \cdot r_{32})}{\sqrt{(1 - r_{12}^2) \cdot (1 - r_{32}^2)}}, \quad (4.9)$$

- korelacija između 2. i 3. varijable uz isključenje utjecaja 1. varijable:

$$\rho_{23.1} = \frac{r_{23} - (r_{21} \cdot r_{31})}{\sqrt{(1 - r_{21}^2) \cdot (1 - r_{31}^2)}}, \quad (4.10)$$

gdje su  $r_{ij}$  odgovarajući koeficijenti korelacije promatranih varijabli.

**Matrica koeficijenata korelacije** između  $k$  varijabli je:

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{21} & 1 & r_{23} & \dots & r_{2k} \\ r_{31} & r_{32} & 1 & \dots & r_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & r_{k3} & \dots & 1 \end{bmatrix}, \quad (4.11)$$

gdje je  $k$  broj promatranih varijabli. S obzirom da za korelaciju vrijedi da je  $r_{ij} = r_{ji}$ , matrica  $R$  je simetrična matrica.

**Koeficijent parcijalne korelacije između 2 varijable, uz isključenje utjecaja ostalih promatranih varijabli** je općenito:

$$\rho_{ij.klm\dots} = -\frac{R_{ij}}{\sqrt{(R_{ii} \cdot R_{jj})}}, \quad (4.12)$$

gdje je  $R_{ij}$  kofaktor, tj. algebarski komplement matrice koeficijenata korelacije  $R$  :

$$R_{ij} = (-1)^{i+j} \cdot M_{ij}, \quad (4.13)$$

a  $M_{ij}$  je odgovarajući minor (subdeterminanta) matrice koeficijenata korelacije  $R$ .



#### Primjer 4.5.

Na fakultetu "E" izvršeno je anonimno istraživanje anketnim upitnikom među studentskom populacijom. Na taj način došlo se do podataka o njihovoj visini u cm, težini u kg, broju djece u obitelji i džeparcu. Potrebno je izračunati linearnu korelaciju između: visine u cm, težine u kg i broja djece u obitelji uz isključenje utjecaja džeparca, te komentirati značajnost izračunatih pokazatelja uz signifikantnost od 5%.

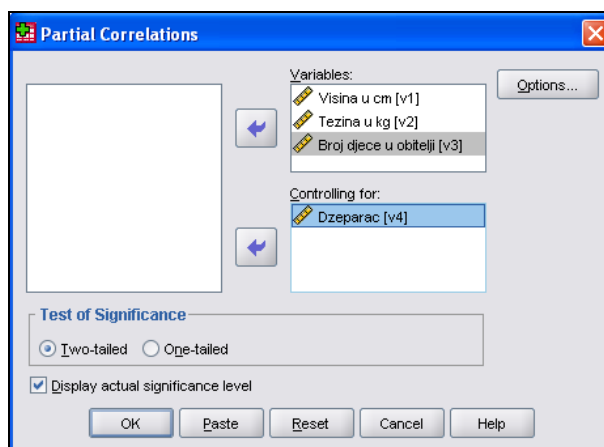


#### Rješenje 4.5.

Da bi se za navedene varijable izračunali koeficijenti parcijalne korelacije između: visine u cm, težine u kg, broja djece u obitelji uz isključenje utjecaja džeparca u programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Correlate** gdje se bira **Partial**.

Slika 4.11.

**Prozor "Partial Correlations" s odabranim varijablama za izračunavanje odgovarajućih parcijalnih koeficijenata korelacije**



*Izvor: Simulirani podaci.*

U otvorenom prozoru biraju se varijable Visina u cm (v1), Težina u kg (v2), Broj djece u obitelji (v3) u: **Variables**. Zatim se varijabla Džeparac (v4) bira u **Controlling for**. Na slici 4.11 prikazan je prozor "Partial Correlations" s odabranim varijablama za izračunavanje odgovarajućih parcijalnih koeficijenata korelacije.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se traženi rezultati koji su prikazani u tablici 4.4.

**Tablica 4.4.**

**Tablica s parcijalnim koeficijentima korelacije između visine, težine i broja djece u obitelji uz isključenje utjecaja visine džeparca zadanog uzorka stanovnika**

Correlations					
Control Variables			Visina u cm	Težina u kg	Broj djece u obitelji
Džeparac	Visina u cm	Correlation	1,000	,796	-,103
		Significance (2-tailed)	.	,000	,183
		df	0	166	166
	Težina u kg	Correlation	,796	1,000	-,147
		Significance (2-tailed)	,000	.	,057
		df	166	0	166
	Broj djece u obitelji	Correlation	-,103	-,147	1,000
		Significance (2-tailed)	,183	,057	.
		df	166	166	0

*Izvor: Simulirani podaci.*

Prema podacima u tablici 4.4 može se zaključiti da je najveća korelacija između visine i težine uzorka od 166 stanovnika na promatranom području uz isključenje utjecaja džeparca ispitanika:  $\hat{\rho}_{12,4} = 0,796$ . To znači pozitivnu vezu između promatranih varijabli, odnosno ako neki stanovnik ima veću težinu, može se očekivati i njegova veća visina (i obratno), uz isključen utjecaj visine džeparca.

Da bi se testirala značajnost izračunatih parcijalnih koeficijenta korelacije postavljaju se hipoteze za svaki koeficijent pojedinačno:

$$H_0 \dots \rho = 0$$

$$H_1 \dots \rho \neq 0$$

Prema podacima u tablici 4.4 može se vidjeti da je samo jedan izračunati koeficijent parcijalne korelacije statistički značajan uz signifikantnost testa od 5% (i uz 1%), jer mu je empirijska signifikantnost  $\alpha^* \approx 0\%$ . To je korelacija  $\hat{\rho}_{12,4} = 0,796$  između visine i težine ispitanika uz isključenje utjecaja džeparca. Parcijalna korelacija između težine ispitanika i broja djece u obitelji uz isključenje utjecaja džeparca  $\hat{\rho}_{23,4} = -0,147$  je značajna uz 10% signifikantnosti testa, jer joj je empirijska signifikantnost  $\alpha^* = 0,057 = 5,7\%$ . Ova korelacija je negativna i niska, što znači da se



kod ispitanika s većom težinom može očekivati manji broj djece u obitelji i obratno, uz isključenje utjecaja džeparca.

## 4.5 Kendallov koeficijent korelacije ranga

Kendallov koeficijent korelacije ranga mjeri stupanj korelacije skupine od  $k \geq 3$  varijabli ranga.

Kod izračunavanja ovog pokazatelja pretpostavlja se da su vrijednosti svih varijabli ranga izražene s prvih  $N$  prirodnih brojeva. Ako su varijable originalno dane u numeričkom obliku, potrebno ih je transformirati u varijable ranga.

$$W = \frac{\sum_{i=1}^N (\bar{R}_i - \bar{R})^2}{\frac{1}{12} \cdot N(N^2 - 1)}, \quad (4.14)$$

gdje je:

$$\bar{R}_i = \frac{\sum_{k=1}^K r_{ki}}{K} \quad (4.15)$$

aritmetička sredina rangova ( $r_{ki}$ ) po svim varijablama  $k$ , za svaku opservaciju  $i$ , a

$$\bar{R} = \frac{\sum_{i=1}^N \sum_{k=1}^K r_{ki}}{K \cdot N} \quad (4.16)$$

aritmetička sredina svih rangova ( $r_{ki}$ ) za sve varijable  $k$  i sve opservacije  $i$ .

Ovaj koeficijent se kreće u intervalu:  $0 \leq W \leq 1$ .

Ako je  $W=0$  to znači potpuno neslaganje rangova promatranih  $K$  varijabli.

Kada se vrijednost ovog pokazatelja od  $0$  približava  $1$  stupanj korelacije promatrane skupine varijabli ranga raste.

Ako je  $W=1$  to znači potpunu podudarnost rangova promatranih  $K$  varijabli.

### **Primjer 4.6.**

Zadani su rangovi za 19 različitih regija prema 3 regionalna pokazatelja:

$r_1$  - društveni proizvod po stanovniku (DP/stan.)

$r_2$  - broj industrijskih poduzeća s pozitivnim poslovanjem (Ind. pod.)

$r_3$  - broj stanovnika s visokom stručnom spremom na 1000 stanovnika (VSS).

Potrebno je izračunati Kendallov koeficijent korelacije ranga između ovih varijabli.

#### **Rješenje 4.6.**

**Tablica 4.5.**

**Rangovi za 19 regija prema 3 regionalna pokazatelja**

Regija	DP/stan. $r_{1i}$	Ind. pod. $r_{2i}$	VSS $r_{3i}$	$\sum_{k=1}^K r_{ki}$	Prosjek retka $\bar{R}_i$	Kvadrat razlike $(\bar{R}_i - \bar{R})^2$
A	2	3	2	7	2,3333	58,7778
B	4	4	3	11	3,6667	40,1111
C	1	2	1	4	1,3333	75,1111
D	7	7	6	20	6,6667	11,1111
E	3	1	4	8	2,6667	53,7778
F	5	6	7	18	6,0000	16,0000
G	6	5	5	16	5,3333	21,7778
H	9	10	9	28	9,3333	0,4444
I	11	11	10	32	10,6667	0,4444
J	8	9	8	25	8,3333	2,7778
K	14	14	13	41	13,6667	13,444
L	10	8	11	29	9,6667	0,1111
M	12	13	14	39	13,000	9,0000
N	13	12	12	37	12,3333	5,4444
O	16	17	16	49	16,3333	40,1111
P	18	18	17	53	17,6667	58,7778
R	15	16	15	46	15,3333	28,4444
S	19	19	19	57	19,0000	81,0000
T	17	15	18	50	16,6667	44,4444
$\Sigma$	190	190	190	-	-	561,1111

*Izvor: Podaci su simulirani.*

Prema podacima iz tablice Kendallov koeficijent korelacije ranga između zadanih varijabli je:

$$W = \frac{\sum_{i=1}^N (\bar{R}_i - \bar{R})^2}{\frac{1}{12} \cdot N(N^2 - 1)} = \frac{561,1111}{\frac{1}{12} \cdot 19(19^2 - 1)} = 0,9844,$$

gdje je aritmetička sredina svih rangova ( $r_{ki}$ ) za sve varijable  $k$  i sve opservacije  $i$ :

$$\bar{R} = \frac{\sum_{i=1}^N \sum_{k=1}^K r_{ki}}{K \cdot N} = \frac{570}{3 \cdot 19} = 10.$$

Izračunati Kendallov koeficijent korelacije ranga ( $W=0,9844$ ) pokazuje da postoji visoki stupanj podudarnosti rangova u 19 odabranih regija prema 3 zadana društveno ekonomska pokazatelja. To bi značilo da regije s visokim društvenim proizvodom po stanovniku imaju veći broj industrijskih poduzeća s pozitivnim poslovanjem, kao i veći broj visoko obrazovanih stanovnika (i obratno).



#### Primjer 4.7.

Na fakultetu "E" izvršeno je anonimno istraživanje anketnim upitnikom među studentskom populacijom. Uzeti su podaci za uzorak od 30 studenta o njihovoj prosječnoj ocjeni u srednjoj školi, prosječnoj ocjeni na I godini studija, te prosječnoj ocjeni na II godini studija.

Potrebno je izračunati Spearmanov koeficijent korelacije ranga između odabranih varijabli te izvršiti testiranje značajnosti dobivenih koeficijenta uz signifikantnost od 5%! Zatim je potrebno izračunati Kendallov koeficijent korelacije ranga između odabranih varijabli te izvršiti testiranje značajnosti dobivenih koeficijenta uz signifikantnost od 5%!



#### Rješenje 4.7.

Da bi se u programskom paketu **SPSS** dobila matrica Spearmanovih i Kendallovih koeficijenata korelacije između odabranih varijabli potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Correlate** gdje se bira **Bivariate**. U otvorenom prozoru biraju se varijable Prosječna ocjena u srednjoj školi (v1), Prosječna ocjena na I godini studija (v2), Prosječna ocjena na II godini studija (v3) u: **Variables**. U **Correlation Coefficients**: treba aktivirati **Spearman** i **Kendall's tau\_b** koeficijente.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobiju se traženi rezultati koeficijenata korelacije između zadanih varijabli, koji su prikazani u tablici 4.6.

Tablica 4.6.

Tablica s koeficijentima korelacije između prosječne ocjene u srednjoj školi, na I godini studija i na II godini studija uzorka studenata

Correlations					
			Prosječna ocjena u srednjoj školi	Prosječna ocjena na I godini	Prosječna ocjena na II godini
Kendall's tau_b	Prosječna ocjena u srednjoj školi	Correlation Coefficient	1,000	,257	,279
		Sig. (2-tailed)		,077	,057
	Prosječna ocjena na I godini	N	30	30	30
		Correlation Coefficient	,257	1,000	,578**
	Prosječna ocjena na II godini	Sig. (2-tailed)	,077		,000
		N	30	30	30
Spearman's rho	Prosječna ocjena u srednjoj školi	Correlation Coefficient	1,000	,340	,355
		Sig. (2-tailed)		,066	,054
	Prosječna ocjena na I godini	N	30	30	30
		Correlation Coefficient	,340	1,000	,736**
	Prosječna ocjena na II godini	Sig. (2-tailed)	,066		,000
		N	30	30	30

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Izvor: Podaci su simulirani.

Da bi se testirala značajnost svakog izračunatog Spearmanovog i Kendallovog koeficijenta korelacije, za svakog od njih se postavljaju hipoteze:

$$H_0 \dots r = 0$$

$$H_1 \dots r \neq 0$$

$$H_0 \dots W = 0$$

$$H_1 \dots W \neq 0$$

Prema podacima u tablici 4.6 može se zaključiti da je samo po jedan izračunati Spearmanov i Kendallov koeficijent korelacije ranga statistički značajan uz signifikantnost testa od 5% (i uz 1%), jer im je empirijska signifikantnost  $\alpha^* \approx 0\%$ . To su korelacije  $r = 0,736$  i  $W = 0,578$  između prosječne ocjene na I i II godini studija. Obje korelacije su pozitivne, što znači da se kod ispitanika s većom prosječnom ocjenom na I godini studija može očekivati veća prosječna ocjena na II godini studija. Ostali izračunati koeficijenti korelacije ranga u tablici 4.6 nisu statistički značajni uz signifikantnost od 5%.

## 4.6 Regresijska analiza

**Zadaća regresijske analize je da pronađe analitičko-matematički oblik veze između jedne ovisne ili regresand varijable i jedne ili više neovisnih ili regresorskih varijabli.**

Osim objašnjavanja prirode ovisnosti promatranih pojava na temelju tog analitičkog oblika može se vršiti predviđanje vrijednosti ovisne varijable pri određenim vrijednostima neovisne-ih varijabli.

### 4.6.1 Jednostruka linearna regresija

U slučaju postojanja samo jedne ovisne ili regresand i samo jedne neovisne ili regresorske varijable kaže se da je to jednostavni, jednostruki ili jednodimenzionalni regresijski model, a regresijska analiza se može postaviti na slijedeći način:

1. **Potpuno, precizno i koncizno definiranje predmeta i ciljeva istraživanja**, te postavljanje osnovnih pretpostavki.
2. **Crtanje dijagrama rasipanja, izbor modela i definiranje varijabli**. Na primjer, **aditivni model**<sup>1</sup>

$$Y = f(X) + e,$$

gdje je: Y- ovisna ili regresand varijabla,

X - neovisna ili regresorska varijabla,

e - slučajna komponenta.

Svaki model ima slučajnu komponentu  $e$ , koja upućuje da veze između pojava u praksi nisu funkcionalne, nego su statističke ili stohastičke, odnosno oko

---

<sup>1</sup> Modeli u praksi ne moraju biti aditivni: na primjer, multiplikativni model je  $Y = f(X) \cdot e$ , gdje je: Y- ovisna ili regresand varijabla, X - neovisna ili regresorska varijabla, e - slučajna komponenta.

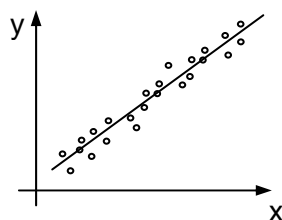
linije konkretnog aditivnog modela postoje pozitivna i/ili negativna odstupanja originalnih vrijednosti.

U ovom koraku bitno je formirati statističko-dokumentacijsku osnovu iz primarnog (direktno) i/ili sekundarnog (literatura) izvora vodeći računa da promatrani podaci budu usporedivi i da njihova usporedba zadovoljava ekonomske kriterije.

**3. Odabir konkretnog regresijskog modela, njegova specifikacija i pretpostavke.** (Na primjer, linearni model:  $Y = \beta_0 + \beta_1 X + e$ .)

**Slika 4.12.**

**Dijagram rasipanja s ucrtanom linijom pravca**



*Izvor: Konstrukcija autora.*

Na slici 4.12 je prikazan dijagram rasipanja koji upućuje na postojanje pozitivne statističke veze između dviju varijabli X i Y. Povlačenjem linije pravca između točaka dijagrama rasipanja pretpostavlja se aditivna linearna veza među varijablama.

**4. Statistička analiza modela: ocjena parametara i pokazatelja reprezentativnosti modela.**

U ovoj fazi regresijske analize ocjenjuju se parametri konkretnog izabranog regresijskog modela, te se računaju odgovarajući pokazatelji reprezentativnosti modela, koji ukazuju na to da li model zadovoljava statističke kriterije.

**5. Testiranje hipoteza o modelu i statističko teorijskih pretpostavki.**

**6a) DA - ako su ispunjene pretpostavke, vrši se sinteza rezultata i donose se sudovi o predmetu istraživanja.**

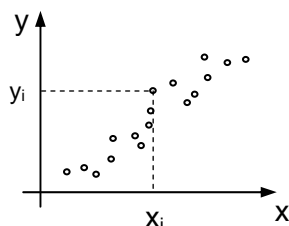
**6b) NE - ako nisu ispunjene pretpostavke: vrši se modifikacija modela i vraća se na korak 2), tj. na izbor novog modela i definiranje varijabli.**

Dakle, regresijskom analizom traže se i ocjenjuju parametri funkcije koja na najbolji mogući način opisuje vezu između varijabli X i Y. Na temelju uzorka parova

vrijednosti varijabli X i Y:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  crta se dijagram rasipanja, koji je prikazan na slici 4.13.

**Slika 4.13.**

#### Dijagram rasipanja između varijabli X i Y

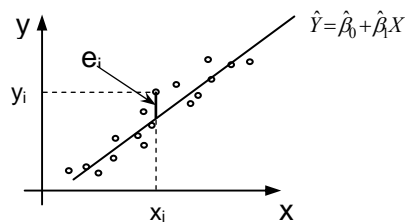


*Izvor: Konstrukcija autora.*

Dijagram rasipanja pokazuje pozitivnu statističku vezu između pojava X i Y.

**Slika 4.14.**

#### Dijagram rasipanja između varijabli X i Y s ucrtanom linijom pravca $\hat{Y}$



*Izvor: Konstrukcija autora.*

Ako se na dijagramu rasipanja povuče **pravac**, on je **općenito oblika**:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X. \quad (4.17)$$

Svaka točka dijagrama rasipanja zadovoljava jednadžbu:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i, \quad (4.18)$$

odnosno **svaka točka  $Y_i$  odstupa od linije pravca za pozitivnu ili negativnu razliku  $e_i$ .**

Regresijska analiza traži parametre  $\hat{\beta}_0$  i  $\hat{\beta}_1$ , tako da pravac  $\hat{Y}$  prolazi između stvarnih točaka promatranih varijabli i da najbolje tumači vezu između njih, odnosno pravac mora biti takav da odstupanja  $e_i$  budu najmanja.

Postoji više različitih metoda za ocjenu ovih parametara, a najčešće korištena metoda je **metoda najmanjih kvadrata**, koja upravo procjenjuje parametre  $\hat{\beta}_0$  i  $\hat{\beta}_1$  tako da odstupanja  $e_i$  budu najmanja.<sup>2</sup> Ona daje najbolje linearne nepristrane ocjene i vrlo je često primjenjivana metoda za ocjenu parametara.

$$Y_i = \hat{Y}_i + e_i \quad (4.18)$$

Odstupanja originalnih vrijednosti od ocijenjenih vrijednosti  $e_i$  mogu biti pozitivna i negativna. Stoga, da se ne bi međusobno poništile te pozitivne i negativne vrijednosti, ova **metoda minimizira sumu kvadrata od  $e_i$** :

$$\min \sum_{i=1}^n e_i^2 = \min \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \min \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2. \quad (4.19)$$

Dakle, traži se minimum kvadrata odstupanja empirijskih (stvarnih) u odnosu na regresijske vrijednosti:

$$f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2. \quad (4.20)$$

Nakon primjene matematičkog postupka traženja minimuma dobiju se dvije jednačbe s dvije nepoznanice, tj. parametri regresijskog modela  $\hat{\beta}_0$  i  $\hat{\beta}_1$ .

$$n \cdot \hat{\beta}_0 + \hat{\beta}_1 \cdot \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \quad (4.21)$$

---

<sup>2</sup> Pri formiranju modela postavljaju se i pretpostavke slučajne greške  $e_i$  (tzv. Gauss-Markovljevi uvjeti):

- I.  $E(e_i) = 0, \forall i$  (očekivanje slučajne greške je nula za svaku opservaciju)
- II.  $E(e_i, e_j) = \sigma_e^2 < +\infty$  za  $i = j$  (homoskedastičnost varijance reziduala, tj. pretpostavlja se da je varijanca reziduala konačna i čvrsta)  
 $= \text{const.}$
- III.  $E(e_i, e_j) = 0, \forall i \neq j$ , tj. (greška je slučajna i nema korelacije između varijabli s  
 $\text{Cov}(e_i, e_j) = 0, \forall i \neq j$  pomakom od  $e_i$ )
- IV.  $E(e_i, X_i) = 0$  (slučajna greška je distribuirana nezavisno od regresorske varijable X)

Vrijedi da je slučajna greška  $e$  distribuirana normalnom distribucijom:  $N(0, \sigma^2 < \infty)$ .



$$\hat{\beta}_0 \cdot \sum_{i=1}^n X_i + \hat{\beta}_1 \cdot \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i \quad (4.22)$$

Sustav uvijek ima rješenja i **vrijedi da je:**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2} \quad i \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad (4.23)$$

gdje su:  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$  i  $\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$ , jednostavne aritmetičke sredine varijabli X i Y.

Konačno je ocijenjeni model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X, \quad (4.24)$$

gdje je  $\hat{\beta}_0$  **konstantni član, tj. očekivana vrijednost zavisne varijable kada je nezavisna varijabla jednaka nuli**: ( $\hat{\beta}_0 = \hat{Y}$  kada je  $X=0$ ). Ovaj parametar interpretira se i kao odsječak na osi ordinata u kojoj regresijski pravac siječe tu os, uz pretpostavku da je apscisa te točke  $X=0$ .

**Regresijski koeficijent  $\hat{\beta}_1$  pokazuje prosječnu promjenu zavisne varijable kada nezavisna varijabla poraste za jedinicu.** Ovaj parametar interpretira se i kao koeficijent smjera, odnosno nagiba regresijskog pravca koji može imati pozitivni i negativni predznak, ovisno o smjeru veze između promatranih varijabli.

Može se postaviti i **suprotna ovisnost u modelu**, na način da je **varijabla X sada ovisna ili regresorska varijabla**:

$$X = \alpha_0 + \alpha_1 Y + e_i. \quad (4.25)$$

Ocjena parametara u ovom slučaju vrši se na jednak način kao kod početnog modela  $\hat{Y}$ , samo što je sada X ovisna varijabla, pa **u izrazima za izračunavanje parametara, X i Y mijenjaju mjesta.**

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n Y_i^2 - n \bar{Y}^2} \quad i \quad \hat{\alpha}_0 = \bar{X} - \hat{\alpha}_1 \bar{Y}. \quad (4.26)$$

Matričnim putem regresijska jednadžba može se napisati:

$$Y = X\hat{\beta} ; \quad (4.27)$$

gdje su matrice:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \quad X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}; \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \quad (4.28)$$

$$\hat{\beta} = (X^T X)^{-1} \cdot (X^T Y), \quad (4.29)$$

gdje su:

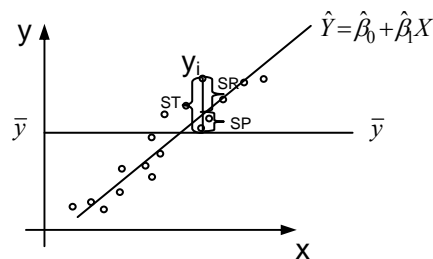
$$(X^T X) = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \quad i \quad (X^T Y) = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}. \quad (4.30)$$

Nakon ocijene parametara regresijskog modela postavlja se pitanje **reprezentativnosti**, odnosno sposobnosti **modela da objasni kretanje ovisne varijable Y uz pomoć odabrane neovisne varijable X**.

U tu svrhu koriste se neki **apsolutni i relativni pokazatelji**. Ovi pokazatelji temelje se na raspodjeli odstupanja vrijednosti ovisne varijable  $Y_i$  u regresijskom modelu od njene aritmetičke sredine  $\bar{Y}$  i njenih očekivanih vrijednosti  $\hat{Y}_i$ .

**Slika 4.15.**

**Dijagram rasipanja između varijabli X i Y s ucrtanom linijom pravca  $\hat{Y}$  i aritmetičkom sredinom  $\bar{Y}$**



*Izvor: Konstrukcija autora.*

Na slici 4.15 je prikazan dijagram rasipanja varijabli X i Y sa ucrtanim ocijenjenim modelom pravca  $\hat{Y}$ . Na slici je označena i aritmetička sredina varijable  $\bar{Y}$ . Pri formiranju suma odgovarajućih odstupanja, zbog već ranije navedenog

razloga, da se ne bi međusobno poništile pozitivne i negativne razlike, računaju se njihovi kvadrati.

$$SP = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_0 \cdot \sum_{i=1}^n Y_i + \hat{\beta}_1 \sum_{i=1}^n X_i Y_i - n \cdot \bar{Y}^2 \quad (4.31)$$

Dakle, **SP je suma kvadrata protumačenog dijela** odstupanja vrijednosti varijable Y od aritmetičke sredine, **odnosno suma kvadrata odstupanja ocijenjenih vrijednosti varijable Y od aritmetičke sredine.**

$$SR = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n Y_i^2 - \hat{\beta}_0 \cdot \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i Y_i \quad (4.32)$$

Dakle, **SR je suma kvadrata neprotumačenog dijela** odstupanja vrijednosti varijable Y od aritmetičke sredine, **odnosno suma kvadrata odstupanja originalnih ili empirijskih vrijednosti varijable Y od ocijenjenih vrijednosti.** Ova odstupanja su u stvari slučajne greške  $e_i$ .

$$ST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n \cdot \bar{Y}^2 \quad (4.33)$$

Dakle, **ST je suma kvadrata ukupnih** odstupanja vrijednosti varijable Y od aritmetičke sredine.

Vrijedi da je:

$$SP + SR = ST, \quad (4.34)$$

što se vidi i na slici 4.15. Ovaj izraz zove se **jednadžba analize varijance i predstavlja temelj analize reprezentativnosti regresijskog modela.**

**Standardna greška regresije je apsolutni pokazatelj reprezentativnosti regresijskog modela, a pokazuje prosječni stupanj varijacije stvarnih vrijednosti ovisne varijable u odnosu na očekivane regresijske vrijednosti:**

$$\hat{\sigma}_{\hat{Y}} = \sqrt{\frac{SR}{n-2}}. \quad (4.35)$$

Navedeni izraz je standardna greška regresije jednostrukog modela. Ovaj pokazatelj izražen je u originalnim jedinicama mjere ovisne varijable Y. Stoga je na temelju standardne greške regresije teško uspoređivati reprezentativnost modela s različitim mjernim jedinicama.

Taj problem eliminira **relativni pokazatelj koeficijent varijacije regresije, koji predstavlja postotak standardne greške regresije od aritmetičke sredine varijable Y:**

$$\hat{V}_{\hat{Y}} = \frac{\hat{\sigma}_{\hat{Y}}}{\bar{Y}} \cdot 100. \quad (4.36)$$

Najmanja vrijednost koeficijenta varijacije je 0%, a najveća nije definirana. **Što je koeficijent varijacije regresijskog modela bliži nuli, to je model reprezentativniji.** Često se uzima dogovorena granica reprezentativnosti od 10%. Dakle ako je koeficijent varijacije manji od 10% kaže se da je model dobar.

**Koeficijent determinacije** je pokazatelj reprezentativnosti regresijskog modela, koji se temelji na analizi varijance. On se definira kao **omjer sume kvadrata odstupanja protumačenih regresijom i sume kvadrata ukupnih odstupanja:**

$$r^2 = \frac{SP}{ST}. \quad (4.37)$$

**Koeficijent determinacije kaže koliko % je sume kvadrata odstupanja vrijednosti varijable Y od aritmetičke sredine protumačeno regresijskim modelom.**

Prema jednadžbi analize varijance, vrijedi da je:

$$r^2 = 1 - \frac{SR}{ST}. \quad (4.38)$$

**Vrijednost koeficijenta determinacije kreće se u intervalu  $0 \leq r^2 \leq 1$ .** Regresijski model je reprezentativniji ako je ovaj pokazatelj bliži 1. Teorijska granica reprezentativnosti modela je 0,9. U praksi je ponekad vrlo teško pronaći varijablu koja dobro objašnjava ovisnu pojavu, pa se ta granica reprezentativnosti spušta i do 0,6.

**Korigirani koeficijent determinacije:**

$$\bar{r}^2 = 1 - \frac{n-1}{n-(k+1)} \cdot (1-r^2), \quad (4.39)$$

je asimptotski nepristrana ocjena koeficijenta determinacije.

### 4.6.2 Regresijski polinom

Ako veza između varijabli  $X$  i  $Y$  nije linearna, za prikazivanje njihove ovisnosti može se upotrijebiti polinom  $k$ -tog stupnja čiji je općeniti oblik:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X + \hat{\beta}_2 \cdot X^2 + \dots + \hat{\beta}_k \cdot X^k \quad (4.40)$$

**Ocjenom parametara metodom najmanjih kvadrata** traži se minimum zbroja kvadrata empirijskih odstupanja u odnosu na regresijske vrijednosti:

$$\begin{aligned} f(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \\ &= \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot X + \dots + \hat{\beta}_k \cdot X^k)]^2 = SR \end{aligned} \quad (4.41)$$

Gdje je za minimum funkcije  $f$  vrijedi da je:

$$\begin{aligned} \frac{\partial f}{\partial \hat{\beta}_0} &= 0; \\ \frac{\partial f}{\partial \hat{\beta}_1} &= 0; \\ &\dots\dots\dots \\ \frac{\partial f}{\partial \hat{\beta}_k} &= 0. \end{aligned} \quad (4.42)$$

**U matričnom obliku** ocjena parametara je:

$$\hat{\beta} = (X^T X)^{-1} \cdot (X^T Y), \quad (4.43)$$

gdje je:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}; \quad X = \begin{bmatrix} 1 & X_1 & X_1^2 & \dots & X_1^k \\ 1 & X_2 & X_2^2 & \dots & X_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & X_n^2 & \dots & X_n^k \end{bmatrix}; \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}. \quad (4.44)$$

**Jednadžbe analize varijance regresijskog polinoma su:**

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \Rightarrow SP = \hat{\beta}^T \cdot (X^T Y) - n \cdot \bar{Y}^2 \quad (4.45)$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \Rightarrow SR = Y^T Y - \hat{\beta}^T \cdot (X^T Y) \quad (4.46)$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 \Rightarrow ST = Y^T Y - n \cdot \bar{Y}^2. \quad (4.47)$$

Interpretacija parametara ovog modela svodi se na objašnjenje slobodnog člana, tj. **parametar  $\hat{\beta}_0$  je konstantni član, tj. očekivana vrijednost zavisne varijable kada je nezavisna varijabla X jednaka nuli.** Ostali parametri u ovom modelu regresijskog polinoma nemaju svoju logičnu interpretaciju.

Vrijedi da je:

$$SP + SR = ST, \quad (4.48)$$

**što kao jednadžba analize varijance predstavlja temelj analize uobičajene reprezentativnosti regresijskog modela.**

### 4.6.3 Eksponencijalna regresija

Ako regresijski model nije aditivan, u regresijskoj analizi se može upotrijebiti multiplikativni model. Jedan od najčešće upotrebljivanih nelinearnih modela je **model jednostavne eksponencijalne regresije, koji je općenito oblika:**

$$\hat{Y} = \hat{\beta}_0 \cdot \hat{\beta}_1^X \quad (4.49)$$

Da bi se za ocjenu parametara upotrijebila **metoda najmanjih kvadrata**, potrebno je početni model logaritamskom transformacijom prevesti u logaritamsko-linearni oblik:

$$\log \hat{Y} = \log \hat{\beta}_0 + \log \hat{\beta}_1 \cdot X \quad (4.50)$$

Traži se minimum sume kvadrata neprotumačenih ili rezidualnih odstupanja:

$$\min SR = \min \sum_{i=1}^n (\log Y_i - \log \hat{Y}_i)^2 = \min \sum_{i=1}^n [\log Y_i - (\log \hat{\beta}_0 + \log \hat{\beta}_1 X_i)]^2 \quad (4.51)$$

Nakon primjene matematičkog postupka traženja minimuma dobije se da je:

$$\log \hat{\beta}_1 = \frac{\sum_{i=1}^n X_i \log Y_i - n \bar{X} \overline{\log Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2} \quad i \quad \log \hat{\beta}_0 = \overline{\log Y} - \log \hat{\beta}_1 \bar{X}, \quad (4.52)$$

Transformacijom:

$$\hat{\beta}_0 = 10^{\log \hat{\beta}_0} \quad i \quad \hat{\beta}_1 = 10^{\log \hat{\beta}_1}, \quad (4.53)$$

se dobiju originalne vrijednosti parametara modela.

**Parametar  $\hat{\beta}_0$  je konstantni član, tj. očekivana vrijednost zavisne varijable kada je nezavisna varijabla jednaka nuli:** ( $\hat{\beta}_0 = \hat{Y}$  kada je  $X=0$ ).

**Parametar  $\hat{\beta}_1$  pokazuje ocjenu tempa ili dinamike kretanja zavisne varijable Y (u %) kada nezavisna varijabla poraste za jedinicu.** Ovaj parametar interpretira se kao ocjena prosječne stope promjene varijable Y, kada varijabla X poraste za jednu jedinicu:

$$\bar{S} = (\hat{\beta}_1 - 1) \cdot 100 \Rightarrow (u \%) \quad (4.54)$$

**Jednadžbe analize varijance** kod modela jednostavne eksponencijalne regresije se transformiraju na sličan način logaritmiranjem.

Vrijedi da je:

$$SP + SR = ST. \quad (4.55)$$

#### 4.6.4 Testiranje hipoteze o značajnosti regresijskog modela kao cjeline

Nulta hipoteza pretpostavlja da su svi parametri regresijskog modela jednaki 0, dok alternativna hipoteza pretpostavlja da postoji barem jedan regresijski parametar različit od nule:

$$\begin{aligned} H_0: \dots \beta_1 = \beta_2 = \dots = \beta_k = 0 \\ H_1: \dots \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k \end{aligned} \quad (4.56)$$

Testiranje se vrši usporedbom **empirijske i tablične vrijednosti F - testa**.

Empirijska vrijednost F-testa računa se pomoću podataka iz tablice analize varijance, tzv. **tablice ANOVA** koja je u svom općem obliku prikazana tablicom 4.7:

**Tablica 4.7.**

**Tablica analize varijance (ANOVA) pri regresijskoj analizi**

Izvor varijacije	Zbroj kvadrata odstupanja	Stupnjevi slobode	Ocjena varijance
protumačeno (SP)	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	k	$SP / k$
neprotumačeno (SR)	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	n-k-1	$SR / (n - k - 1)$
ukupno (ST)	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	n-1	

*Izvor: Konstrukcija autora prema teorijskim postavkama.*

gdje je:

$k$  - broj regresorskih varijabli u modelu,

$n$  - broj opservacija.

Da bi se izvršilo navedeno testiranje pomoću veličina iz tablice 4.7 računa se empirijska i tablična vrijednost F-testa:

$$F^* = \frac{SP / k}{SR / (n - k - 1)}, \quad (4.57)$$

koju je potrebno usporediti s tabličnom vrijednosti uz odgovarajuću razinu signifikantnosti  $\alpha$  i stupnjevima slobode  $df_1$  i  $df_2$  (Tablice D1 i D2).

$$F_{tab}^{[\alpha]}[df_1=k; df_2=(n-k-1)]. \quad (4.58)$$

Ako je:

$$F^* = \frac{SP / k}{SR / (n - k - 1)} < F_{tab} \Rightarrow H_0, \quad (4.59)$$

**prihvaća se početna hipoteza**  $H_0$  tj. zaključuje se da ocijenjeni regresijski model nije statistički značajan, dok se u suprotnom početna hipoteza odbacuje tj. zaključuje se da je ocijenjeni regresijski model statistički značajan.



Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha$  pomoću  $F^*$  (Tablice): ako je  $\alpha^* < 5\% \Rightarrow H_1$ , dok u suprotnom vrijedi hipoteza  $H_0$ .

#### 4.6.5 Testiranje hipoteze o značajnosti regresijskog parametra

Nulta hipoteza pretpostavlja da je parametar kojeg testiramo iz promatranog regresijskog modela jednak 0, tj. da nije statistički značajan, dok alternativna hipoteza pretpostavlja da je testirani regresijski parametar različit od nule, tj. da je statistički značajan:

$$\begin{aligned} H_0: \dots \beta_j &= 0 \\ H_1: \dots \beta_j &\neq 0 \end{aligned} \quad (4.60)$$

Testiranje se vrši usporedbom **empirijske i tablične vrijednosti  $t$  - testa**.

Empirijska vrijednost  $t$ -testa računa se:

$$t^* = \frac{\hat{\beta}_j}{Se(\hat{\beta}_j)}, \quad (4.61)$$

gdje je:

$\hat{\beta}_j$  - vrijednost ocijenjenog parametra,

$Se(\hat{\beta}_j)$  - standardna greška ocijenjenog parametra, koja se izračunava:

$$Se(\hat{\beta}_j) = \hat{\sigma}_{\hat{Y}} \cdot \sqrt{s_{jj}}, \quad (4.62)$$

gdje  $s_{jj}$  predstavlja vrijednost odgovarajućeg dijagonalnog elementa<sup>3</sup> u matrici  $(X^T X)^{-1}$ .

<sup>3</sup> Za jednostruku linearnu regresiju vrijedi da je:

$$S_{11} = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{1}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}.$$

Tablična vrijednost  $t$  - testa se uz određenu signifikantnost  $\alpha$  i stupnjeve slobode  $df$  određuje pomoću tablica Studentove distribucije:

$$t_{tab(df=n-k-1)}^{\alpha/2} \quad (4.63)$$

Ako je:  $t^* < t_{tab} \Rightarrow H_0$  **prihvaća se početna hipoteza**  $H_0$  tj. zaključuje se da ocijenjeni regresijski parametar nije statistički značajan, dok se u suprotnom početna hipoteza odbacuje tj. zaključuje se da je ocijenjeni regresijski parametar statistički značajan.

Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha$  pomoću  $t^*$  (Tablica B): ako je  $\alpha^* < 5\% \Rightarrow H_1$ , dok u suprotnom vrijedi hipoteza  $H_0$ .

#### 4.6.6 Granice intervala pouzdanosti za regresijski parametar $\beta_j$

Intervalna procjena regresijskog parametra  $\beta_j$  uz odgovarajući nivo pouzdanosti procjene  $(1 - \alpha)$  je:

$$\Pr\{\hat{\beta}_j - t \cdot Se(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t \cdot Se(\hat{\beta}_j)\} = 1 - \alpha \quad (4.64)$$

gdje je:

$t$  - odgovarajuća vrijednost iz tablica Studentove distribucije,

$1 - \alpha$  - odgovarajući nivo pouzdanosti procjene,

$Se(\hat{\beta}_j)$  - standardna greška ocijenjenog parametra, koja se izračunava prema 4.62.

#### 4.6.7 Točkasta ocjena prognostičke vrijednosti $\hat{Y}_0$

**Točkasta ocjena** prognostičke vrijednosti  $\hat{Y}_0$  računa se tako da se u ocijenjeni model uvrste konkretne vrijednosti regresorskih tj. neovisnih varijabli:

$$\hat{Y}_0 = X_0 \cdot \hat{\beta}, \quad (4.66)$$

gdje je:

$$X_0 = [1 \quad X_i]. \quad (4.67)$$

Odnosno računa se:

$$\hat{Y}_0 = \hat{Y}(X_1, X_2, X_3, \dots). \quad (4.68)$$

#### 4.6.8 Intervalna procjena prognostičke vrijednosti $\hat{Y}_0$

Intervalna procjena prognostičke vrijednosti  $\hat{Y}_0$  je točkasta ocjena prognostičke vrijednosti korigirana za veličinu greške:

$$\Pr\{\hat{Y}_0 - t \cdot Se(\hat{Y}_0) < Y_0 < \hat{Y}_0 + t \cdot Se(\hat{Y}_0)\} = 1 - \alpha, \quad (4.69)$$

gdje je:

$\hat{Y}_0$  - točkasta ocjena prognostičke vrijednosti,

$t$  - odgovarajuća vrijednost iz tablica Studentove distribucije i,

$$Se(\hat{Y}_0) = \hat{\sigma}_{\hat{Y}} \cdot \sqrt{1 + X_0(X^T X)^{-1} X_0^T}, \quad (4.70)$$

je standardna greška prognostičke vrijednosti.



##### Primjer 4.8.

Potrebno je za varijable BDP (u mil. EUR) i indeks industrijske proizvodnje (2000=100) od 1997. do 2007. u Republici Hrvatskoj (izvor: **Eurostat**; <http://epp.eurostat.ec.europa.eu>) izračunati parametre jednostruke linearne regresije s BDP-om kao regresorskom varijablom.



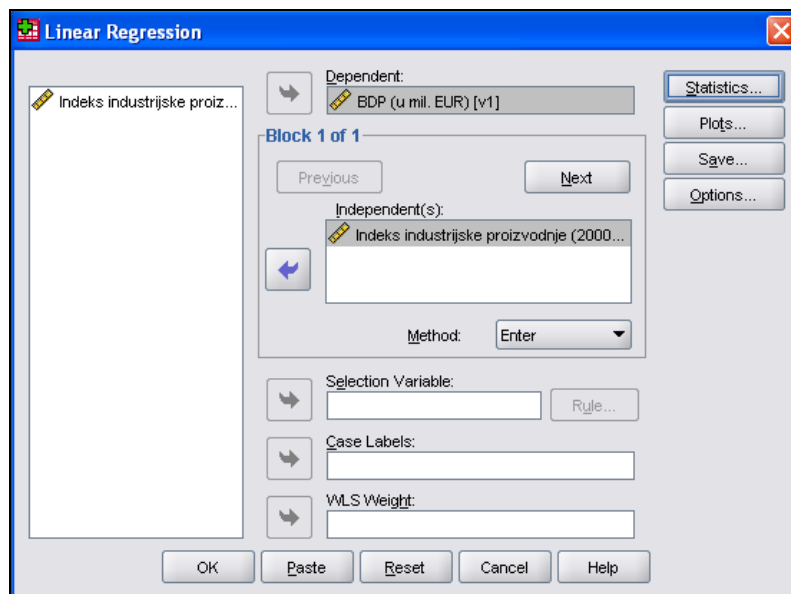
##### Rješenje 4.8.

Da bi se u programskom paketu **SPSS** dobila ocjena parametara linearnog regresijskog modela u kojem BDP (u mil. EUR) ovisi o indeksu industrijske

proizvodnje potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Regression** gdje se bira **Linear**. U otvorenom prozoru bira se varijabla BDP (u mil. EUR) (v1), u: **Dependent** i Indeks industrijske proizvodnje (v2) u: **Independent(s)**. Odabrane varijable u prozoru **Linear Regression** prikazane su na slici 4.16.

Slika 4.16.

**Prozor Linear Regression s odabranim varijablama za ocjenu parametara odgovarajućeg linearnog regresijskog modela**



Izvor: Simulirani podaci.

U **Statistics** treba aktivirati: **Estimates**, **Confidence Intervals**, **Model fit**, **Continue**. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi ocijenjeni regresijski model s potrebnom regresijskom dijagnostikom.

Tablica 4.8.

**Osnovni podaci o ocijenjenom modelu s BDP (u mil. EUR) kao zavisnom varijablom**

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999 <sup>a</sup>	.998	.998	291,27902

a. Predictors: (Constant), Indeks industrijske proizvodnje (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici 4.8 izračunate su vrijednosti koeficijenta korelacije  $R = 0,999$ , koji pokazuje jaku i pozitivnu linearnu vezu između varijabli BDP (u mil. EUR) i indeksa industrijske proizvodnje (2000=100) u R Hrvatskoj. Koeficijent determinacije (R Square) je  $r^2 = 0,998$ , što znači da je ocijenjenim regresijskim modelom protumačeno 99,8% sume kvadrata ukupnih odstupanja zavisne varijable od njene aritmetičke sredine. Protumačenost ocijenjenog modela je jako visoka. Korigirani koeficijent determinacije (Adjusted R Square) je  $\bar{r}^2 = 0,998$  i standardna greška ocijenjene regresije je  $\hat{\sigma}_{\hat{y}} = 291,279$ .

**Tablica 4.9.**

**Tablica ANOVA ocijenjenog regresijskog modela**

ANOVA <sup>b</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4,510E8	1	4,510E8	5315,466	,000 <sup>a</sup>
	Residual	763591,221	9	84843,469		
	Total	4,517E8	10			

a. Predictors: (Constant), Indeks industrijske proizvodnje (2000=100)

b. Dependent Variable: BDP (u mil. EUR)

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenog modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema tablici 4.9 empirijska vrijednost F-testa je:  $F^* = \frac{SP/k}{SR/(n-k-1)} = 5315,466$ , koju

je potrebno usporediti s tabličnom vrijednosti uz odgovarajuću razinu signifikantnosti  $\alpha = 5\%$  (Tablice D1 i D2):  $F_{tab}^{[\alpha=5\%]}[df_1=1; df_2=9] = 5,117$ .

Vrijedi da je:  $F^* > F_{tab} \Rightarrow H_1$ , pa se odbacuje početna hipoteza  $H_0$  tj. zaključuje se da je ocijenjeni regresijski model statistički značajan.

Prema tablici ANOVA vrijedi da je  $\alpha^* \approx 0\%$ , što je manje od 5%, pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

**Tablica 4.10.**

**Ocijenjeni linearni regresijski model gdje s BDP (u mil. EUR) ovisi o indeksu industrijske proizvodnje (2000=100)**

Coefficients <sup>a</sup>							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-25408,182	703,331		-36,125	,000	-26999,228	-23817,136
Indeks industrijske proizvodnje (2000=100)	448,482	6,151	,999	72,907	,000	434,566	462,397

a. Dependent Variable: BDP (u mil. EUR)

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici 4.10 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

Analitički izraz konačnog ocijenjenog modela je:  $\hat{y}_i = -25408,2 + 448,482 \cdot x_i$ .

Parametar  $\hat{\beta}_0 = -25408,182$  predstavlja očekivani BDP ( $\hat{Y}$ ) u slučaju da indeks industrijske proizvodnje ( $X$ ) iznosi nula. Ovaj parametar nema uvijek ekonomski logično značenje.

Parametar uz nezavisnu varijablu  $x$ , tj.  $\hat{\beta}_1 = 448,482$  pokazuje da se može očekivati porast BDP-a za 448,48 mil. EUR-a ako indeks industrijske proizvodnje poraste za 1.

Da bi se izvršilo testiranje značajnosti pojedinačnih parametara modela potrebno je postaviti hipoteze:

$$H_0 \dots \beta_j = 0$$

$$H_1 \dots \beta_j \neq 0$$

Prema tablici 4.10 empirijska vrijednost t-testa za parametar  $\hat{\beta}_0$  je  $t^* = -36,125$ , a empirijska signifikantnost je  $\alpha^* \approx 0\%$ , što je manje od 5%, pa se može donijeti zaključak o odbacivanju početne hipoteze  $H_0$ , tj. zaključuje se da je parametar  $\hat{\beta}_0$  statistički značajan.

Empirijska vrijednost t-testa za parametar  $\hat{\beta}_1$  je  $t^* = 72,907$ , a empirijska signifikantnost je  $\alpha^* \approx 0\%$ , što je manje od 5%, pa se može donijeti zaključak o odbacivanju početne hipoteze  $H_0$ , tj. zaključuje se da je ocijenjeni regresijski parametar statistički značajan.

Standardizirani regresijski koeficijenti pokazuju relativni utjecaj pojedinih nezavisnih varijabli na promatranu zavisnu varijablu:  $\hat{b}_k = \hat{\beta}_k \frac{\sigma_{xk}}{\sigma_y}$ .

Ocijenjeni regresijski model u standardiziranom obliku je:  $\hat{y}_i = 0,999 \cdot x_i$ .

Standardizirana vrijednost parametra pokazuje za koliko će se standardnih devijacija promijeniti zavisna varijabla, ako se nezavisna varijabla promijeni za 1 standardnu devijaciju.

Intervalna procjena ocijenjenih parametara uz 95% pouzdanosti za parametar  $\hat{\beta}_0$  je:

$$\Pr\{-26999,228 < \hat{\beta}_0 < -23817,136\} = 95\%$$

i za parametar  $\hat{\beta}_1$  je:

$$\Pr\{434,566 < \hat{\beta}_1 < 462,397\} = 95\%$$

što znači da ako indeks industrijske proizvodnje poraste za 1, da se može uz 95% pouzdanosti očekivati porast BDP-a u intervalu između 434,57 mil. EUR i 462,40 mil. EUR.



#### Primjer 4.9.

Na osnovu zadanih podataka o BDP-u, zaposlenima i izvozu od 1983. do 2004. u Republici Hrvatskoj potrebno je izračunati BDP po zaposlenom.

Potrebno je za varijable BDP po zaposlenom i izvoz nacrtati dijagram rasipanja, izračunati linearnu korelaciju, te izračunati parametre jednostruke linearne regresije s varijablom BDP po zaposlenom kao regresand varijablom.



#### Rješenje 4.9.

Najprije je potrebno formirati novu varijablu BDP po zaposlenom.

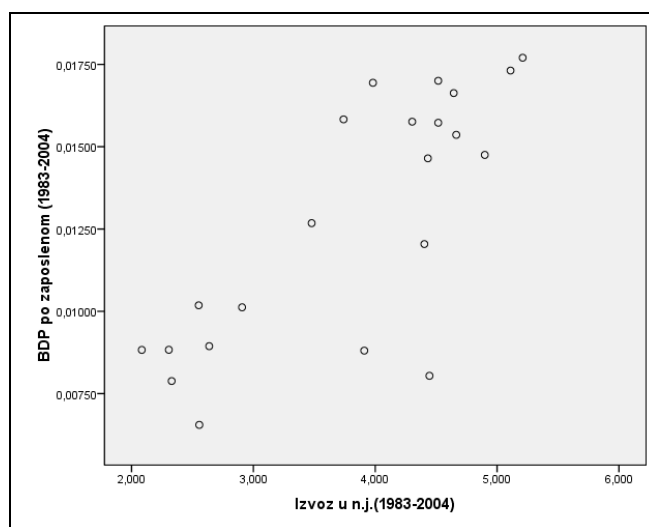
U programskom paketu **SPSS** potrebno je odabrati: **Transform; Compute; u Target variable:** upisati v4; u **Numeric Expression:** v1/v2; **OK.** (U prostor **Label** u **Variable view** može se upisati da je nova varijabla v4: BDP po zaposlenom.)

U programskom paketu **SPSS** potrebno je na glavnom izborniku izabrati ikonu **Graphs**, a na njezinom padajućem izborniku **Legacy Dialogs** i na pomoćnom izborniku **Scatter/Dot** gdje se bira grafikon **Simple Scatter**. U otvorenom prozoru definira se varijabla BDP po zaposlenom (v4) u: **Y Axis**, a Izvoz (v3) u: **X Axis**.

Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi dijagram rasipanja, koji je prikazan na slici 4.17.

**Slika 4.17.**

**Dijagram rasipanja BDP-a po zaposlenom i izvoza u RH**



Izvor: SLJH, RH.

Može se zaključiti da postoji pozitivna veza između visine BDP-a po zaposlenom i izvoza u R Hrvatskoj u promatranom razdoblju.

Da bi se u programskom paketu **SPSS** dobila ocjena parametara linearnog regresijskog modela u kojem BDP po zaposlenom ovisi o izvozu potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Regression** gdje se bira **Linear**. U otvorenom prozoru bira se varijabla BDP po zaposlenom (v4), u: **Dependent** i Izvoz (v3) u: **Independent(s)**. U **Statistics** treba aktivirati: **Estimates, Confidence Intervals, Model fit; Continue**. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi ocijenjeni regresijski model s potrebnom regresijskom dijagnostikom.

U tablici 4.11 izračunate su vrijednosti koeficijenta korelacije  $R = 0,784$ , koji pokazuje jaku i pozitivnu linearnu vezu između varijabli BDP-a po zaposlenom i izvoza u R Hrvatskoj. Koeficijent determinacije (R Square) je  $r^2 = 0,614$ , što znači da je ocijenjenim regresijskim modelom protumačeno 61,4% sume kvadrata ukupnih odstupanja zavisne varijable od njene aritmetičke sredine. Protumačenost ocijenjenog modela je nije jako visoka, ali se u praksi mogu koristiti jer je teško pronaći model koji dobro tumače vezu između odabranim varijablama u praksi. Korigirani



koeficijent determinacije (Adjusted R Square) je  $\bar{r}^2 = 0,595$  i standardna greška ocijenjene regresije je  $\hat{\sigma}^2 = 0,0239$ .

**Tablica 4.11.**

**Osnovni podaci o ocijenjenom modelu s BDP po zaposlenom kao zavisnom varijablom**

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,784 <sup>a</sup>	,614	,595	,00239447
a. Predictors: (Constant), Izvoz u n.j.(1983-2004)				

Izvor: SLJH, RH.

**Tablica 4.12.**

**Tablica ANOVA ocijenjenog regresijskog modela**

ANOVA <sup>b</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,000	1	,000	31,799	,000 <sup>a</sup>
	Residual	,000	20	,000		
	Total	,000	21			
a. Predictors: (Constant), Izvoz u n.j.(1983-2004)						
b. Dependent Variable: BDP po zaposlenom (1983-2004)						

Izvor: SLJH, RH.

U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenog modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema tablici 4.12 empirijska vrijednost F-testa je:  $F^* = \frac{SP/k}{SR/(n-k-1)} = 31,799$ , koju

je potrebno usporediti s tabličnom vrijednosti uz odgovarajuću razinu signifikantnosti  $\alpha = 5\%$  (Tablice D1 i D2):  $F_{tab}^{[\alpha=5\%]}[df_1=1; df_2=20] = 4,351$ .

Vrijedi da je:  $F^* > F_{tab} \Rightarrow H_1$ , pa se odbacuje početna hipoteza  $H_0$  tj. zaključuje se da je ocijenjeni regresijski model statistički značajan.

Prema tablici ANOVA vrijedi da je  $\alpha^* \approx 0\%$ , što je manje od 5%, pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

**Tablica 4.13.**

**Ocijenjeni linearni regresijski model gdje s BDP po zaposlenom ovisi o izvozu**

Coefficients <sup>a</sup>							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	,002	,002		,843	,409	-,003	,006
Izvoz u n.j.(1983-2004)	,003	,001	,784	5,639	,000	,002	,004

a. Dependent Variable: BDP po zaposlenom (1983-2004)

Izvor: SLJH, RH.

U tablici 4.13 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

Analitički izraz konačnog ocijenjenog modela je:  $\hat{y}_i = 0,002 + 0,003 \cdot x_i$ .

Parametar  $\hat{\beta}_0 = 0,002$  predstavlja očekivani BDP po zaposlenom ( $\hat{Y}$ ) u slučaju da izvoz ( $X$ ) iznosi nula. Ovaj parametar nema uvijek ekonomski logično značenje.

Parametar uz nezavisnu varijablu  $x$ , tj.  $\hat{\beta}_1 = 0,003$  pokazuje da se može očekivati porast BDP-a po zaposlenom 0,003 n.j. ako izvoz poraste za 1 n.j.

Da bi se izvršilo testiranje značajnosti pojedinačnih parametara modela potrebno je postaviti hipoteze:

$$H_0 \dots \beta_j = 0$$

$$H_1 \dots \beta_j \neq 0$$

Prema tablici 4.13 empirijska vrijednost t-testa za parametar  $\hat{\beta}_0$  je  $t^* = 0,843$ , a empirijska signifikantnost je  $\alpha^* = 40,9\%$ , što je veće od 5%, pa se može donijeti zaključak o prihvatanju početne hipoteze  $H_0$ , tj. zaključuje se da parametar  $\hat{\beta}_0$  nije statistički značajan.

Empirijska vrijednost t-testa za parametar  $\hat{\beta}_1$  je  $t^* = 5,639$ , a empirijska signifikantnost je  $\alpha^* \approx 0\%$ , što je manje od 5%, pa se može donijeti zaključak o odbacivanju početne hipoteze  $H_0$ , tj. zaključuje se da je ocijenjeni regresijski parametar statistički značajan.

Ocijenjeni regresijski model u standardiziranom obliku je:  $\hat{y}_i = 0,784 \cdot x_i$ .

Standardizirana vrijednost parametra pokazuje za koliko će se standardnih devijacija promijeniti zavisna varijabla, ako se nezavisna varijabla promijeni za 1 standardnu devijaciju.

Intervalna procjena ocijenjenih parametara uz 95% pouzdanosti za statistički značajan parametar  $\hat{\beta}_1$  je:

$$\Pr\{0,002 < \hat{\beta}_1 < 0,004\} = 95\%,$$

što znači da ako izvoz poraste za 1 n.j., da se može uz 95% pouzdanosti očekivati porast BDP-a po zaposlenom u intervalu između 0,002 n.j. i 0,004 n.j.

## 4.7 Višestruka (multipla) regresija

Kod modela višestruke ili multiple regresije jedna zavisna (ili regresand) varijabla ovisi o  $k \geq 2$  nezavisnih (ili regresorskih) varijabli.

Grafičko prikazivanje ovdje nije jednostavno jer se radi o tzv. "hiperravninama".

Općeniti oblik modela je:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot X_2 + \dots + \hat{\beta}_k \cdot X_k \quad (4.70)$$

Ocjena parametara je slična kao kod regresijskog polinoma k-tog stupnja (na desnoj strani jednadžbe nalazi se više različitih regresorskih varijabli, a ne polinomi jedne regresorske varijable).

**Ocjenom parametara metodom najmanjih kvadrata** traži se minimum zbroja kvadrata empirijskih odstupanja u odnosu na regresijske vrijednosti:

$$\begin{aligned} f(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \\ &= \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \dots + \hat{\beta}_k \cdot X_k)]^2 = SR \end{aligned} \quad (4.71)$$

Gdje je za minimum funkcije  $f$  vrijedi da je:

$$\frac{\partial f}{\partial \hat{\beta}_0} = 0;$$

$$\frac{\partial f}{\partial \hat{\beta}_1} = 0; \quad (4.72)$$

.....

$$\frac{\partial f}{\partial \hat{\beta}_k} = 0.$$

**U matričnom obliku** ocjena parametara je:

$$\hat{\beta} = (X^T X)^{-1} \cdot (X^T Y), \quad (4.73)$$

gdje je:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}; \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}; \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}. \quad (4.74)$$

**Jednadžbe analize varijance višestruke (multiple) regresije:**

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \Rightarrow SP = \hat{\beta}^T \cdot (X^T Y) - n \cdot \bar{Y}^2 \quad (4.75)$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \Rightarrow SR = Y^T Y - \hat{\beta}^T \cdot (X^T Y) \quad (4.76)$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 \Rightarrow ST = Y^T Y - n \cdot \bar{Y}^2 \quad (4.77)$$

Parametar  $\hat{\beta}_0$  je konstantni član, tj. očekivana vrijednost zavisne varijable u originalnim jedinicama mjere kada su sve nezavisne varijable jednake nuli: ( $\hat{\beta}_0 = \hat{Y}$  kada je  $X_j=0$ ).

Regresijski koeficijenti  $\hat{\beta}_j$  pokazuju prosječnu promjenu zavisne varijable u originalnim jedinicama mjere kada odgovarajuća nezavisna varijabla poraste za 1 jedinicu, uz uvjet da su sve ostale neovisne varijable neizmijenjene.

Vrijedi da je:

$$SP + SR = ST, \quad (4.78)$$

što predstavlja jednadžbu analize varijance i predstavlja temelj analize reprezentativnosti regresijskog modela.

Standardna greška regresije je apsolutni pokazatelj reprezentativnosti regresijskog modela, a pokazuje prosječni stupanj varijacije stvarnih vrijednosti ovisne varijable u odnosu na očekivane regresijske vrijednosti.

$$\hat{\sigma}_{\hat{Y}} = \sqrt{\frac{SR}{n-k-1}} \quad (4.79)$$

Ovaj izraz je standardna greška regresije modela. Izražen je u originalnim jedinicama mjere ovisne varijable Y. Stoga je na temelju standardne greške regresije teško uspoređivati reprezentativnost modela s različitim mjernim jedinicama.

Taj problem eliminira **relativni pokazatelj koeficijent varijacije regresije, koji predstavlja postotak standardne greške regresije od aritmetičke sredine varijable Y.**

$$\hat{V}_{\hat{Y}} = \frac{\hat{\sigma}_{\hat{Y}}}{\bar{Y}} \cdot 100 \quad (4.80)$$

Najmanja vrijednost koeficijenta varijacije je 0%, a najveća nije definirana. **Što je koeficijent varijacije regresijskog modela bliži nuli, to je model reprezentativniji.** Često se uzima dogovorena granica reprezentativnosti od 10%. Dakle ako je koeficijent varijacije manji od 10% kaže se da je model dobar.

**Koeficijent multiple determinacije za model višestruke regresije:**

$$r^2 = \frac{SP}{ST} \quad (4.81)$$

**Korigirani koeficijent determinacije:**

$$\bar{r}^2 = 1 - \frac{n-1}{n-(k+1)} \cdot (1-r^2), \quad (4.82)$$

je asimptotski nepristrana ocjena koeficijenta determinacije.

#### 4.7.1 Metode odabira varijabli u modelu višestruke regresije

Može se primijeniti više metoda:

**a) FORWARD metoda:**

- na početku nema ni jedne varijable u modelu,
- prva varijabla koja ulazi u model je ona s najvećim koeficijentom korelacije s regresand varijablom,
- za  $F$ -test  $\Rightarrow$  najčešće se nivo signifikantnosti  $\alpha$  fiksira na 5%  $\Rightarrow$  što ne smije biti premašeno,
- ako je " $X_1$ " varijabla ušla u model, između ostalih se bira ona koja ima najveći koeficijent parcijalne korelacije s regresand varijablom,
- testira se  $F$ -testom.....
- postupak se ponavlja dok sve varijable, koje zadovoljavaju, ne budu uvrštene u regresijski model,
- ako je previše varijabli zadovoljilo, može se zahtijevati da samo prvih " $m$ " varijabli bude uključeno u model.

**b) BACKWARD metoda:**

- na početku se u model uključuju sve varijable,
- postavljaju se uvjeti za izlazak:
  1. minimalna vrijednost  $F$ -testa,
  2. maksimalna signifikantnost  $\alpha$  za isključenje je 10%.
- izlazak počinje od onih varijabli koje imaju najmanji koeficijent parcijalne korelacije s regresand varijablom i od njih se testiraju kriteriji za izlazak iz modela,
- postupak se ponavlja sve dok god ima varijabli koje udovoljavaju kriterijima za izlazak.

**c) STEPWISE metoda:**

- je kombinacija FORWARD i BACKWARD metode i u praksi se najčešće koristi,
- prva varijabla se bira kao kod FORWARD metode,

- druga se bira ona s najvećim koeficijentom parcijalne korelacije s regresand varijablom,
- STEPWISE metoda se nastavlja kao kod BACKWARD metode ispitujući da li uključena varijabla udovoljava kriterijima za izlazak,
- da se ne bi dogodilo stalno "uključivanje-isključivanje" nivo signifikantnosti  $\alpha$  za ulaz se može fiksirati na nižoj razini od  $\alpha$  za izlazak iz modela,
- NA KRAJU  $\Rightarrow$  potrebno je ispitati karakteristike slučajne greške "e" kod izabranog modela.

**d) ENTER metoda:**

- sve varijable ulaze odjednom u model (forsirani ulaz).

**e) TEST metoda:**

- zadan je podskup regresorskih varijabli,
- testira se niz mogućnosti pomoću kriterija koeficijenta determinacije ( $r^2$ ).

#### 4.7.2 Problem međuovisnosti regresorskih varijabli

Kod višestrukih regresijskih modela **treba biti ispunjena pretpostavka da su regresorske varijable međusobno nezavisne**. Ako to **nije ispunjeno** kaže se da postoji problem **kolinearnosti dviju**, odnosno **multikolinearnosti više regresorskih varijabli**.

U praksi se uvijek javlja manja ili veća ovisnost regresorskih varijabli, te je potrebno statistički utvrditi jačinu multikolinearnosti i ocijeniti ozbiljnost problema.

U slučaju potpune ovisnosti problem ocjene parametara postaje nerješiv.

Ovaj problem se često javlja kod analize vremenskih nizova.

U regresijskom modeliranju vrijedi pretpostavka o međusobnoj nezavisnosti regresorskih varijabli u matrici X, reda  $[n \times k]$ , tj. ni jedan vektor-redak matrice X se ne može izraziti kao linearna kombinacija ostalih vektor-redaka te matrice.

U suprotnom, tj. u slučaju postojanja takve linearne kombinacije ( $X^T X$ ) matrica postaje singularna, njena determinanta je nula, pa nema jedinstvenog rješenja tj. ocjena parametara u modelu. (Ili se dogodi toliko visoki stupanj ovisnosti da

matrica  $(X^T X)$  prilikom invertiranja postaje preosjetljiva na greške numeričkog računanja i zaokruživanja, pa se takva matrica onda naziva tzv. "zlobna matrica" tj. "ill conditioned matrix".)

U empirijskim istraživanjima, tj. u praksi, ne postoji takva potpuna ovisnost, ali je često ona približna linearnoj ovisnosti, što u statistici i ekonometriji predstavlja veliki problem ocjene i interpretacije parametara.

U takvom slučaju dijagonalne vrijednosti u matrici  $(X^T X)^{-1}$  postaju velike, što uvjetuje velike standardne pogreške ocijenjenih parametara:

$$Se(\hat{\beta}_j) = \hat{\sigma}_{\hat{Y}} \cdot \sqrt{s_{jj}}, \quad (4.83)$$

gdje  $s_{jj}$  predstavlja vrijednost odgovarajućeg dijagonalnog elementa u matrici  $(X^T X)^{-1}$ .

To uvjetuje malu vrijednost  $t^*$ - omjera što bi navelo na zaključak da parametar  $\hat{\beta}_j$  - nije statistički značajan, odnosno da je:

$$t^* = \frac{\hat{\beta}_j}{Se(\hat{\beta}_j)} < t_{tab}. \quad (4.84)$$

Problem utvrđivanja multikolinearnosti, praćenja i eliminiranja nije do danas riješen u statističkoj teoriji.

Među standardnim pokazateljima multikolinearnosti u programskim paketima su faktor inflacije varijance  $VIF$  (Variance Inflation Factor) ili ekvivalentni pokazatelj  $TOL$  (Tolerance):

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \dots, k; \quad (4.85)$$

$$TOL_j = \frac{1}{VIF_j} = 1 - R_j^2. \quad (4.86)$$

Ozbiljan problem multikolinearnosti je prisutan ako je  $R_j^2 > 0,8$ , odnosno  $VIF_j > 5$ , ili ekvivalentno  $TOL_j < 0,2$ .

U slučaju visoke korelacije regresorske varijable  $X_j$  s ostalim regresorskim varijablama, što rezultira s visokim koeficijentom determinacije  $R_j^2 \approx 1$  (u modelu multiple regresije u kojem je  $j$ -ta regresorska varijabla zavisna, a preostale



regresorske varijable su nezavisne). Zbog toga dolazi do povećanja tj. inflacije varijance od  $\hat{\beta}_j$ :

$$\text{var}(\hat{\beta}_j) = \sigma^2 (X'X)^{-1}_{jj} = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \cdot (1 - R_j^2)}, \quad (4.87)$$

pa odatle i naziv "faktor inflacije varijance".

#### 4.7.2.1 *Farrar-Glauberov test za utvrđivanje postojanja problema multikolinearnosti*

Postavljaju se hipoteze:

$H_0$ .....ne postoji multikolinearnost

$H_1$ ..... postoji multikolinearnost

Nulta ili početna hipoteza kod ovog testiranja pretpostavlja da u ocijenjenom regresijskom modelu nije prisutan problem multikolinearnosti, odnosno da je matrica  $R$  (matrica koeficijenata korelacije između regresorskih varijabli) jedinična, tj. da su vektori regresorskih varijabli izraženi u standardnim vrijednostima ortogonalni.

Alternativna hipoteza tvrdi suprotno, odnosno da je problem multikolinearnosti prisutan u promatranom regresijskom modelu.

Testiranje se vrši usporedbom **empirijske i tablične vrijednosti  $\chi^2$  - testa**. Empirijska vrijednost  $\chi^2 *$  se računa:

$$\chi^2 * = - \left[ n - 1 - \frac{1}{6}(2k + 5) \right] \ln \det R, \quad (4.88)$$

gdje je:

$n$  - broj opservacija,

$k$  - broj regresorskih varijabli u modelu,

$\det R$  - determinanta matrice koeficijenata korelacije između regresorskih varijabli.

**Determinanta matrice koeficijenata korelacije** općenito je:

$$\det R = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{21} & 1 & r_{23} & \dots & r_{2k} \\ r_{31} & r_{32} & 1 & \dots & r_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & r_{k3} & \dots & 1 \end{vmatrix}, \quad (4.89)$$

gdje je  $k$  broj promatranih varijabli. S obzirom da za korelaciju vrijedi da je  $r_{ij} = r_{jk}$ , matrica  $R$  je simetrična matrica.

Tablična vrijednost  $\chi^2$  - testa se uz određenu signifikantnost  $\alpha$  i stupnjeve slobode  $df$  određuje pomoću tablica Hi-kvadrat distribucije:

$$\chi^2_{tab}[\alpha]_{df=\frac{1}{2}k(k+1)} \quad (4.90)$$

Ako je:  $\chi^2 * > \chi^2_{tab} \Rightarrow H_1$  odbacuje se početna hipoteza  $H_0$ , tj. zaključuje se da u ocijenjenom regresijskom modelu postoji problem multikolinearnosti.

Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha$  pomoću  $\chi^2 *$  (Tablice C1 i C2): ako je  $\alpha < 5\% \Rightarrow H_1$ , dok u suprotnom vrijedi hipoteza  $H_0$ .

#### 4.7.2.2 Rješenje problema multikolinearnosti

1. Ako postoji problem multikolinearnosti ocjenjuje se regresijski model gdje je jedna od regresorskih varijabli:

- $X_j$  - regresand (ili ovisna) varijabla,
- preostale  $(k-1)$ ,  $X$  varijable su regresorske varijable.

-Za taj model se računa koeficijent determinacije  $r^2$  i  $F$ -test:

$H_0 \dots X_j$  nije multikolnearna s ostalim regresorskim varijablama

$H_1 \dots X_j$  je multikolnearna s ostalim regresorskim varijablama

U tom slučaju je empirijska vrijednost  $F$ -testa:

$$F^* = \frac{r^2_{X_j \cdot X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_k} / (k-1)}{(1 - r^2_{X_j \cdot X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_k}) / (n-k)} \quad (4.91)$$

Tablična vrijednost  $F$ - testa se uz određenu signifikantnost  $\alpha$  i stupnjeve slobode  $df$  određuje pomoću tablica  $F$ - distribucije:

$$F_{tab}^{[\alpha]}[df_1=k-1; df_2=n-k] \cdot \quad (4.92)$$

Ako je:  $F^* > F_{tab} \Rightarrow H_1$  odbacuje se početna hipoteza  $H_0$ , tj. zaključuje se da u ocijenjenom regresijskom modelu postoji problem multikolinearnosti vezan za varijablu  $X_j$ .

2. Ako postoji problem multikolinearnosti mogu se računati i koeficijenti parcijalne korelacije među regresorskim varijablama modela. Ako je taj koeficijent značajan ( $t$ -test) prihvaća se pretpostavka o postojanju multikolinearnosti između te dvije varijable.
  3. Ako postoji problem multikolinearnosti dodavanjem novih podataka može se smanjiti neka standardna greška.
- Ne postoji jedinstvena metoda za rješenje problema multikolinearnosti i samo **rješenje ovisi o konkretnom slučaju**, na primjer:
1. - povećanje uzorka, tj. broja opservacija, što je u praksi često nemoguće, posebno u slučaju vremenskih serija;
  2. - korištenje eksternih ocjena, - npr. informacije o parametrima "a priori";
  3. - isključenje iz regresijskog modela neke od regresorskih varijabli koje pridonose pojavi problema multikolinearnosti;
  4. - primjena tzv. "ridge regression" - uvodi se vrijednost  $(k+1)$ , gdje je  $k$ -konstanta kojom se množe dijagonalni elementi matrice varijanci-kovarijanci, pomoću čega se dobiju ocjene  $\hat{\beta}_{j(k)}$ . Dakle postoji  $k > 0$ , za koje je suma kvadrata pogrešaka za  $\hat{\beta}_{j(k)}$  manja od sume kvadrata pogrešaka za  $\hat{\beta}_j$ . Na taj način se postiže stabilnost sustava.
  5. - kod regresorskih varijabli se formiraju prve diferencije (najčešće kod analize vremenskih nizova), što smanjuje multikolinearnost.



#### Primjer 4.10.

Za razdoblje od siječnja 1996. godine do svibnja 2006. godine ispitivana je ovisnost ukupne industrije u RH o ukupnom izvozu, ukupnom uvozu, PDV-u, kamatnim stopama, novčanoj masi M1 i stopi registrirane nezaposlenosti. Stepwise metodom potrebno je ocijeniti odgovarajući regresijski model.



#### Rješenje 4.10.

Da bi se u programskom paketu **SPSS** dobila ocjena parametara višestrukog linearnog regresijskog modela u kojem se ispituje ovisnost *Ukupne industrije o ukupnom izvozu, ukupnom uvozu, PDV-u, kamatnoj stopi, novčanoj masi M1 i stopi registrirane nezaposlenosti* potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Regression** gdje se bira **Linear**. U otvorenom prozoru bira se varijabla *Ukupna industrija* u: **Dependent** i *ukupni izvoz, ukupni uvoz, PDV, kamatne stope, novčana masa M1, stopa registrirane nezaposlenosti* u: **Independent(s)**. U **Statistics** treba aktivirati: **Estimates, Confidence Intervals, Covariance matrix, Model fit, Collinearity diagnostics; Continue**. Među metodama za odabir varijabli bira se **Stepwise metoda**. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi ocijenjeni regresijski model s potrebnom regresijskom dijagnostikom.

Konačni regresijski model sastoji od tri regresorske varijable: *ukupni uvoz, PDV i stopa registrirane nezaposlenosti* budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U tablici 4.14 je vidljivo kako se konačni regresijski model sastoji od tri regresorske varijable: uvoza, PDV-a i stope reg. nezaposlenosti, budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U prvom koraku uvoz je prva varijabla za koju je ispitivan kriterij ulaska, jer je promatrana zavisna varijabla u najvišoj korelaciji sa ukupnom industrijom.

U drugom koraku ispituje se kriterij ulaska druge po redu nezavisne varijable s najvišim koeficijentom parcijalne korelacije. To je varijabla PDV. Varijabla PDV također ulazi u model.

Kada se u modelu nađu barem dvije nezavisne varijable, može se ispitivati kriterij izlaska varijable iz modela u trećem koraku. Očito u ovom primjeru varijable koje su ušle u model ne zadovoljavaju kriterij izlaska (*p-vrijednost* veća ili jednaka 10%), stoga ostaju u regresijskom modelu.

Tablica 4.14

Postupak Stepwise metode pri izboru parametara za ocijenjeni model s ukupnom industrijom kao zavisnom varijablom

Variables Entered/Removed <sup>a</sup>			
Model	Variables Entered	Variables Removed	Method
1	Uvoz ukupno	.	Stepwise (Criteria: Probability-of-F-to-enter ≤ .050, Probability-of-F-to-remove ≥ .100).
2	PDV	.	Stepwise (Criteria: Probability-of-F-to-enter ≤ .050, Probability-of-F-to-remove ≥ .100).
3	Stopa reg. nez.	.	Stepwise (Criteria: Probability-of-F-to-enter ≤ .050, Probability-of-F-to-remove ≥ .100).

a. Dependent Variable: Industrija ukupno (2000=100)

U trećem koraku ispituje se kriterij ulaska sljedeće po redu nezavisne varijable s najvišim koeficijentom parcijalne korelacije. To je varijabla stopa reg. nezaposlenosti, koja također ulazi u model.

U sljedećem koraku, koji je ujedno i posljednji, niti jedna od preostalih varijabli ne zadovoljava kriterij ulaska. Stoga *Stepwise* metoda izbora završava s tri nezavisne varijable u regresiji u kojoj se ukupna industrija promatra kao zavisna varijabla.

Tablica 4.15.

Osnovni podaci o ocijenjenom modelu s ukupnom industrijom kao zavisnom varijablom

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,937 <sup>a</sup>	,878	,877	4,38349
2	,953 <sup>b</sup>	,907	,906	3,83586
3	,957 <sup>c</sup>	,915	,913	3,69368

a. Predictors: (Constant), Uvoz ukupno  
b. Predictors: (Constant), Uvoz ukupno, PDV  
c. Predictors: (Constant), Uvoz ukupno, PDV, Stopa reg. nez.

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici 4.15 izračunati su osnovni pokazatelji o ocijenjenom višestrukome modelu.

Za 1. model:

$$r_1 = 0,937, r_1^2 = 0,878, \bar{r}_1^2 = 0,877, \hat{\sigma}_{\hat{y}_1} = 4,38349.$$

Za 2. model:

$$r_2 = 0,953, r_2^2 = 0,907, \bar{r}_2^2 = 0,906, \hat{\sigma}_{\hat{y}_2} = 3,83586.$$

Za 3. konačni model:

$$r_3 = 0,957, r_3^2 = 0,915, \bar{r}_3^2 = 0,913, \hat{\sigma}_{\hat{y}_3} = 3,69368.$$

Vrijednosti koeficijenta korelacije je  $r_3 = 0,957$ , i on pokazuje jaku i pozitivnu linearnu vezu između varijabli modela. Koeficijent multiple determinacije (R Square) je  $r_3^2 = 0,915$ , što znači da je ocijenjenim regresijskim modelom protumačeno 91,5% sume kvadrata ukupnih odstupanja zavisne varijable od njene aritmetičke sredine. Protumačenost ocijenjenog modela je jako visoka, što ukazuje na dobru reprezentativnost. Korigirani koeficijent determinacije (Adjusted R Square) je  $\bar{r}_3^2 = 0,913$  i standardna greška ocijenjene regresije je  $\hat{\sigma}_{\hat{y}_3} = 3,69368$ .

**Tablica 4.16.**

**Tablica ANOVA ocijenjenog regresijskog modela**

ANOVA <sup>d</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15220,873	1	15220,873	792,135	,000 <sup>a</sup>
	Residual	2113,651	110	19,215		
	Total	17334,524	111			
2	Regression	15730,717	2	7865,358	534,556	,000 <sup>b</sup>
	Residual	1603,807	109	14,714		
	Total	17334,524	111			
3	Regression	15861,051	3	5287,017	387,518	,000 <sup>c</sup>
	Residual	1473,473	108	13,643		
	Total	17334,524	111			

a. Predictors: (Constant), Uvoz ukupno

b. Predictors: (Constant), Uvoz ukupno, PDV

c. Predictors: (Constant), Uvoz ukupno, PDV, Stopa reg. nez.

d. Dependent Variable: Industrija ukupno (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenih modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema tablici 4.16 empirijska vrijednost F-testa za konačni model je:

$$SP_3 = 15861,051, \quad SR_3 = 1473,473, \quad ST_3 = 17334,524, \quad F_3^* = 387,518.$$

Konačno vrijedi da je  $\alpha_3^* \approx 0 < \alpha = 5\% \Rightarrow H_1$ , pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

**Tablica 4.17.**

**Ocijenjeni linearni regresijski model s ukupnom industrijom kao zavisnom varijablom**

Coefficients <sup>a</sup>										
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	70,436	1,427		49,344	,000	67,607	73,264		
	Uvoz ukupno	5,856E-6	,000	,937	28,145	,000	,000	,000	1,000	1,000
2	(Constant)	65,551	1,500		43,712	,000	62,579	68,523		
	Uvoz ukupno	4,196E-6	,000	,671	12,497	,000	,000	,000	,294	3,400
	PDV	7,658E-6	,000	,316	5,886	,000	,000	,000	,294	3,400
3	(Constant)	75,692	3,585		21,115	,000	68,586	82,797		
	Uvoz ukupno	4,198E-6	,000	,672	12,987	,000	,000	,000	,294	3,400
	PDV	7,608E-6	,000	,314	6,073	,000	,000	,000	,294	3,401
	Stopa reg. nez.	-,523	,169	-,087	-3,091	,003	-,859	-,188	1,000	1,000
a. Dependent Variable: Industrija ukupno (2000=100)										

a. Dependent Variable: Industrija ukupno (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici 4.17 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

Analitički izraz konačnog modela (3) je:

$$\hat{y}_i = 75,692 + 0,0000042 \cdot X_{1i} + 0,00000761 \cdot X_{2i} - 0,523 \cdot X_{3i}$$

Parametar  $\hat{\beta}_0 = 75,692$  predstavlja očekivani ukupnu industriju ( $\hat{Y}$ ) u slučaju da sve regresorske varijable poprime vrijednost nula. Ovaj parametar nema uvijek ekonomski logično značenje.

Parametar uz nezavisnu varijablu  $X_1$ , tj.  $\hat{\beta}_1 = 0,0000042$  pokazuje da se može očekivati porast ukupne industrije za 0,0000042 ako uvoz poraste za 1 jedinicu uz c.p. (*ceteris paribus*), tj. uz uvjet da se ne mijenjaju ostale varijable u modelu.

Parametar uz nezavisnu varijablu  $X_2$ , tj.  $\hat{\beta}_2 = 0,00000761$  pokazuje da se može očekivati porast ukupne industrije za 0,00000761 ako PDV poraste za 1 jedinicu uz c.p. (*ceteris paribus*), tj. uz uvjet da se ne mijenjaju ostale varijable u modelu.

Parametar uz nezavisnu varijablu  $X_3$ , tj.  $\hat{\beta}_3 = -0,523$  pokazuje da se može očekivati pad ukupne industrije za 0,523 ako stopa reg. nezaposlenosti poraste za 1 jedinicu uz c.p. (*ceteris paribus*), tj. uz uvjet da se ne mijenjaju ostale varijable u modelu.

Regresijski model u standardiziranom obliku je:  $\hat{y}_i = 0,672 \cdot x_{1i} + 0,314 \cdot x_{2i} - 0,087 \cdot x_{3i}$ .

Standardizirana vrijednost parametra pokazuje za koliko će se standardnih devijacija promijeniti zavisna varijabla, ako se nezavisna varijabla promijeni za 1 standardnu devijaciju. Na temelju toga, može se zaključiti da na ukupnu industriju u RH najveći relativni utjecaj ima prva varijabla u konačno ocijenjenom modelu, tj. uvoz, jer je uz tu varijablu najveća apsolutna vrijednost standardiziranog koeficijenta.

Da bi se izvršilo testiranje značajnosti pojedinačnih parametara modela potrebno je postaviti hipoteze:

$$H_0 \dots \beta_j = 0$$

$$H_1 \dots \beta_j \neq 0$$

Empirijska signifikantnost za parametar  $\hat{\beta}_0$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_1$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_2$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_3$  je  $\alpha^* = 0,3\%$ .

Sve empirijske signifikantnosti za regresijske parametre  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$  su manje od  $\alpha = 5\%$  i može se zaključiti da su svi parametri statistički značajni.

Potvrdu o nepostojanju problema multikolinearnosti u tablici 4.16 za sve parametre konačnog modela daju vrijednosti faktora inflacije varijance ( $VIF_j$ ) i njihove recipročne vrijednosti ( $TOL_j$ ):

$$VIF_1 = 3,4 \Rightarrow VIF_1 < 5, \text{ i } TOL_1 = 0,294 \Rightarrow TOL_1 > 20\%$$

$$VIF_2 = 3,401 \Rightarrow VIF_2 < 5, \text{ i } TOL_2 = 0,294 \Rightarrow TOL_2 > 20\%$$

$$VIF_3 \approx 1,000 \Rightarrow VIF_3 < 5, \text{ i } TOL_3 \approx 1,000 \Rightarrow TOL_3 > 20\%.$$

Drugim riječima, za sva 3 parametra regresije (tj.  $\hat{\beta}_1, \hat{\beta}_2$  i  $\hat{\beta}_3$ ) faktori inflacije varijance su manji od 5, a postotak tolerancije je veći od 20%, što potvrđuje da niti jedna regresorska varijable ne uvjetuje problem multikolinearnosti.



U tablici ocijenjenih regresijskih koeficijenata 4.16 može se vidjeti i intervalna procjena statistički značajnih parametara uz 95% pouzdanosti.



#### Primjer 4.11.

Za razdoblje od siječnja 1996. godine do svibnja 2006. godine ispitivana je ovisnost izvoza u RH o ukupnom uvozu, PPI indeksu cijena, PDV-u, kamatnim stopama, zaposlenim u pravnim osobama, novčanoj masi M1 i ukupnoj industriji. Stepwise metodom potrebno je ocijeniti odgovarajući regresijski model.



#### Rješenje 4.11.

Da bi se u programskom paketu **SPSS** dobila ocjena parametara višestrukog linearnog regresijskog modela u kojem se ispituje ovisnost *ukupnog izvoza* o *ukupnom uvozu*, *PPI indeksu cijena*, *PDV-u*, *kamatnoj stopi*, *zaposlenim u pravnim osobama*, *novčanoj masi M1* i *ukupnoj industriji* potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Regression** gdje se bira **Linear**. U otvorenom prozoru bira se varijabla *ukupni izvoz* a u: **Dependent** i *ukupni uvoz*, *PPI indeks cijena*, *PDV*, *kamatne stope*, *zaposleni u pravnim osobama*, *novčana masa M1* i *ukupna industrija* u: **Independent(s)**. U **Statistics** treba aktivirati: **Estimates**, **Confidence Intervals**, **Covariance matrix**, **Model fit**, **Collinearity diagnostics**; **Continue**. Među metodama za odabir varijabli bira se **Stepwise metoda**. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi ocijenjeni regresijski model s potrebnom regresijskom dijagnostikom.

Konačni regresijski model sastoji od pet regresorskih varijabli: *ukupni uvoz*, *PPI cijena*, *zap. u pravnim osobama*, *M1* i *industrija ukupno* budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

Tablica 4.18.

#### Ocijenjeni linearni regresijski model s ukupnim izvozom kao zavisnom varijablom

5	(Constant)	-9586023,007	1449932,340		-6,611	,000	-1,246E7	-6713167,936		
	Uvoz uk.	,101	,051	,250	1,965	,052	,000	,202	,075	13,250
	PPI cijene (2000=100)	52400,043	8750,076	,552	5,989	,000	35062,889	69737,197	,144	6,945
	Zap. u prav. os.	4,940	1,320	,223	3,743	,000	2,325	7,555	,345	2,897
	M1 u mil. kn	-29,833	8,913	-,355	-3,347	,001	-47,492	-12,173	,108	9,217
	Ind. ukupno (2000=100)	20660,146	7325,714	,326	2,820	,006	6145,182	35175,109	,091	10,952

a. Dependent Variable: Izvoz uk.

Izvor: <http://epp.eurostat.ec.europa.eu>.

Konačni regresijski model je dobar po svoj regresijskoj dijagnostici, ali kako se može vidjeti u tablici 4.18 utvrđen je problem multikolinearnosti. Najveći  $VIF_j$  u konačnom modelu je onaj uz varijablu ukupnog uvoza:

$$VIF_1 = 13,25 \Rightarrow VIF_1 > 5, \text{ i } TOL_1 = 0,075 \Rightarrow TOL_1 < 20\%$$

pa je regresija ponovljena bez te varijable.

U programskom paketu **SPSS** ponavlja se **Stepwise** metoda, ali bez varijable uvoza:

- varijable *PPI indeks cijena*, *PDV*, *kamatne stope*, *zaposleni u pravnim osobama*, *novčana masa M1* i *ukupna industrija* u: **Independent**. U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

Konačni regresijski model sada se sastoji od četiri regresorske varijable: **industrija ukupno**, **PPI cijena**, **zap. u pravnim osobama** i **M1** budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

**Tablica 4.19.**

**Ocijenjeni linearni regresijski model s ukupnim izvozom kao zavisnom varijablom**

4	(Constant)	-1,033E7	1417409,033		-7,287	,000	-1,314E7	-7520621,978		
	Ind. ukupno (2000=100)	31621,953	4808,075	,500	6,577	,000	22096,288	41147,617	,217	4,601
	PPI cijene (2000=100)	59123,367	8154,715	,823	7,250	,000	42967,406	75279,328	,170	5,883
	Zap. u prav. os.	4,453	1,313	,201	3,392	,001	1,052	7,054	,358	2,795
	M1 u mil. kn	-27,736	8,960	-,330	-3,096	,002	-45,488	-9,985	,110	9,085

a. Dependent Variable: Izvoz uk.

Izvor: <http://epp.eurostat.ec.europa.eu>.

Konačni regresijski model je dobar po svoj regresijskoj dijagnostici, ali kako se može vidjeti u tablici 4.19 utvrđen je problem multikolinearnosti. Najveći  $VIF_j$  u konačnom modelu je onaj uz varijablu M1:

$$VIF_4 = 9,085 \Rightarrow VIF_4 > 5, \text{ i } TOL_4 = 0,11 \Rightarrow TOL_4 < 20\%$$

pa je regresija opet ponovljena bez navedene varijable.

U programskom paketu **SPSS** ponavlja se **Stepwise** metoda, ali bez varijable M1:

- varijable *PPI indeks cijena*, *PDV*, *kamatne stope*, *zaposleni u pravnim osobama* i *ukupna industrija* u: **Independent**. U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

Konačni regresijski model sada se sastoji od tri regresorske varijable: **ukupna industrija**, **PPI cijena** i **zaposlenost u pravnim osobama** budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

Tablica 4.20.

**Osnovni podaci o ocijenjenom modelu s ukupnim izvozom kao zavisnom varijablom**

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,871 <sup>a</sup>	,758	,756	3,97327E5
2	,915 <sup>b</sup>	,837	,834	3,27466E5
3	,920 <sup>c</sup>	,846	,842	3,19752E5

a. Predictors: (Constant), Ind. ukupno (2000=100)  
b. Predictors: (Constant), Ind. ukupno (2000=100), PPI cijene (2000=100)  
c. Predictors: (Constant), Ind. ukupno (2000=100), PPI cijene (2000=100), Zap. u prav. os.

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici 4.20 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

Za 3. konačni model:

$$r_3 = 0,920, \quad r_3^2 = 0,846, \quad \bar{r}_3^2 = 0,842, \quad \hat{\sigma}_{\hat{y}_3} = 319751,991.$$

Vrijednosti koeficijenta korelacije je  $r_3 = 0,920$ , i on pokazuje jaku i pozitivnu linearnu vezu između varijabli modela. Koeficijent multiple determinacije (R Square) je  $r_3^2 = 0,846$ , što znači da je ocijenjenim regresijskim modelom protumačeno 84,6% sume kvadrata ukupnih odstupanja zavisne varijable od njene aritmetičke sredine. Protumačenost ocijenjenog modela je jako visoka, što ukazuje na dobru reprezentativnost. Korigirani koeficijent determinacije (Adjusted R Square) je  $\bar{r}_3^2 = 0,842$  i standardna greška ocijenjene regresije je  $\hat{\sigma}_{\hat{y}_3} = 319751,991$ .

U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenih modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema tablici 4.21 empirijska vrijednost F-testa za konačni model je:

$$SP_3 = 6 \cdot 10^{13}, \quad SR_3 = 1 \cdot 10^{13}, \quad ST_3 = 8 \cdot 10^{13}, \quad F_3^* = 209,104,$$

Konačno vrijedi da je  $\alpha_3^* \approx 0 < \alpha = 5\% \Rightarrow H_1$ , pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

**Tablica 4.21.**

**Tablica ANOVA ocijenjenog regresijskog modela**

ANOVA <sup>d</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5,748E13	1	5,748E13	364,100	,000 <sup>a</sup>
	Residual	1,831E13	116	1,579E11		
	Total	7,579E13	117			
2	Regression	6,346E13	2	3,173E13	295,898	,000 <sup>b</sup>
	Residual	1,233E13	115	1,072E11		
	Total	7,579E13	117			
3	Regression	6,414E13	3	2,138E13	209,104	,000 <sup>c</sup>
	Residual	1,166E13	114	1,022E11		
	Total	7,579E13	117			

a. Predictors: (Constant), Ind. ukupno (2000=100)  
b. Predictors: (Constant), Ind. ukupno (2000=100), PPI cijene (2000=100)  
c. Predictors: (Constant), Ind. ukupno (2000=100), PPI cijene (2000=100), Zap. u prav. os.  
d. Dependent Variable: Izvoz uk.

Izvor: <http://epp.eurostat.ec.europa.eu>.

**Tablica 4.22.**

**Ocijenjeni linearni regresijski model s ukupnim izvozom kao zavisnom varijablom**

Coefficients <sup>a</sup>									
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance VIF
1	(Constant)	-2835785,229	308658,538		-9,187	,000	-3447122,354	-2224448,103	
	Ind. ukupno (2000=100)	55112,483	2888,284	,871	19,081	,000	49391,872	60833,093	1,000 1,000
2	(Constant)	-4726409,319	358890,229		-13,170	,000	-5437301,790	-4015516,848	
	Ind. ukupno (2000=100)	30470,355	4068,665	,481	7,489	,000	22411,112	38529,597	,342 2,921
	PPI cijene (2000=100)	45574,921	6102,557	,480	7,468	,000	33486,929	57662,912	,342 2,921
3	(Constant)	-7361399,969	1082707,762		-6,799	,000	-9506235,696	-5216564,222	
	Ind. ukupno (2000=100)	25118,639	4484,672	,397	5,601	,000	16234,538	34002,741	,269 3,723
	PPI cijene (2000=100)	41766,181	6140,021	,440	6,802	,000	29602,848	53929,515	,322 3,102
	Zap. u prav. os.	3,376	1,313	,152	2,572	,011	,776	5,977	,385 2,598

a. Dependent Variable: Izvoz uk.

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici 4.22 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

Analitički izraz konačnog modela (3) je:

$$\hat{y}_i = -7361400 + 25118,639 \cdot X_{1i} + 41766,181 \cdot X_{2i} + 3,376 \cdot X_{3i}$$

Parametar predstavlja očekivani ukupni izvoz industrije ( $\hat{Y}$ ) u slučaju da sve regresorske varijable poprima vrijednost nula. Ovaj parametar nema uvijek ekonomski logično značenje.

Parametar uz nezavisnu varijablu  $X_1$ , tj.  $\hat{\beta}_1 = 25118,64$  pokazuje da se može očekivati porast izvoza za 25118,64 jedinice ako industrija poraste za 1 jedinicu uz c.p. (*ceteris paribus*), tj. uz uvjet da se ne mijenjaju ostale varijable u modelu.

Parametar uz nezavisnu varijablu  $X_2$ , tj.  $\hat{\beta}_2 = 41766,18$  pokazuje da se može očekivati porast izvoza za 41766,18 jedinica ako PPI cijena poraste za 1 jedinicu uz c.p. (*ceteris paribus*), tj. uz uvjet da se ne mijenjaju ostale varijable u modelu.

Parametar uz nezavisnu varijablu  $X_3$ , tj.  $\hat{\beta}_3 = 3,376$  pokazuje da se može očekivati porast izvoza za 3,376 jedinica ako zaposlenost poraste za 1 jedinicu uz c.p. (*ceteris paribus*), tj. uz uvjet da se ne mijenjaju ostale varijable u modelu.

Regresijski model u standardiziranom obliku je:  $\hat{y}_i = 0,397 \cdot x_{1i} + 0,440 \cdot x_{2i} + 0,152 \cdot x_{3i}$

Može se zaključiti da na ukupni izvoz u RH najveći relativni utjecaj ima varijabla  $X_2$  PPI cijena (2000=100), jer je uz tu varijablu najveća apsolutna vrijednost standardiziranog koeficijenta.

Da bi se izvršilo testiranje značajnosti pojedinačnih parametara modela potrebno je postaviti hipoteze:

$$H_0 \dots \beta_j = 0$$

$$H_1 \dots \beta_j \neq 0$$

Empirijska signifikantnost za parametar  $\hat{\beta}_0$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_1$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_2$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_3$  je  $\alpha^* = 1,1\%$ .

Sve empirijske signifikantnosti za regresijske parametre  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$  su manje od  $\alpha = 5\%$  i može se zaključiti da su svi parametri statistički značajni.

Potvrdu o nepostojanju problema multikolinearnosti u tablici 4.22 za sve parametre konačnog modela daju vrijednosti faktora inflacije varijance ( $VIF_j$ ) i njihove recipročne vrijednosti ( $TOL_j$ ):

$$VIF_1 = 3,723 \Rightarrow VIF_1 < 5, \text{ i } TOL_1 = 0,269 \Rightarrow TOL_1 > 20\%$$

$$VIF_2 = 3,102 \Rightarrow VIF_2 < 5, \text{ i } TOL_2 = 0,322 \Rightarrow TOL_2 > 20\%$$

$$VIF_3 = 2,598 \Rightarrow VIF_3 < 5, \text{ i } TOL_3 = 0,385 \Rightarrow TOL_3 > 20\%.$$

Drugim riječima, za sva 3 parametra regresije (tj.  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  i  $\hat{\beta}_3$ ) faktori inflacije varijance su manji od 5, a postotak tolerancije je veći od 20%, što potvrđuje da niti jedna regresorska varijable ne uvjetuje problem multikolinearnosti.

U tablici ocijenjenih regresijskih koeficijenata 4.22 može se vidjeti i intervalna procjena statistički značajnih parametra uz 95% pouzdanosti.



#### Primjer 4.12.

Za razdoblje od siječnja 1996. godine do svibnja 2006. godine ispitivana je ovisnost  $I_t$  izvoza u RH (2000=100) o ukupnoj industriji (2000=100), PPI indeksu cijena (2000=100) i  $I_t$  zaposlenih u pravnim osobama (2000=100) (**NAPOMENA: zadane varijable su iz prethodnog primjera 4.11, ali su sve transformirane u bazne indekse po jednakoj bazi!**). Stepwise metodom potrebno je ocijeniti odgovarajući regresijski model.



#### Rješenje 4.12.

Da bi se u programskom paketu **SPSS** dobila ocjena parametara višestrukog linearnog regresijskog modela u kojem se ispituje ovisnost  $I_t$  ukupnog izvoza o  $I_t$  ukupnoj industriji, PPI indeksu cijena,  $I_t$  zaposlenim u pravnim osobama potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Regression** gdje se bira **Linear**. U otvorenom prozoru bira se varijabla  $I_t$  ukupni izvoz u: **Dependent** i  $I_t$  ukupna industrija, PPI indeks cijena,  $I_t$  zaposleni u pravnim osobama u: **Independent(s)**. U **Statistics** treba aktivirati: **Estimates, Confidence Intervals, Covariance matrix, Model fit, Collinearity diagnostics; Continue**. Među metodama za odabir varijabli bira se **Stepwise metoda**. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi ocijenjeni regresijski model s potrebnom regresijskom dijagnostikom.

Konačni regresijski model opet se sastoji od tri iste, ali transformirane regresorske varijable:  $I_t$  ukupna industrija, PPI cijena i  $I_t$  zaposlenost u pravnim osobama budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika, koja je jednaka konačnom modelu iz prethodnog **primjera 4.11**, samo su konkretni ocijenjeni parametri drugačiji. Dakle, jednaki su podaci o determinaciji modela, kao i svi relativni pokazatelji u tablici ANOVA.

**Tablica 4.23.**

**Ocijenjeni linearni regresijski model s  $I_t$  izvoza kao zavisnom varijablom**

3	(Constant)	-240,143	35,320		-6,799	,000	-310,111	-170,174		
	Ind. ukupno (2000=100)	,819	,146	,397	5,601	,000	,530	1,109	,269	3,723
	PPI cijene (2000=100)	1,362	,200	,440	6,802	,000	,966	1,759	,322	3,102
	Iz zaposleni (2000=100)	1,160	,451	,152	2,572	,011	,267	2,053	,385	2,598
a. Dependent Variable: It izvoz (2000=100)										

Izvor: <http://epp.eurostat.ec.europa.eu>.

Prema tablici 4.23 analitički izraz ovog modela je:

$$\hat{y}_i = -240,143 + 0,819 \cdot X_{1i} + 1,362 \cdot X_{2i} - 1,160 \cdot X_{3i}$$

Po svim karakteristikama ovaj model jednak je onom iz prethodnog **primjera 4.11**. Naime, ovdje su varijable u originalnim vrijednostima transformirane u bazne indekse, a poznato je da bazni indeksi čuvaju originalne odnose među vrijednostima promatrane pojave.

Model u standardiziranom obliku isto je identičan onom iz **primjera 4.11**. Jednaki su i faktori inflacije varijance ( $VIF_j$ ) i njihove recipročne vrijednosti ( $TOL_j$ ), na temelju čega se zaključuje da u ocijenjenom modelu nije prisutan problem multikolinearnosti.



#### **Primjer 4.13.**

Za razdoblje od siječnja 1996. godine do svibnja 2006. godine ispitivana je ovisnost PPI cijena u RH (2000=100) o kamatnim stopama na kun. kred. s val. klauz. u %, i  $I_t$  Izvoza (2000=100). Stepwise metodom potrebno je ocijeniti odgovarajući regresijski model. Na kraju analize potrebno je sve varijable iz ocijenjenog regresijskog modela prikazati na zajedničkom dijagramu rasipanja.



#### **Rješenje 4.13.**

Da bi se u programskom paketu **SPSS** dobila ocjena parametara višestrukog linearnog regresijskog modela u kojem se ispituje ovisnost *PPI indeksa cijena o kamatnim stopama na kun. kred. s val. klauz. u %,  $I_t$  izvoza* potrebno je na glavnom izborniku izabrati ikonu **Analyze**, a na njezinom padajućem izborniku **Regression** gdje se bira **Linear**. U otvorenom prozoru bira se varijabla *PPI indeks cijena* u: **Dependent** i *kamatne stope na kun. kred. s val. klauz. u %,  $I_t$  izvoza* u: **Independent(s)**. U **Statistics** treba aktivirati: **Estimates, Confidence Intervals, Covariance matrix, Model fit, Collinearity diagnostics; Continue**. Među metodama za odabir varijabli bira se **Stepwise metoda**. Klikom na ikonu **OK** u **Output**-u programa **SPSS** dobije se traženi ocijenjeni regresijski model s potrebnom regresijskom dijagnostikom.

Konačni regresijski model sastoji od dvije regresorske varijable: *kamatne stope na kun. kred. s val. klauz. u %* i  *$I_t$  izvoza*, budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

**Tablica 4.24.**

**Osnovni podaci o ocijenjenom modelu s PPI cijena kao zavisnom varijablom**

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,883 <sup>a</sup>	,779	,777	4,00242
2	,926 <sup>b</sup>	,858	,856	3,21985

a. Predictors: (Constant), Kam. st. na kun. kred. s val. klauz. u %  
 b. Predictors: (Constant), Kam. st. na kun. kred. s val. klauz. u %,  $I_t$  izvoz (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici 4.24 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

Za konačni model:

$$r_2 = 0,926, \quad r_2^2 = 0,858, \quad \bar{r}_2^2 = 0,856, \quad \hat{\sigma}_{\hat{y}_2} = 3,21985.$$

Vrijednosti koeficijenta korelacije je  $r_2 = 0,926$ , i on pokazuje jaku i pozitivnu linearnu vezu između varijabli modela. Koeficijent multiple determinacije (R Square) je  $r_2^2 = 0,858$ , što znači da je ocijenjenim regresijskim modelom protumačeno 85,8% sume kvadrata ukupnih odstupanja zavisne varijable od njene aritmetičke sredine. Protumačenost ocijenjenog modela je jako visoka, što ukazuje na dobru reprezentativnost. Korigirani koeficijent determinacije (Adjusted R Square) je  $\bar{r}_2^2 = 0,856$  i standardna greška ocijenjene regresije je  $\hat{\sigma}_{\hat{y}_2} = 3,21985$ .



U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenih modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema tablici 4.25 empirijska vrijednost F-testa za konačni model je:

$$SP_2 = 7219,689, \quad SR_2 = 1192,257, \quad ST_2 = ST_1 = 8411,947, \quad F_2^* = 348,19,$$

Konačno vrijedi da je  $\alpha_3^* \approx 0 < \alpha = 5\% \Rightarrow H_1$ , pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

**Tablica 4.25.**

**Tablica ANOVA ocijenjenog regresijskog modela**

ANOVA <sup>o</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6553,697	1	6553,697	409,110	,000 <sup>a</sup>
	Residual	1858,250	116	16,019		
	Total	8411,947	117			
2	Regression	7219,689	2	3609,845	348,190	,000 <sup>b</sup>
	Residual	1192,257	115	10,367		
	Total	8411,947	117			

a. Predictors: (Constant), Kam. st. na kun. kred. s val. klauz. u %  
b. Predictors: (Constant), Kam. st. na kun. kred. s val. klauz. u %, It izvoz (2000=100)  
c. Dependent Variable: PPI cijene (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

**Tablica 4.26.**

**Ocijenjeni linearni regresijski model s ukupnim izvozom kao zavisnom varijablom**

Coefficients <sup>a</sup>									
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics
		B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	117,880	1,010		116,708	,000	115,880	119,881	
	Kam. st. na kun. kred. s val. klauz. u %	-1,591	,079	-,883	-20,226	,000	-1,747	-1,436	1,000
2	(Constant)	95,417	2,918		32,698	,000	89,636	101,197	
	Kam. st. na kun. kred. s val. klauz. u %	-,934	,104	-,518	-9,022	,000	-1,140	-,729	,374
	It izvoz (2000=100)	,149	,019	,460	8,015	,000	,112	,185	,374

a. Dependent Variable: PPI cijene (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici 4.26 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

Analitički izraz konačnog modela (2) je:  $\hat{y}_i = 95,417 - 0,934 \cdot X_{1i} + 0,149 \cdot X_{i2}$

Parametar  $\hat{\beta}_0 = 95,417$  predstavlja očekivani PPI cijena ( $\hat{Y}$ ) u slučaju da sve regresorske varijable poprime vrijednost nula. Ovaj parametar nema uvijek ekonomski logično značenje.

Parametar uz nezavisnu varijablu  $X_1$ , tj.  $\hat{\beta}_1 = -0,934$  pokazuje da se može očekivati smanjenje PPI za 0,934 ako kamatne stope poraste za 1 % uz c.p. (*ceteris paribus*), tj. uz uvjet da se izvoz ne mijenja.

Parametar uz nezavisnu varijablu  $X_2$ , tj.  $\hat{\beta}_2 = 0,149$  pokazuje da se može očekivati porast PPI za 0,149 ako  $I_t$  izvoza poraste za 1 jedinicu uz c.p. (*ceteris paribus*), tj. uz uvjet da se PPI ne mijenjaju.

Regresijski model u standardiziranom obliku je:  $\hat{y}_i = -0,518 \cdot x_{1i} + 0,460 \cdot x_{i2}$

Može se zaključiti da kamatne stope imaju veći relativan utjecaj na indeks cijena od izvoza, jer je uz tu varijablu veća apsolutna vrijednost standardiziranog koeficijenta.

Da bi se izvršilo testiranje značajnosti pojedinačnih parametara modela potrebno je postaviti hipoteze:

$$H_0 \dots \beta_j = 0$$

$$H_1 \dots \beta_j \neq 0$$

Empirijska signifikantnost za parametar  $\hat{\beta}_0$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_1$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_2$  je  $\alpha^* \approx 0\%$ .

Sve empirijske signifikantnosti za regresijske parametre  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  su manje od  $\alpha = 5\%$  i može se zaključiti da su svi parametri statistički značajni.

Potvrdu o nepostojanju problema kolinearnosti (multikolinearnosti) daju i vrijednosti faktora inafacije varijance ( $VIF_j$ ) i njihove recipročne vrijednosti ( $TOL_j$ ):

$$VIF_1 = 2,677 \Rightarrow VIF_1 < 5, \text{ i } TOL_1 = 0,374 \Rightarrow TOL_1 > 20\%$$

$$VIF_2 = 2,677 \Rightarrow VIF_2 < 5, \text{ i } TOL_2 = 0,374 \Rightarrow TOL_2 > 20\%$$

Drugim riječima, za oba parametra regresije (tj.  $\hat{\beta}_1$  i  $\hat{\beta}_2$ ) faktori inflacije varijance su manji od 5, a postotak tolerancije je veći od 20%, što potvrđuje da niti jedna regresorska varijable ne uvjetuje problem multikolinearnosti.

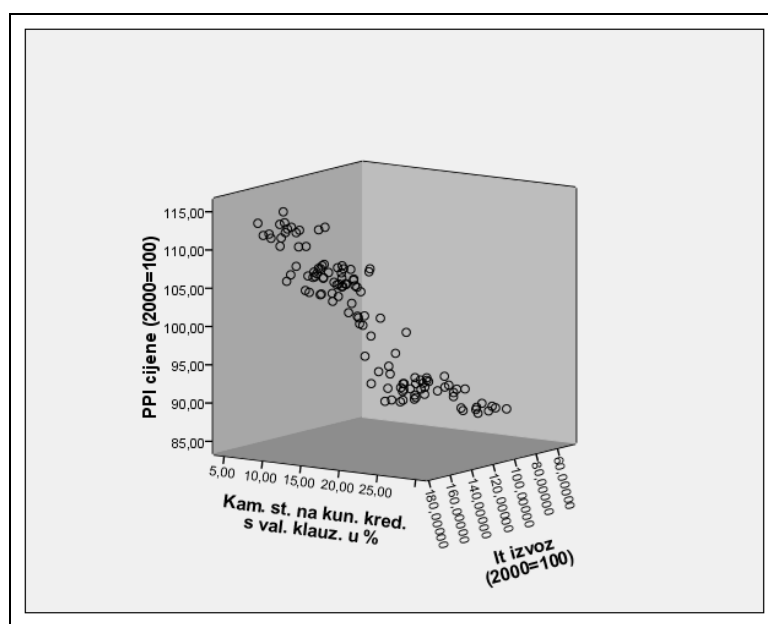
U tablici ocijenjenih regresijskih koeficijenata može se vidjeti i intervalna procjena statistički značajnih parametra uz 95% pouzdanosti.

Na kraju analize su zavisna varijabla, te dvije nezavisne varijable prikazane dijagramom rasipanja u tri dimenzije.

Da bi se u programskom paketu **SPSS** konstruirao trodimenzionalni dijagram rasipanja potrebno je na glavno izborniku odabrati: **Graphs; Legacy Dialogs**, a na pomoćnom izborniku **Scatter/Dot**, gdje je potrebno aktivirati: **3-D Scatter**. Klikom na **Define** na odgovarajuće osi treba prebaciti varijable: **Y Axis:** PPI cijene; **X Axis:** Kamatne stope; **Z Axis:**  $I_t$  Izvoz.

Slika 4.18.

**Trodimenzijski dijagram rasipanja između promatranih varijabli u regresijskom modelu**



Izvor: <http://epp.eurostat.ec.europa.eu>.

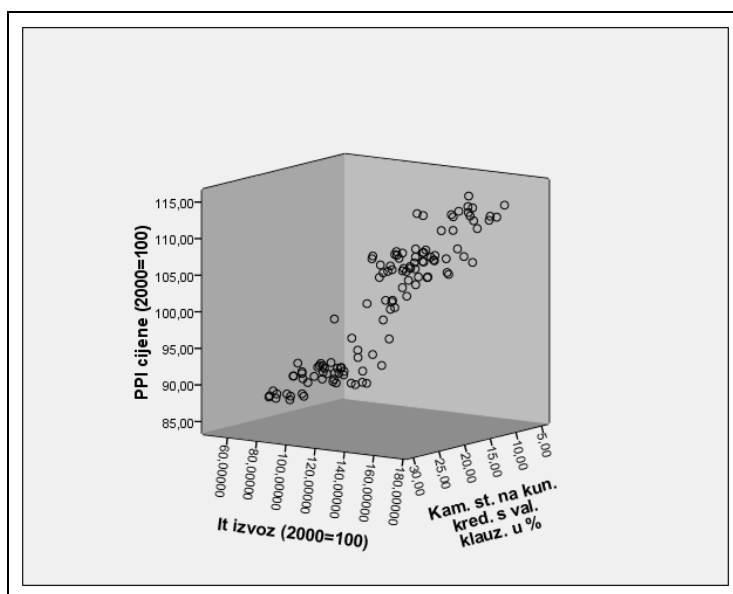
Na slici 4.18 prikazan je trodimenzionalni dijagram rasipanja između varijabli PPI cijena, kamatnih stopa i  $I_t$  izvoza. Može se vidjeti da između varijabli PPI cijena i

kamatnih stopa u RH postoji negativna veza, u skladu s predznakom regresijskog parametra  $\hat{\beta}_1$ .

Kada se prikaz na slici 4.18 okrene da bi se grafički utvrdio karakter veze između varijabli PPI cijena i  $I_t$  izvoza, u SPSS-u je potrebno ponoviti sličan postupak, samo se promatrane varijable na drugi način prebacuju na odgovarajuće osi: **Y Axis:** PPI cijene; **X Axis:**  $I_t$  Izvoz; **Z Axis:** Kamatne stope.

**Slika 4.19.**

**Trodimenzijski dijagram rasipanja između promatranih varijabli u regresijskom modelu**



Izvor: <http://epp.eurostat.ec.europa.eu>.

Na temelju trodimenzionalnog dijagrama rasipanja na slici 4.19 može se vidjeti da između varijabli PPI cijena i  $I_t$  izvoza u RH postoji pozitivna veza u skladu s predznakom regresijskog parametra  $\hat{\beta}_2$ .

## 4.8 Regresijsko modeliranje u uvjetima narušenih osnovnih pretpostavki

### 4.8.1 Problem heteroskedastičnosti varijance reziduala

**Slučajna greška**, odnosno **reziduali** u regresijskoj analizi su odstupanja stvarnih vrijednosti regresand varijable  $Y_i$  od ocijenjenog modela  $\hat{Y}_i$ :

$$e_i = (Y_i - \hat{Y}_i) \quad (4.93)$$

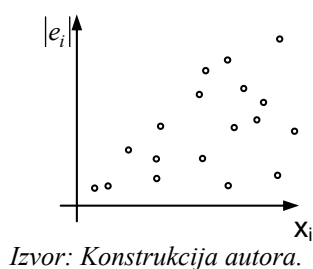
Gauss-Markovi uvjeti za slučajnu grešku u regresijskoj analizi koji se odnose na varijancu reziduala su:

- $Cov(e_i, e_j) = 0, \forall i \neq j$ , ( $\Rightarrow$  nema korelacije između varijabli s pomakom od  $e_i$ ),
- $Cov(e_i, e_j) = \sigma_e^2, \forall i = j$ , ( $\Rightarrow$  varijanca reziduala je konačna i konstanta, tj. homoskedastičnost varijance reziduala).
- $Cov(x_i, e_i) = 0, \forall i$ , ( $\Rightarrow$  regresorska varijabla je nestohastična; varijanca reziduala treba biti konstantna i ne smije korelirati s regresorskom varijablom; kod višestruke regresije to se odnosi na korelaciju između slučajne greške i svake regresorske varijable).

**Heteroskedastičnost varijance reziduala** podrazumijeva varijancu slučajne greške koja u svakoj opservaciji u uzorku izgleda kao da ne potječe iz iste populacije, tj. nije nužno ista za svako "i", što je prikazano na slici 4.20.

Slika 4.20.

**Dijagram rasipanja između apsolutnih reziduala i regresorske varijable**



U takvom slučaju:

- a) ocjene parametara metodom najmanjih kvadrata su neefikasne;
- b) procjene standardnih grešaka su pogrešne, tj. podcijenjene, pa je  $t^*$ -test precijenjen što pogrešno dovodi do zaključka o značajnosti parametara  $\hat{\beta}_j$ .

#### **4.8.1.1 Spearmanov koeficijent korelacije ranga pri utvrđivanju postojanja problema heteroskedastičnosti varijance reziduala**

Postavljaju se hipoteze, gdje nulta hipoteza  $H_0$  pretpostavlja da je vrijednost Spearmanovog koeficijenta korelacije ranga između regresorske varijable i apsolutnih reziduala jednaka 0, tj. da ne postoji problem heteroskedastičnosti varijance reziduala u ocijenjenom modelu. Suprotna hipoteza  $H_1$  pretpostavlja da je vrijednost Spearmanovog koeficijenta korelacije ranga između regresorske varijable i apsolutnih reziduala različit od 0, tj. da u ocijenjenom modelu postoji problem heteroskedastičnosti varijance reziduala.

$$H_0: r_s = 0$$

$$H_1: r_s \neq 0$$

Testiranje se može napraviti usporedbom **empirijske i tablične vrijednosti Z - testa** (ili **t - testa** za manji uzorak):

$$\left| Z^* = \frac{\hat{r}_s}{Se(r_s)} \right| > Z_{tab}, \Rightarrow H_1 \quad (4.94)$$

tj. postoji problem heteroskedastičnosti varijance reziduala.)

gdje je:

$$Se(r_s) = \sqrt{\frac{1}{n-1}}, \quad n > 30, \quad (4.95)$$

$$Se(r_s) = \sqrt{\frac{1-\hat{r}^2}{n-2}}, \quad n \leq 30, \quad (4.96)$$

a  $Z_{\frac{1-\alpha}{2}}$  je odgovarajuća vrijednost  $Z$  testa iz tablica površina ispod normalne krivulje (Tablica).

Kod malog uzorka, umjesto  $Z$  koristi se vrijednost  $t$  - iz Studentove distribucije ( $t_{\frac{\alpha}{2}, df=n-2}$ ) - (Tablica).

**Spearmanov koeficijent korelacije ranga ( $\hat{r}_s$ ) je u ovom slučaju mjera korelacije ranga između regresorske varijable i apsolutnih reziduala:**

$$\hat{r}_{s(X_i, |e_i|)} = 1 - \frac{6 \cdot \sum_{i=1}^N d_i^2}{N^3 - N}, \quad (4.97)$$

gdje je:

$N$  - broj parova vrijednosti varijabli  $X_i$  i  $e_i$ ,

$d_i = r(X_i) - r(|e_i|)$  - razlika rangova vrijednosti varijabli  $X_i$  i  $e_i$ .

Svakoj vrijednosti  $X_i$  i  $e_i$  dodjeljuje se rang iskazan prvim  $N$  prirodnim brojevima. Pri tome se rangiranje može započeti rangom 1, počevši od najmanje vrijednosti ili počevši od najveće vrijednosti. Pri tom se rangiranje mora provesti na jednak način i za  $X_i$  i  $e_i$ . Ako se javi **više jednakih vrijednosti u jednom nizu** mora im se dodijeliti jednak rang na način da **se izračuna aritmetička sredina njihovih rangova**.

Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $Z^*$  (Tablica) (ili  $t^*$  za mali uzorak - Tablica A):

- ako je  $\alpha^* > 5\% \Rightarrow H_0$ , tj. donosi se zaključak da Spearmanov koeficijent korelacije ranga između regresorske varijable i apsolutnih reziduala nije statistički značajan, tj. u ocijenjenom regresijskom modelu nije prisutan problem heteroskedastičnosti varijance reziduala.

#### 4.8.1.2 Goldfeld-Quandtov test pri utvrđivanju postojanja problema heteroskedastičnosti varijance reziduala

Kod ovog testiranja potrebno je sve opservacije poredati po veličini vrijednosti varijable  $X_i$ .

Zatim se odvoji regresija prvih  $n_1'$  i posljednjih  $n_2'$  opservacija. (Praksa je pokazala da ako je:  $n = 30 \Rightarrow n_1' = n_2' = 11$  i/ili  $n = 60 \Rightarrow n_1' = n_2' = 22$ , odnosno, poželjno je uzeti nešto više od 1/3 prvih i posljednjih podataka.)

Testiranje se vrši  $F$ -testom. Postavljaju se hipoteze:

$H_0$ .....ne postoji problem heteroskedastičnosti

$H_1$ ..... postoji problem heteroskedastičnosti

$$F^* = \frac{SR_2 / (n_2' - k - 1)}{SR_1 / (n_1' - k - 1)}; \quad F_{tab}^{[\alpha]}[df_1 = (n_1' - k - 1); df_2 = (n_2' - k - 1)], \quad (4.98)$$

gdje je:

$k$  - broj regresorskih varijabli u modelu,

$n_1'$  - prvi dio opservacija,

$n_2'$  - posljednji dio opservacija,

Suma kvadrata reziduala prvih  $n_1'$  opservacija je:

$$SR_1 = \sum_{i=1}^{m_1'} e_i^2 = \sum_{i=1}^{m_1'} (Y_i - \hat{Y}_i)^2. \quad (4.99)$$

Suma kvadrata reziduala posljednjih  $n_2'$  opservacija je:

$$SR_2 = \sum_{i=(n-n_2'+1)}^n e_i^2 = \sum_{i=(n-n_2'+1)}^n (Y_i - \hat{Y}_i)^2. \quad (4.100)$$

Ako je:

$$F^* = \frac{SR_2 / (n_2' - k - 1)}{SR_1 / (n_1' - k - 1)} < F_{tab} \Rightarrow H_0,$$

**prihvaća se početna hipoteza  $H_0$**  tj. zaključuje se da u ocijenjenom regresijskom modelu ne postoji problem heteroskedastičnosti varijance reziduala.



Testiranje se može izvršiti i izračunavanjem granične signifikantnosti  $\alpha^*$  pomoću  $F^*$  (Tablice):

ako je  $\alpha^* < 5\% \Rightarrow H_0$ , što potvrđuje da u ocijenjenom regresijskom modelu ne postoji problem heteroskedastičnosti varijance reziduala.

#### 4.8.1.3 Rješenje problema heteroskedastičnosti varijance reziduala

- 1) Ako je standardna devijacija reziduala ( $\sigma_i$ ) poznata, **svaka opservacija se s njom dijeli**, te tada reziduali u modelu postaju izraženi u obliku:  $(e_i / \sigma_i)$ . U tom slučaju njihova populacijska varijanca je:

$$E\left\{\frac{e_i}{\sigma_i}\right\}^2 = \frac{1}{\sigma_i^2} E(e_i)^2 = \frac{1}{\sigma_i^2} (\sigma_i^2) = 1. \quad (4.101)$$

Dakle, svaka opservacija će imati rezidual iz populacije s varijancom 1, pa je u tom slučaju ispunjen uvjet homoskedastičnosti varijance.

Model:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + e_i, \quad (4.102)$$

dijeljenjem s  $\sigma_i$  postaje:

$$\frac{Y_i}{\sigma_i} = \frac{\beta_0}{\sigma_i} + \beta_1 \cdot \frac{X_i}{\sigma_i} + \frac{e_i}{\sigma_i}, \quad (4.103)$$

odnosno:

$$Y' = \beta_0 v + \beta_1 x' + e', \quad (4.104)$$

gdje je:

$$v_i = \frac{1}{\sigma_i}, \quad e'_i = \frac{e_i}{\sigma_i}. \quad (4.105)$$

Uz pretpostavku da je standardna devijacija proporcionalna svakoj opservaciji rezultati će biti zadovoljavajući. Model je bez slobodnog člana, a regresorske varijable su:  $V$  i  $X'$ , te se efikasne ocjene dobiju s nepristranim

standardnim pogreškama. Može se reći da se problem u stvari svodi na vaganu regresiju, jer je  $\frac{1}{\sigma_i}$  najveća, kada je  $\sigma_i$  najmanja.

- 2) Problem heteroskedastičnosti varijance reziduala može se riješiti i **korištenjem nelinearnih funkcija**.

Na primjer, ako je početni model u nelinearnoj formi sljedeći:

$$Y = \beta_0 \cdot X^{\beta_1} \cdot e, \quad (4.105)$$

gdje reziduali povećavaju ili smanjuju vrijednost regresand varijable u slučajnoj proporciji.

Transformacijom ovog modela u log-linearni oblik dobije se homoskedastična varijanca reziduala.

#### 4.8.2 Problem autokorelacije reziduala

**Slučajna greška**, odnosno **reziduali** u regresijskoj analizi su odstupanja stvarnih vrijednosti regresand varijable  $Y_i$  od ocijenjenog modela  $\hat{Y}_i$ :

$$e_i = (Y_i - \hat{Y}_i). \quad (4.106)$$

Gauss-Markovljevi uvjeti kako je već rečeno odnose na slučajnu grešku i jedan od njih glasi:

- $Cov(e_i, e_j) = 0, \forall i \neq j, (\Rightarrow \text{nema korelacije između varijabli s pomakom od } e_i),$

Promatranjem reziduala ocijenjenog modela može se utvrditi je li taj uvjet ispunjen.

U slučaju da navedeni uvjet nije ispunjen kod ocijenjenog regresijskog modela:

- I. ocjene parametara su nepristrane, ali više nisu efikasne,
- II. ocjene parametara sadrže grešku,
- III. ocjena varijance je potcjenjena i standardne greške parametara su također potcjenjene, što može dovesti po pogrešnog zaključka o značajnosti parametara,

IV. vrijednost F-testa nije valjana.

Autokorelacija reziduala se najčešće javlja pri analizi vremenskih nizova. Ako je interval između mjerenja opsevacija veći, tada je problem manje izražen.

Može se reći da reziduali sadrže utjecaje svih drugih relevantnih varijabli koje nisu sadržane u modelu, a možda su sttistički značajne pri objašnjavanju regresand varijable. U tom slučaju se kaže da je model pogrešno specificiran.

Autokorelacija reziduala prvog reda se može opisati kao Markovljev stohastički proces prvog reda:

$$e_i = \rho \cdot e_{i-1} + u_i, \quad (4.107)$$

gdje je:

$u_i$  - slučajna komponenta tog procesa.

Ako parametar  $\rho$  uz grešku s pomakom nije statistički značajan onda problem autokorelacije reziduala nije prisutan, a ako je taj parametar statistički značajno različit od nule, problem autokorelacije je prisutan, tj. greške s pomakom su korelirane.

Problem se grafički može prikazati dijagramom rasipanja na kojem je  $\hat{e}_{i-1}$  na apscisi, a  $\hat{e}_i$  na ordinati. Prema rasporedu točaka dijagrama rasipanja na klasičan način može se pričati o pozitivnoj i negativnoj korelaciji ili odsutnosti autokorelacije reziduala.

#### **4.8.2.1 Durbin-Watsonov test pri utvrđivanju postojanja problema autokorelacije reziduala**

U praksi je najčešće korišten Durbin-Watsonov test:

$$D-W = d = \frac{\sum_{i=2}^n (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^n \hat{e}_i^2}, \quad (4.108)$$

Tablična vrijednost za  $d_L$  i  $d_U$  ovog testa traži se na temelju broja opservacija  $n$  i broja regresorskih varijabli  $k$ .

Zaključak se na temelju dobivene vrijednosti  $D - W = d$  donosi na sljedeći način:

pozitivna autokorelacija	$d_L$	nije moguće donijeti sud	$d_U$	nema autokorelacije	$(4 - d_U)$	nije moguće donijeti sud	$(4 - d_L)$	negativna autokorelacija
--------------------------	-------	--------------------------	-------	---------------------	-------------	--------------------------	-------------	--------------------------

Vrijednost ovog testa kreće se između 0 i 4. Kada mu je vrijednost oko 2 problem autokorelacije nije prisutan.

Nedostatak ovog testa je što postoji područje inkonkluzivnosti (područje gdje nije moguće donijeti zaključak). Ne može se primijeniti:

- ako matrica  $X$  sadrži vrijednosti varijabli koje su stohastične,
- ako je neka regresorska varijabla ustvari regresand varijabla s vremenskim pomakom
- ako su vrijednosti varijabli u modelu transformirane kao diferencije prvog ili višeg reda.



#### Primjer 4.14.

Na temelju 22 opažanja (u razdoblju od 1983. godine do 2004. godine) ispitivana je ovisnost investicija (u mil. USD) o izvozu (u mil. USD), bruto domaćem proizvodu (u mil. USD). Uz ocjenu parametara **Stepwise metodom** i standardnu regresijsku dijagnostiku potrebno je izvršiti **testiranje problema multikolinearnosti, heteroskedastičnosti varijance reziduala i autokorelacije reziduala** (uz  $\alpha = 1\%$ ).



#### Rješenje 4.14.

Nakon odabira varijabli modela i **Stepwise metode** za ocjenu parametara u **SPSS-u**:

- u **Statistics** potrebno je aktivirati: **Estimates, Confidence Intervals, Covariance matrix, Model fit, Collinearity diagnostics, DW; Continue;**

- u **Plots** potrebno je aktivirati: **Histogram, Normal probability plot; Continue;**

- u **Save** potrebno je aktivirati: **Residuals: Unstandardized; Continue; OK.**

Konačni regresijski model sastoji od dvije regresorske varijable: **bruto domaći proizvod i izvoz**, budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U tablici 4.27 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

Tablica 4.27.

Osnovni podaci o ocijenjenom modelu s investicijama kao zavisnom varijablom

Model Summary <sup>c</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,845 <sup>a</sup>	,714	,700	1,02550	
2	,885 <sup>b</sup>	,783	,760	,91647	1,016

a. Predictors: (Constant), BDP (u mil. USD)  
b. Predictors: (Constant), BDP (u mil. USD), Izvoz (u mil. USD)  
c. Dependent Variable: Investicije (u mil. USD)

Izvor: www.dzs.hr

Pokazatelji multiple korelacije i determinacije modela u tablici 4.26 pokazuju reprezentativnost.

Prema podacima u tablici *Model Summary* može se izvršiti i **testiranje autokorelacije reziduala** pomoću Durbin-Watsonovog testa.  $DW = 1,016$  na temelju čega se prema tablicama testa (E1 i E2) može utvrditi da je:

$$d_L = 0,92, d_U = 1,28, (4 - d_U) = 2,72, (4 - d_L) = 3,09.$$

U skladu s tim:  $d_L < DW < d_U$ ; pa se ne može donijeti sud (inkonkluzivnost) o postojanju autokorelacije rezidualnih odstupanja.

Tablica 4.28.

Tablica ANOVA ocijenjenog regresijskog modela

ANOVA <sup>c</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	52,488	1	52,488	49,910	,000 <sup>a</sup>
	Residual	21,033	20	1,052		
	Total	73,521	21			
2	Regression	57,563	2	28,782	34,267	,000 <sup>b</sup>
	Residual	15,958	19	,840		
	Total	73,521	21			

a. Predictors: (Constant), BDP (u mil. USD)

b. Predictors: (Constant), BDP (u mil. USD), Izvoz (u mil. USD)

c. Dependent Variable: Investicije (u mil. USD)

Izvor: www.dzs.hr

U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenih modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema tablici 4.28 empirijska vrijednost F-testa za konačni model je:  $F_2^* = 34,267$ .

Konačno vrijedi da je  $\alpha_3^* \approx 0 < \alpha = 5\% \Rightarrow H_1$ , pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

**Tablica 4.29.**

**Ocijenjeni linearni regresijski model s investicijama kao zavisnom varijablom**

Coefficients <sup>a</sup>									
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-2,061		-2,643	,016	-3,688	-,434		
	BDP (u mil. USD)	,296	,845	7,065	,000	,208	,383	1,000	1,000
2	(Constant)	-3,086		-3,799	,001	-4,785	-1,386		
	BDP (u mil. USD)	,218	,624	4,461	,000	,116	,321	,585	1,710
	Izvoz (u mil. USD)	,634	,344	2,458	,024	,094	1,174	,585	1,710

a. Dependent Variable: Investicije (u mil. USD)

Izvor: [www.dzs.hr](http://www.dzs.hr)

U tablici 4.29 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

Analitički izraz konačnog modela (2) je:  $\hat{y}_i = -3.086 + 0.218 \cdot X_{1i} + 0.634 \cdot X_{i2}$

Tumačenje regresijskih parametara i u ovom modelu vrši se na standardan način.

Regresijski model u standardiziranom obliku je:  $\hat{y}_i = 0,624 \cdot x_{1i} + 0,344 \cdot x_{i2}$

Može se zaključiti da BDP ima veći relativan utjecaj na investicije od izvoza, jer je uz varijablu BDP veća vrijednost standardiziranog koeficijenta. Naime porast BDP-a za 1 standardnu devijaciju uvjetuje porast investicija za 0,642 standardnih devijacija uz *c.p.* Porast izvoza za 1 standardnu devijaciju uvjetuje porast investicija za 0,344 standardnih devijacija uz *c.p.*

Da bi se izvršilo testiranje značajnosti pojedinačnih parametara modela potrebno je postaviti hipoteze:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Empirijska signifikantnost za parametar  $\hat{\beta}_0$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_1$  je  $\alpha^* \approx 0\%$ .

Empirijska signifikantnost za parametar  $\hat{\beta}_2$  je  $\alpha^* \approx 0\%$ .

Sve empirijske signifikantnosti za regresijske parametre  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  su manje od  $\alpha = 5\%$  i može se zaključiti da su svi parametri statistički značajni.

Potvrdu o nepostojanju **problema kolinearnosti (multikolinearnosti)** daju vrijednosti faktora inflacije varijance ( $VIF_j$ ) i njihove recipročne vrijednosti ( $TOL_j$ ):

$$VIF_1 = 1,71 \Rightarrow VIF_1 < 5, \text{ i } TOL_1 = 0,585 \Rightarrow TOL_1 > 20\%$$

$$VIF_2 = 1,71 \Rightarrow VIF_2 < 5, \text{ i } TOL_2 = 0,585 \Rightarrow TOL_2 > 20\%$$

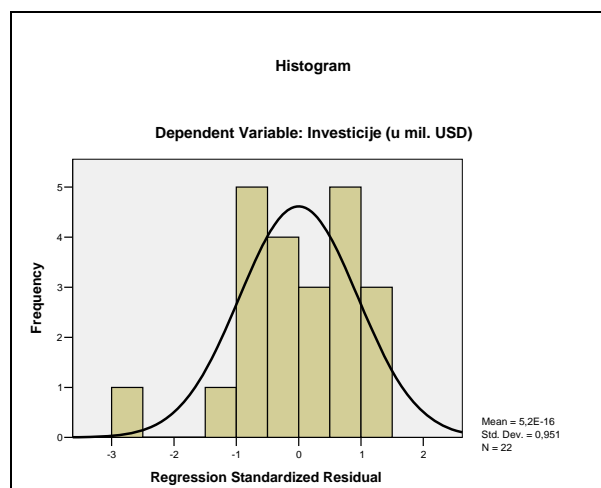
Drugim riječima, za oba parametra regresije (tj.  $\hat{\beta}_1$  i  $\hat{\beta}_2$ ) faktori inflacije varijance su manji od 5, a postotak tolerancije je veći od 20%, što potvrđuje da niti jedna regresorska varijable ne uvjetuje problem multikolinearnosti.

U tablici 4.29 ocijenjenih regresijskih koeficijenata može se vidjeti i intervalna procjena statistički značajnih parametra uz 95% pouzdanosti.

Da bi se utvrdilo **jesu li reziduali normalno distribuirani** potrebno je prikazati histogram standardiziranih reziduala.

**Slika 4.21.**

**Histogram standardiziranih reziduala**



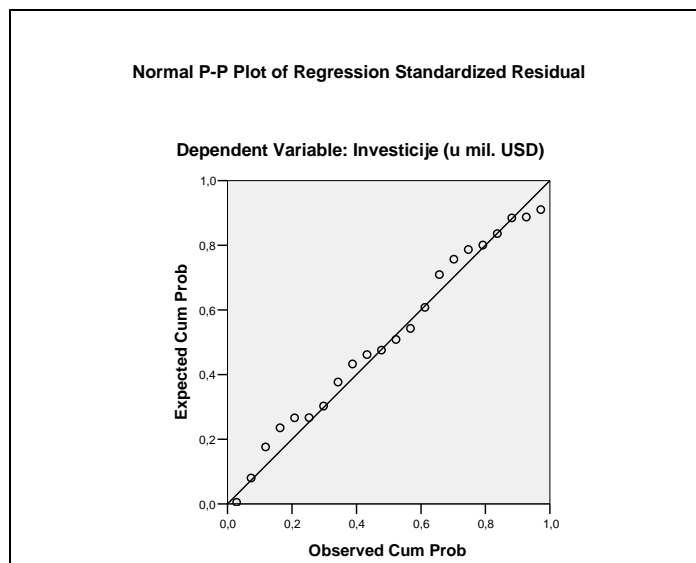
Izvor: [www.dzs.hr](http://www.dzs.hr)

Prema slici 4.21 može se vidjeti da su reziduali  $e_i$  normalno distribuirani s očekivanjem jednakim nuli i standardnom devijacijom približno jednakom jedinici.

Isto se može zaključiti iz grafikona na slici 4.22 na kojem su ucrtane vrijednosti opaženih i očekivanih vjerojatnosti, kada bi reziduali bili normalno distribuirani (*Normal P-P Plot of Regression Standardized Residual*). Pretpostavku o normalnosti reziduala potvrđuje i funkcija distribucije opaženih vjerojatnosti, koja gotovo ne odstupa od očekivane funkcije vjerojatnosti kada bi rezidualna odstupanja bila normalno distribuirana (dijagonalna linija na grafikonu).

**Slika 4.22.**

**P-P grafikon normalno distribuiranih rezidula**



Izvor: [www.dzs.hr](http://www.dzs.hr)

**Problem heteroskedastičnosti varijance reziduala** testira se neparametrisjickim testom i to pomoću Spearmanovog koeficijenta korelacije ranga. Ovaj test se temelji na korelaciji ranga između apsolutnih vrijednosti reziduala i izabranih regresorskih varijabli.

U SPSS-u potrebno je za reziduala (koji su automatskim postupkom regresijske analize formirani u bazi podataka) izračunati njihove apsolutne vrijednosti.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Transform; Compute**, u **Target Variable:** "absres". U **Numeric Expression:** abs(RES\_1); **OK**.

Na navedeni način u bazi podataka formira se niz apsolutnih reziduala. Zatim se računaju Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli BDP-a i izvoza.



Na glavnom izborniku SPSS-a potrebno je odabrati: **Analyze; Correlate; Bivariate**. U **Variables** se prebace: **BDP, Izvoz, absres**. Potrebno je aktivirati koeficijent: **Spearman, OK**.

Da bi se testirala značajnost izračunatih Spearmanovih koeficijenta korelacije postavljaju se hipoteze:

$$H_0: \dots r_S = 0$$

$$H_1: \dots r_S \neq 0$$

**Tablica 4.30.**

**Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli**

Correlations					
			Apsolutni reziduali	BDP (u mil. USD)	Izvoz (u mil. USD)
Spearman's rho	Apsolutni reziduali	Correlation Coefficient	1,000	,149	-,103
		Sig. (2-tailed)	.	,510	,647
		N	22	22	22
	BDP (u mil. USD)	Correlation Coefficient	,149	1,000	,668**
		Sig. (2-tailed)	,510	.	,001
		N	22	22	22
	Izvoz (u mil. USD)	Correlation Coefficient	-,103	,668**	1,000
		Sig. (2-tailed)	,647	,001	.
		N	22	22	22

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Izvor: [www.dzs.hr](http://www.dzs.hr)

Prema rezultatima iz tablice 4.30 može se vidjeti da je empirijska signifikantnost koeficijentata korelacije  $\alpha^*_1 = 0,51$  i  $\alpha^*_2 = 0,647$ , pa se za oba slučaja može zaključiti da je  $\alpha^* > 5\%$  i da se prihvaća početna hipoteza da korelacija nije statistički značajna. To znači da u ocijenjenom modelu ne postoji problem heteroskedastičnosti varijance reziduala.

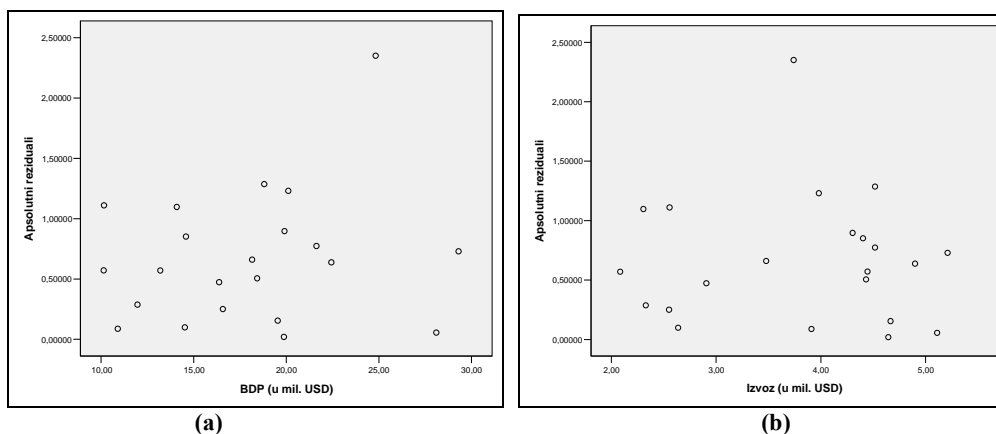
Na kraju su konstruirani dijagrami rasipanja između apsolutnih reziduala i regresorskih varijabli.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Graphs; Legacy Dialogs; Scatter/Dot**, gdje treba aktivirati: **Simple Scatter**. U **Define** se prebace odgovarajuće varijable: u **Y Axes: Apsolutni reziduali**; u **X Axes: BDP, (Izvoz)**; **OK**.

Na temelju dobivenih dijagrama rasipanja, koji su prikazani na slici 4.23 (a) i (b) može se vidjeti da u ocijenjenom regresijskom modelu nije izražen problem heteroskedastičnosti varijance reziduala.

### Slika 4.23.

Dijagrami rasipanja apsolutnih reziduala s regresorskim varijablama BDP (a) i izvoza (b)



Izvor: [www.dzs.hr](http://www.dzs.hr)



### Primjer 4.15

U razdoblju od siječnja 1996. godine do svibnja 2006. godine ispitivana je ovisnost ukupne industrije u RH (2000=100) o ukupnom uvozu, PDV-u i stopi registrirane nezaposlenosti. Uz ocjenu parametara **Stepwise metodom** i standardnu regresijsku dijagnostiku potrebno je izvršiti **testiranje problema multikolinearnosti, heteroskedastičnosti varijance reziduala i autokorelacije reziduala**.



### Rješenje 4.15.

Nakon odabira varijabli modela i **Stepwise metode** za ocjenu parametara u **SPSS-u**:

- u **Statistics** potrebno je aktivirati: **Estimates, Confidence Intervals, Covariance matrix, Model fit, Collinearity diagnostics, DW; Continue;**

- u **Plots** potrebno je aktivirati: **Histogram, Normal probability plot; Continue;**

- u **Save** potrebno je aktivirati: **Residuals: Unstandardized; Continue; OK.**

Konačni regresijski model sastoji od tri regresorske varijable: **ukupni uvoz, PDV i stopa registrirane nezaposlenosti** budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U tablici 4.31 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

Tablica 4.32.

## Osnovni podaci o ocijenjenom modelu s industrijom kao zavisnom varijablom

Model Summary <sup>d</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,940 <sup>a</sup>	,883	,882	4,38062	
2	,954 <sup>b</sup>	,910	,909	3,85206	
3	,958 <sup>c</sup>	,918	,916	3,69707	1,727

a. Predictors: (Constant), Uvoz uk.  
b. Predictors: (Constant), Uvoz uk., PDV  
c. Predictors: (Constant), Uvoz uk., PDV, Stopa reg. nez.  
d. Dependent Variable: Ind. ukupno (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Pokazatelji multiple korelacije i determinacije modela u tablici 4.31 pokazuju reprezentativnost.

Prema podacima u tablici **Model Summary** može se izvršiti testiranje autokorelacije reziduala pomoću Durbin-Watsonovog testa.  $DW = 1,721$  na temelju čega se prema tablicama testa može utvrditi da je za  $\alpha = 5\%$ :

$$d_L = 1,693, d_U = 1,774, (4 - d_U) = 2,226, (4 - d_L) = 2,307.$$

U skladu s tim:  $d_L < DW < d_U$ ; pa se ne može donijeti sud (inkonkluzivnost) o postojanju autokorelacije rezidualnih odstupanja.

$$\text{Za } \alpha = 1\%: d_L = 1,482, d_U = 1,604, (4 - d_U) = 2,396, (4 - d_L) = 2,518.$$

U skladu s tim:  $d_U < DW < (4 - d_U)$ ; pa se može zaključiti da u ocijenjenom modelu nije prisutan problem autokorelacije rezidualnih odstupanja.

U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenih modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema rezultatima u output-u vrijedi da je  $\alpha_3^* \approx 0 < \alpha = 5\% \Rightarrow H_1$ , pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

U tablici 4.33 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

Analitički izraz konačnog modela (2) je:

$$\hat{y}_i = 75,692 + 0,0000042 \cdot X_{1i} + 0,00000761 \cdot X_{2i} - 0,523 \cdot X_{3i}$$

Tumačenje regresijskih parametara i u ovom modelu vrši se na standardan način.

Regresijski model u standardiziranom obliku je:  $\hat{y}_i = 0,678 \cdot x_{1i} + 0,306 \cdot x_{2i} - 0,089 \cdot x_{3i}$ .

**Tablica 4.33.**

**Ocijenjeni linearni regresijski model s industrijom kao zavisnom varijablom**

Coefficients <sup>a</sup>									
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	70,218	1,407	49,909	,000	67,430	73,006		
	Uvoz uk.	5,895E-6	,000	28,937	,000	,000	,000	1,000	1,000
2	(Constant)	65,330	1,497	43,626	,000	62,363	68,298		
	Uvoz uk.	4,269E-6	,000	12,819	,000	,000	,000	,289	3,456
	PDV	7,555E-6	,000	5,792	,000	,000	,000	,289	3,456
3	(Constant)	75,920	3,582	21,195	,000	68,820	83,019		
	Uvoz uk.	4,253E-6	,000	13,307	,000	,000	,000	,289	3,456
	PDV	7,528E-6	,000	6,014	,000	,000	,000	,289	3,456
	Stopa reg. nez.	-,543	,168	-3,228	,002	-,877	-,210	,999	1,001

a. Dependent Variable: Ind. ukupno (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Sve empirijske signifikantnosti za regresijske parametre  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$  su manje od  $\alpha = 5\%$  i može se zaključiti da su svi parametri statistički značajni.

Potvrdu o nepostojanju **problema kolinearnosti (multikolinearnosti)** daju vrijednosti faktora inflacije varijance ( $VIF_j$ ) i njihove recipročne vrijednosti ( $TOL_j$ ):

$$VIF_1 = 3,4 \Rightarrow VIF_1 < 5, \text{ i } TOL_1 = 0,294 \Rightarrow TOL_1 > 20\%$$

$$VIF_2 = 3,401 \Rightarrow VIF_2 < 5, \text{ i } TOL_2 = 0,294 \Rightarrow TOL_2 > 20\%$$

$$VIF_3 \approx 1,000 \Rightarrow VIF_3 < 5, \text{ i } TOL_3 \approx 1,000 \Rightarrow TOL_3 > 20\%.$$

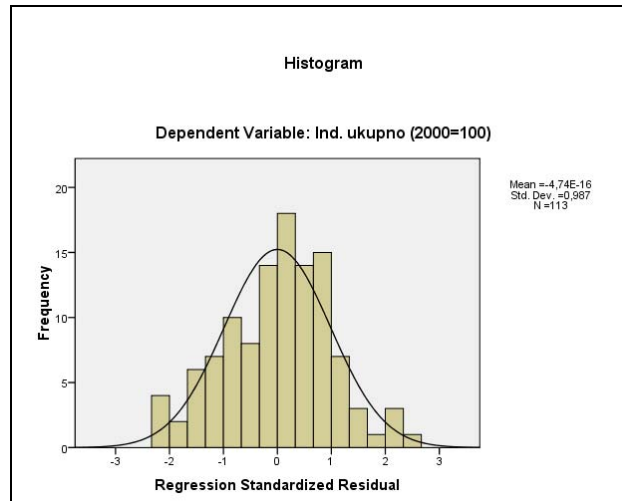
Drugim riječima, za sva 3 parametra regresije (tj.  $\hat{\beta}_1, \hat{\beta}_2$  i  $\hat{\beta}_3$ ) faktori inflacije varijance su manji od 5, a postotak tolerancije je veći od 20%, što potvrđuje da niti jedna regresorska varijable ne uvjetuje problem multikolinearnosti.

Da bi se utvrdilo **jesu li reziduali normalno distribuirani** potrebno je prikazati histogram standardiziranih reziduala.

Prema slici 4.24 histograma može se vidjeti da su reziduali  $e_i$  normalno distribuirani s očekivanjem jednakim nuli i standardnom devijacijom približno jednakom jedinici.

Slika 4.24.

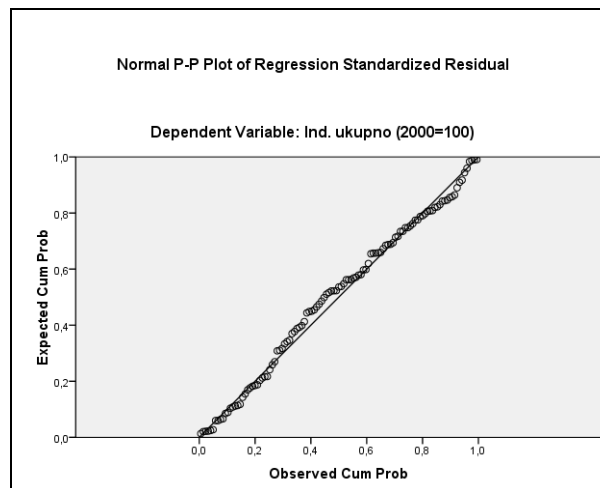
### Histogram standardiziranih reziduala



Izvor: <http://epp.eurostat.ec.europa.eu>.

Slika 4.25.

### P-P grafikon normalno distribuiranih rezidula



Izvor: <http://epp.eurostat.ec.europa.eu>.

Isto se može zaključiti iz grafikona na kojem su ucrtane vrijednosti opaženih i očekivanih vjerojatnosti, kada bi reziduali bili normalno distribuirani (**Normal P-P Plot of Regression Standardized Residual**). Pretpostavku o normalnosti reziduala

povrđuje i funkcija distribucije opaženih vjerojatnosti, koja gotovo ne odstupa od očekivane funkcije vjerojatnosti kada bi rezidualna odstupanja bila normalno distribuirana (dijagonalna linija na grafikonu).

**Problem heteroskedastičnosti varijance reziduala** testira se neparametrisjistikim testom i to pomoću Spearmanovog koeficijenta korelacije ranga.

U SPSS-u potrebno je za rezidualne (koji su automatskim postupkom regresijske analize formirani u bazi podataka) izračunati njihove apsolutne vrijednosti.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Transform; Compute**, u **Target Variable:** "absres". U **Numeric Expression:** abs(RES\_1); **OK**.

Na navedeni način u bazi podataka formira se niz apsolutnih reziduala. Zatim se računaju Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli ukupnog uvoza, PDV, stope reg. nezaposlenosti.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Analyze; Correlate; Bivariate**. U **Variables** se prebace: **absres, ukupni uvoz, PDV, stopa reg. nezaposlenosti**. Potrebno je aktivirati koeficijent: **Spearman**, **OK**.

**Tablica 4.34.**

**Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli**

Correlations						
Spearman's rho	Apsolutni reziduali	Correlation Coefficient	Apsolutni reziduali	Uvoz uk.	PDV	Stopa reg. nez.
		Sig. (2-tailed)	1,000	,051	,028	,098
		N	113	,594	,767	,302
	Uvoz uk.	Correlation Coefficient		1,000	,870**	,044
		Sig. (2-tailed)		,594	,000	,646
		N		113	125	113
	PDV	Correlation Coefficient		,028	1,000	,018
		Sig. (2-tailed)		,767	,000	,849
		N		113	125	113
	Stopa reg. nez.	Correlation Coefficient		,098	,044	1,000
		Sig. (2-tailed)		,302	,646	,849
		N		113	113	113

\*\*. Correlation is significant at the 0.01 level (2-tailed).

Izvor: <http://epp.eurostat.ec.europa.eu>.

Da bi se testirala značajnost izračunatih Spearmanovih koeficijenta korelacije postavljaju se hipoteze:

$$H_0: \dots r_S = 0$$

$$H_1: \dots r_S \neq 0$$

Prema rezultatima iz tablice outputa može se vidjeti da je empirijska signifikantnost koeficijenata korelacije  $\alpha^*_1 = 0,594$ ,  $\alpha^*_2 = 0,767$  i  $\alpha^*_3 = 0,302$  pa se za sve slučajeve

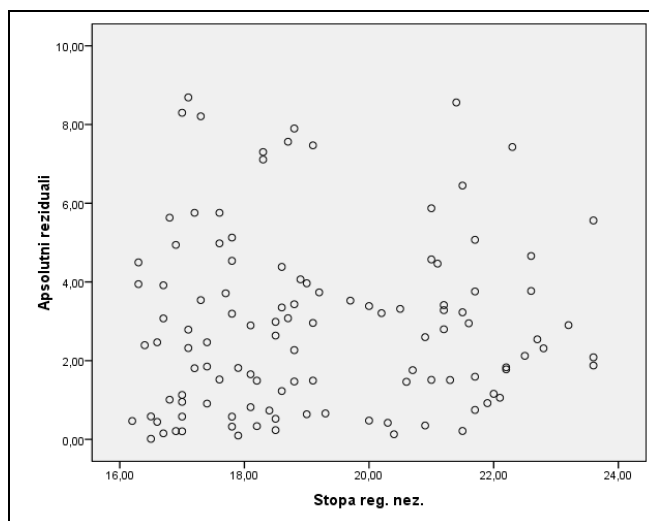
može zaključiti da je  $\alpha^* > 5\%$  i da se prihvaća početna hipoteza da korelacija nije statistički značajna. To znači da u ocijenjenom modelu ne postoji problem heteroskedastičnosti varijance reziduala.

Na kraju su konstruirani dijagrami rasipanja između apsolutnih reziduala i regresorskih varijabli.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Graphs; Legacy Dialogs; Scatter/Dot**, gdje treba aktivirati: **Simple Scatter**. U **Define** se prebace odgovarajuće varijable: u **Y Axes: Apsolutni reziduali**; u **X Axes: ukupni uvoz, (PDV), (stopa reg. nezaposlenosti)**; OK.

Slika 4.26.

**Dijagram rasipanja apsolutnih reziduala s regresorskom varijablom stopa reg. nezaposlenosti**



Izvor: <http://epp.eurostat.ec.europa.eu>.

Na temelju dobivenih dijagrama rasipanja u output-u **SPSS**-a, od kojih je onaj s regresorskom varijablom stopa registrirane nezaposlenosti prikazan na slici 4.26, može se vidjeti da u ocijenjenom regresijskom modelu nije izražen problem heteroskedastičnosti varijance reziduala.



#### Primjer 4.16.

U razdoblju od siječnja 1996. godine do svibnja 2006. godine ispitivana je ovisnost  $I_t$  izvoza u RH (2000=100) o ukupnoj industriji (2000=100), PPI indeksu cijena

(2000=100) i  $I_t$  zaposlenih u pravnim osobama (2000=100). Uz ocjenu parametara **Stepwise metodom** i standardnu regresijsku dijagnostiku potrebno je izvršiti **testiranje problema multikolinearnosti, heteroskedastičnosti varijance reziduala i autokorelacije reziduala**.



#### Rješenje 4.16.

Nakon odabira varijabli modela i **Stepwise metode** za ocjenu parametara u **SPSS-u**:

- u **Statistics** potrebno je aktivirati: **Estimates, Confidence Intervals, Covariance matrix, Model fit, Collinearity diagnostics, DW; Continue;**

- u **Plots** potrebno je aktivirati: **Histogram, Normal probability plot; Continue;**

- u **Save** potrebno je aktivirati: **Residuals: Unstandardized; Continue; OK.**

1. **Konačni regresijski model** se sastoji od tri regresorske varijable:  **$I_t$  ukupna industrija, PPI cijena i  $I_t$  zaposlenost u pravnim osobama** budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

Analitički izraz ovog modela je:

$$\hat{y}_i = -240,143 + 0,819 \cdot X_{1i} + 1,362 \cdot X_{2i} - 1,160 \cdot X_{3i}$$

Po svim karakteristikama ovaj model zadovoljava uvjete reprezentativnosti, ali je utvrđen **problem heteroskedastičnosti varijance reziduala**. Spearmanovi koeficijenti korelacije ranga između apsolutnih vrijednosti reziduala i izabranih regresorskih varijabli su statistički značajni.

Ovaj test se temelji na korelaciji ranga između apsolutnih vrijednosti reziduala i izabranih regresorskih varijabli.

U SPSS-u potrebno je za reziduala (koji su automatskim postupkom regresijske analize formirani u bazi podataka) izračunati njihove apsolutne vrijednosti.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Transform; Compute**, u **Target Variable: "absres1"**. U **Numeric Expression: abs(RES\_1); OK.**

Na navedeni način u bazi podataka formira se niz apsolutnih reziduala. Zatim se računaju Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli ukupne industrije (2000=100), PPI indeksa cijena (2000=100) i  $I_t$  zaposlenih u pravnim osobama (2000=100).



Na glavnom izborniku SPSS-a potrebno je odabrati: **Analyze; Correlate; Bivariate**. U **Variables** se prebace: **absres, ukupna industrija (2000=100), PPI indeks cijena (2000=100) i I<sub>t</sub> zaposlenih u pravnim osobama (2000=100)**. Potrebno je aktivirati koeficijent: **Spearman, OK**.

Da bi se testirala značajnost izračunatih Spearmanovih koeficijenta korelacije postavljaju se hipoteze:

$$H_0 \dots r_S = 0$$

$$H_1 \dots r_S \neq 0$$

**Tablica 4.35.**

**Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli**

Correlations						
			Apsolutni reziduali 1	Ind. ukupno (2000=100)	PPI cijene (2000=100)	I <sub>t</sub> zaposleni (2000=100)
Spearman's rho	Apsolutni reziduali 1	Correlation Coefficient	1,000	,181*	,175	,204*
		Sig. (2-tailed)		,050	,059	,027
		N	118	118	118	118
	Ind. ukupno (2000=100)	Correlation Coefficient	,181*	1,000	,802**	,791**
		Sig. (2-tailed)	,050		,000	,000
		N	118	125	118	125
	PPI cijene (2000=100)	Correlation Coefficient	,175	,802**	1,000	,676**
		Sig. (2-tailed)	,059	,000		,000
		N	118	118	118	118
	I <sub>t</sub> zaposleni (2000=100)	Correlation Coefficient	,204*	,791**	,676**	1,000
		Sig. (2-tailed)	,027	,000	,000	
		N	118	125	118	125

\*, Correlation is significant at the 0.05 level (2-tailed).

\*\*, Correlation is significant at the 0.01 level (2-tailed).

Izvor: <http://epp.eurostat.ec.europa.eu>.

Prema rezultatima iz tablice 4.35 može se vidjeti da je empirijska signifikantnost koeficijenata korelacije  $\alpha^*_1 = 0,05$ ,  $\alpha^*_2 = 0,059$  i  $\alpha^*_3 = 0,027$ , pa se za prvu i treću varijablu može zaključiti da je  $\alpha^* < 5\%$  i da se prihvaća alternativna hipoteza da je korelacija statistički značajna. To znači da u ocijenjenom modelu postoji problem heteroskedastičnosti varijance reziduala.

**2.** Ako se iz analize izbaci varijabla ***I<sub>t</sub> zaposlenost u pravnim osobama*** koja je **Stepwise metodom** posljednja ušla u model i ponovi procedura dobije se konačni regresijski model sa dvije regresorske varijable: *I<sub>t</sub> ukupna industrija, PPI cijena* budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

Analitički izraz ovog modela je:

$$\hat{y}_i = -154,184 + 0,994 \cdot X_{1i} + 1,487 \cdot X_{2i}$$

Po svim karakteristikama ovaj model zadovoljava uvjete reprezentativnosti, ali je utvrđen **problem heteroskedastičnosti varijance reziduala**. Spearmanovi koeficijenti korelacije ranga između apsolutnih vrijednosti reziduala i izabranih regresorskih varijabli su statistički značajni. Nakon provedene procedure u **SPSS**-u za izračun apsolutnih reziduala te izračunavanja Spearmanovih koeficijenata korelacije ranga između apsolutnih reziduala i regresorskih varijabli ukupne industrije (2000=100), PPI indeksa cijena (2000=100), dobiveni su rezultati.

Da bi se testirala značajnost izračunatih Spearmanovih koeficijenata korelacije postavljaju se hipoteze:

$$H_0 \dots r_S = 0$$

$$H_1 \dots r_S \neq 0$$

**Tablica 4.36.**

**Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli**

Correlations					
			Apsolutni reziduali 2	Ind. ukupno (2000=100)	PPI cijene (2000=100)
Spearman's rho	Apsolutni reziduali 2	Correlation Coefficient	1,000	,188*	,190*
		Sig. (2-tailed)		,041	,039
		N	118	118	118
	Ind. ukupno (2000=100)	Correlation Coefficient	,188*	1,000	,802**
		Sig. (2-tailed)	,041		,000
		N	118	125	118
	PPI cijene (2000=100)	Correlation Coefficient	,190*	,802**	1,000
		Sig. (2-tailed)	,039	,000	
		N	118	118	118

\*. Correlation is significant at the 0.05 level (2-tailed).

\*\*. Correlation is significant at the 0.01 level (2-tailed).

Izvor: <http://epp.eurostat.ec.europa.eu>.

Prema rezultatima iz tablice 4.36 može se vidjeti da je empirijska signifikantnost koeficijenata korelacije  $\alpha^*_1 = 0,041$  i  $\alpha^*_2 = 0,039$  pa se za obje varijable može zaključiti da je  $\alpha^* < 5\%$  i da se prihvaća alternativna hipoteza da je korelacija statistički značajna. To znači da u ocijenjenom modelu postoji problem heteroskedastičnosti varijance reziduala.

**3.** Ako se iz analize osim varijable ***I<sub>t</sub> zaposlenost u pravnim osobama*** izbaci i varijabla ***PPI cijena*** i ponovi procedura dobije se konačni regresijski model sa jednom regresorskom varijablom: *I<sub>t</sub> ukupna industrija* budući da je ona zadovoljila kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

Analitički izraz ovog modela je:

$$\hat{y}_i = -104,120 + 1,917 \cdot X_{1i}$$

Po svim karakteristikama ovaj model zadovoljava uvjete reprezentativnosti, ali je utvrđen **problem autokorelacije reziduala**.

U tablici 4.37 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

**Tablica 4.37.**

**Osnovni podaci o ocijenjenom modelu s izvozom kao zavisnom varijablom**

Model Summary <sup>b</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,884 <sup>a</sup>	,781	,780	13,70130026	1,252
a. Predictors: (Constant), Ind. ukupno (2000=100)					
b. Dependent Variable: It izvoz (2000=100)					

Izvor: <http://epp.eurostat.ec.europa.eu>.

Pokazatelji multiple korelacije i determinacije modela u tablici 4.37 pokazuju reprezentativnost.

Prema podacima u tablici **Model Summary** može se izvršiti i **testiranje autokorelacije reziduala** pomoću Durbin-Watsonovog testa.  $DW = 1,252$  na temelju čega se prema tablicama testa (E1 i E2) može utvrditi da je:

$$d_L = 1,654, d_U = 1,694, (4 - d_U) = 2,306, (4 - d_L) = 2,346.$$

U skladu s tim:  $DW < d_L$ ; pa se uz signifikantnost testa od 5% može donijeti sud o postojanju pozitivne autokorelacije reziduala.

**4. Ako se iz analize iz početnog modela (1.) izbaci samo varijabla *PPI cijena* i ponovi procedura dobije se konačni regresijski model sa dvije regresorske varijable: *I<sub>t</sub> ukupna industrija* i *I<sub>t</sub> zaposlenost u pravnim osobama* budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).**

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

Analitički izraz ovog modela je:

$$\hat{y}_i = -260,521 + 1,505 \cdot X_{1i} + 1,987 \cdot X_{2i}$$

Po svim karakteristikama ovaj model zadovoljava uvjete reprezentativnosti, ali je utvrđen **problem autokorelacije reziduala**.

U tablici 4.38 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu. Pokazatelji multiple korelacije i determinacije modela pokazuju reprezentativnost.

Tablica 4.38.

## Osnovni podaci o ocijenjenom modelu s izvozom kao zavisnom varijablom

Model Summary <sup>c</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,884 <sup>a</sup>	,781	,780	13,70130026	
2	,896 <sup>b</sup>	,803	,800	13,05333965	1,334

a. Predictors: (Constant), Ind. ukupno (2000=100)  
b. Predictors: (Constant), Ind. ukupno (2000=100), It zaposleni (2000=100)  
c. Dependent Variable: It izvoz (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Prema podacima u tablici **Model Summary** može se izvršiti testiranje autokorelacije reziduala pomoću Durbin-Watsonovog testa.  $DW = 1,334$  na temelju čega se prema tablicama testa (E1 i E2) može utvrditi da je:

$$d_L = 1,634, d_U = 1,715, (4 - d_U) = 2,285, (4 - d_L) = 2,366.$$

U skladu s tim:  $DW < d_L$ ; pa se uz signifikantnost testa od 5% može donijeti sud o postojanju pozitivne autokorelacije reziduala.

5. Ako se iz analize iz **početnog modela (1.)** izbacila varijabla ***I<sub>t</sub> ukupna industrija*** i ponovi procedura dobije se konačni regresijski model sa dvije regresorske varijable: ***PPI cijena*** i ***I<sub>t</sub> zaposlenost u pravnim osobama*** budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U tablici 4.39 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

Tablica 4.39.

## Osnovni podaci o ocijenjenom modelu s izvozom kao zavisnom varijablom

Model Summary <sup>c</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,871 <sup>a</sup>	,758	,756	12,97338328	
2	,897 <sup>b</sup>	,804	,800	11,72768318	2,412

a. Predictors: (Constant), PPI cijene (2000=100)  
b. Predictors: (Constant), PPI cijene (2000=100), It zaposleni (2000=100)  
c. Dependent Variable: It izvoz (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Pokazatelji multiple korelacije i determinacije modela u tablici 4.39 pokazuju reprezentativnost.

Prema podacima u tablici **Model Summary** može se izvršiti **testiranje autokorelacije reziduala** pomoću Durbin-Watsonovog testa.  $DW = 2,412$  na temelju čega se prema tablicama testa (E1 i E2) može utvrditi da je za  $\alpha = 5\%$ :

$$d_L = 1,634, d_U = 1,715, (4 - d_U) = 2,285, (4 - d_L) = 2,366.$$

U skladu s tim:  $DW > (4 - d_L)$ ; pa se ne može donijeti sud o postojanju negativne autokorelacije rezidualnih odstupanja.

$$\text{Za } \alpha = 1\%: d_L = 1,502, d_U = 1,582, (4 - d_U) = 2,418, (4 - d_L) = 2,499.$$

U skladu s tim:  $(4 - d_L) < DW < (4 - d_U)$ ; pa se ne može zaključiti o prisutnosti problema autokorelacije rezidualnih odstupanja u ocijenjenom modelu jer  $DW$  vrijednost upada u interval područja inkonkluzivnosti.

U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenih modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema rezultatima u output-u vrijedi da je  $\alpha_3^* \approx 0 < \alpha = 5\% \Rightarrow H_1$ , pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

U tablici 4.40 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

**Tablica 4.40.**

**Ocijenjeni linearni regresijski model s izvozom kao zavisnom varijablom**

Coefficients <sup>a</sup>									
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance VIF
1	(Constant)	-168,237	14,035		-11,987	,000	-196,034	-140,440	
	PPI cijene (2000=100)	2,696	,141	,871	19,058	,000	2,416	2,976	1,000 1,000
2	(Constant)	-336,074	34,729		-9,677	,000	-404,866	-267,281	
	PPI cijene (2000=100)	2,019	,183	,652	11,058	,000	1,657	2,381	,490 2,039
	It zaposleni (2000=100)	2,332	,449	,306	5,191	,000	1,442	3,222	,490 2,039

a. Dependent Variable: It izvoz (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Analitički izraz konačnog modela (2) je:

$$\hat{y}_i = -336,074 + 2,019 \cdot X_{1i} + 2,332 \cdot X_{2i}$$

Tumačenje regresijskih parametara i u ovom modelu vrši se na standardan način.

Regresijski model u standardiziranom obliku je:  $\hat{y}_i = 0,652 \cdot x_{i1} + 0,306 \cdot x_{i2}$ .

Sve empirijske signifikantnosti za regresijske parametre  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  su manje od  $\alpha = 5\%$  i može se zaključiti da su svi parametri statistički značajni.

Potvrdu o nepostojanju **problema kolinearnosti (multikolinearnosti)** daju vrijednosti faktora inflacije varijance ( $VIF_j$ ) i njihove recipročne vrijednosti ( $TOL_j$ ):

$$VIF_1 = 2,039 \Rightarrow VIF_1 < 5$$

$$VIF_2 = 2,039 \Rightarrow VIF_2 < 5.$$

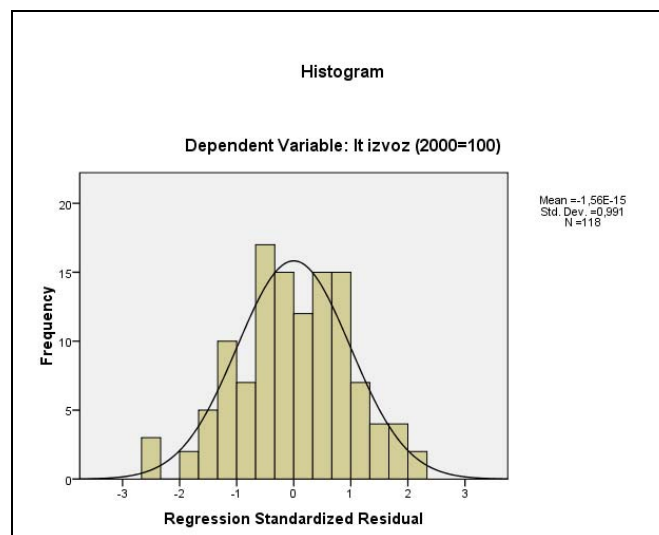
Drugim riječima, za oba parametra regresije (tj.  $\hat{\beta}_1$  i  $\hat{\beta}_2$ ) faktori inflacije varijance su manji od 5, (a postotak tolerancije je veći od 20%,) što potvrđuje da niti jedna regresorska varijabla ne uvjetuje problem multikolinearnosti.

Da bi se utvrdilo **jesu li reziduali normalno distribuirani** potrebno je prikazati histogram standardiziranih reziduala.

Prema slici 4.28 histograma u outputu može se vidjeti da su reziduali  $e_i$  normalno distribuirani s očekivanjem jednakim nuli i standardnom devijacijom približno jednakom jedinici.

**Slika 4.28.**

**Histogram standardiziranih reziduala**

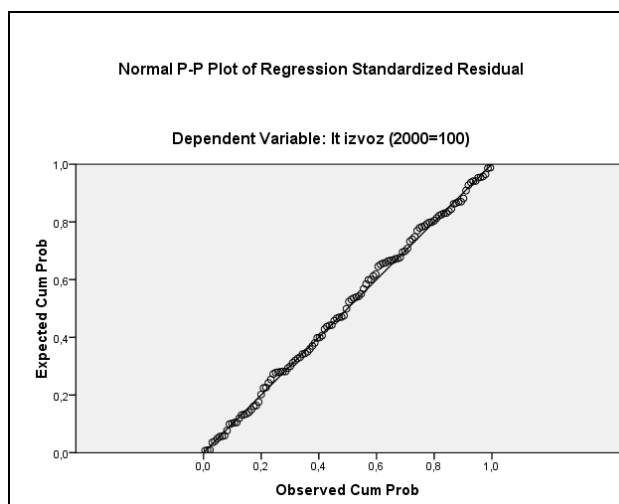


Izvor: <http://epp.eurostat.ec.europa.eu>.

Isto se može zaključiti iz grafikona na slici 4.29 na kojem su ucrtane vrijednosti opaženih i očekivanih vjerojatnosti, kada bi reziduali bili normalno distribuirani (*Normal P-P Plot of Regression Standardized Residual*). Pretpostavku o normalnosti reziduala potvrđuje i funkcija distribucije opaženih vjerojatnosti, koja gotovo ne odstupa od očekivane funkcije vjerojatnosti kada bi rezidualna odstupanja bila normalno distribuirana (dijagonalna linija na grafikonu).

Slika 4.29.

#### P-P grafikon normalno distribuiranih rezidua



Izvor: <http://epp.eurostat.ec.europa.eu>.

**Problem heteroskedastičnosti varijance reziduala** testira se neparametrisjickim testom i to pomoću Spearmanovog koeficijenta korelacije ranga.

U SPSS-u potrebno je za rezidualne (koji su automatskim postupkom regresijske analize formirani u bazi podataka) izračunati njihove apsolutne vrijednosti.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Transform; Compute**, u **Target Variable:** "absres5". U **Numeric Expression:** abs(RES\_5); **OK**.

Na navedeni način u bazi podataka formira se niz apsolutnih reziduala (za 5. ocijenjeni model). Zatim se računaju Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli PPI cijena i  $I_t$  zaposlenih.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Analyze; Correlate; Bivariate**. U **Variables** se prebace: **absres5, PPI cijena i  $I_t$  zaposlenih**. Potrebno je aktivirati koeficijent: **Spearman, OK**.

**Tablica 4.41.**

**Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorskih varijabli**

Correlations			Apsolutni reziduali 5	PPI cijene (2000=100)	It zaposleni (2000=100)
Spearman's rho	Apsolutni reziduali 5	Correlation Coefficient	1,000	,176	,119
		Sig. (2-tailed)		,057	,198
		N	118	118	118
	PPI cijene (2000=100)	Correlation Coefficient	,176	1,000	,676**
		Sig. (2-tailed)	,057		,000
		N	118	118	118
	It zaposleni (2000=100)	Correlation Coefficient	,119	,676**	1,000
		Sig. (2-tailed)	,198	,000	
		N	118	118	125

\*\*. Correlation is significant at the 0.01 level (2-tailed).

Izvor: <http://epp.eurostat.ec.europa.eu>.

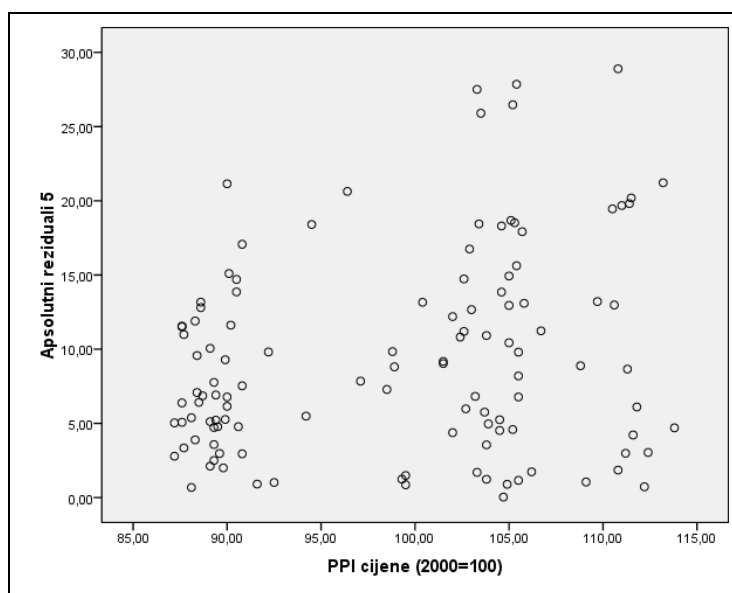
Da bi se testirala značajnost izračunatih Spearmanovih koeficijenta korelacije postavljaju se hipoteze:

$$H_0 \dots r_S = 0$$

$$H_1 \dots r_S \neq 0$$

**Slika 4.30.**

**Dijagram rasipanja apsolutnih reziduala s regresorskom varijablom PPI cijena**



Izvor: <http://epp.eurostat.ec.europa.eu>.



Prema rezultatima iz tablice outputa može se vidjeti da je empirijska signifikantnost koeficijenata korelacije  $\alpha^*_1 = 0,057$  i  $\alpha^*_2 = 0,198$  pa se za sve slučajeve može zaključiti da je  $\alpha^* > 5\%$  i da se prihvaća početna hipoteza da korelacija nije statistički značajna. To znači da u ocijenjenom modelu ne postoji problem heteroskedastičnosti varijance reziduala.

Na kraju su konstruirani dijagrami rasipanja između apsolutnih reziduala i regresorskih varijabli.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Graphs; Legacy Dialogs; Scatter/Dot**, gdje treba aktivirati: **Simple Scatter**. U **Define** se prebace odgovarajuće varijable: u **Y Axes: Apsolutni reziduali 5**; u **X Axes: PPI cijena (I<sub>t</sub> zaposlenost u pravnim osobama)**; **OK**.

Na temelju dobivenih dijagrama rasipanja u output-u **SPSS**-a, od kojih je onaj s regresorskom varijablom PPI cijena (2000=100) prikazan na slici 4.30, može se vidjeti da u ocijenjenom regresijskom modelu nije izražen problem heteroskedastičnosti varijance reziduala.



#### Primjer 4.17.

U razdoblju od siječnja 1996. godine do svibnja 2006. godine ispitivana je ovisnost PPI cijena u RH (2000=100) o kamatnim stopama na kun. kred. s val. klauz. u %, i I<sub>t</sub> izvoza (2000=100). Uz ocjenu parametara **Stepwise metodom** i standardnu regresijsku dijagnostiku potrebno je izvršiti **testiranje problema multikolinearnosti, heteroskedastičnosti varijance reziduala i autokorelacije reziduala**.



#### Rješenje 4.17.

Nakon odabira varijabli modela i **Stepwise metode** za ocjenu parametara u **SPSS-u**:

- u **Statistics** potrebno je aktivirati: **Estimates, Confidence Intervals, Covariance matrix, Model fit, Collinearity diagnostics, DW**; **Continue**;

- u **Plots** potrebno je aktivirati: **Histogram, Normal probability plot**; **Continue**;

- u **Save** potrebno je aktivirati: **Residuals: Unstandardized**; **Continue**; **OK**.

1. Konačni regresijski model se sastoji od dvije regresorske varijable: kamatne stope na kun. kred. s val. klauz. u %, i I<sub>t</sub> Izvoz budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

Analitički izraz ovog modela je:

$$\hat{y}_i = 95,417 - 0,934 \cdot X_{1i} + 0,149 \cdot X_{2i}$$

Po svim karakteristikama ovaj model zadovoljava uvjete reprezentativnosti, ali je utvrđen **problem autokorelacije reziduala**.

U tablici 4.42 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

**Tablica 4.42.**

**Osnovni podaci o ocijenjenom modelu s PPI kao zavisnom varijablom**

Model Summary <sup>c</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,883 <sup>a</sup>	,779	,777	4,00242	
2	,926 <sup>b</sup>	,858	,856	3,21985	,834

a. Predictors: (Constant), Kam. st. na kun. kred. s val. klauz. u %  
b. Predictors: (Constant), Kam. st. na kun. kred. s val. klauz. u %, It izvoz (2000=100)  
c. Dependent Variable: PPI cijene (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Pokazatelji multiple korelacije i determinacije modela u tablici 4.42 pokazuju reprezentativnost.

Prema podacima u tablici **Model Summary** može se izvršiti i **testiranje autokorelacije reziduala** pomoću Durbin-Watsonovog testa.  $DW = 0,834$  na temelju čega se prema tablicama testa (E1 i E2) može utvrditi da je:

$$d_L = 1,634, d_U = 1,715, (4 - d_U) = 2,285, (4 - d_L) = 2,366.$$

U skladu s tim:  $DW < d_L$ ; pa se uz signifikantnost testa od 5% može donijeti sud o postojanju pozitivne autokorelacije reziduala.

**2.** Ako se iz analize iz početnog modela izbaci varijabla **I, Izvoz konačni regresijski model se sastoji od jedne regresorske varijable: kamatne stope na kun. kred. s val. klauz. u %**, budući da je ona zadovoljila kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

Analitički izraz ovog modela je:

$$\hat{y}_i = 117,88 - 1,591 \cdot X_{1i}$$

Po svim karakteristikama ovaj model zadovoljava uvjete reprezentativnosti, ali je utvrđen **problem autokorelacije reziduala**.

U tablici 4.43 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

Tablica 4.43.

## Osnovni podaci o ocijenjenom modelu s PPI kao zavisnom varijablom

Model Summary <sup>b</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,883 <sup>a</sup>	,779	,777	4,00242	,109

a. Predictors: (Constant), Kam. st. na kun. kred. s val. klauz. u %  
b. Dependent Variable: PPI cijene (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Pokazatelji multiple korelacije i determinacije modela u tablici 4.43 pokazuju reprezentativnost.

Prema podacima u tablici *Model Summary* može se izvršiti i testiranje autokorelacije reziduala pomoću Durbin-Watsonovog testa.  $DW = 0,109$  na temelju čega se prema tablicama testa (E1 i E2) može utvrditi da je:

$$d_L = 1,654, d_U = 1,694, (4 - d_U) = 2,306, (4 - d_L) = 2,346.$$

U skladu s tim:  $DW < d_L$ ; pa se uz signifikantnost testa od 5% može donijeti sud o postojanju pozitivne autokorelacije reziduala.

3. Ako se iz analize iz početnog modela izbací varijabla *kamatne stope na kun. kred. s val. klauz. u % konačni regresijski model se sastoji od jedne regresorske varijable:  $I_t$  Izvoz*, budući da je ona zadovoljila kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

U tablici 4.44 izračunati su osnovni pokazatelji o ocijenjenom višestrukom modelu.

Tablica 4.44.

## Osnovni podaci o ocijenjenom modelu s izvozom kao zavisnom varijablom

Model Summary <sup>b</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,871 <sup>a</sup>	,758	,756	4,18966	1,534

a. Predictors: (Constant), It izvoz (2000=100)  
b. Dependent Variable: PPI cijene (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Pokazatelji korelacije i determinacije modela u tablici 4.44 pokazuju reprezentativnost.

Prema podacima u tablici **Model Summary** može se izvršiti testiranje autokorelacije reziduala pomoću Durbin-Watsonovog testa.  $DW = 1,534$  na temelju čega se prema tablicama testa (E1 i E2) može utvrditi da je za  $\alpha = 1\%$ :

$$d_L = 1,522, d_U = 1,562, (4 - d_U) = 2,438, (4 - d_L) = 2,478.$$

U skladu s tim:  $d_L < DW < d_U$ ; pa se ne može donijeti sud o postojanju negativne autokorelacije rezidualnih odstupanja.

$$\text{Za } \alpha = 1\%: d_L = 1,502, d_U = 1,582, (4 - d_U) = 2,418, (4 - d_L) = 2,499.$$

U skladu s tim:  $d_L < DW < d_U$ ; pa se ne može zaključiti o prisutnosti problema autokorelacije rezidualnih odstupanja u ocijenjenom modelu jer  $DW$  vrijednost upada u interval područja inkonzistentnosti.

U tablici ANOVA prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenih modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema rezultatima u output-u vrijedi da je  $\alpha_3^* \approx 0 < \alpha = 5\% \Rightarrow H_1$ , pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

U tablici 4.45 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

**Tablica 4.45.**

#### Ocijenjeni linearni regresijski model s izvozom kao zavisnom varijablom

Coefficients <sup>a</sup>									
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1									
(Constant)	71,230	1,500		47,483	,000	68,259	74,201		
It izvoz (2000=100)	,281	,015	,871	19,058	,000	,252	,310	1,000	1,000

a. Dependent Variable: PPI cijene (2000=100)

Izvor: <http://epp.eurostat.ec.europa.eu>.

Analitički izraz konačnog modela je:

$$\hat{y}_i = 71,230 + 0,281 \cdot X_{1i}$$

Tumačenje regresijskih parametara i u ovom modelu vrši se na standardan način.

Regresijski model u standardiziranom obliku je:  $\hat{y}_i = 0,871 \cdot x_{1i}$ .

Sve empirijske signifikantnosti za regresijske parametre  $\hat{\beta}_0, \hat{\beta}_1$  su manje od  $\alpha = 5\%$  i može se zaključiti da su svi parametri statistički značajni.

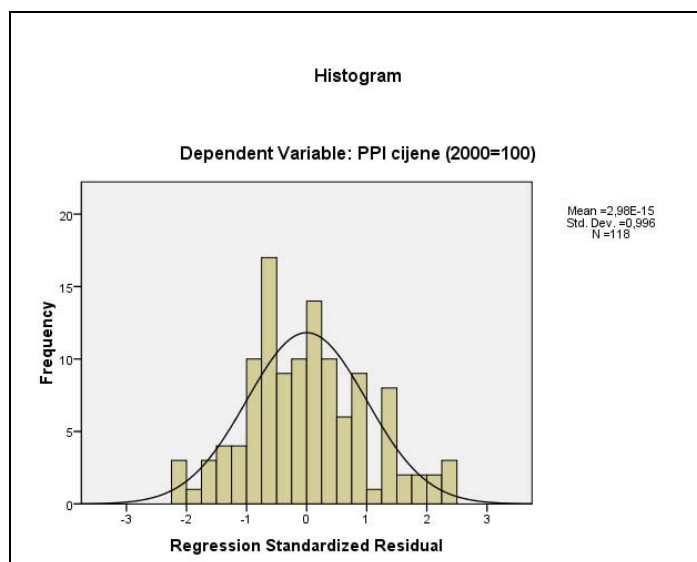
**Problem kolinearnosti (multikolinearnosti)** u ovom *jednostrukom modelu* (s jednom regresorskom varijablom) ne postoji.

Da bi se utvrdilo **jesu li reziduali normalno distribuirani** potrebno je prikazati histogram standardiziranih reziduala.

Prema slici 4.31 histograma u outputu može se vidjeti da su reziduali  $e_i$  normalno distribuirani s očekivanjem jednakim nuli i standardnom devijacijom približno jednakom jedinici.

**Slika 4.31.**

**Histogram standardiziranih reziduala**

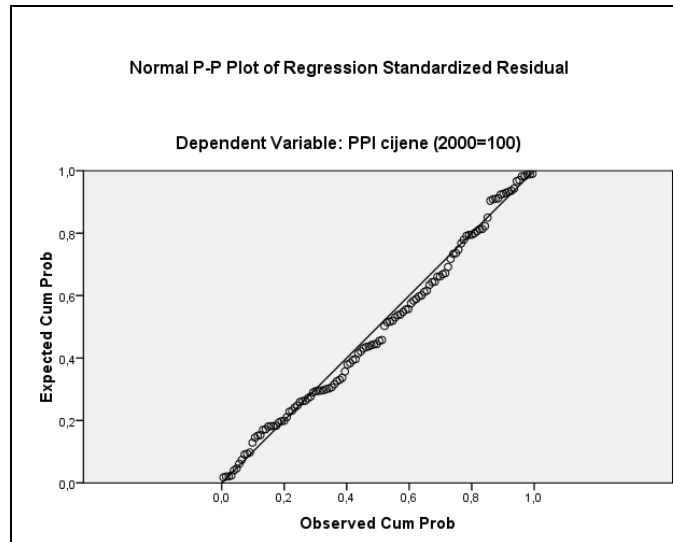


Izvor: <http://epp.eurostat.ec.europa.eu>.

Isto se može zaključiti iz grafikona na slici 4.32 kojem su ucrtane vrijednosti opaženih i očekivanih vjerojatnosti, kada bi reziduali bili normalno distribuirani (***Normal P-P Plot of Regression Standardized Residual***). Pretpostavku o normalnosti reziduala potvrđuje i funkcija distribucije opaženih vjerojatnosti, koja gotovo ne odstupa od očekivane funkcije vjerojatnosti kada bi rezidualna odstupanja bila normalno distribuirana (dijagonalna linija na grafikonu).

Slika 4.32

**P-P grafikon normalno distribuiranih rezidula**



Izvor: <http://epp.eurostat.ec.europa.eu>.

**Problem heteroskedastičnosti varijance reziduala** testira se neparametrisjnim testom i to pomoću Spearmanovog koeficijenta korelacije ranga.

U SPSS-u potrebno je za reziduala (koji su automatskim postupkom regresijske analize formirani u bazi podataka) izračunati njihove apsolutne vrijednosti.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Transform; Compute**, u **Target Variable:** "absres3. U **Numeric Expression:** abs(RES\_3) **OK**.

Na navedeni način u bazi podataka formira se niz apsolutnih reziduala (za 3 ocijenjeni model). Zatim se računaju Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorske varijable  $I_t$  izvoza.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Analyze; Correlate; Bivariate**. U **Variables** se prebace: **absres3 i  $I_t$  izvoza**. Potrebno je aktivirati koeficijent: **Spearman, OK**.

Da bi se testirala značajnost izračunatog Spearmanovog koeficijenta korelacije postavljaju se hipoteze:

$$H_0: \dots r_S = 0$$

$$H_1: \dots r_S \neq 0$$

Prema rezultatima iz tablice 4.46 može se vidjeti da je empirijska signifikantnost koeficijenta korelacije  $\alpha^* = 0,750$  pa se može zaključiti da je  $\alpha^* > 5\%$  i da se prihvaća početna hipoteza da korelacija nije statistički značajna. To znači da u ocijenjenom modelu ne postoji problem heteroskedastičnosti varijance reziduala.

**Tablica 4.46.**

**Spearmanovi koeficijenti korelacije između apsolutnih reziduala i regresorske varijable**

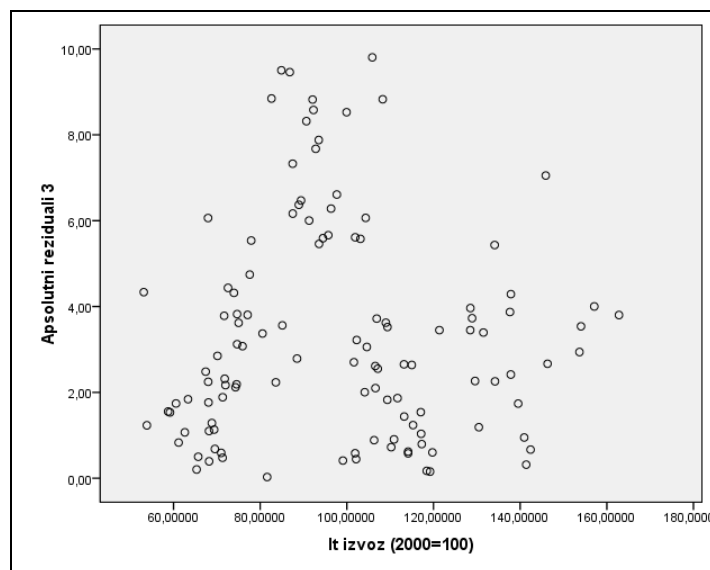
Correlations				
	Apsolutni reziduali 3	It izvoz (2000=100)		
Spearman's rho	Apsolutni reziduali 3	Correlation Coefficient	1,000	,030
		Sig. (2-tailed)	.	,750
		N	118	118
	It izvoz (2000=100)	Correlation Coefficient	,030	1,000
		Sig. (2-tailed)	,750	.
		N	118	125

Izvor: <http://epp.eurostat.ec.europa.eu>.

Na kraju je konstruiran dijagram rasipanja između apsolutnih reziduala i regresorske varijable.

**Slika 4.32.**

**Dijagram rasipanja apsolutnih reziduala s regresorskom varijablom  $I_t$  izvoz**



Izvor: <http://epp.eurostat.ec.europa.eu>.

Na glavnom izborniku SPSS-a potrebno je odabrati: **Graphs; Legacy Dialogs; Scatter/Dot**, gdje treba aktivirati: **Simple Scatter**. U **Define** se prebace odgovarajuće varijable: u **Y Axes: Apsolutni reziduali 3**; u **X Axes: I<sub>t</sub> izvoz**; **OK**.

Na temelju dobivenog dijagrama rasipanja između apsolutnih reziduala i regresorske varijable I<sub>t</sub> izvoz na slici 4.32, može se vidjeti da u ocijenjenom regresijskom modelu nije izražen problem heteroskedastičnosti varijance reziduala.

## 4.9 Nenumeričke dummy varijable i poslovne prognoze

Varijable koje se uključuju u regresijski model najčešće su numeričke varijable. Ponekad se u model pored numeričkih varijabli uključuju kvalitativne varijable. Kvalitativna varijabla u modelu se može pojaviti kao zavisna ili kao nezavisna varijabla.

Regresijski model u kojem je zavisna varijabla numerička, a koji kao nezavisne varijable uz numeričke sadrži jednu ili više kvalitativnih varijabli, analizira se na jednak način kao i standardni regresijski model (Modeli koji sadrže dummy varijablu kao zavisnu su logit i probit modeli).

Dummy (indikator ili binarna) varijabla u regresijskoj analizi je umjetno konstruirana varijabla, čije se vrijednosti ne mogu numerički izraziti. Takva varijabla je najčešće rezultat postojanja ili nepostojanja nekog fenomena/događaja.

Ako taj fenomen ne postoji dummy varijabla poprima vrijednost 0, a ako taj fenomen postoji dummy varijabla poprima vrijednost 1. Ponekad je moguće dummy varijabli dodjeliti i neke druge vrijednosti, npr. 0,25; 0,5 i sl.

U tom smislu, dummy varijabla sa može lako uključiti u regresijski model. Regresijski parametar uz dummy varijablu predstavlja utjecaj spomenutog fenomena/događaja na zavisnu ili regresand varijablu.

Ako je u regresijski model potrebno uključiti više dummy varijabli, treba naglasiti da one nisu korelirane, pa stoga ne uvjetuju postojanje problema multikolinearnosti u regresijskom modelu.

Dummy varijable se u model mogu uključiti iz različitih razloga:

- efekti vremena (npr. broj turista je različit u zimskim - ili ljetnim mjesecima; proizvodnja je različita u ovisnosti o tome je li u privredi kriza - ili nema krize),



- efekti prostora (npr. pokazatelji uspješnosti poduzeća su različiti u hrvatskim - ili stranim poduzećima, ekonomski pokazatelji su različiti u razvijenim - ili nerazvijenim zemljama),
- kvalitativne varijable (npr. plaće u nekom poduzeću mogu ovisiti i o spolu muškom - ili ženskom, mogu ovisiti i o stručnoj spremi visokoj - ili srednjoj),
- šire grupiranje numeričkih varijabli (npr. mogoće je kombinirati BDP i nezaposlenost, ali takva varijabla više nije numerička varijabla).

Za svaku kvalitativnu varijablu broj uvedenih dummy varijabli u model je za 1 manji od broja modaliteta promatrane varijable.

#### 4.9.1 Dummy varijable konstantnog člana

Može se pretpostaviti da se regresijski modeli za različite modalitete kvalitativne varijable razlikuju samo u konstantnom članu  $\hat{\beta}_0$ , a da im je nagib (tj. kod jednostruke regresije  $\hat{\beta}_1$ ) jednak.

Na primjer, ispituje se regresijska veza između kretanja plaća u jednom poduzeću u ovisnosti o prihodu poduzeća, ali pretpostavka je da se plaće razlikuju u ovisnosti o tome je li zaposlenik ima visoku stručnu spremu ili nema. Znači da je nagib modela  $\hat{\beta}_1$  jednak, ali da je konstanta  $\hat{\beta}_0^2$  veća za model onih zaposlenika koji imaju visoku stručnu spremu od  $\hat{\beta}_0^1$  onih zaposlenika koji nemaju visoku stručnu spremu.

Općenito su regresijske jednadžbe koje treba procijeniti za svaku od 2 kategorije stručnih sprema (visoka i one koje nisu visoka stručna sprema):

$$Y = \begin{cases} \beta_0^1 + \beta_1 X + e \\ \beta_0^2 + \beta_1 X + e \end{cases} \quad (4.109)$$

Ove dvije jednadžbe se mogu spojiti u jedinstvenu jednadžbu:

$$Y = \beta_0^1 + (\beta_0^2 - \beta_0^1)D + \beta_1 X + e, \quad (4.110)$$

gdje je:

$$D = \begin{cases} 1 & \text{za grupu 2} \\ 0 & \text{za grupu 1} \end{cases} \quad (4.111)$$

Varijabla D je dummy regresorska varijabla, a parametar uz D mjeri razlike između konstantnih članova:

$$Y_i = \beta_0^1 + \beta_1 X_i + (\beta_0^2 - \beta_0^1) D_i + e_i, \quad i = 1, 2, \dots, n. \quad (4.112)$$

Model se u ovisnosti o vrijednost dummy varijable (binarne: 0 ili 1) može raščlaniti na dvije regresijske jednadžbe:

$$Y_i = \beta_0^1 + \beta_1 X_i + (\beta_0^2 - \beta_0^1) \cdot 0 + e_i, \quad i = 1, 2, \dots, n_1, \quad (4.113)$$

$$Y_i = \beta_0^1 + \beta_1 X_i + e_i, \quad i = 1, 2, \dots, n_1, \quad (4.114)$$

$$Y_i = \beta_0^1 + \beta_1 X_i + (\beta_0^2 - \beta_0^1) \cdot 1 + e_i, \quad i = 1, 2, \dots, n_2, \quad (4.115)$$

$$Y_i = \beta_0^2 + \beta_1 X_i + e_i, \quad i = 1, 2, \dots, n_2. \quad (4.116)$$

Može se vidjeti da obje regresijske jednadžbe imaju isti koeficijent smjera  $\hat{\beta}_1$ , a razlikuju se samo u konstantnom članu (odsječku na osi ordinata).

Kao prednost sa statističkog aspekta ovakvog modela je u većem uzorku ( $n = n_1 + n_2$ ) čime se eliminira problem sa stupnjevima slobode.

Ako je broj modaliteta kvalitativne regresorske varijable veći od 2, u model treba uključiti veći broj dummy varijabli. Na primjer, ako se želi istražiti više od 2 kategorije stručnih sprema (visoka, viša, srednja i osnovno obrazovanje). Tada se za  $k$  grupa formira  $k$  regresijskih jednadžbi:

$$Y = \begin{cases} \beta_0^1 + \beta_1 X + e & \text{za grupu 1} \\ \beta_0^2 + \beta_1 X + e & \text{za grupu 2} \\ \dots\dots\dots \\ \beta_0^k + \beta_1 X + e & \text{za grupu } k \end{cases} \quad (4.117)$$

i one se mogu napisati:

$$Y_i = \beta_0^1 + (\beta_0^2 - \beta_0^1) D_1 + (\beta_0^3 - \beta_0^1) D_2 + (\beta_0^k - \beta_0^1) D_{k-1} + \beta_1 X_i + e_i, \quad i = 1, 2, \dots, n \quad (4.118)$$

gdje je:

$$D = \begin{cases} 1 & \text{za grupu } j+1 \\ 0 & \text{za ostale grupe} \end{cases} \quad (4.119)$$

Kombiniranje  $k$  regresorskih varijabli moguće je zbog pretpostavke da je koeficijent smjera  $\hat{\beta}_1$  za sve regresorske varijable jednak i da su slučajne varijable  $e$  identično distribuirane za sve grupe.

Za regresijske jednadžbe s konstantnim članom broj dummy varijabli je uvijek za 1 manji od broja modaliteta kvalitativne varijable. U suprotnom, ako bi broj dummy varijabli bio jednak broju modaliteta kvalitativne varijable u regresiji bi se dogodio problem potpune multikolinearnosti i ne bi bilo moguće riješiti model. Bilo koji od modaliteta kvalitativne varijable se može uzeti kao bazni.

Za  $k$  modaliteta, ako se npr. 1. izabere kao bazni, vrijedi da je:

$$\beta_0^1 \equiv \beta_0 \quad \beta_0^2 \equiv \beta_0 + (\beta_0^2 - \beta_0^1) \quad \dots \quad \beta_0^k \equiv \beta_0 + (\beta_0^k - \beta_0^1) \quad (4.120)$$

što znači da je konstanta za cijeli uzorak regresije  $\beta_0$  ustvari konstanta za 1. modalitet kvalitativne varijable. Konstanta za drugi uzorak je zbroj konstante  $\beta_0$  i koeficijenta uz  $D_1$ . Konstanta za  $k$ -ti uzorak je zbroj konstante  $\beta_0$  i koeficijenta uz  $D_{k-1}$ .

Za regresiju bez konstantnog člana za svaki modalitet kvalitativne varijable se definira dummy varijabla, a regresijski koeficijenti uz  $D_i$  su konstantni članovi jednadžbi za pojedine modalitete.

#### 4.9.2 Dummy varijable za promjene u nagibu

Dummy varijable se u regresijske modele mogu uvesti i da bi se dozvolile promjene u nagibima. Ako su dvije regresijske jednadžbe za 2 modaliteta kvalitativne varijable:

$$\begin{aligned} Y_{i1} &= \beta_0^1 + \beta_1^1 X_{i1} + e_{i1} & i &= 1, 2, \dots, n_1 \\ Y_{i2} &= \beta_0^2 + \beta_1^2 X_{i2} + e_{i2} & i &= 1, 2, \dots, n_2 \end{aligned} \quad (4.121)$$

mogu se napisati:

$$\begin{aligned} Y_{i1} &= \beta_0^1 + (\beta_0^2 - \beta_0^1) \cdot 0 + \beta_1^1 X_{i1} + (\beta_1^2 - \beta_1^1) \cdot 0 + e_{i1} & i = 1, 2, \dots, n_1 \\ Y_{i2} &= \beta_0^1 + (\beta_0^2 - \beta_0^1) \cdot 1 + \beta_1^1 X_{i1} + (\beta_1^2 - \beta_1^1) \cdot X_{i2} + e_{i2} & i = 1, 2, \dots, n_2 \end{aligned} \quad (4.121)$$

tj.

$$Y_i = \beta_0^1 + (\beta_0^2 - \beta_0^1) \cdot D_1 + \beta_1^1 X_i + (\beta_1^2 - \beta_1^1) \cdot D_2 + e_i \quad i = 1, 2, \dots, n \quad (4.122)$$

gdje je:

$$D_1 = \begin{cases} 1 & \text{za mod 2} \\ 0 & \text{za mod 1} \end{cases} \quad (4.123)$$

$$D_2 = \begin{cases} X_2 & \text{za mod 2} \\ 0 & \text{za mod 1} \end{cases} \quad (4.124)$$

Parametar uz  $D_1$  mjeri razliku u konstantnim članovima, a parametar uz  $D_2$  mjeri razliku u nagibima. Ovdje vrijedi pretpostavka da su slučajne varijable u polaznim regresijskim modelima nezavisne i identično distribuirane.

### 4.9.3 Sezonske dummy varijable

U regresijski model dummy varijable se mogu uvesti i da bi se uočio utjecaj sezonskog faktora. Ako regresijska jednadžba sadrži konstantni član, tada je broj dummy varijabli za 1 manji od broja jediničnih intervala promatranja (za npr. kvartalni utjecaj uvode se 3 dummy varijable, za mjesečni utjecaj uvodi se 11 dummy varijabli).

Ako su dani kvartalni podaci o regresand varijabli  $Y$ , regresijska jednadžba se može napisati kao:

$$Y = \beta_0 + \beta_1 X + \delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + e, \quad (4.125)$$

gdje je  $X$  odabrana regresorska varijabla (kod analize vremenskih nizova može biti i vrijeme), a dummy varijable su:

$$D_1 = \begin{cases} 1 & \text{za 1. kvrt.} \\ 0 & \text{za ost. kvrt.} \end{cases} \quad D_2 = \begin{cases} 1 & \text{za 2. kvrt.} \\ 0 & \text{za ost. kvrt.} \end{cases} \quad D_3 = \begin{cases} 1 & \text{za 3. kvrt.} \\ 0 & \text{za ost. kvrt.} \end{cases} \quad (4.126)$$

Ovdje je 4. kvartal odabran kao bazno razdoblje (a mogao je proizvoljno biti odabran bilo koji).

Ako su dani godišnji podaci, u regresijskom modelu bit će 11 dummy varijabli:

$$D_i = \begin{cases} 1 & \text{za } i\text{-ti mjsc.} \\ 0 & \text{za ost. mjsce.} \end{cases} \quad i = 1, 2, \dots, 11 \quad (4.127)$$



#### **Primjer 4.18.**

U razdoblju od siječnja 1996. godine do svibnja 2006. godine ispitivana je ovisnost noćenja turista u RH o gotovom novcu u mil. kn i sezonskoj kvartalnoj komponenti. Uz ocjenu parametara **Stepwise metodom** potrebno je testirati standardnu regresijsku dijagnostiku.



#### **Rješenje 4.18.**

Nakon odabira varijabli modela (ovdje je pretpostavka da je utjecaj 3. kvartala sadržan u konstanti modela):

*Nocnja turista* u: **Dependent** i

**1. kvartal ( $D_1$ ), 2. kvartal ( $D_2$ ), 4. kvartal ( $D_4$ ) i gotov novac u mil. kn.** u: **Independent**

i **Stepwise metode** za ocjenu parametara u **SPSS-u**:

- u **Statistics** potrebno je aktivirati: **Estimates, Confidence Intervals, Covariance matrix, Model fit, Collinearity diagnostics, DW; Continue;**

- u **Plots** potrebno je aktivirati: **Histogram, Normal probability plot; Continue;**

- u **Save** potrebno je aktivirati: **Unstandardized Predicted Value; Continue; OK.**

Konačni regresijski model sastoji od 4 regresorske varijable:  **$D_1$ ,  $D_2$ ,  $D_4$  i gotov novac u mil. kn** budući da su one zadovoljile kriterij ulaska (empirijska razina signifikantnosti je manja ili jednaka 5%).

U Outputu se dobije ocjena parametara i ostala regresijska dijagnostika.

U tablici 4.47 izračunati su osnovni pokazatelji o ocijenjenom modelu s dummy varijablama.

Pokazatelji multiple korelacije i determinacije modela pokazuju njegovu reprezentativnost.

Prema podacima u tablici 4.47 može se izvršiti testiranje autokorelacije reziduala pomoću Durbin-Watsonovog testa:  $DW = 2,087$ , što vodi zaključku da u modelu nije prisutan problem autokorelacije reziduala, jer je vrijednost DW približno 2.

Tablica 4.47

**Osnovni podaci o ocijenjenom modelu s noćenjem turista kao zavisnom varijablom**

Model Summary <sup>a</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,379 <sup>a</sup>	,144	,137	4117,120	
2	,621 <sup>b</sup>	,385	,375	3503,367	
3	,832 <sup>c</sup>	,692	,685	2487,778	
4	,852 <sup>d</sup>	,725	,716	2361,850	2,087

a. Predictors: (Constant), Prvi kvartal  
b. Predictors: (Constant), Prvi kvartal, Cetrvti kvartal  
c. Predictors: (Constant), Prvi kvartal, Cetrvti kvartal, Drugi kvartal  
d. Predictors: (Constant), Prvi kvartal, Cetrvti kvartal, Drugi kvartal, Gotov novac u mil. kn  
e. Dependent Variable: Nocenja turista

Izvor: <http://epp.eurostat.ec.europa.eu>.

Tablica 4.48

**Tablica ANOVA za ocijenjeni model s noćenjem turista kao zavisnom varijablom**

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4E+008	1	350392907,7	20,671	,000 <sup>a</sup>
	Residual	2E+009	123	16950677,75		
	Total	2E+009	124			
2	Regression	9E+008	2	468974897,1	38,210	,000 <sup>b</sup>
	Residual	1E+009	122	12273577,68		
	Total	2E+009	124			
3	Regression	2E+009	3	562150935,1	90,830	,000 <sup>c</sup>
	Residual	7E+008	121	6189036,905		
	Total	2E+009	124			
4	Regression	2E+009	4	441481449,2	79,142	,000 <sup>d</sup>
	Residual	7E+008	120	5578337,283		
	Total	2E+009	124			

a. Predictors: (Constant), Prvi kvartal

b. Predictors: (Constant), Prvi kvartal, Cetvrti kvartal

c. Predictors: (Constant), Prvi kvartal, Cetvrti kvartal, Drugi kvartal

d. Predictors: (Constant), Prvi kvartal, Cetvrti kvartal, Drugi kvartal, Gotov novac u mil. kn

e. Dependent Variable: Nocenja turista

Izvor: <http://epp.eurostat.ec.europa.eu>.

U tablici ANOVA 4.48 prikazani su podaci o protumačenim, neprotumačenim i ukupnim odstupanjima ocijenjenih modela, te vrijednost F-testa s empirijskom signifikantnosti.

Nakon postavljanja hipoteza o značajnosti modela:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0 \quad j = 1, 2, \dots, k$$

prema rezultatima u output-u vrijedi da je  $\alpha_3^* \approx 0 < \alpha = 5\% \Rightarrow H_1$ , pa se može potvrditi zaključak da je ocijenjeni regresijski model statistički značajan.

U tablici 4.48 prikazane su vrijednosti ocijenjenih parametara, njihove standardne greške, empirijski t-omjeri i procjene parametara uz nivo pouzdanosti od 95%.

**Tablica 4.48.**

**Ocijenjeni linearni regresijski model s noćenjem turista kao zavisnom varijablom**

4	(Constant)	7217,147	734,963		9,820	,000	5761,972	8672,322		
	Prvi kvartal	-8934,319	598,407	-,892	-14,930	,000	-10119,123	-7749,516	,641	1,559
	Četvrti kvartal	-8884,170	610,349	-,860	-14,556	,000	-10092,619	-7675,721	,657	1,523
	Drugi kvartal	-6849,833	600,842	-,677	-11,400	,000	-8039,458	-5660,208	,649	1,541
	Gotov novac u mil. kn	,280	,074	,181	3,774	,000	,133	,426	,991	1,009

a. Dependent Variable: Nocenja turista

Izvor: <http://epp.eurostat.ec.europa.eu>.

Analitički izraz konačnog modela je:

$$\hat{y}_i = 7217,15 - 8934,32 \cdot D_1 - 8884,17 \cdot D_4 - 6849,83 \cdot D_3 + 0,28 \cdot X_{1i}$$

U ocijenjenom regresijskom modelu konstanta, tj. parametar  $\hat{\beta}_0$  predstavlja noćenja turista u trećem kvartalu uz *c.p.* Parametar  $\hat{\beta}_1$  nalazi se uz sezonsku dummy varijablu  $D_1$  i predstavlja pad noćenja turista u svakom promatranom prvom kvartalu uz *c.p.* Parametar  $\hat{\beta}_2$  nalazi se uz sezonsku dummy varijablu  $D_4$  i predstavlja pad noćenja turista u svakom promatranom četvrtom kvartalu uz *c.p.* Parametar  $\hat{\beta}_3$  nalazi se uz sezonsku dummy varijablu  $D_3$  i predstavlja pad noćenja turista u svakom promatranom trećem kvartalu uz *c.p.* Parametar  $\hat{\beta}_4$  nalazi se uz regresorsku varijablu  $X_{1i}$  i pokazuje da se kada gotov novac poraste za 1 mil. kn, očekuje porast noćenja turista za 0,28 jedinica uz *c.p.*

Sve empirijske signifikantnosti za regresijske parametre  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$  i  $\hat{\beta}_4$  su manje od  $\alpha = 5\%$  i može se zaključiti da su svi parametri statistički značajni.

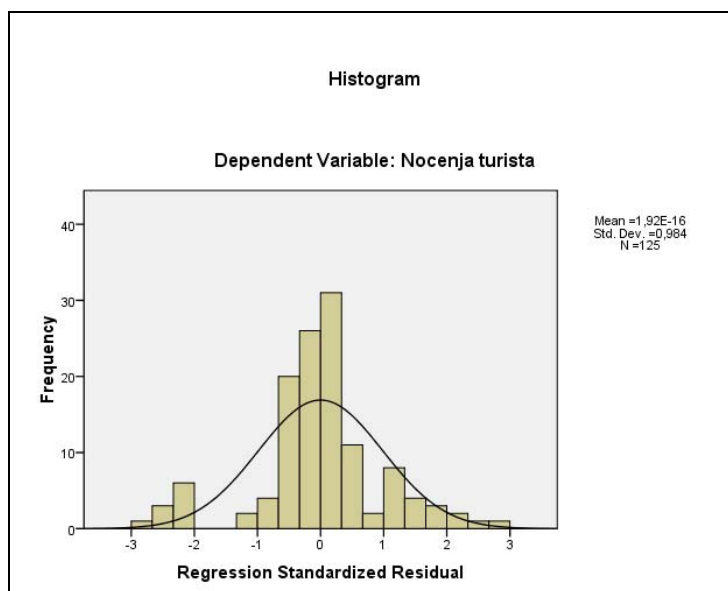
Potvrdu o nepostojanju **problema kolinearnosti (multikolinearnosti)** daju vrijednosti faktora inaflacije varijance ( $VIF_j$ ) jer su svi manji od 5.

Da bi se utvrdilo **jesu li reziduali normalno distribuirani** potrebno je prikazati histogram standardiziranih reziduala.

Prema slici 4.33 histograma u outputu može se vidjeti da su reziduali  $e_i$  normalno distribuirani s očekivanjem jednakim nuli i standardnom devijacijom približno jednakom jedinici.

**Slika 4.33.**

**Histogram standardiziranih reziduala**



Izvor: <http://epp.eurostat.ec.europa.eu>.

Isto se može zaključiti iz grafikona u **Outputu** na kojem su ucrtane vrijednosti opaženih i očekivanih vjerojatnosti, kada bi reziduali bili normalno distribuirani (***Normal P-P Plot of Regression Standardized Residual***). Pretpostavku o normalnosti reziduala potvrđuje i funkcija distribucije opaženih vjerojatnosti, koja gotovo ne odstupa od očekivane funkcije vjerojatnosti kada bi rezidualna odstupanja bila normalno distribuirana (dijagonalna linija na grafikonu).

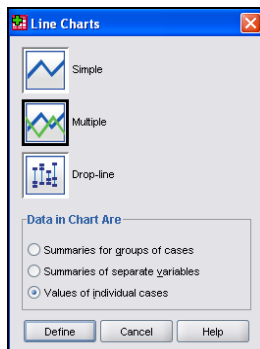
Da bi se u outputu dobio linijski grafikon vrijednosti regresand varijable i očekivanih vrijednosti regresand varijable potrebno je u **SPSS-u** odabrati **Graphs; Legacy**



**Dialogs; Line**, gdje treba aktivirati: **Multiple** (u **Data in Chart Are** aktivirati: **Values of individual cases**), kako je prikazano na slici 4.34.

Slika 4.34.

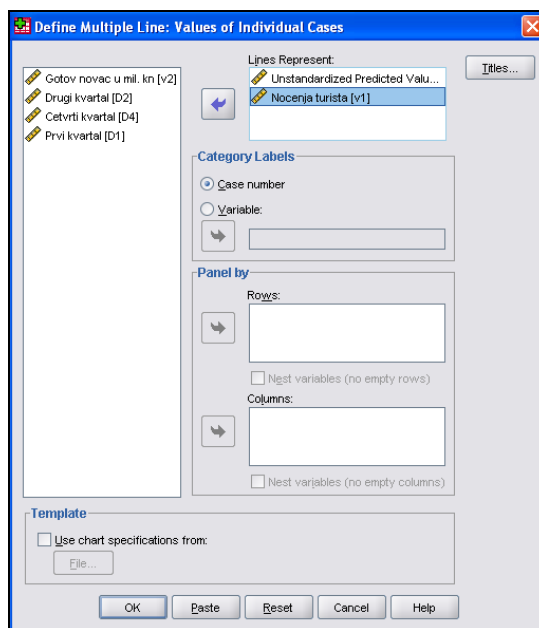
#### Prozor Line Charts u SPSS-u za konstrukciju linijskog grafikona



Izvor: Prema SPSS-u.

Slika 4.35.

#### Prozor Define Multiple Line u SPSS-u za konstrukciju višestrukog linijskog grafikona

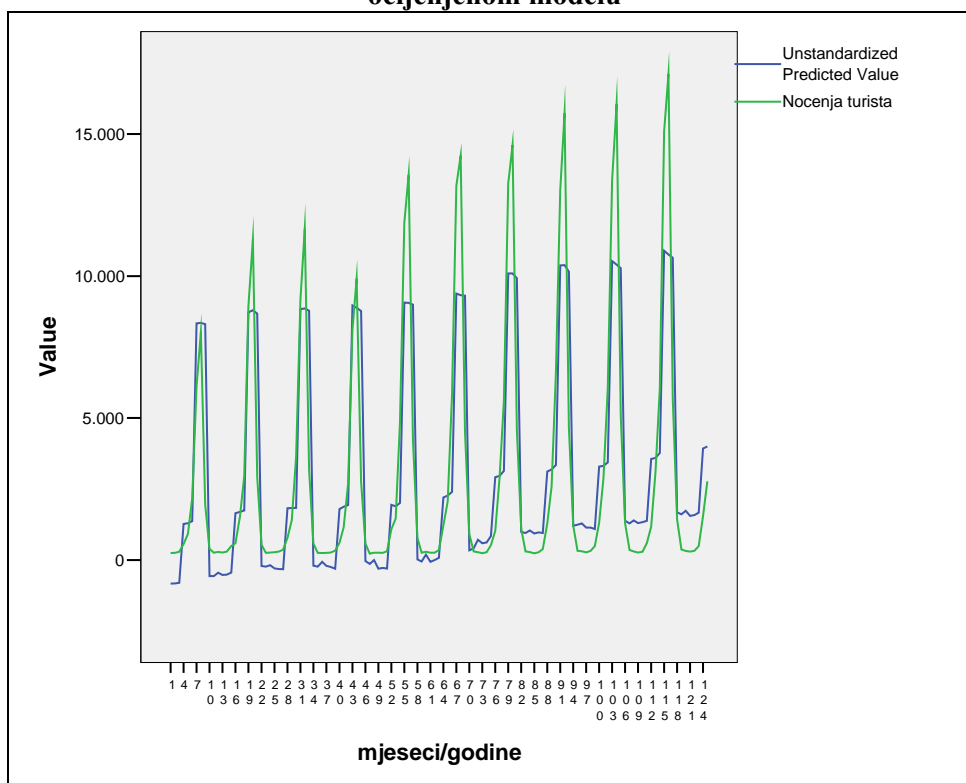


Izvor: <http://epp.eurostat.ec.europa.eu>.

Na slici 4.35 prikazan je prozor **Define Multiple Line: Values of Individual Cases** za konstrukciju višestrukog linijskog grafikona gdje su u **Lines Represent:** odabrane varijable *Unstandardized Predicted Value* i *Noćenja turista*. Klikom na **OK** u **Output-u** se dobije grafikon prikazan na slici 4.36.

Slika 4.36.

**Linijski grafikon noćenja turista i očekivanih noćenja turista prema ocijenjenom modelu**



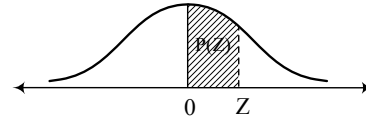
Izvor: <http://epp.eurostat.ec.europa.eu>.

Prema grafikonu može se zaključiti da varijabla noćenja turista ima jaku sezonsku komponentu, po kojoj se može uočiti porast broja turista u ljetnim mjesecima, tj. u trećem kvartalu svake godine.

U ovom primjeru, da bi se dobila još bolja prilagodba modela, mogu se umjesto kvartalnih, odabrati i mjesečne dummy varijable za proizvoljno odabranih 11 mjeseci (onaj preostali 12. mjesec sadržan je u konstanti regresije).

## 5. TABLICE ODABRANIH STATISTIČKIH DISTRIBUCIJA

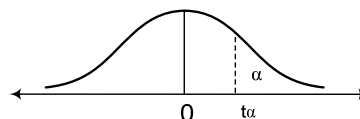
### A. Površine ispod normalne krivulje



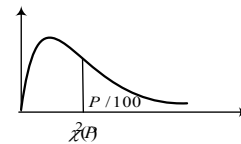
	0	1	2	3	4	5	6	7	8	9
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1,0	34134	34375	34614	34849	35083	35314	35543	35769	35993	36214
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1,3	40320	40490	40658	40824	40988	41149	41308	41466	41621	41774
1,4	41924	42073	42220	42364	42507	42647	42785	42922	43056	43189
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49896	49900
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983
4,0	499968	499970	499971	499972	499973	499974	499975	499976	499977	499978
4,5	499997	499997	499997	499997	499997	499997	499997	499998	499998	499998
5,0	500000	500000	500000	500000	500000	500000	500000	500000	500000	500000

*Napomena: Ispred svakog broja u polju tablice je decimalna točka.*

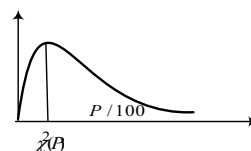
## B. Kritične vrijednosti t, Studentove distribucije



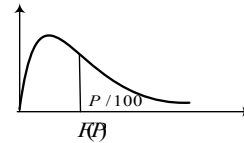
df/P	40	30	25	20	15	10	5	2,5	1	0,5	0,1	0,05
1	0,325	0,727	1,000	1,376	1,963	3,078	6,314	12,71	31,82	31,83	318,3	636,6
2	0,289	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	10,21	12,92
4	0,271	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	5,894	6,869
6	0,265	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,262	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	0,257	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	0,256	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,421	3,689
28	0,256	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,396	3,660
30	0,256	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	0,255	0,530	0,682	0,853	1,054	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	0,255	0,529	0,682	0,852	1,052	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	0,255	0,529	0,681	0,852	1,052	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	0,255	0,529	0,681	0,851	1,051	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	0,255	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,232	3,460
100	0,254	0,526	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626	3,174	3,390

**C1. Kritične vrijednosti Hi-kvadrat distribucije (za  $P \leq 0,05$ )**

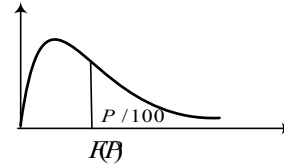
df	0,05	0,1	0,5	1	2	5	10	20	30	40	50
1	12,12	10,83	7,88	6,63	5,02	3,84	2,706	1,642	1,0742	0,7083	0,4549
2	15,20	13,82	10,60	9,21	7,38	5,99	4,605	3,219	2,4079	1,8326	1,3863
3	17,73	16,27	12,84	11,34	9,35	7,81	6,251	4,642	3,665	2,946	2,366
4	20,00	18,47	14,86	13,28	11,14	9,49	7,779	5,989	4,878	4,045	3,357
5	22,11	20,51	16,75	15,09	12,83	11,07	9,236	7,289	6,064	5,132	4,351
6	24,10	22,46	18,55	16,81	14,45	12,59	10,645	8,558	7,231	6,211	5,348
7	26,02	24,32	20,28	18,48	16,01	14,07	12,017	9,803	8,383	7,283	6,346
8	27,87	26,12	21,95	20,09	17,53	15,51	13,362	11,030	9,524	8,351	7,344
9	29,67	27,88	23,59	21,67	19,02	16,92	14,684	12,242	10,656	9,414	8,343
10	31,42	29,59	25,19	23,21	20,48	18,31	15,987	13,442	11,781	10,473	9,342
11	33,14	31,26	26,76	24,73	21,92	19,68	17,28	14,63	12,90	11,53	10,34
12	34,82	32,91	28,30	26,22	23,34	21,03	18,55	15,81	14,01	12,58	11,34
13	36,48	34,53	29,82	27,69	24,74	22,36	19,81	16,98	15,12	13,64	12,34
14	38,11	36,12	31,32	29,14	26,12	23,68	21,06	18,15	16,22	14,69	13,34
15	39,72	37,70	32,80	30,58	27,49	25,00	22,31	19,31	17,32	15,73	14,34
16	41,31	39,25	34,27	32,00	28,85	26,30	23,54	20,47	18,42	16,78	15,34
17	42,88	40,79	35,72	33,41	30,19	27,59	24,77	21,61	19,51	17,82	16,34
18	44,43	42,31	37,16	34,81	31,53	28,87	25,99	22,76	20,60	18,87	17,34
19	45,97	43,82	38,58	36,19	32,85	30,14	27,20	23,90	21,69	19,91	18,34
20	47,50	45,31	40,00	37,57	34,17	31,41	28,41	25,04	22,77	20,95	19,34
21	49,01	46,80	41,40	38,93	35,48	32,67	29,62	26,17	23,86	21,99	20,34
22	50,51	48,27	42,80	40,29	36,78	33,92	30,81	27,30	24,94	23,03	21,34
23	52,00	49,73	44,18	41,64	38,08	35,17	32,01	28,43	26,02	24,07	22,34
24	53,48	51,18	45,56	42,98	39,36	36,42	33,20	29,55	27,10	25,11	23,34
25	54,95	52,62	46,93	44,31	40,65	37,65	34,38	30,68	28,17	26,14	24,34
26	56,41	54,05	48,29	45,64	41,92	38,89	35,56	31,79	29,25	27,18	25,34
27	57,86	55,48	49,65	46,96	43,19	40,11	36,74	32,91	30,32	28,21	26,34
28	59,30	56,89	50,99	48,28	44,46	41,34	37,92	34,03	31,39	29,25	27,34
29	60,73	58,30	52,34	49,59	45,72	42,56	39,09	35,14	32,46	30,28	28,34
30	62,16	59,70	53,67	50,89	46,98	43,77	40,26	36,25	33,53	31,32	29,34
40	76,10	73,40	66,77	63,69	59,34	55,76	51,81	47,27	44,16	41,62	39,34
50	89,56	86,66	79,49	76,15	71,42	67,50	63,17	58,16	54,72	51,89	49,33
60	102,70	99,61	91,95	88,38	83,30	79,08	74,40	68,97	65,23	62,13	59,33
100	153,16	149,45	140,17	135,81	129,56	124,34	118,50	111,67	106,91	102,95	99,33

**C2. Kritične vrijednosti Hi-kvadrat distribucije (za  $P > 0,05$ )**

df/P	60	70	75	80	90	95	97,5	99	99,5
1	0,275	0,148	0,102	0,064	0,016	0,004	0,00098	0,00016	0,00004
2	1,022	0,713	0,575	0,446	0,211	0,103	0,05064	0,02010	0,01002
3	1,869	1,424	1,213	1,005	0,584	0,352	0,216	0,115	0,072
4	2,753	2,195	1,923	1,649	1,064	0,711	0,484	0,297	0,207
5	3,656	3,000	2,675	2,343	1,610	1,145	0,831	0,554	0,412
6	4,570	3,828	3,455	3,070	2,204	1,635	1,237	0,872	0,676
7	5,493	4,671	4,255	3,822	2,833	2,167	1,690	1,239	0,989
8	6,423	5,527	5,071	4,594	3,490	2,733	2,180	1,647	1,344
9	7,357	6,393	5,899	5,380	4,168	3,325	2,700	2,088	1,735
10	8,295	7,267	6,737	6,179	4,865	3,940	3,247	2,558	2,156
11	9,237	8,148	7,584	6,989	5,578	4,575	3,816	3,053	2,603
12	10,182	9,034	8,438	7,807	6,304	5,226	4,404	3,571	3,074
13	11,129	9,926	9,299	8,634	7,041	5,892	5,009	4,107	3,565
14	12,078	10,821	10,165	9,467	7,790	6,571	5,629	4,660	4,075
15	13,030	11,721	11,037	10,307	8,547	7,261	6,262	5,229	4,601
16	13,983	12,624	11,912	11,152	9,312	7,962	6,908	5,812	5,142
17	14,937	13,531	12,792	12,002	10,085	8,672	7,564	6,408	5,697
18	15,893	14,440	13,675	12,857	10,865	9,390	8,231	7,015	6,265
19	16,850	15,352	14,562	13,716	11,651	10,117	8,907	7,633	6,844
20	17,809	16,266	15,452	14,578	12,443	10,851	9,591	8,260	7,434
21	18,768	17,182	16,344	15,445	13,240	11,591	10,283	8,897	8,034
22	19,729	18,101	17,240	16,314	14,041	12,338	10,982	9,542	8,643
23	20,690	19,021	18,137	17,187	14,848	13,091	11,689	10,196	9,260
24	21,652	19,943	19,037	18,062	15,659	13,848	12,401	10,856	9,886
25	22,616	20,867	19,939	18,940	16,473	14,611	13,120	11,524	10,520
26	23,579	21,792	20,843	19,820	17,292	15,379	13,844	12,198	11,160
27	24,544	22,719	21,749	20,703	18,114	16,151	14,573	12,878	11,808
28	25,509	23,647	22,657	21,588	18,939	16,928	15,308	13,565	12,461
29	26,475	24,577	23,567	22,475	19,768	17,708	16,047	14,256	13,121
30	27,442	25,508	24,478	23,364	20,599	18,493	16,791	14,953	13,787
40	37,134	34,872	33,660	32,345	29,051	26,509	24,433	22,164	20,707
50	46,864	44,313	42,942	41,449	37,689	34,764	32,357	29,707	27,991
60	56,620	53,809	52,294	50,641	46,459	43,188	40,482	37,485	35,534
100	95,808	92,129	90,133	87,945	82,358	77,929	74,222	70,065	67,328

**D1. Kritične vrijednosti F-distribucije (za  $P = 0,05$ )**

$\nu_1$	1	2	3	4	5	6	7	8	9	10	12	24	$\infty$
$\nu_2 = 1$	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	249,1	254,2
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,45	19,49
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,638	8,53
4	7,709	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,774	5,63
5	6,608	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,527	4,37
6	5,987	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,841	3,67
7	5,591	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,410	3,23
8	5,318	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,115	2,93
9	5,117	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	2,900	2,71
10	4,965	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,737	2,54
11	4,844	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,609	2,41
12	4,747	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,505	2,30
13	4,667	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,420	2,21
14	4,600	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,349	2,14
15	4,543	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,288	2,07
16	4,494	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,235	2,02
17	4,451	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,190	1,97
18	4,414	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,150	1,92
19	4,381	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,114	1,88
20	4,351	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,082	1,85
21	4,325	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,054	1,82
22	4,301	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,028	1,79
23	4,279	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,005	1,76
24	4,260	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	1,984	1,74
25	4,242	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	1,964	1,72
26	4,225	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	1,946	1,70
27	4,210	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	1,930	1,68
28	4,196	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	1,915	1,66
29	4,183	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	1,901	1,65
30	4,171	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	1,887	1,63
40	4,085	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,793	1,52
50	4,034	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,95	1,737	1,45
60	4,001	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,700	1,40
120	3,920	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,608	1,27
$\infty$	3,851	3,00	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,76	1,528	1,11

**D2. Kritične vrijednosti F-distribucije (za  $P = 0,01$ )**

	$\nu_1 = 1$	2	3	4	5	6	7	8	9	10	12	24	$\infty$
$\nu_2 = 1$	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6234	6363
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,42	99,46	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,60	26,14
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	13,93	13,47
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,888	9,466	9,032
6	13,75	10,92	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,718	7,313	6,891
7	12,25	9,55	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,469	6,074	5,660
8	11,26	8,65	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,667	5,279	4,869
9	10,56	8,02	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257	5,111	4,729	4,321
10	10,04	7,56	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,706	4,327	3,920
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539	4,397	4,021	3,613
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296	4,155	3,780	3,372
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191	4,100	3,960	3,587	3,176
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030	3,939	3,800	3,427	3,015
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,666	3,294	2,880
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780	3,691	3,553	3,181	2,764
17	8,400	6,112	5,185	4,669	4,336	4,101	3,927	3,791	3,682	3,593	3,455	3,083	2,664
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,597	3,508	3,371	2,999	2,577
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,523	3,434	3,297	2,925	2,501
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368	3,231	2,859	2,433
21	8,017	5,780	4,874	4,369	4,042	3,812	3,640	3,506	3,398	3,310	3,173	2,801	2,372
22	7,945	5,719	4,817	4,313	3,988	3,758	3,587	3,453	3,346	3,258	3,121	2,749	2,317
23	7,881	5,664	4,765	4,264	3,939	3,710	3,539	3,406	3,299	3,211	3,074	2,702	2,268
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363	3,256	3,168	3,032	2,659	2,223
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,217	3,129	2,993	2,620	2,182
26	7,721	5,526	4,637	4,140	3,818	3,591	3,421	3,288	3,182	3,094	2,958	2,585	2,144
27	7,677	5,488	4,601	4,106	3,785	3,558	3,388	3,256	3,149	3,062	2,926	2,552	2,109
28	7,636	5,453	4,568	4,074	3,754	3,528	3,358	3,226	3,120	3,032	2,896	2,522	2,077
29	7,598	5,420	4,538	4,045	3,725	3,499	3,330	3,198	3,092	3,005	2,868	2,495	2,047
30	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	3,067	2,979	2,843	2,469	2,019
40	7,314	5,178	4,313	3,828	3,514	3,291	3,124	2,993	2,888	2,801	2,665	2,288	1,819
50	7,171	5,057	4,199	3,720	3,408	3,186	3,020	2,890	2,785	2,698	2,563	2,183	1,698
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718	2,632	2,496	2,115	1,617
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559	2,472	2,336	1,950	1,401
$\infty$	6,660	4,626	3,801	3,338	3,036	2,820	2,657	2,529	1,889	2,339	2,203	1,810	1,159



## E1. Kritične vrijednosti Durbin-Watsonovog pokazatelja (za $P = 0,05$ )

n	k'=1		k'=2		k'=3		k'=4		k'=5	
	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	-----	-----	-----	-----	-----	-----	-----	-----
7	0.700	1.356	0.467	1.896	-----	-----	-----	-----	-----	-----
8	0.763	1.332	0.559	1.777	0.367	2.287	-----	-----	-----	-----
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	-----	-----
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780
150	1.720	1.747	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802
200	1.758	1.779	1.748	1.789	1.738	1.799	1.728	1.809	1.718	1.820

\*k' - broj regresorskih varijabli u modelu, bez konstante

## E2. Kritične vrijednosti Durbin-Watsonovog pokazatelja (za $P = 0,01$ )

	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.390	1.142	-----	-----	-----	-----	-----	-----	-----	-----
7	0.435	1.036	0.294	1.676	-----	-----	-----	-----	-----	-----
8	0.497	1.003	0.345	1.489	0.229	2.102	-----	-----	-----	-----
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	-----	-----
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645
26	1.072	1.222	1.000	1.311	0.928	1.410	0.855	1.517	0.782	1.635
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611
30	1.134	1.264	1.070	1.339	1.006	1.421	0.941	1.510	0.877	1.606
31	1.147	1.274	1.085	1.345	1.022	1.425	0.960	1.509	0.897	1.601
32	1.160	1.283	1.100	1.351	1.039	1.428	0.978	1.509	0.917	1.597
33	1.171	1.291	1.114	1.358	1.055	1.432	0.995	1.510	0.935	1.594
34	1.184	1.298	1.128	1.364	1.070	1.436	1.012	1.511	0.954	1.591
35	1.195	1.307	1.141	1.370	1.085	1.439	1.028	1.512	0.971	1.589
36	1.205	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.987	1.587
37	1.217	1.322	1.164	1.383	1.112	1.446	1.058	1.514	1.004	1.585
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.584
39	1.237	1.337	1.187	1.392	1.137	1.452	1.085	1.517	1.033	1.583
40	1.246	1.344	1.197	1.398	1.149	1.456	1.098	1.518	1.047	1.583
45	1.288	1.376	1.245	1.424	1.201	1.474	1.156	1.528	1.111	1.583
50	1.324	1.403	1.285	1.445	1.245	1.491	1.206	1.537	1.164	1.587
55	1.356	1.428	1.320	1.466	1.284	1.505	1.246	1.548	1.209	1.592
60	1.382	1.449	1.351	1.484	1.317	1.520	1.283	1.559	1.248	1.598
65	1.407	1.467	1.377	1.500	1.346	1.534	1.314	1.568	1.283	1.604
70	1.429	1.485	1.400	1.514	1.372	1.546	1.343	1.577	1.313	1.611
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.586	1.340	1.617
80	1.465	1.514	1.440	1.541	1.416	1.568	1.390	1.595	1.364	1.624
85	1.481	1.529	1.458	1.553	1.434	1.577	1.411	1.603	1.386	1.630
90	1.496	1.541	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.641
100	1.522	1.562	1.502	1.582	1.482	1.604	1.461	1.625	1.441	1.647
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725

\* $k'$  - broj regresorskih varijabli u modelu, bez konstante

## 6. LITERATURA

1. Aczel, A.D. (1999). Complete Business Statistics, fourth edition, New York: Irving McGraw-Hill.
2. Allen, M. V. and Myddelton, D. R. (1992). Essential Management Accounting. London: Prentice Hall.
3. Anderberg, M. R. (1973). Cluster Analysis for Applications. New York: Academic Press.
4. Anderson, D. R. et al. (2002). Statistics for Business and Economics. 8th. Edt. St Paul: West.
5. Anderson, D. R. et al. (1998). Quantitative Methods for Business. 7th. Edt. South-Western College Publishing.
6. Anderson, R. E., Black, W., Hair, J. F. and Tatham, R. L. (1998). Multivariate Data Analysis. Prentice Hall.
7. Anderson, T. W., 2003 An Introduction to Multivariate Statistical Analysis (Wiley Series in Probability and Statistics), Wiley-Interscience,
8. Arnold, J. A. and Hope, A. J. B. (1990). Accounting for Management Decisions. London: Prentice Hall.
9. Babbie, E. Halley, F. Zaino, J. (2000). Adventures in Social Research. Data Analysis Using SPSS<sup>X</sup> for Windows 95/98. USA: Pine Forge Press.
10. Bahovec, V. and Erjavec, N (2009). Uvod u ekonometrijsku analizu. Zagreb. Element.
11. Basilevsky, A. T. (1994). Statistical Factor Analysis and Related Methods: Theory and Applications. John Wiley and Sons Inc.
12. Bešter, M. i Bregar, L. (1986). Ekonomska statistika. Ljubljana. Ekonomska fakulteta.
13. Biljan-August, M, Pivac, S. and Štambuk, A. (2006). Upotreba statistike u ekonomiji, Sveučilište u Rijeci, Ekonomski fakultet Rijeka, ([www.efri.hr](http://www.efri.hr)).
14. Blejec, M. (1976). Statističke metode za ekonomiste. Ljubljana: Ekonomska fakulteta.
15. Bollerslev, T (1986). Generalized Autoregressive Conditional Heteroscedasticity, Journal of Econometrics, 31, pp 307-327.

16. Box, G. E. P. and Pierce, D. A. (1970). Distribution of Residual Autocorrelations in Autoregressive Moving Average Time Series Models, *Journal of the American Statistical Association*, 82, pp 276-282.
17. Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, 2<sup>nd</sup> edition, Holden-Day, San Francisco.
18. Brooks, C. (1997). GARCH Modelling in Finance: A Review of the Software Options, *Economics Journal*, 107 (443), pp 1271-1276.
19. Brooks, C. (2002). *Introductory econometrics for finance*. Cambridge: University Press.
20. Canavos, G. C. and Miller, D. M. (1993). *An Introduction to Modern Business Statistics*. Belmont. Wadsworth.
21. Charniak, E. (1993). *Statistical Language Learning*. Cambridge and Massachusetts: A Bradford Book, The MIT Press.
22. Chiang, A. C. (1994). *Osnovne metode matematičke ekonomije*. Zagreb: Mate.
23. Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.
24. Conover, W. J. (1980). *Practical Nonparametric Statistics*. New York: Wiley.
25. Dickey, D. and Fuller, W. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, 79, pp 355-367.
26. Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis, Methods and Applications*. New York: Wiley.
27. Dodge, M., Kinata, C. i Stison, C. (1997). *Kako koristiti Microsoft Excel 97*, Zagreb: Znak.
28. Draper, N. and Smith, H. (1981). *Applied Regression Analysis*. 2nd Edt. New York: Wiley.
29. Enders, W. (2004). *Applied Econometric Time Series*, 2nd edition, John Wiley & Sons, Inc., London.
30. EUROSTAT, URL: [http:// www.epp.eurostat.ec.europa.eu](http://www.epp.eurostat.ec.europa.eu)
31. Everit, B. S. (1996). *Making Sense of Statistics in Psychology*, Oxford: Oxford University Press.
32. Feller, W. (1971). *An Introduction to Probability Theory and its Applications*. Vol. I and Vol. II, New York: Wiley.

33. Fisher, R. A. (1950). Statistical Methods for Research Workers. 11 th Edt. Edinburgh: Oliver and Boyd.
34. Fisher, R. A. (1993). Statistical Methods, Eksperimental Design and Scientific Inference. Oxford: Oxford University Press.
35. Frye, C. (2003). Microsoft Excel 2002 Korak po korak. Zagreb: Algoritam.
36. Fulton, J. (1997). Vodič kroz Excel 97. Zagreb: Znak.
37. Gujarati, D. N. (2003). Basic Econometrics, 4th Ed., McGraw-Hill/Irwin.
38. Hadživuković, S. (1975). Tehnika metoda uzorka. Beograd: Naučna knjiga.
39. Halmi, A. (2003). Multivarijantna analiza u društvenim znanostima. Alinea. Zagreb.
40. Heij, C. et. al. (2004). Econometric Methods with Applications in Business and Economics, New York : Oxford University Press.
41. Horvatić, K. (2004). Linearna algebra, Zagreb: Golden marketing-Tehnička knjiga.
42. Johnson, R. A. and Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. London: Prentice Hall.
43. Jolliffe, I.T. (2002). Principal Component Analysis. Berlin. New York: Springer.
44. Kaplan, R. S. (1984). The Evolution of Management Accounting. The Accounting Review, 59 (3), pp 95-101.
45. Karatzas, I. and Shreve, S E. (2001). Methods of Mathematical Finance, Berlin. New York: Springer.
46. Kendall, M. and Stuart, A. (1973). The Advanced Theory of Statistics. Vol I. London: Griffin.
47. Kendall, M. and Stuart, A. (1974). The Advanced Theory of Statistics. Vol II. London: Griffin.
48. Kendall, M. and Stuart, A. (1976). The Advanced Theory of Statistics. Vol III. London: Griffin.
49. Kish, L. (1965). Survey Sampling. New York Wiley.
50. Kmenta, J. (1997). Počela ekonometrije. Zagreb: MATE d.o.o.
51. Kolesarić, V. i Petz. B. (2003). Statistički rječnik. Tumač statističkih pojmova. Naklada Slap.

52. Kurepa, S. (1979). Konačno dimenzionalni vektorski prostori i primjene, Zagreb, Sveučilišna naklada Liber.
53. de Levie, R. (2004). Advanced Excel for Scientific Data Analysis. Oxford: Oxford University Press.
54. Lucey, T. (1989). Management Information Systems. London: DP Publication LTD.
55. Maddala, G. S. (2002). Introduction to Econometrics, 3rd Ed., Chichester, John Wiley & Sons.
56. Martić, Lj. (1986). Mjere nejednakosti i siromaštva. Zagreb. Birotehnika.
57. Mason, D. R. and Lind, D. A. (1993). Statistical Techniques in Business and Economics. Boston, Massachusetts: IRVIN Publishing.
58. Metre, J. G. and Gilbreath, G. H. (1983). Statistics for Business and Economics. Plano. Texas: Business Publications.
59. Milton, J. S. et al. (1986). Introduction to Statistics. Lexington: D. C. Heath.
60. Momirović, K. (1998). Uvod u analizu nominalnih varijabli. Metodološke sveske 2. Ljubljana: JUS.
61. Mood, A. M. and Graybill, F. P. (1963). Introduction to the Theory of Statistics. New York: McGraw-Hill.
62. Mott, G. (1991). Management Accounting for Decision Makers. London: Pitman Publishing.
63. Neter, J. et al. (1993). Applied Statistics. 4th Edt. Boston: Allyn and Bacon.
64. Newbold, P. (1991). Statistics for Business and Economics. Englewood Cliffs: Prentice-Hall.
65. Pauše, Ž. (1993). Uvod u matematičku statistiku. Zagreb: Školska knjiga.
66. Pauše, Ž. (2003). Vjerojatnost, informacija, stohastički procesi. Zagreb: Školska knjiga.
67. Pavlič, I. (1971). Statistička teorija i primjena. Zagreb: Tehnička knjiga.
68. Petz, B., (2004). Osnovne statističke metode za nematematičare. Zagreb: Naklada Slap.
69. Pivac S., Aljinović Z., Tomić-Plazibat N. (2010). Risk Assessment of Transition Economies by Multivariate and Multicriteria Approaches, Panoeconomicus, No 3, pp 283-300.

70. Pivac S., Bodrožić I. and Jurun E., (2007). Chi-Square Versus Proportions Testing - Case Study on Tradition in Croatian Brand, Proceedings of the 9<sup>th</sup> International Symposium on Operational Research SOR'07, Slovenia, pp 415-421.
71. Pivac, S. and Jurun, E. (2005). Parameter Estimation in Excel, Proceedings of the 28<sup>th</sup> International Convention MIPRO 2005. Computers in Education, Opatija, pp 168-173.
72. Pivac, S. Jurun, E. and Jujnović, I. (2006). Ocjena parametara nelinearnih funkcija u programskom paketu Statistica na primjeru nezaposlenosti, Proceedings of the 29<sup>th</sup> International Convention MIPRO 2006. Computers in Education, Opatija, pp 241-245.
73. Pivac, S. and Rozga, A. (2006). Statistika za sociološka istraživanja, Sveučilište u Splitu, Filozofski fakultet, Split.
74. Pivac, S. and Rozga, A. (2008). Statističke analize socioloških istraživanja, Redak, Split.
75. Pivac, S. and Šego, B. (2005). Statistika, udžbenik sa zbirkom zadataka za IV razred srednje Ekonomske škole, Zagreb: Alkascript.
76. Rozga, A. (2003). Statistika za ekonomiste, Sveučilište u Splitu. Ekonomski fakultet Split.
77. Rozga, A. i Grčić, B. (2003). Poslovna statistika, Ekonomski fakultet Split.
78. Seber, G. A. F. Alan, J. L. (2003). Linear Regression Analysis. London: Wiley.
79. Seddighi, H. R., Lawler, K. A. and Katos, A. V. (2006). Econometrics, A practical approach, London and New York: Routledge.
80. Seplaki, L. (1991). Atorneys' Dictionary and Handbook of Economics and Statistics, New York: Professional Horizons Press.
81. Serdar, V. i Šošić, I. (1994). Uvod u statistiku. Zagreb: Školska knjiga.
82. Siegel, A. F. (1994). Practical Business Statistics. Boston, Massachusetts: IRVIN Publishing.
83. Studenmund, A. H. (2006). Using Econometrics, A Practical Guide. Boston, New York: Pearson International Edition.
84. Šošić, I. (1983). Metode statističke analize. Zagreb: Sveučilišna naklada Liber.
85. Šošić, I. (1985). Zbirka zadataka iz osnova statistike. Zagreb: Sveučilišna naklada Liber.

86. Šošić, I. and Serdar, V. (2002). Uvod u statistiku. XII. izdanje, Zagreb: Školska knjiga.
87. Šošić, I. (2004). Primijenjena statistika. Zagreb: Školska knjiga.
88. Tenjović, L. (2002). Statistika u psihologiji. Beograd: Centar za primenjenu psihologiju.
89. Terrell, D. (1992). Business Statistics for Management and Economics. Boston: Houghton Mifflin Company.
90. The Central Bureau of Statistics. Republic of Croatia(2006). [http:// www.dzs.hr](http://www.dzs.hr)
91. Vujković, T. (1976). Ekonometrijske metode i tehnike. Zagreb: Informator.
92. Žarković, S. S. (1965). Sampling Methods and Censuses. Rim:FAO
93. Weisberg, H.F., Krosnick, J.A. and Bowen, B.D. (2005). Survey research, Polling and Data Analysis. Third Edition. USA: Ohio State University. Assessment Systems Corporation.
94. Whigham, D. (1998). Quantitative Business Methods Using Excel. Oxford University Press.
95. Wonnacott, T. H. and Wonnacott, R. J. (1990). Introductory Statistics for Business and Economics (fourth edition). New York: Wiley.
96. The Central Bureau of Statistics. Republic of Croatia: URL: [http:// www.dzs.hr](http://www.dzs.hr)
97. The Eurostat: URL: <http://epp.eurostat.ec.europa.eu>