Dr Miloš Bešić

STATISTIKA U DRUŠTVENIM I POLITIČKIM ISTRAŽIVANJIMA

- PRIRUČNIK -

TESTIRANJE TEORIJA
ili
Čemu sva ta statistika?

Svekoliko znanje u nauci oblikovano je naučnim teorijama. Teorije predstavljaju skup logički povezanih simbola koji daju **objašnjenje** o tome šte se dešava u neposrednom iskustvu. Teorije su zapravo skup pojmova različitog stepena opštosti, pri čemu odnos između pojmova opisuje i objašnjava događaje u samoj stvarnosti. Teorije nisu manje ili više istinite, one mogu biti samo manje ili više korisne. Teorije nastaju u dugom periodu u svakoj nauci. One pružaju osnov za naučna istraživanja. Da bi sproveli naučno istraživanje nužno je da formulišemo **konceptualni okvir** koji predstavlja prilagođavanje teorije konkretnim društvenim uslovima u kojima se istraživanje odvija. U procesu izgradnje konceptualnog okvira uobičajene su sledeće faze:

- Koristimo indukciju kako bi na osnovu posmatranja činjenica formulisali pretpostavke
- Koristimo dedukciju kako bi forulisali predikcije
- Testiramo ova predviđanja na način što ih uporežujemo sa novim opservacijama
- Dopunjujemo ili preurežujemo naše pretpostavke kako bi one bile konzistentne sa rezultatima obaljenih opservacija

Osnovni elementi svake teorije jesu pojmovi. Koncepti jesu reči ili simboli koji reprezentuju neku ideju. Svaki koncept se sastoji iz:

- termin tj. sama reč
- ekstenzija tj. klasa pojava koja je pojmom obuhvaćena
- referans tj. svojstva koja tu klasu pojava karakterišu
- značenje tj. odnos između termina i denotata (stvarnosti)

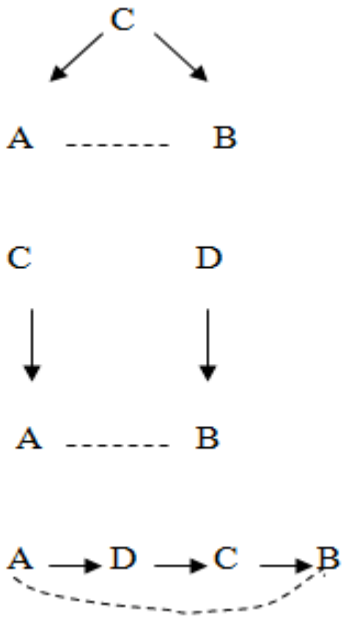
Budući da se teorija bavi stvarnošću, pojmovi u osnovi obuhvataju **klasu objekata** koji mogu biti predmet posmatranja. Dakle, u empirijskom smislu, pojmovi se svode na **opservacije**, što znači da je moguće čulima posmatrati određene karakteristike fenomena. Konsekventno, uslov za realizaciju empirijskog istraživanja jeste precizna identifikacija **empirijskog referansa** koji su pojmovima obuhvaćeni. Teorija i pojmovi su korisni ako nam pomažu da objasnimo stvarnost. U okviru teorije se uspostavljaju veze i odnosi između pojmova, jer je to jedini način da se opisuju i objašnjavaju veze i odnosi u predmetnoj stvarnosti. Ovi iskazi, koji omogućuju uspostavljanje veza između pojmova zovu se **propozicije**. U osnovi, propozicije predviđaju dve ključne moguće veze između pojmova a to su odnosi **kovarijacije** i **kauzaliteta**.

Odnosi kovarijacije nam govore o tome da dva (ili više) pojmova se menjaju zajedno. Kako se jedan povećava (ili smanjuje) i drugi se povećava (ili smanjuje). Ovi odnosi nam ništa ne govore o tome šta je uzrok a šta posledica. Npr. mi možemo utvrditi da postoji odnos kovarijacije između partijske identifikacije i glasanja, tj. da

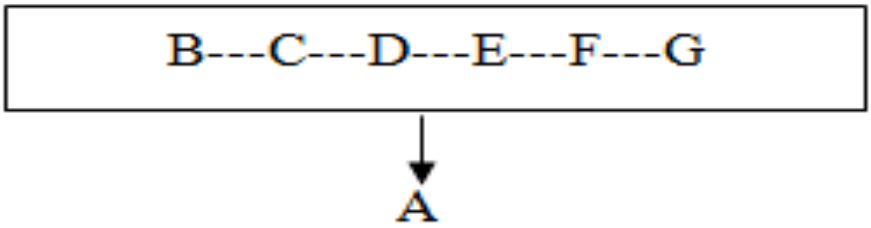
se u stvarnosti visok nivo partijske identifikacije i izlazak na birališta dešavaju zajedno. Drugim rečima, odnosi kovarijacije kažu da povećanje stepena partijske identifikacije i povećanje izlaznosti 'idu zajedno' ali nam ne govore ništa o uzročno-posledičnim vezama, ili preciznije, ova propozicija nam ne govori da je partijska identifikacija 'uzrok' glasanja.

Kauzalni odnosi postoje onda kada promene u jednom konceptu proizvode promene u drugom ili drugim konceptima. Npr., propozicija koja ima kauzalni karakter između dva pojma kada glasi: 'Što je veči stepen partijske identifikacije, veća je i verovatnoća izlaska na birališta'. Dakle, definisana je uzročno posledična veza između dva pojma na način da propozicija kaže da partijska identifikacija jeste uzrok glasanja. Osećaj snažne identifikacije, po ovoj propoziciji, jeste osnov da neko izađe na birališta i glasa. Međutim, verovatnoća glasanja ne utiče na nivo partijske identifikacije, tj. ova propozicija je kauzalna zato što ne kaže da je moguć obrnut odnos između uzroka i posledice.

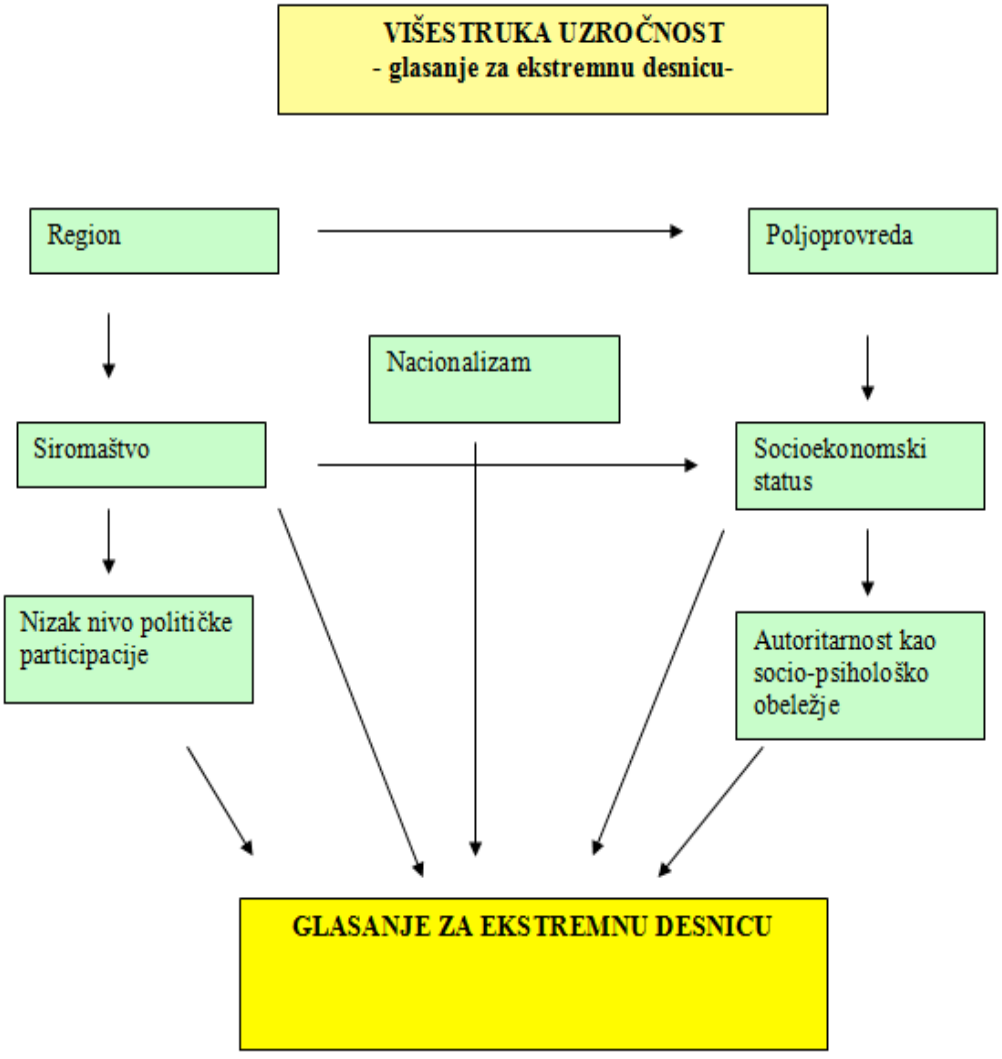
Prividna povezanost je neretka situaciju u kojoj se npr. A i B zaista pojavljuju zajedno, ali između njih ne postoji niti kovarirajući niti kauzalni odnos. Ovi odnosi najčešće jesu proizvod sticaja okolnosti i oni značajno zamagljuju sliku kada otkrivamo uzročno posledične odnose . Npr. ako A i B jesu povezano prividno onda je je to proizvod različitih mogućnosti od kojih ćemo grafički prikazati nekoliko uobičajenih:



U društvenim naukam se retko govori o uzrocima i posledicama na način da uvek i samo jedan 'uzrok' proizvodi uvek i samo jednu 'posledicu'. Ono što je praksa u društvenim naukama jeste takozvana **višestruki kauzalitet** (multiple causation). Naime, u društvu najčešće je slučaj da veći broj društvenih pojava u međusobnoj interakciji jesu uzrok nekoj konkretnoj posledici. Drugim rečima, nije moguće govoriti u kategorijama jednog jedinog uzroka, već govorimo o jednom lancu uzročnosti koji podrazumeva čitav niz uslova da bi se posledica pojavila:



Višestruki kauzalitet primer



Propozicije mogu izražavati veze koje mogu biti **pozitivne** i **negativne**. Ovo važi i za kovarirajuće i kauzalne odnose. Dakle, pojmovi se mogu kretati proporcionalno ili obrnuto proporcionalno. Primer **pozitivne povezanosti** bi bio: 'Što je veći stepen obrazovanja glasača, veća je verovatnoća da će oni glasati'. Primer negativne povezanosti bi bio: 'Što je veći stepen obrazovanje glasača, manja verovatnoća da će oni glasati'

Drugim rečima, teorija mora da specificira da li je reč o pozitivnim (+) ili negativnim (-) odnosima između pojmova.

Testiranje teorija ima za cilj procenu vrednosti objašnjenja koje sama teorija nudi. Teorije su apstrakcije koje objašnjavaju stvarnost, i ključni faktor testiranja teorije jeste uspostavljanje 'saradnje' između apstraktnog i konkretnog sveta. Teorije, koje predstavljaju skup pojmova, pretpostavki i propozicija, nikada ne mogu biti konačno potvrđene ili opovrgnute.

U procesu testriranje teorije, uloga hipoteza je **nezamenljiva**. Hipoteza je **sredstvo** koje ima funkciju testiranja teorije. Hipotezu je nužno oblikovati na način da ona odgovara konkretnim uslovima jer je ovo jedini način da se obezbedni konceptualna adekvatnost i konzistentnost između teorije i opservacija.

Hipoteze predstavljaju iskaze o tome šta mi mislimo o tome šta se događa u samom društvenom životu. Hipoteze nam govore o tome šta u svarnosti trebamo očekivati kada na adekvatan način organizujemo opservacije. Hipoteze su deklarativne rečenice u kojima se izražava očekivanje o povezanosti između pojava koje su obuhvaćene konceptima. Evo nekoliko primera:

Što je veći (duži, viši, manji...) ***x****, to je veći (viši, duži, manji...)* ***y****.*

ili konkretno na primeru:

Što je veći stepen obrazovanja glasača, to je veća izlaznosti na izborima.

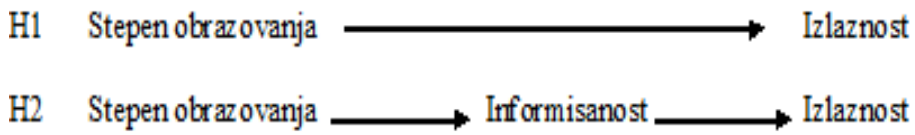
Ova hipoteza izražava **odnos kovarijacije** između pojava. Ovo zato što hipoteza ne govori o tome da je stepen obrazovanja 'uzrok' veće izlaznost na izborima, već jednostavno izražava 'povezanost' između dve pojave. Ukoliko želimo da obezbedimo empirijsku evidenciju za potvrdu naše hipoteze, nužno je da sa nivoa teorije 'siđemo' u svarnost. Ovo znači da moramo na jedan metodičan način da organizujemo opservacije, a za ovo je važno da usvojimo koncept **varijable**. Varijable, kao što znamo, imaju različite metrijske karakteristike, ali sve one imaju određen broj vrednosti. Limitiran broj vrednosti varijable nam omogućava da 'posmatramo' i merimo one aspekte stvarnosti koje su predmet našeg naučnog interesovanja.

Kako bi omogućili empirijsko testiranje teorije posredstvom hipoteza nužno je da pojmove transformišemo u varijable. Pritom, neke pojmove možemo relativno lako transformisati u empirijski prihvatljive jedinice posmatranja (varijable) dok je sa drugim pojmovima to teže. Npr. pojam 'pluralizam' je veoma važan u političkoj teoriji, ali je tranformacija ovog koncepta u varijable prilično složena, obzirom da nije sasvim jasno koje sve empirijske aspekte ovaj pojam podrazumeva. Ukoliko želimo da ovaj pojam uključimo u bilo koju hipotezu, nužno je da obavimo složen posao transformacije ovog pojma u nekoliko varijabli koje imaju jasan i precizan empirijski referans. Tako npr. možemo uzeti kao jednu od varijabli 'broj organizovanih

interesnih skupina' u jednom društvu kao jednu od varijabli koja reprezentuje pluralizam. Kako bi bili uvereni da ova varijabla (broj organizovanih interesnih skupina), doista predstavlja pojam 'pluralizma', ona bi morala biti u korelaciji sa još nekoliko varijabli koje predstavljaju empirijske opservacije istog pojma ('broj civilnih inicijativa' ili 'uvažavanje stavova manjine od strane većine u parlamentu' npr.). Na ovaj način, na osnovama povezanosti nekoliko varijabli, koje empirijski obuhvataju one aspekte pojava koje konceptualno predstavljaju pluralizam, mi povećavamo stepen pouzdanosti koji opravdava naše merenje pluralizma.

U procesu testiranja hipoteze nužno je varijable staviti u odnos tzv. **zavisne** i **nezavisne varijable.** Ono što znamo jeste da svaka varijabla ima više od dve vrednosti i da ove vrednosti mogu da variraju. U cilju testiranja hipoteza, dve varijable se moraju dovesti u vezu na način da promena vrednosti na jednoj varijabli odgovara promena vrednosti na drugoj varijabli. Pri tome, ona varijabla čije se vrednosti menjaju kao ishod promena vrednosti na drugoj varijabli zove se **zavisna varijabla.** Ovo zato što vrednosti ove varijable 'zavise' od vrednosti druge varijable. Naspram tome, varijabla koja utiče na promene vrednosti druge varijable zove se **nezavisna varijabla**. Ovo zato, što su promene vrednosti na ovoj varijabli uslov promena vrednosti na drugoj varijabli. Pri tom, i ovo je jako važno, **da li je jedna varijabla zavisna ili nezavisna zavisi od dizajna istraživanja i hipoteza**. Drugim rečima, ne postoji univerzalni vanteorijski kriterijum na osnovu koga neke varijable jesu zavisne a druge nezavisne.

Ukoliko želimo da precizno validiramo odnos između dve varijable ključno je važno da vodimo računa o tzv. **posredujućim varijablama.** Odnos između zavisne i nezavisne varijable neretko je posredovan nekom trećom varijablom. Ove varijable u osnovi ukazuju na prirodu veze između zavisne i nezavisne varijable, i ovo je ključni razlog zbog kojeg se mi moramo baviti ovim varijablama. Identifikacija posredujućih varijabli produbljuju naše razumevanje o prirodi veze između dva fenomena, i na taj način pružaju dodatne informacije na osnovu kojih je naše objašnjenje kvalitetnije. U našem primeru veze između obrazovanja i glasanja, npr. posredujuća varijabla može biti 'informisanost o političkim zbivanjima'. To bi shematski izgledalo ovako:



Dakle, recimo da je pretpostavljena veza između stepena obrazovanja kao nezavisne i izlaznosti kao zavisne varijable potvrđena (H1: potvrđeno). Međutim, daljim ispitivanjem, smo utvrdili da postoji veza između informisanosti i izlaznosti, kao i veza između stepena obrazovanja i informisanosti. Tako smo došli do saznanja da je informisanost o političkim dešavanjima posredujuća varijabla između nezavisne i zavisne varijable. Time smo došli do veoma važnih informacija koje značajno doprinose kvalitetu našeg potencijalnog objašnjenja, naime, identifikacijom posredujuće varijable mi možemo reći da je uzrok veće izlaznosti, pre svega, stepen političke informisanosti, a ne stepen obrazovanja po sebi. Međutim, jednako smo utvrdili da od stepena obrazovanja zavisi stepen političke informisanosti i time je ova tročlana relacija uspostavljena na način da smo identifikovali vezu između nezavisne

i zavisne varijable između kojih se nalazi posredujuća varijabla. Sada, na osnovu novog saznanja možemo postaviti sledeću predikciju:

Stepen obrazovanja će imati pozitivnu korelaciju sa izlaznošću u situaciji kada je visok nivo političke informisanosti.

Za razliku od posredujućih varijabli, neretko, na prirodu i vezu između zavisne i nezavisne varijable utiče tzv. antecedentna varijabla. Pod ovim se razume da nezavisna varijabla u osnovi 'zavisi' od neke treće varijable, i ukoliko je ovo slučaj, ovu treću varijablu zovemo antecedentnom varijablom. Npr., uzmimo hipotezu da od stepena identifikacije sa partijom zavisi verovatnoća izlaska na birališta. Identifikujući 'partijsku identifikaciju' kao nezavisnu i 'izlaznost' kao zavisnu varijablu, možemo formulisati hipotezu i izraziti je na sledeći način:

Partijska identifikacija ispitanika ⟶ *Izlaznost*

Sada, pretpostavimo da smo na osnovu empirijske evidencije potvrdili vezu između nezavisne i zavisne varijable, ali jednako, pretpostavimo da smo pronašli vezu između partijske identifikacije naših ispitanika i partijske identifikacije njihovih roditelja. Drugim rečima, partijska identifikacija kao varijabla, zavisi od partijske identifikacije roditelja naših ispitanika. Tako možemo uspostaviti sledeću relaciju:

Partijska identifikacija roditelja → *Partijska identifikacija ispitanika* → *Izlaznost*

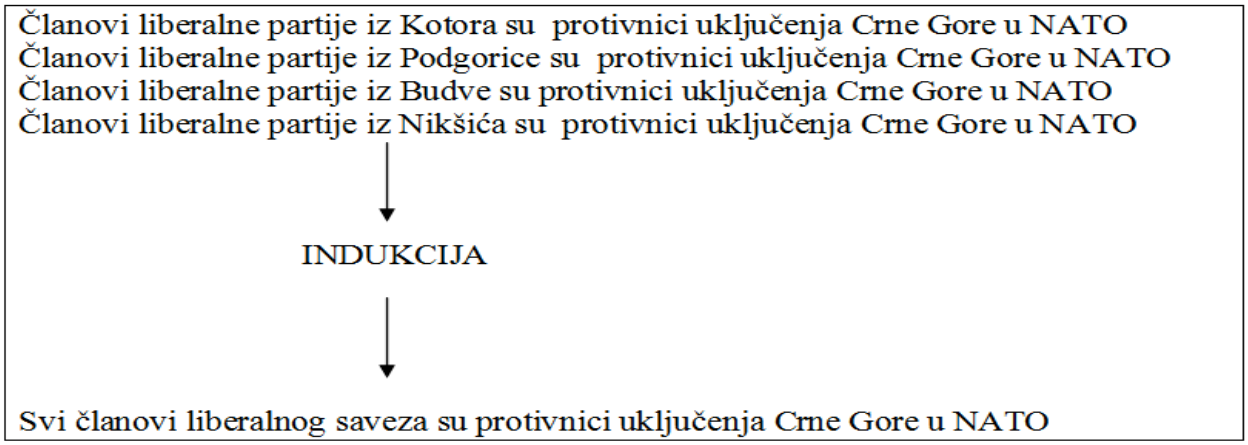
U ovoj situaciji, kažemo, da je partijska identifikacija roditelja **antecedentna varijabla**. Dakle, identifikacijom antecedentne varijable mi smo dodatno doprineli kvalitetu našeg objašnjenja analizirajući veze između pojava. Dalje, onog trenutka kada identifikujemo vezu između antecedentne, nezavisne i zavisne varijable, ukoliko želimo da dokažemo pretpostavljeni kauzalni lanac mi moramo na osnovu empirijske evidencije da dokažemo da ukoliko ne postoji visoka partijska identifikacija roditelja, neće biti ni visoke partijske identifikacije glasača. Sve u svemu, otrivanje antecedentnih i posredujućih varijabli veoma je važan posao u otkrivanju **kauzalnog lanca** koji je veoma važan sa stanovišta našeg istraživačkog pitanja. Prema tome, hipoteze predstavljaju ključni instrument za testiranje teorija. One omogućavaju posredovanje između teorijskih premisa i iskustva, na takav način da se testiraju veze i odnosi između onih jedinica opservacija koje predstavljaju empirijski referans za koncepte same teorije.

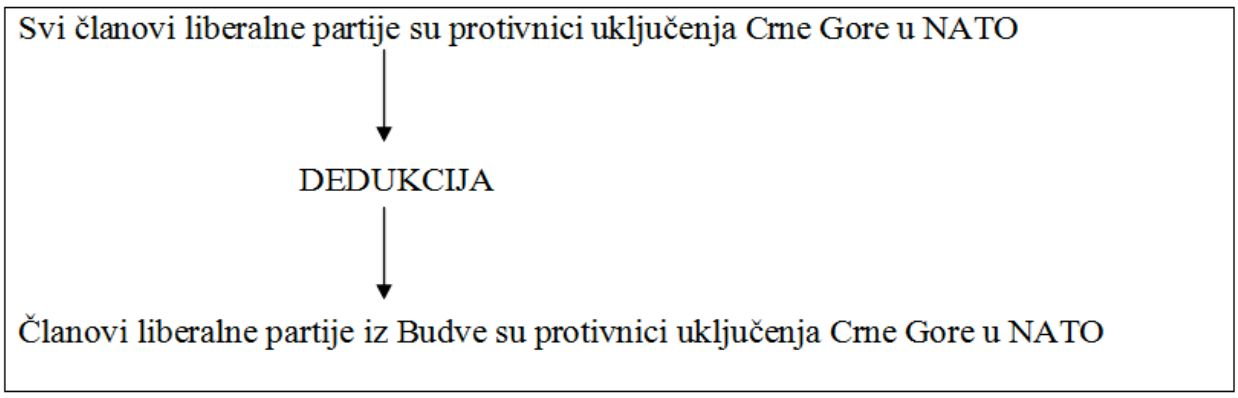
Do hipoteza se dolazi induktivnim ili deduktivnim putem. Da li ćemo pritom koristiti indukciju ili dedukciju, zavisi od toga do kog smo stepenika došli u istraživačkom procesu. Ukoliko smo još uvek u fazi formulisanja teorije po principu pokušaja i pogrešaka, onda je verovatno najbolji način da formulišemo hipoteze metod **induktivne generalizacije**. Npr. na osnovu analize sekundarnih podataka uvideli smo da izlaznost na izbore zavisi od ekonomske razvijenosti regiona. Na osnovu ovoga možemo postaviti hipotezu koja upravo dovodi u vezu odnos između ekonomske razvijenosti (nezavisna varijabla) i izlaznosti na izborima (zavisna

varijabla). Kasnije, ako buduće opservacije i empirijska evidencija idu u prilog postavljenoj hipotezi, mi ćemo imati više poverenja da postoji veza između ekonomske razvijenosti i glasanja. Međutim, sve dok ne generišemo teoriju koja objašnjava ***zašto*** postoji veza između ekonomske razvijenosti i izlaznosti mi ne možemo koristiti povezanost između ovih činjenica kao objašnjenje za političku participaciju.

Onog trenutka kada definišemo teoriju na način da povežemo varijable u jedan koherentan sistem, postaje moguće da izvedemo hipoteze iz te teorije koristeći *dedukciju*. Važno je imati u vidu da je dedukcija proces u kome su već poznate informacije sadržane u teoriji stavljaju u eksplicitnu formu. Mi na osnovu dedukcije ne možemo saznati ništa novo o vezama i odnosima među pojavama. Indukciju koristimo, prema tome, da bi hipoteze koje smo precizno konceptualno definisali poslužile kao instrument za testiranje same teorije, ili tačnije, za testiranje informacija o povezanosti između pojava koje su sadržane u hipotezi. Dedukciju prevashodno koristimo kako bi pojasnili implikacije koje se mogu izvesti iz teorije, a nakon ovih pojašnjenja postaje moguće precizno formulisanje hipoteza.

Primer:



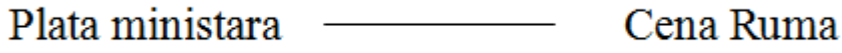


Pored direktnog načina testiranja teorije, moguće je ovaj postupak obaviti i na indirektan način. Ovaj način podrazumeva preces u kome formulišemo i testiramo **alternativne suprotne hipoteze**. Jedan isti događaj je moguće objasniti na više načina

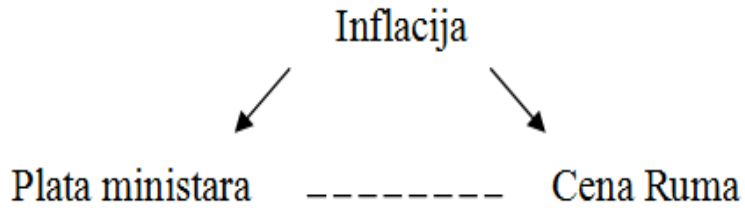
Nekada objašnjenja jesu konzistenta jedna sa drugima, pri čemu više od jednog objašnjenja može biti ispravno. Međutim, objašnjenja koja se nude mogu biti opozitna, ili, drugim rečima, ako je jedno objašnjenje tačno, drugo nužno nije tačno i

obrnuto. To znači da mi možemo formulisati hipoteze koje nude objašnjenja koja su međusobno isključiva. Ova procedura podrazmeva da pored osnovne formulišemo i **alternativnu hipotezu** a to je ona hipoteza čije potvrđivanje isključuje mogućnost da je prvotna hipoteza tačna. Ove hipoteze su alternativne zato što one na drugačijim osnovama nude objašnjenje događaja. One su, dalje, suprotne zato što ne može biti da su istinite obe hipoteze, jedna drugu isključuje

Recimo npr. da smo obavili eksplorativno istraživanje kojim smo ustanovili da postoji statistička povezanost između cene ruma i povećanja plate ministara. To izgleda ovako:



Ključna alternativna hipoteza, u ovom slučaju, bila je da su i cena ruma i plata ministara jesu posledica delovanja nekog trećeg faktora, npr. inflacije:



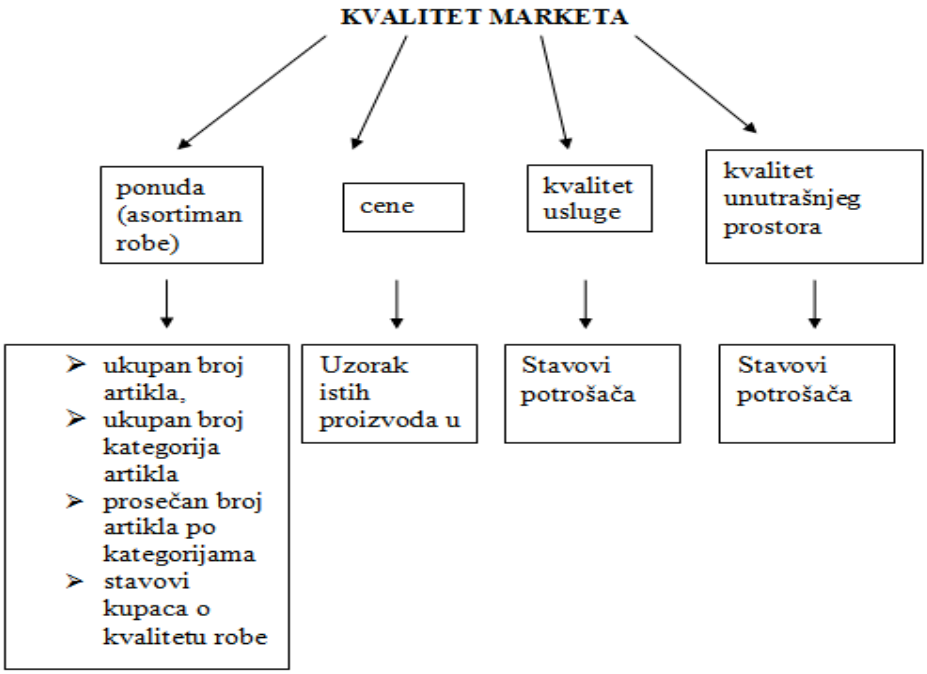
Ako potvrdimo alternativnu hipotezu, to znači da će povezanost između cene ruma i plate ministara da nestane ako ovu vezu 'kontrolišemo' inflacijom. Ovo je statistička procedura koja je relativno jednostavna i koja podrazumeva ispitivanje odnosa između dve varijable; u ovom slučaju cene ruma i plate u situaciji kada držimo konstantnu vrednost inflacije. Prema tome, potvrđivanje alternativne hipoteze osporiće prvotnu korelaciju koju smo postavili i ponudiće validno objašnjenje za događaj o kome je reč

Uslov da se istraživanje realizuje jeste da naši koncepeti budu pripremljeni na takav način da je moguće u stvarnosti identifikovati empirijske opservacije. Ovo znači da za svaki pojam mi moramo naći empirijske indikatore koji jesu 'predstavnici' naših koncepata. Ovako se u najkraćem može opisati proces i značaj **operacionalizacije**. Operacionalizovati, prema tome, znači pronaći fenomene i aspekte fenomena koji su empirijski po svom karakteru i koje je kao takve moguće svesti na kvantitativna obeležja. U isto vreme, empirijski indikatori u osnovi moraju biti jasno i nedvosmisleno povezani sa pojmovima.

U našem gradu postoji veliki broj marketa. Neki tvrde da su Carine najbolje, drugi da je Voli a treći preferiraju male markete. Način da se rasprava okonča jeste formiranje jednog pojma koji radno možemo nazvati 'kvalitet marketa'. Ukoliko želimo da zaista uporedimo markete, nužno je da operacionalizujemo pojam *kvalitet marketa.*

Ovo znači da moramo pronaći konkretne empirijske indikatore koji se mogu kvantifikovati a koji na jedan validan način 'predstavljaju' pojam 'kvalitet marketa'.

Moguća operacionalizacija:

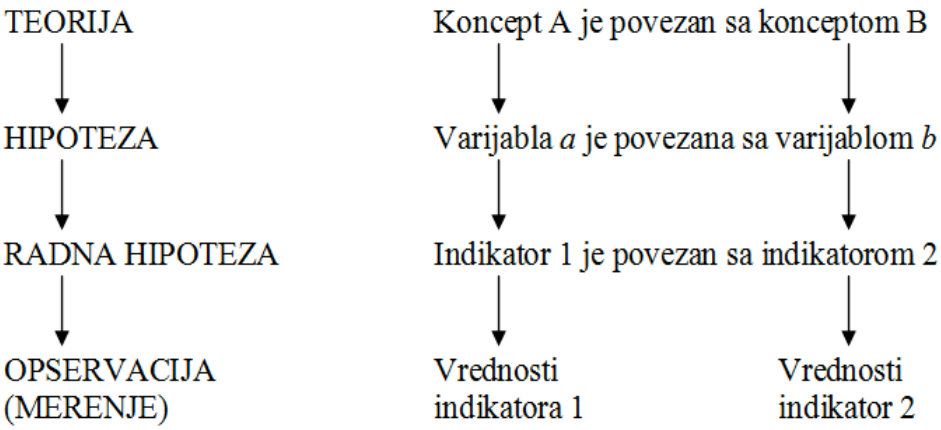


Zadatak operacionalizacije jeste da apstraktne pojmove prevedemo na empirijske indikatore koji se mogu kvantifikovati. Operacionalizacija je jedna od ključnih faza u istraživačkom procesu. Ovaj postupak je jedini način da apstraktni karakter teorije prevedemo na empirijski jezik i na taj način ispitamo naše hipoteze. Tri su ključna elementa procesa operacionalizacije. Prvo, to je sama **operacionalizacija** koja predstavlja izbor fenomena koji predstavljaju koncepte a mogu biti predmet posmatranja. Drugo, specifikacija svih koraka koje je potrebno preduzeti kako bi se realizovale opsrervacija naziva se **instrumentalizacija**. Treće, primenom instrumenta slučajevi dobijaju numeričke vrednosti i ovim se realizuje **merenje**. Rezultati merenja jesu naša konačna evidencija na osnovu koje donosimo odluke i odgovaramo na istraživačko pitanje

Evo primera iz prirodnik nauka. Recimo da hoćemo da istražimo uticaj hemikalije X na rast pšenice. ***Rast*** *je* apstraktan pojam koji se ne može kao takav meriti, i koji se prema tome mora prevesti na neku varijablu koju je moguće direktno posmatrati i meriti uticaj nove hemikalije upravo na ovu varijablu. Varijabla koja meri rast je npr. ***postignuta visina pšenice.*** Ukoliko želimo precizno da kvantifikujemo visinu pšenice, onda moramo da razvijemo instrument koji ima jasne metrijske karakteristike i koji je u stanju da precizno izrazi vrednosti same varijable. Za ovu svrhu su nam potrebni indikatori a u konkretnoj situaciji, indikator za visinu pšenice biće visina u centimetrima. Mi na osnovu samog instrumenta pripisujemo određene numeričke vrednosti varijabli. Sada je svakako mnogo jednostavnije korišćenjem ovog instrumenta da uporedimo visinu pšenice u slučajevima kada smo koristili hemikaliju x i visinu pšenice gde nismo koristili ovu hemikaliju. Poređenje vrednosti ove dve varijable će nam mnogo preciznije govoriti o tome da li i koliko

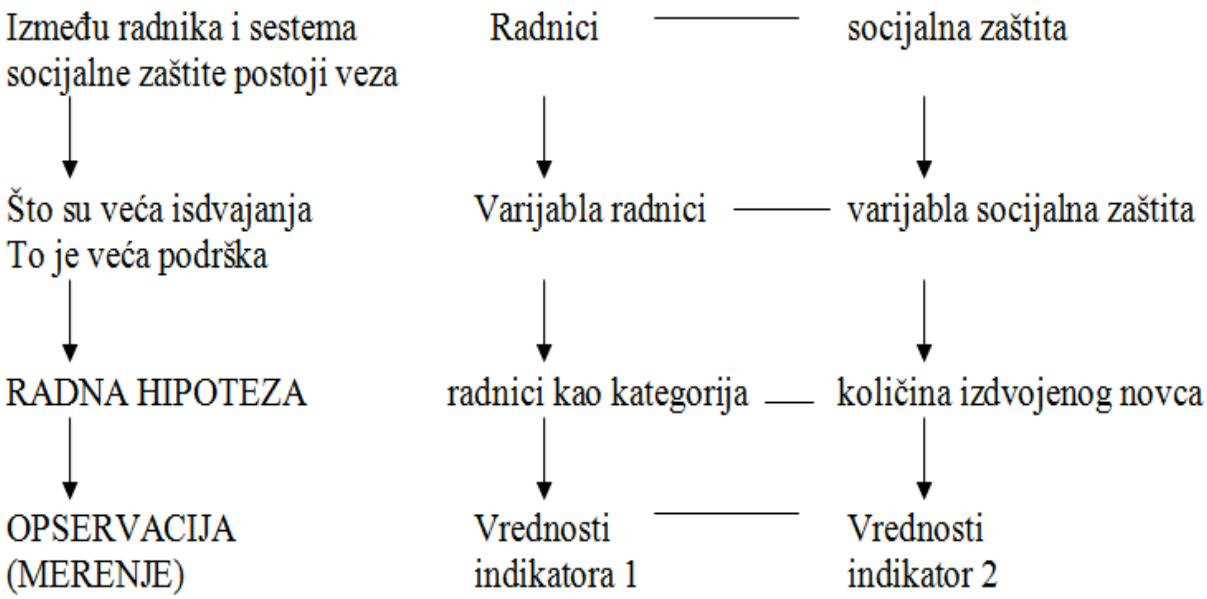
hemikalija x ima uticaj na rast pšenice. Pod opservacijama u istraživanju podrazumevamo proces primene instrumenta koji poseduje određene metrijske karakteristike sa ciljem da pripišemo vrednosti nekoj karakteristici pojave koja je predmet našeg istraživanja. U društvenim naukama, opseravacija je znači korišćenje instrumetna kako bi merili različite oblike ponašanja. **Operacionalizacija** i **merenje** jesu ključni elementi istraživanja. Mi nikada u praksi ne možemo da poredimo pojmove. Jedini način da to bude moguće jeste da poredimo vrednosti na varijablama koje su rezultat operacionalizacije i primene instrumenata koji imaju određene merne karakteristike. Ono što poredimo, prema tome, nisu koncepti, već njihovi **indikatori.** Komparacija može biti precizna ako i samo ako sami indikatori predstavljaju empirijsko ogledalo samih koncepata koji su predmet našeg merenja. Ukoliko koncepti nisu valjano operacionalizovani, međusobne veze i odnosi između indikatora neće precizno reflektovati odnose između koncepata koje ti indikatori reprezentuju. Kao rezultat, možemo doći do pogrešnih zaključaka koji imaju teorijski karakter.

SHEMA:



Da bi hipoteze 'oživele' i da bi obavile svoju epistemološku funkciju, moraju biti postavljeni u jedan operativan oblik. Ovakve hipoteze nazivamo **radnim hipotezama**. Radna hipoteza pretpostavlja povezanost između indikatora varijabli. Opservacije koje u osnovi jesu proces merenja, ispituju odnos između vrednosti indikatora obe varijable.

Evo jednog primera, pretpostavimo teorijski da bi radnici podržali onaj sistem socijalne zaštite koji podrazumeva veći procenat budžeta za socijalna davanja. Pretpostavimo zatim da je donesen nov Zakon u Skupštini koji podrazumeva veći udeo u budžetu za socijalna davanja. Teorijski, jednako, pretpostavljamo da preduzetnici i vlasnici kapitala ne podržavaju ovaj Zakon, budući da veći socijalni izdaci podrazumevaju veće poreze koje vlasnici kapitala ne žele da plate. Način da testiramo, dakle, hipotezu da radnici podržavaju novi sistem socijalnog budžetiranja je sledeći:



Važno je znati da proces operacionalizacije počiva na *redukciji*. Ovo znači da mi *moramo da simplifikujemo pojmove kako bi ih preveli na empirijski merljive indikatore*. Budući da indikatori nisu u stanju da reflektuju celinu koncepata, proističe da neka značenja koncepata ostaju van istraživačkog procesa, i ona nisu predmet merenja. Zadatak 'dobre' operacionalizacije jeste da ovi gubici budu što manji. Od ovog procesa bitno zavisi vrednost indikatora i kvalitet samog merenja.

Početni korak u pronalaženju valjanih indikatora za merenje neke društvene pojave jeste formulisanje **operacionalih definicija** koje treba da nam omoguće identifikaciju samih indikatora. To znači da pojmove koje kanimo da operacionalizujemo moramo definisati na jedan operacionalan način koji nam omogućava da identifikujemo same indikatore. Kako bi bile korisne, a to znači da moraju biti validne i pouzdane, operacionalne definicije nam moraju reći veoma precizno šta i kako treba da radimo kako bi uspeli da odredimo koje kvantitativne vrednosti bi morale biti povezane sa varijablom u svakom pojedinom slučaju. Tri su osnovna razloga zbog kojih su nam potrebne precizne operacionalne definicije:

- **Prvo**, mi želimo drugima da saopštimo tačno šta je to što smo uradili u procesu merenja, jer je ovo jedini način da oni mogu da procene naš rad i eventualno ponove studiju kako bi verifikovali naše istraživanje u drugim uslovima
- **Drugo**, budući da je u istraživanje uključen veći broj istraživača, samo precizno definisanje će obezbediti da svi istraživači sprovode identičnu proceduru, a ovo je fundamentalan uslov za preciznost merenja
- **Treće**, precizni i detaljni iskazi o tome kako smo operacionalizovali varijable će pomoći nama samima da valjano procenimo dobijene rezultate i da se obračunamo sa suprotstavljenim objašnjenjima i nalazima drugih istraživača koji su došli do drugačijih rezultata u odnosu na naše

U procesu formulisanja operacionalnih definicija, neophodno je da pismenim putem opišemo sve postupke i procedure koje smo koristili sa ciljem definisanja merila koje ćemo upotrebiti u istraživanju. Ovo nije samo zato da bi smo ostavili zapis svih koraka koje smo preduzeli u istraživačkom procesu, već prvenstveno zato

što na ovaj način imamo priliku da kritički promislimo sve poteze, uočimo slabosti, i smanjimo eventualne greške i propuste koji se mogu pojaviti. Glavni cilj operacionalih definicija jeste precizno uspostavljanje veze između koncepata i stvarnosti, ili tačnije, jasna delimitacija empirijskog referansa. Npr. Ako istražujemo ***stavove radnika o sistemu socijalne zaštite***, npr., u ovoj situaciji imamo dva ključna pojma koji se moraju operacionalno definisati. Prvo, to je pojam *radnika*
Operacionalna definicija mora jasno da odredi na koje to ljude koje označavamo radnicima mi zaista mislimo. Da li ćemo uzeti radnike bez obzira na stručnu spremu? Da li ćemo da uzmemo radnike koji rade i u privatnmom i u javnom sektoru? Da li se pod radnicima podrazumevaju oni koji obavljaju radničke poslove, ili oni koji su radnici po zanimanju bez obzira koji posao obavljaju? Da li ćemo uzeti i radnike koji su zapošljeni i one koji su nezapošljeni? Ako odlučimo da uzmemo u obzir i zapošljene i nezapošljene radnike, da li ćemo da pravimo razliku između nezapošljenih koji traže posao u struci i one koji ne žele da se zaposle kao radnici? Kako ćemo se odnosit prema radnicima koji su recimo trenutno tehnološki višak i primaju socijalnu naknadu? Šta ćemo sa radnicima koji trenutno završavaju neku višu školu, da li ćemo ih tretirati jednako kao ostale radnike? Drugim rečima, **šta tačno i precizno podrazumevamo pod radnicima a šta ne podrazumevamo imajući u vidu sve moguće situacije u stvarnosti. To je zadatak operacionalih definicija**.

U istraživačkoj praksi, jedan indikator najčešće nije dovoljan da isrcpi sve dimenzije i značenja jednog pojma u našem primeru iz prirodnih nauka veoma brzo bi zaključili da *postignuta visina pšenice* mereno u *centimetrima* nije dovoljan indikator za *rast pšenice*. Naime, rast podrazumeva *težinu zrna pšenice* a ne samo visinu. Ovaj problem je u društvenim naukama još izraženiji, ili tačnije, koncepti u društvenim i političkim naukama su po pravilu složeniji i imaju **veći broj dimenzija** (aspekata) i značenja. Prema tome, jedan indikator jednostavno nije dovoljan za operacionalizaciju ovih koncepata već je potrebno da se identifikuju sve dimenzije i značenja a zatim da se izabere veći broj indikatora koji zajedno operacionalizuju koncept koji je u pitanju. Na primer, ako pojam *demokratije* operacionalizujemo samo posredstvom jedne dimenzije: održavanje izbora, veoma brzo ćemo zaključiti da ćemo veliki broj nedemokratskih autoritarnih režima podvesti pod demokratske sisteme.

Evo primera operaconalizacije socio-ekonomskog statusa, naime ovo je pojam za koji koristimo više indikatora:

- Ukupan prihod svih članova domaćinstva
- Obrazovanje
- Zanimanje
- Posedovanje stambenog prostora koji je u porodičnom vlasništvu
- Površina stana/ kuće
- Broj članova domaćinstva
- Upravljačka ovlašćenja

U istraživačkoj praksi, ovaj pojam se operacionalizuje preko sledećih varijabli:

- Prihod izražen intervalnom skalom u valuti
- Broj završenih godina školovanja
- ISCO88 kodovi za zanimanja
- Vlasnik vs. Nije vlasnik stambenog prostora

- Broj kvadratnih metara po članu domaćinstva
- Dvovalentno: Ima-nema upravljačka ovlašćenja
- Broj ljudi nad kojima ima upravljačka ovlašćenja

Evo prikaza u tabelama koji govori o tome kakav je socioekonomski status crnogorskih građana kada se primeni ovakav metodološki pristup:

PRIHOD

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No income	48	3,2	3,4	3,4
	Less than 100 EUR	183	12,1	12,9	16,3
	100 to 149 EUR	173	11,4	12,2	28,6
	150 to 199 EUR	137	9,1	9,7	38,3
	200 to 249 EUR	171	11,3	12,1	50,4
	250 to 299 EUR	132	8,7	9,3	59,7
	300 to 349 EUR	122	8,0	8,6	68,3
	350 to 399 EUR	70	4,6	4,9	73,2
	400 to 449 EUR	114	7,5	8,0	81,3
	500 to 549 EUR	90	5,9	6,3	87,6
	550 to 599 EUR	44	2,9	3,1	90,7
	more then 600 EUR	131	8,7	9,3	100,0
	Total	1414	93,3	100,0	
Missing	Don't know/ Refused/ No answer	102	6,7		
Total		1516	100,0		

V115. If you supervise anyone, how many people do you supervise?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	19	1,2	9,2	9,2
	2	21	1,4	10,1	19,3
	3	10	,7	5,0	24,3
	4	16	1,0	7,6	31,8
	5	15	1,0	7,3	39,1
	6	18	1,2	8,5	47,7
	7	4	,3	2,0	49,7
	8	7	,4	3,3	52,9
	9	2	,1	1,1	54,0
	10	11	,7	5,1	59,1
	12	4	,2	1,8	60,9
	13	1	,0	,3	61,2
	14	1	,1	,7	61,9
	15	7	,5	3,4	65,3
	16	2	,1	,9	66,2
	17	3	,2	1,3	67,4
	18	2	,1	,7	68,2
	20	11	,8	5,6	73,7
	22	1	,1	,4	74,1
	25	1	,1	,5	74,6
	30	8	,5	3,8	78,4
	35	2	,1	,8	79,1
	40	2	,1	,9	80,1
	50	5	,4	2,6	82,7
	60	1	,1	,5	83,2
	90	1	,0	,3	83,5
	98	1	,1	,7	84,2
	99	10	,7	5,1	89,3
	100	3	,2	1,5	90,8
	150	2	,2	1,1	91,9
	200	3	,2	1,3	93,3
	250	1	,1	,4	93,6
	300	2	,2	1,1	94,8
	386	1	,1	,7	95,4
	400	1	,1	,5	95,9
	700	1	,1	,7	96,6
	790	1	,0	,3	96,9
	More than 996	6	,4	3,1	100,0
	Total	206	13,6	100,0	
Missing	Don't know	7	,5		
	Refusal	1303	85,9		
	Total	1310	86,4		
Total		1516	100,0		

SUpervizija

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	246	16,2	24,9	24,9
	No	739	48,7	75,1	100,0
	Total	984	64,9	100,0	
Missing	no answer	531	35,1		
Total		1516	100,0		

V113. Respondent's occupational group for respondent's main job

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Higher level manager or administrator/ not an owner	10	,6	1,0	1,0
	Middle level manager or administrator	13	,9	1,3	2,3
	Lower level manager or administrator	15	1,0	1,5	3,8
	Politician	2	,1	,2	4,0
	Entrepreneur	32	2,1	3,3	7,3
	Higher professionals/ positions requiring university degree	115	7,6	11,7	19,0
	Technicians in health, production, and sciences, nurses	38	2,5	3,8	22,8
	Technicians in education, humanities, social sciences, etc.	16	1,0	1,6	24,4
	Supervisors of clerks	31	2,0	3,1	27,5
	Clerks and office workers on higher level of qualification	173	11,4	17,6	45,1
	Clerks and office workers on lower level of qualification	16	1,1	1,6	46,7
	Self-employed	18	1,2	1,8	48,5
	Foremen, supervisors of manual workers, lower technicians	21	1,4	2,1	50,6
	Skilled workers	301	19,8	30,6	81,2
	Semi-skilled workers	59	3,9	6,0	87,2
	Unskilled workers	110	7,2	11,1	98,3
	Farmers	16	1,1	1,7	100,0
	Total	982	64,8	100,0	
Missing	Never had a full-time occupation	20	1,3		
	Don't know/ Refused	513	33,9		
	Total	533	35,2		
Total		1516	100,0		

education_cat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Elementary	338	22,3	23,1	23,1
	Secondary III level	280	18,5	19,2	42,3
	Secondary IV level	518	34,2	35,4	77,7
	Higher	327	21,6	22,3	100,0
	Total	1463	96,5	100,0	
Missing	Don't know	53	3,5		
Total		1516	100,0		

V166. How many persons live in this household?

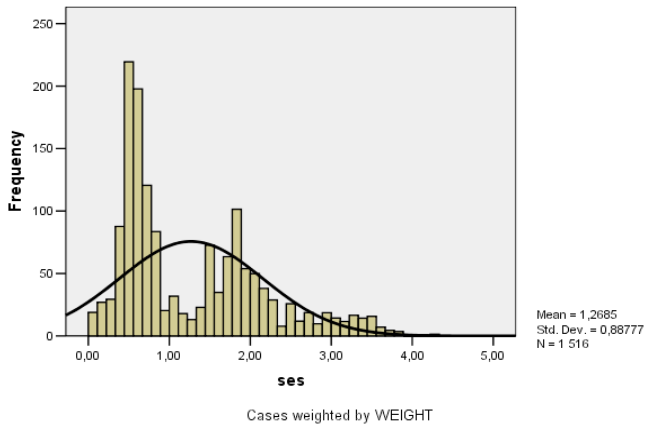
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	123	8,1	8,2	8,2
	2	192	12,7	12,8	21,0
	3	230	15,2	15,3	36,3
	4	390	25,7	25,9	62,2
	5	290	19,2	19,3	81,5
	6	177	11,7	11,8	93,3
	7	59	3,9	3,9	97,2
	8	19	1,2	1,3	98,5
	9	8	,5	,5	99,0
	10	7	,5	,5	99,5
	11	6	,4	,4	99,8
	14	2	,1	,1	100,0
	22	1	,0	,0	100,0
	Total	1503	99,2	100,0	
Missing	refuse	12	,8		
Total		1516	100,0		

SOCIO-EKONOMSKI SKOR KOJI INTEGRIŠE SVE GORNJE VARIJABLE:

Descriptives

			Statistic	Std. Error
ses	Mean		1,2685	,02280
	95% Confidence Interval for Mean	Lower Bound	1,2238	
		Upper Bound	1,3133	
	5% Trimmed Mean		1,2088	
	Median		,8333	
	Variance		,788	
	Std. Deviation		,88777	
	Minimum		,00	
	Maximum		4,26	
	Range		4,26	
	Interquartile Range		1,33	
	Skewness		,878	,063
	Kurtosis		-,088	,126

Histogram



Cases weighted by WEIGHT

U nastavku kursa moramo da naučimo da postavljamo i testiramo hipoteze. Budući da smo naučili jezik varijabli, sada ćemo naučiti da ispitujemo veze i odnose među varijablama. No, da bi ovaj zadatak mogli da izvršimo važno je da naučimo još neke operacije kako bi pripremili naše varijable za testiranje hipoteza. Jedna od ključnih operacija u ovom smislu jeste transformacija varijabli. Postupak transformacije varijabli podrazumeva primenu različitih procedura, od kojih su neke jednostavnije a druge složenije, a posredstvom kojih se po određenom kriterijumu originalne vrednosti transformišu u neke druge vrednosti. Razlog za transformaciju varijabli jesu potrebe da testiramo hipoteze, a svaka hipoteza zahteva da varijable imaju određene karakteristike. Prema tome, transformacija varijabli je metodološki postupak prolagođavanja varijabli za potrebe testiranja hipoteza.

Jedan od najučestalijih načina transformacije varijabli jeste formiranje tzv. 'dummy' varijabli. Ove varijable predstavljaju varijable koje imaju samo dve vrednosti. To znači da mi originalnu varijablu transformišemo u dummy varijablu na način da novonastala varijabla ima samo dve vrednosti. Evo jednog primera:

Obrazovanje

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Bez obrazovanja	8	,8	,8	,8
	Osnovno obrazovanje	111	10,9	11,0	11,8
	Srednje Obrazovanje	603	59,0	59,4	71,1
	Vi{e obrazovanje	165	16,2	16,3	87,4
	Visoko obrazovanje	128	12,5	12,6	100,0
	Total	1016	99,3	100,0	
Missing	Bez odgovora	7	,7		
Total		1022	100,0		

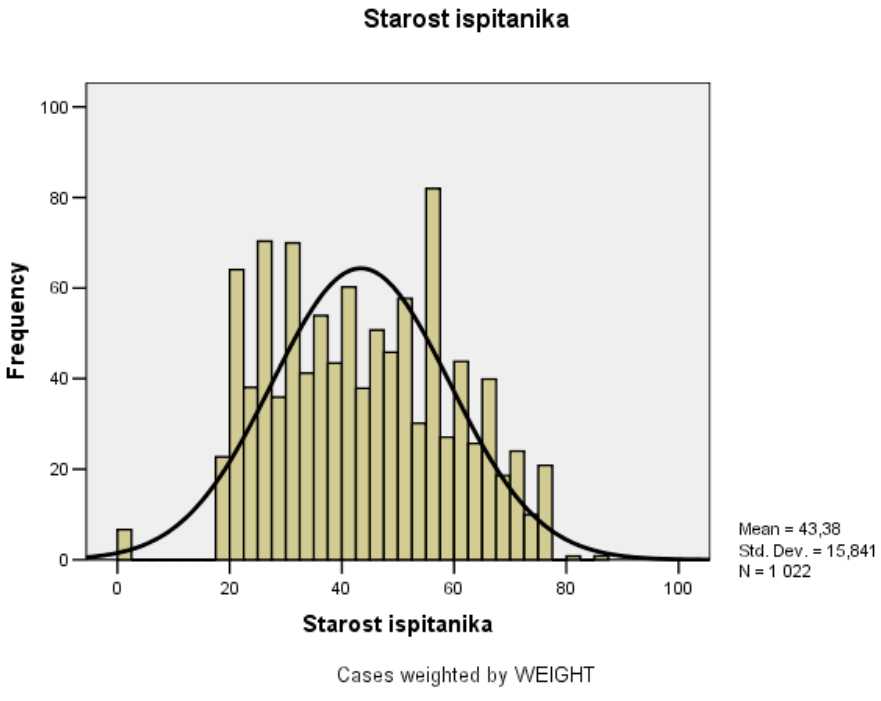
obrayovanje_dummy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ostalo	894	87,5	87,5	87,5
	fakultet	128	12,5	12,5	100,0
	Total	1022	100,0	100,0	

Drugi uobičajen način jeste kada transformišemo originalnu varijablu koja operiše sa intervalnom skalom u varijablu koja operiše sa ordinalnom skalom. Evo primera:

Starost ispitanika

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Bez odgovora	7	,7	,7	,7
	18	4	,4	,4	1,0
	19	19	1,9	1,9	2,9
	20	26	2,6	2,6	5,4
	21	21	2,1	2,1	7,5
	22	17	1,7	1,7	9,1
	23	21	2,1	2,1	11,2
	24	17	1,6	1,6	12,9
	25	24	2,3	2,3	15,2
	26	25	2,5	2,5	17,6
	27	21	2,1	2,1	19,7
	28	17	1,7	1,7	21,4
	29	19	1,8	1,8	23,3
	30	26	2,6	2,6	25,8
	31	21	2,1	2,1	27,9
	32	23	2,2	2,2	30,1
	33	25	2,4	2,4	32,5
	34	17	1,6	1,6	34,1
	35	27	2,6	2,6	36,7
	36	16	1,5	1,5	38,3
	37	12	1,1	1,1	39,4
	38	22	2,1	2,1	41,5
	39	22	2,1	2,1	43,7
	40	29	2,9	2,9	46,5
	41	16	1,5	1,5	48,1
	42	15	1,5	1,5	49,5
	43	14	1,4	1,4	51,0
	44	23	2,3	2,3	53,3
	45	20	1,9	1,9	55,2
	46	16	1,6	1,6	56,7
	47	15	1,5	1,5	58,2
	48	26	2,6	2,6	60,8
	49	20	1,9	1,9	62,7
	50	27	2,6	2,6	65,3
	51	9	,9	,9	66,3
	52	21	2,1	2,1	68,3
	53	15	1,5	1,5	69,9
	54	15	1,4	1,4	71,3
	55	34	3,3	3,3	74,6
	56	28	2,7	2,7	77,3
	57	20	2,0	2,0	79,3
	58	18	1,8	1,8	81,1
	59	9	,9	,9	82,0
	60	17	1,7	1,7	83,6
	61	4	,4	,4	84,0
	62	22	2,2	2,2	86,2
	63	17	1,6	1,6	87,9
	64	9	,9	,9	88,8
	65	19	1,9	1,9	90,6
	66	11	1,0	1,0	91,7
	67	10	1,0	1,0	92,7
	68	10	1,0	1,0	93,6
	69	9	,8	,8	94,5
	70	13	1,2	1,2	95,7
	71	10	1,0	1,0	96,7
	72	1	,1	,1	96,8
	73	7	,6	,6	97,5
	74	3	,3	,3	97,8
	75	14	1,4	1,4	99,2
	76	5	,5	,5	99,7
	77	1	,1	,1	99,8
	82	1	,1	,1	99,9
	87	1	,1	,1	100,0
	Total	1022	100,0	100,0	

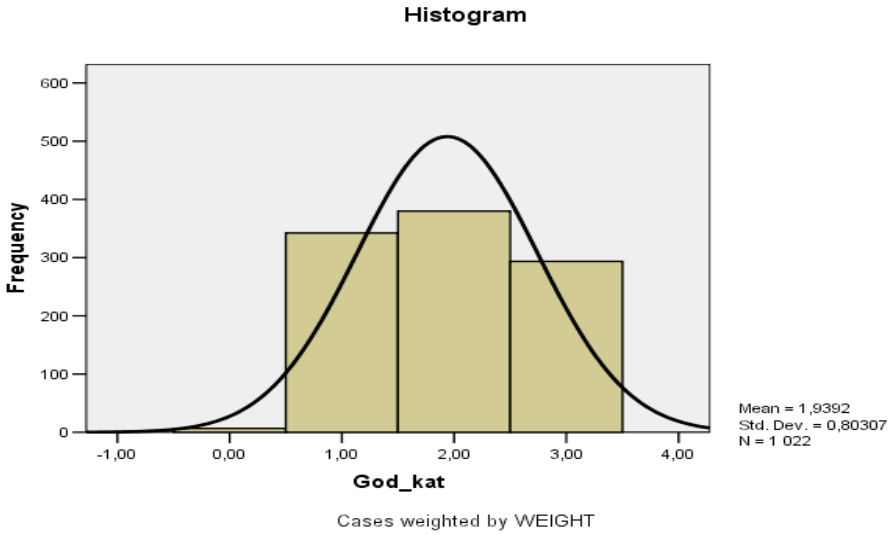


Cases weighted by WEIGHT

Tranformacija u novu varijablu:

God_kat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	7	,7	,7	,7
	18 do 34 god	342	33,5	33,5	34,1
	35 do 54 god	380	37,2	37,2	71,3
	preko 55	294	28,7	28,7	100,0
	Total	1022	100,0	100,0	



Deskriptivna statistika I

Kvantitativan pristup je moguć samo ukoliko je naučno mišljenje usmereno na spoznaju sveta neposrednog iskustva. Postoje tri osnovna elementa koji su uslov primene kvantitativnih metoda i to su:

1. **Objekti -** Stvarnost se sastoji iz objekata. Objekti su bilo koji empirijski predmeti (entiteti) koji imaju fizičke karakteristike i koji su dostupni našim čulima. Oni predstavljaju **izvor posmatranja.** Budući da je proučavanje sveta koji nas okružuje zasnovan na čulima, to znači da proučavajući neku klasu pojava mi moramo identifikovati objekte koji sačinjavaju tu klasu. Na taj način se identifikuju jedinice posmatranja iliti **analitičke jedinice**

2. **Varijable** - Objekti se sastoje iz određenih karakteristika, tačnije, objekt nije ništa drugo do skup karakteristika (atributa). Same karakteristike koje svaki pojedini objekt poseduje mogu biti različite, odnosno variraju u zavisnosti od svakog pojedinačnog objekta. Ove karakteristike tj. svojstva objekta koje mogu imati dve ili više vrednosti nazivaju se **varijablama.**

3. **Skale** - Priroda samih vrednosti i odnos između vrednosti od kojih su varijable sačinjene mogu biti različite. Zbog ovih razlika, postoji nekoliko tipova **skala** tačnije, skale predstavljaju **numeričku reprezentaciju** vrednosti varijabli. Skale, prema tome, predstavljaju set brojeva u svom finalnom obliku. Budući da se svaka varijabla sastoji iz različitih vrednosti, pri čemu vrednosti imaju svoju numeričku reprezentaciju, to znači da svaka varijabla potencijalno predstavlja jedan niz brojeva.

Osnovna tipologija skala uključuje sledeća četiri tipa:

- **Nominalne skale** - Veliki broj objekata poseduje karakteristike koje se razlikuju **kvalitativno**, npr.: krvna grupa, pol/rod, zanimanje....Skale, kao nužan element svake varijable, klasifikuju karakteristike tako da svakom od kvaliteta pridodaju jedan broj. Na ovaj način, kvaliteti dobijaju numerička svojstva. Nominalne skale, koje su prema tome kvalitativne i neretko se u literaturi označavaju kao kategorijalne ili kategorijske (jer u stvari predstavljaju klasifikaciju po kategorijama), su specifične, dakle, prema tome što između brojeva koji predstavljaju kvalitete nije moguće uspostaviti bilo kakav redosled u numeričkom smislu, tj. nije moguće između numeričkih vrednosti utvrditi bilo kakav kvantitativan odnos. Npr: Krvne grupe A, B, AB i O u konceptu varijable dobijaju vrednosti A=1, B=2, AB=3 i 0= 4. Da li postoji bilo kakav numerički odnos između brojeva 1, 2, 3, i 4 koji reprezentuju krvne grupe? NE, ne postoji, brojevi su samo numeracija kategorija

- **Ordinalne** - Ukoliko vrednosti na varijabli mogu da se stave u određen smisleni redosled (rang), onda je karakteristika o kojoj je reč operacionalizovana posredstvom **ordinalne skale.** Numeričke vrednosti

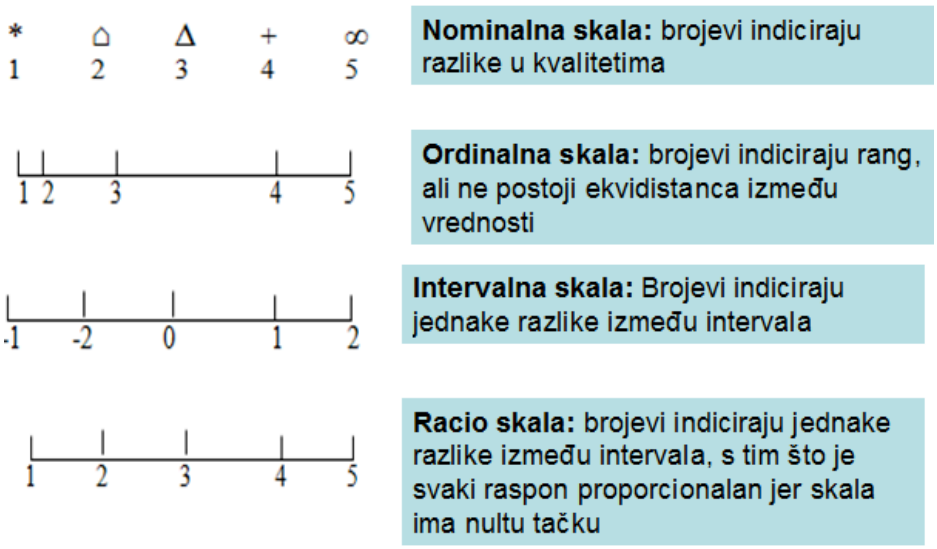
na ordinalnoj skali indiciraju **hijerarhiju** nivoa date varijable. Osnovna osobina ordinalnih varijabli jeste **tranzitivnost**, tj. ako objekat A ima veću vrednost na ordinalnoj skali od objekta B, a objekt B ima veću vrednost od objekta C, onda nužno proističe da objekt A ima veću vrednost na datoj skali od objekta C. Limitirajuća karakteristika ordinalne skale jeste činjenica da nije moguće doneti bilo kakav zaključak o **stepenu** razlika između vrednosti na skali. Drugim rečima, *jednake razlike između ordinalnih vrednosti* ne znače i *jednako kvantitativno tumačenje* tih razlika. Zbog ovog limitirajućeg faktora, ordinalne skale, isto kao i nominalne, se kvalifikuju kao **nemetrijske** skale. Npr. kategorizacija ispitanika u istraživanju po godinama: od 18-34 =1, od 35-54 = 2 i preko 55 god = 3. Dakle, iako su intervali utvrđeni, u praksi ispitanik koji ima 35 godina, na ovoj skali dobiće vrednost 2 a ispitanik koji ima 33 godina dobiće vrednost 1, iako je između njih razlika u starosti samo 2 godine. Ova razlika između opisanih vrednosti hipotetičkih ispitanika (2-1) je veoma različita u odnosu na razliku između ispitanika koji ima 52 godine (dakle na skali ima vrednost 2) i ispitanika koji ima 22 godina (na ordinalnoj skali vrednost 1). U ovom drugom slučaju razlika između ispitanika je 30 godina?! Dakle, i u prvom i u drugom slučaju, razlike na ordinalnoj skali su 2-1, s tim što se u prvom slučaju dva ispitanika razlikuju u starosti samo 2 godine, a u drugom 20 godina.

- **Intervalne** - Intervalne skale su numeričkog (metrijskog) karaktera, tj. jednake razlike između vrednosti na skali imaju jednako značenje. Npr. razlika između ispitanika koji ima 59 godina i onog koji ima 54 godine, je identična kao i razlika između ispitanika koji ima 24 i onog koji ima 19 godina (dakle razlika u oba slučaja je 5). Zbog svojih metrijskih karakteristika, varijable koje su operacionalizovane posredstvom intervalne skale jesu **kvantitativne** varijable. Limitacija intervalnih skala jeste u tome što one ne poseduju apsolutnu nulu. Drugim rečima ne postoji proporcija (ratios) između vrednosti u smislu upoređenja tih vrednosti sa apsolutnom nulom. Pošto nema apsolutne nule, kod intervalnih skala vrednosti nemaju apsolutni smisao, već je reč samo o odnosu između vrednosti na skali koji je uvek relativan. Npr. na skali Farenhajt razlika između 80 F i 90 F je ista kao i razlika između 50 F i 60 F. Međutim, ne može se reći da je 80 F duplo više temperatura od 40F, zato što nema apsolutne nule. Ovo se jasno vidi kad Farenhajte transformišemo u Celzijuse. Naime, u Celzijusima 80F = 26,7C a 40F = 4,4C. Odnos između 80F i 40F je 2:1 dok je odnos između 26,7C i 4,4C 6:1 (tačnije 6,1). Dakle, na jednoj skali je ista razlika u temeraturi dupla a na drugoj skali ta ista razlika šestodupla. Ovo se dešava samo zato što ova skala nema apsolutnu nulu. Intervalne skale svakako u društvenim i političkim naukama imaju veliku upotrebljivost. Naime, one su najčešće upravo zato što se varijable koje mere određeno ponašanje ili stav formiraju na osnovu nekoliko drugih varijabli, a dobijena varijabla upravo ima karakter intervalne skale, u smislu, da postoje intervali i ekvidistanca između vrednosti ali ne postoji apsolutna nula. Veliki broj

skala u naukama o ponašanju jesu *kvazi-intervalne*, npr. skala za merenje IQ. Problem sa ovim skalama jeste u tome što npr. u ovom slučaju ne može se reći da je razlika između skorova 90 i 100 jednaka kao razlika između skorova 100 i 110. Otud se za ove skale kaže da su *kvazi...*

- **Racio** - Racio skale poseduju sve karakteristike kao i intervalne skale s tim što ove skale imaju i apsolutnu nulu. Drugim rečima, imaju karakteristiku da proporcija između vrednosti ima jednako značenje u odnosu na apsolutnu 0. Npr. lestvica prihoda je racio skala jer ima apsolutnu nulu (onaj ko nema nikakvih prihoda). Prema tome, sve razlike između vrednosti unutar skale imaju jednake proporcije i jasan odnos prema nultoj tački. U društvenim naukama se neretko koristi jedan specifičan tip racio skale tzv. **skala učestalosti**. Tako se veliki broj varijabli može operacionalizovati posredstvom ove skale, koja ima sve prednosti intervalnog merenja. Npr. uspešnost u prodaji može se meriti kao broj prodatih komada, umesto da se prodaja izražava u eurima ili dolarima. Odsustvo iz škole može se meriti preko broja odsutnih itd. Skala učestalosti je prema tome karakteristična jer ne poseduje **jedinice merenja**, već jednostavno se bazira na **prebrojavanju** tj. učestalosti nekog događaja. Drugim rečima, vrednost 35 na skali učestalosti nije dvosmislena kao što može biti vrednosti 35 na monetarnoj lestvici u nekoj valuti. Drugi tip racio skale koji takođe ne poseduje jedinice je skala **procenata**. O ovoj skali će jako puno biti reči u narednim poglavljima tako da je nećemo posebno ovde elaborirati

Skale – vizuelni prikaz:



U notiranju i prikazivanju varijabli u statistici često se koriste **slova** kao simboli. Npr. mi možemo govoriti o varijabli ***x*** pri čemu ovim slovom označavamo bilo koju varijablu (godine, pol, prihod, obrazovanje...). Po konvenciji slova koja se koriste za varijable nalaze se na kraju alfabeta, te su to najčešće: ***x, y, z*** ali se mogu sresti i ***u, v, w.*** Takođe se koriste ponekad i velika slova ***U, V, W, X ,Y, Z***, u zavisnosti od autora i preferencija. Ponekad se mogu sresti i druga slova kao oznake za

varijable, ali obično je raspon od ***r*** do ***z.*** Posmatrane vrednosti na datoj varijabli mogu takođe biti označene istim slovima kao i varijabla ali u tom slučaju moramo koristiti subskript (donju oznaku ili index) kako bi razlikovali dobijene vrednosti od različitih objekata koji su predmet analize. Npr. Varijabla ***X*** može imati vrednosti: X_1, X_2, X_3 itd. Pri tome, X_1 je vrednost Objekta 1 na toj varijabli, X_2 je vrednost Objekta 2 na toj varijabli, X_3 je vrednost Objekta 3 na toj varijabli itd sve do ***n*** – tog objekta koji bi imao vrednost X_n na datoj varijabli. Prema tome, mi možemo uopšteno da specificiramo vrednost na varijabli kao X_i, pri čemu ***i*** može biti bilo koji broj 1,2,3,4.....n. Na taj način razlikujemo varijablu ***x*** od vrednosti na toj varijabli X_i. Evo kako izgleda osnovni format za prikupljanje podataka o jednoj varijabli za veći broj objekata:

PRIMER		FORMAT	
Individue	Godine	Objekti	Vrednosti na varijabli ***x***
Goran Djoković	45	O_1	x_1
Andreja Milunović	56	O_2	x_2
Veselin Petrović	37	O_3	x_3
....	....	....	....
....	....	....	....
....	....	....	....
Petar Petrović	63	O_n	x_n

U praksi, mi se gotovo uvek interesujemo za veći broj karakteristika klase pojava (objekata) koje su predmet istraživanja. Prema tome, za valjan opis neke klase objekata potreban je i veći broj varijabli. Npr. ako istražujemo zaposlene, nama su potrebni podaci o godinama, radnom stažu, plati itd za svakog ispitanika u uzorku. Forma matrice za datoteku koja sadrži podatke o većem broju ispitanika sa većim brojem varijabli je s toga složenija, ali je princip isti, naime, redovi su rezervisani za ispitanike (objekat) a kolone za karakteristike (varijable). Pošto je u multivarijantnoj matrici prisutan veliki broj varijabli, nije uobičajeno da se za njih koriste različita slova npr. ***u,v,w,x,z,y***, jednostavno zato što za veliki broj varijabli nećemo imati dovoljno slova (alfabet je ograničen). Zato se umesto različitih slova za varijable koriste numerički supskripti: X_1, X_2, X_3. Prema tome, treba voditi računa kada numerički supskripti notifikuju vrednosti na varijabalama a kada sami numerički supskripti notifikuju različite varijable. Međutim, pošto smo numeričke supskripte već koristili za označavanje vrednosti, to znači da za oznaku različitih varijabli i vrednosti moramo koristiti **duple supskripte** za svaki pojedini objekt u uzorku. Tako npr. X_{32} predstavlja vrednosti Objekta 3 na varijabli X_2. Dalje, supskript ***k,*** ima sprecifično značenje. Naime, njime se označava broj varijabli koje učestvuju u

modelu. Ponekad se umesto ***k*** koristi ***m.*** Sa druge strane supskript ***n*** se koristi za označavanje ukupnog broja opservacija u nekom setu podataka
Konačno, opservacije u multivarijantnom setu podataka možemo označiti X_{ij} pri čemu prvi supskript označava Objekte a drugi supskript označava varijable. Evo primera multivarijantne matrice sa većim brojem varijabli i objekata:

PRIMER					***FORMAT***				
Individue	VARIJABLE				Objekti	VARIJABLE			
	Godine	Radni staž	Prihod	....		x_1	x_2	x_3	x_k
Goran Djoković	45	22	750	....	O_1	x_{11}	x_{12}	x_{13}	x_{1k}
Andreja Milunović	56	33	450	....	O_2	x_{21}	x_{22}	x_{23}	x_{2k}
Veselin Petrović	37	9	290	....	O_3	x_{31}	x_{32}	x_{33}	x_{3k}
....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....
Petar Petrović	63	45		320	O_n	x_{n1}	x_{n2}	x_{n3}	x_{nk}

Najosnovniji oblik statističke analize podataka jeste merenje razlika između objekata koje se mogu naći na jednoj varijabli. Na taj način se opisuju razlike u karakteristikama objekata koje postoje unutar jedne klase objekata (pojava). S obzirom na to da se ovakva analiza bazira na ispitivanju vrednosti objekata na **jednoj** varijabli, ovaj tip statističke analize se naziva ***univarijantna statistika.*** Univarijantna statistika spada u kategoriju ***deskriptivne statistike***, tačnije, to je onaj tip statističke analize koji nema za cilj da objašnjava veze i odnose među varijablama,već ima za cilj da ***opiše*** jednu ili više varijabli.

Najosnovniji oblik statističke analize podataka jeste ***distribucija učestalosti*** ili distribucija frekvencija kako se to uobičajeno naziva (frequency distribution).
Distribucija frekvencije predstavlja **prebrojavanje učestalosti** svake od vrednosti koja je definisana varijablom. Drugim rečima, distribucija učestalosti govori o tome **koliko često se svaka vrednost na jednoj varijabli pojavljuje** unutar seta objekata koji su predmet merenja.

Tabela frekvencije predstavlja prosto prebrojavanje broja slučajeva unutar svake od definisanih vrednosti varijabe. Ova učestalost se naziva frekvancijom pa otud naziv **tabela frekvencije** ili **distribucija frekvencije:**

		Učestalost
	Poljoprivrednik	21
	Radnik	200
	Službenik	110
	Tehničar	49
	Stručnjak	33
	Rukovodilac	11
	Privatnik-Vlasnik	45
	Učenik-student	67
	Penzioner	222
	Domaćica	103
	Nezaposlen	158
	Total	1018
	Bez odgovora	5
Total		1022

Algoritam za frekvencije je:

$$f_j = \sum_{j=1}^{N} w_j k_j \qquad j = 1,\ 2,\ \ldots,\ NV$$

Algoritam za frekvencije pri čemu je:
$k_i = 1$ ako $X_i = X_j$; u suprotnom $k_i = 0$

X_k	vrednosti varijable za slučaj ***k***
W	suma svih slučajeva
W_k	ponder za slučaj ***k***
NV	broj pretpostavljenih distinktivnih vrednosti
N	broj slučajeva

Gde X_j predstavlja j-tu najveću distinktivnu vrednost varijable X

Kada se učestalost opserviranih vrednosti jednostavno podeli sa ukupnim brojem opservacija i pomnoži sa brojem 100 dobija se ***relativna frekvencija,*** koja ima za cilj da pokaže koji je procenat svake od kategorija učestalosti.

		Učestalost	%
	Poljoprivrednik	21	2,1
	Radnik	200	19,5
	Službenik	110	10,7
	Tehničar	49	4,8
	Stručnjak	33	3,2
	Rukovodilac	11	1,0
	Privatnik-Vlasnik	45	4,4
	Učenik-student	67	6,6
	Penzioner	222	21,7
	Domaćica	103	10,1
	Nezaposlen	158	15,5
	Total	1018	99,6
	Bez odgovora	5	,4
Total		1022	100,0

Algoritam za frekvencije je:

$$Rf_j = \left(\frac{f_j}{W'} \right) \times 100$$

Pri čemu je

$$W' = \sum_{j=1}^{NV} f_j$$

Suma svih kategorija

Često želimo da znamo kolika je učestalost vrednosti **iznad**, a koliko **ispod** određene vrednosti. Ovaj podatak naročito je koristan kada je skala intervalna. To je svrha ***kumulativne frekvencije***. Ovaj tip frekvencije nam govori o tome koliko je opservacija procentualno dio određene vrednosti uključujući i tu vrednost (kolona 4). SPSS kalkuliše i **validni procenat**, a ovo je relativna frekvencija koja kao reper za ukupan broj opseravcija uzima ukupan broj slučajeva umanjen za osperavcije koje su programom definisane kao nedostajuće vrednosti. Ovaj podatak može u nekim situacijama biti jako koristan kao što ćemo kasnije videti.

		Učestalost	%	Validni %	Kumulativni %
Valid	Poljoprivrednik	21	2,1	2,1	2,1
	Radnik	200	19,5	19,6	21,7
	Službenik	110	10,7	10,8	32,5
	Tehničar	49	4,8	4,8	37,3
	Stručnjak	33	3,2	3,2	40,5
	Rukovodilac	11	1,0	1,1	41,5
	Privatnik-Vlasnik	45	4,4	4,4	46,0
	Učenik-student	67	6,6	6,6	52,6
	Penzioner	222	21,7	21,8	74,3
	Domaćica	103	10,1	10,1	84,5
	Nezaposlen	158	15,5	15,5	100,0
	Total	1018	99,6	100,0	
Missing	Bez odgovora	5	,4		
Total		1022	100,0		

Alg

$$Cf_j = \sum_{i=1}^{j} f_i$$

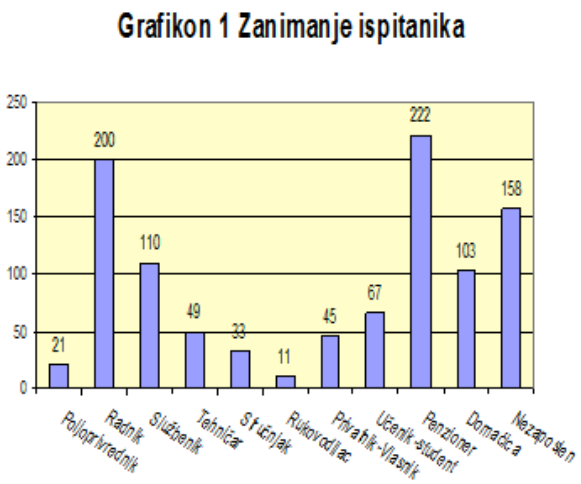
Pravila za tabele frekvencije su:

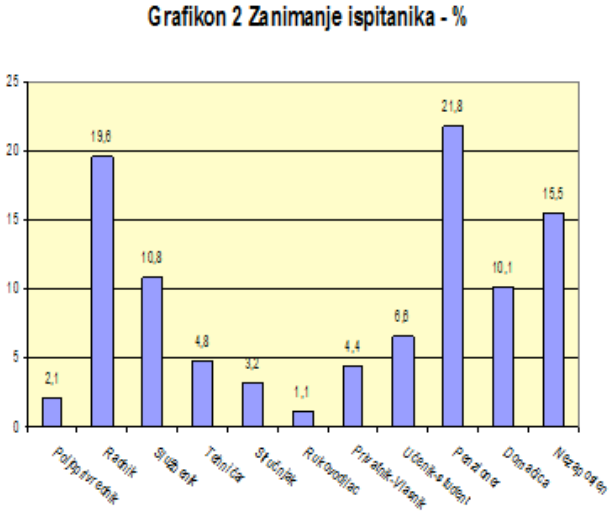
- Prikazati distribuciju svih vrednosti.
- **Obavezno** pored relativne frekvencije dati i podatak o učestalosti u fizičkim brojevima. Neretko, istraživači, koji operišu sa malim uzorcima daju podatke samo u procentima i na taj način skrivaju da se iza nekog procenta (npr. 20%) kriju samo nekoliko ispitanika. Ovo je neprihvatljivo, jer relativna frekvencija i njen smisao imaju specifičnu težinu samo kada iza procenata stoji 'pristojan' broj slučajeva. Inače, statistika počiva na zakonu 'velikih brojeva' i kada je uzorak jako mali, onda je standardna statistička greška obično velika.
- U interpretaciji u izveštaju biti pažljiv, precizan i voditi računa da se svi podaci komentarišu jasno i nedvosmisleno.
- U interpretaciji podataka na osnovu tabele frekvencije, uvek poštovati princip da se vrednosti interpretiraju od najučestalijih do onih koje imaju najmanju frekvenciju.
- Ne zaboraviti da tabele frekvencije imaju cilj da ***opišu*** Vaš set podataka i s toga voditi računa da se na osnovu njih ne donose prerano zaključci koji se ne mogu videti u samoj tabeli frekvencije.
- U interpretaciji relativnih frekvencija (procenata) voditi računa o standardnoj statističkoj greški uzorka.

Veoma često, sa ciljem da se što bolje opišu podaci, tačnije, da se što jasnije razume distribucija vrednosti na nekoj varijabli, u statistici se koriste grafički prikazi. Grafički prikazi su u savremenoj metodologiji postali nezamenljiva oruđa za jasno prikazivanje, opisivanje, razumevanje i analiziranje podataka. Moderna statistika je razvila čitav niz veoma specifičnih grafičkih načina za prikazivanje podataka, od kojih su neki veoma složeni i sofisticirani. Opšta je tendencija da se najsloženije multivarijantne metode, koje je je jako teško razumeti posredstvom tabela i matematičkih formula, prikazuju grafički.

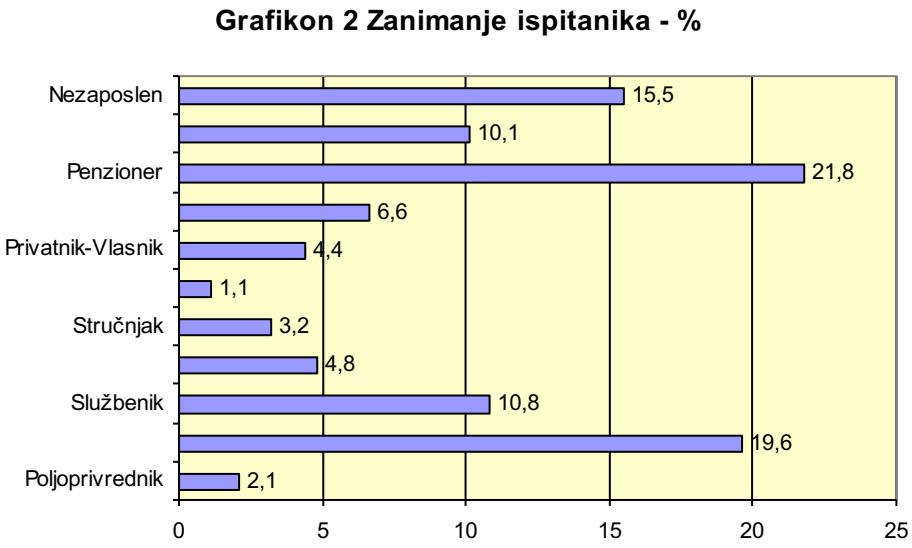
Jedan od najčešćih metoda grafičkog prikazivanja podataka jeste **histogram** koji se najčešće u programima naziva 'bar chart'. Barovi predstavljaju 'gredice'. Histogram prikazuje distribuciju tako što se na horizontalnoj osi nalaze vrednosti varijable a na vertikalnoj osi brojčane oznake u odnosu na koju se posmatra visina svakog bara (gredice). Visina svakog bara pokazuje relativnu frekvenciju vrednosti koju taj bar predstavlja. Uobičajeno je da se na svakom baru daju vrednosti bilo frekvencija bilo relativnih frekvencija:

Histogram
graf. 1: distribucija frekvencije, graf. 2: distribucija relativne frekvencije



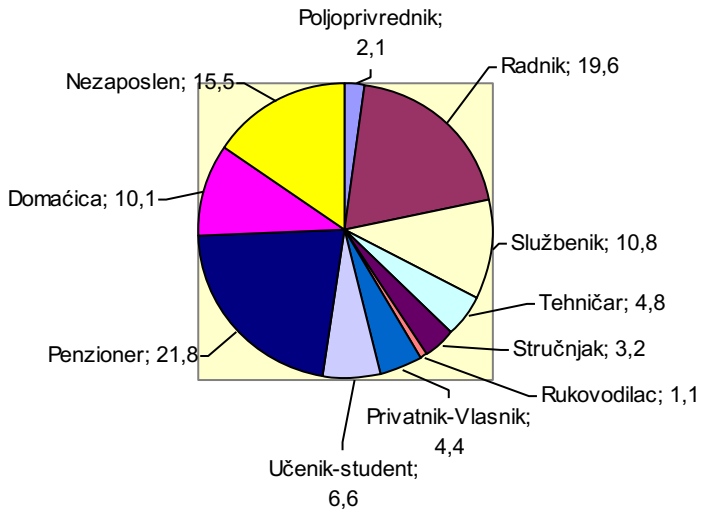


Neretko se histogram rotira za 90 stepeni (ulevo) tako da horizontalna osa sada stoji vertikalno a vertikalna horizontalno. Dakle, u ovoj transformaciji, vrednosti su date na vertikalnoj osi a na horizontalnoj mogu se videti reperi za distribuciju. U praksi, bar chart ovog tipa se pokazao preglednijim ako je u pitanju distribucija na kategorijalnim varijablama, jednostavno iz estetskih razloga jer je veliki broj nomina za vrednosti lakše situirati po horizontali.



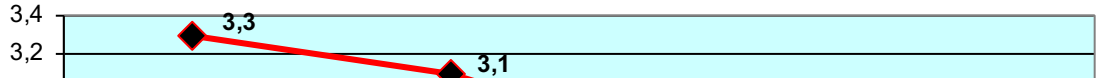
Drugi tip grafičkog prikazivanja podataka koji se često koristi jeste 'pita'. Pita se koristi ***isključivo sa kvalitativnim varijablama***, tj. onim varijablama čije vrednosti imaju karakter nominalne skale. Ovo zato što vrednosti na ovim varijabalam nemaju numerički karakter te ih je najbolje prikazati u obliku pite budući da između vrednosti ne postoji nikakav poredak zasnovan na brojevima. Pita je grafikon kružnog oblika u kome svako parče pite predstavlja relativnu frekvenciju vrednosti kvalitativne varijable. Pita se koristi **uvek i isključivo** kada se prikazuju relativne frekvencije i kada je skala nominalnog tipa. Površina koju svaka vrednost zauzima izračunava se tako što se relativna frekvencije neke vrednosti pomnoži sa 360, a ovo je ukupna površina svakog kruga. Dobijena vrednost se zatim podeli sa 100. Tako npr. ako je relativna frekvencija 21,8% (u našem slučaju penzioneri), onda je površina koju će ova vrednost da zauzme na piti = 78,5 stepeni.

Grafikon 3 Zanimanje ispitanika - %



Poligon frekvencije je treći tip grafičkog prikazivanja podataka. Ovaj tip grafikona se koristi najčešće u tri slučaja: prvo, kada se želi izraziti **trend**, tačnije, distribucija vrednosti u različitim vremenskim periodima, drugo, kada je distribucija vrednosti gradualnog tipa, i treće, kada se želi proceniti tip distribucije (o ovom trećem pričaćemo detaljno na sledećem predavanju). Kad je o 'liniji' reč, ovaj grafikon takođe na horizontalnoj osi sadrži vrednosti, a na vertikalnoj referentne bojeve. Takođe je uobičajeno, da se vrednosti na liniji brojčano iskazuju, osim kada je svrha 'linije' da grafički prikaže tip distribucije.

Grafikon 4 Povjerenje u vladu - Trend



Neka pravila za grafičko prikazivanje podataka:

1. ***Nije nužno stvari prikazivati grafički***: Neretko se grafikoni uvrštavaju u Izveštaje iz estetskih razloga, a da pritom ni na koji način ne doprinose svom cilju, boljoj deskripciji podataka.
2. Izaberite onaj ***tip grafikona*** koji na najbolji mogući način ispunjava svoju eksplanatornu funkciju.
3. Nemojte distribuciju vrednosti neke varijable da prikazujete i tabelarno i grafički: Ovo je redudantno i nema nikakvog razloga da se dupliraju informacije. Drugim rečima, odaberite, distribuciju frekvencija ćete ***prikazati ILI tabelarno ILI grafički***.
4. Pri izboru grafikon ILI tabela, vodite računa o tome da ***tabela po pravilu sadrži više informacija*** dok je grafikon pregledniji i jednostavniji za deskripciju podataka.
5. Pri izboru grafikona vodite računa ***ko je primalac informacije***. Naime, stručna publika preferira tabele dok javnost preferira grafički prikaz.

Osnovni elementi koje svaki grafikon mora da sadrži su:

- **Broj grafikona** (Grafikon br x). Ovaj broj zavisi od redosleda, I to redosleda u poglavlju (Grafikon br. 2.5. – peti grafikon u drugom poglavlju)
- **Naslov**: Ovaj naslov mora da odgovara suštini prikaza. Izbegnite deskriptivne naslove tipa 'Distribucija frekvencije zanimanja' i koristite ili opisne naslove ili skraćenice npr. 'Zanimanje',
- Uz naslov se mora **naznačiti** da li su vrednosti date u frekvencijama ili relativnim frekvencijama,
- Ukoliko koristite u grafikonu relativne frekvencije, poželjno je da se ispod grafikona predstavi broj ispitanika na način: **N – 1200** (npr.)
- **Proverite** da li svaki bar, parče pite ili linija prikazuju frekvencije koje ste želeli,
- **Pogledajte** i proverite horizontalnu i vertikalnu osu da li sadrže nomine za vrednosti odnosno numeričke repere. U slučaju vertikalne ose, vodite računa o rasponu vrednosti,
- Uvek **protumačite** grafikon sami pre nego što ga tumačite drugima. Ako ovo izbegnete, neretko pred publikom ćete primetiti da grafikonu fale neki elementi.

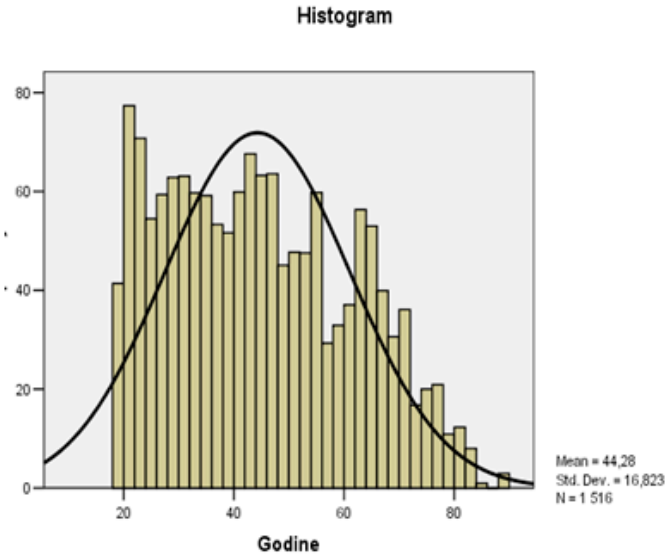
Deskriptivna statistika II

Evo jednog primera distribucije vrednosti na varijabli ’starost ispitanika’ sa statisticima koje moramo razumeti:

Statistics

Godine

N	Valid	1516
	Missing	0
Mean		44,28
Std. Error of Mean		,432
Median		43,00
Mode		21
Std. Deviation		16,823
Variance		283,018
Skewness		,317
Std. Error of Skewness		,063
Kurtosis		-,896
Std. Error of Kurtosis		,126
Range		72
Minimum		18
Maximum		90
Sum		67110
Percentiles	25	30,00
	50	43,00
	75	57,56
	90	68,00



Osnovni koncepti koje moramo savladati jesu **mere centralne tendencije**. Varijable sadrže veći broj vrednosti, naročito ako je reč o intervalnim ili racio skalama. Tabele frekvencije, tabele relativne frekvencije i grafikoni koji prikazuju distribuciju po kategorijama nam svakako pomožu da bolje razumemo distribuciju nekih karakteristika objekata koje nas interesuju. Međutim, u statistici postoje načini da se celokupna distribucija izrazi na jedan sumaran način, ili tačnije da ukupnu distribuciju izrazimo jednim brojem, na taj način postižemo da se na najkoncizniji mogući način iskaže distribucija vrednosti o kojoj je reč. Ovo je ključna svrha upotrebe **mera centralne tendencije** u društvenim naukama. Mere centralne tendencije, prema tome predstavljaju određene statističke parametre koji opisuju način na koji se sve vrednosti jednog numeričkog niza (varijable) 'centriraju' i iskazuju posredstvom određene (jedne) numeričke vrednosti. Obzirom da mere centralne tendencije veliki broj različitih vrednosti iskazuju sumarno, one počivaju na **redukcionizmu**, ili drugim rečima nužno je da u ovom procesu sintetizacije mi gubimo određeno bogatstvo u pogledu količine informacija zarad potrebe da distribucija na varijabli iskažemo singularnim numeričkim izrazom.

Jedna od mera centralne tendencije koja se najčešće upotrebljava jeste **aritmetička sredina**. Ona je ujedno i najlakša za razumevanje obzirom da se neretko koristi u svakodnevnom životu (najčeće koristimo reč 'prosek' da izrazimo upravo aritmetičku sredinu). Dakle, aritmetička sredina predstavlja prosečnu vrednost nekog kontinuiranog niza brojeva. U društvenim naukama se, takođe, aritmetička sredina često koristi kako bi se izrazile određene prosečne karakteristike populacije ili uzorka.

Aritmetička sredina nekog seta kvantitativnih podataka (numeričkog niza) jeste suma svih vrednosti podeljena sa ukupnim brojem objekata od kojih se set podataka sastoji

Aritmetička sredina se izračunava jednostavnom formulom:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Napomena: *Simbol* $\bar{x}$ *se koristi kada se označava aritmetička sredina uzorka, dok se simbol* μ *koristi kada je reč o populaciji*

Npr.. ako kožemo da su ispitanici u našem uzorku stari u proseku 44,5 godina (uzorkom su obuhvaćeni svi punoletni stanovnici Crne Gore), to znači da smo mi sabrali godine svih ispitanika u uzorku i podelili sa ukupnim brojem ispitanika. Dva su ključna faktora od kojih zavisi preciznost ovog podatka (aritmetičke sredine):

1. Od veličine uzorka, naime, što je veći uzorak i aritmetička sredina je preciznija
2. Od varijabilnosti tj. ospega (rasprostranjenosti), naime, što je veća varijablnost vrednosti, to je aritmetička sredina manje pouzdana

Medijana predstavlja drugačiju meru centralne tendencije u odnosu na aritmetičku sredinu. Medijana,takođe, jeste neka vrsta prosečne vrednosti, ali kad upotrebljavamo medijanu u komunikaciji ne možemo koristiti reč prosek, već jednostavno reč 'medijana'.

Medijana nekog kvantitativnog seta podataka jeste srednji broj u situaciji kada se sve vrednosti poređaju od najniže do najviše ili obrnuto

Ukoliko je niz brojeva neparan, onda je medijana broj u sredini. Ukoliko je broj paran, onda je medijana srednja vrednost srednja dva broja.

Primer.1. Ako se varijabla sastoji od 7 brojeva 5,7,4,5,20,6 i 2 onda se medijana izračunava:

2,4,5,**5,** 6,7,20 M = 5 (broj u sredini posmatrano s krajeva)

Primer.2. Ukoliko je pak varijabla sa parnim brojem brojeva (n=6) , npr. varijabla 4,5,5,6,7,20 onda se medijana izračunava:

4,5,**5,6**,7,20 M = (5+6)/2 = 5.5

U nekim situacijama medijana je bolja mera centralne tendencije u odnosu na aritmetičku sredinu. Ovo zato što je medijana manje senzitivna na ekstremno male i ekstremno velike vrednosti (u malopređašnjem primeru br.1, medijana je identična i ukoliko se iz niza brojeva isključi '20' kao ekstremno velika vrednost. Neke varijable u društvenim istraživanjima su naročito pogodne za medijanu, npr. prihod iskazan na intervanoj skali. Budući da u društvu uvek postoji mali sloj jako bogatih ljudi, ova činjenica bi ozbiljno uticala na aritmetičku sredinu, ali ne i na medijanu. Medijana je, takođe, pogodna kada se koriste lestvice procene, a i zgodnija je za interpretaciju.

Modus najčešća vrednost koja se pojavljuje u jednom setu brojeva. Npr. ako je set brojeva:

3,4,6,1,**8,8**,9,3,4,6,**8**,2,3,**8,8**,0,9,**8**,4,5,6,**8**,3,3,4,7,**8**,9,**8**,0,**8**,5,**8**,

Onda je modus = 8, dakle, broj koji se najviše puta pojavio u nizu. U zavisnosti od prirode podataka modus može biti manje ili više korisna mera centralne tendencije. U našem primeru da u uzorku ima najviše ispitanika koji imaju 21 godinu, taj podatkak nam i ne govori mnogo. Međutim ako pogledamo grafik primetićemo da je distribucija neravnomerna upravo u ovom sektoru, tačnije kod mlađe kategorije ispitanika. U društvenim istraživanjima ima situacija kada je modus sasvim koristan. Recimo, u ovim istraživanjima se neretko koriste likertove lestvice procene koje imaju vrednosti npr.: veome dobro, dobro, loše i veoma loše. U ovakvim situacijama modus nam daje vredniju informaciju u poređenju sa aritmetičkom sredinom i medijanom tj. jednostavno nam govori koju od ovih vrednosti su ispitanici izabrali najviše puta. U političkim istraživanjima javnog mnjenja neretko se koristi **prosečna ocena** (aritmetička sredina) kao indikator preferencije stavova građana o nekom pitanju, instituciji ili pojedincima. Tipičan primer je korišćenje prosečne ocene za institucije i političare. Evo da pogledamo podatke jednog istraživanja koje govori o rejtingu političara na osnovu prosečne ocene:

Descriptive Statistics

	N	Mean	Std. Deviation
Zeljko STURANOVIC	915	3,14	1,416
Nebojsa MEDOJEVIC	891	3,10	1,406
Milo DJUKANOVIC	919	3,09	1,684
Filip VUJANOVIC	928	2,99	1,468
Gordana DJUROVIC	852	2,86	1,514
Andrija MANDIC	874	2,51	1,504
Vujica LAZOVIC	768	2,48	1,341
Ranko KRIVOKAPIC	919	2,45	1,503
Srdjan MILIC	802	2,27	1,310
Miodrag ZIVKOVIC	883	2,18	1,258
Predrag POPOVIC	873	2,16	1,315
Ranko KADIC	843	2,15	1,370
Emilo LABUDOVIC	874	2,04	1,334
Zoran ZIZIC	863	2,01	1,280
Ferhat DINOSA	864	1,88	1,216
Mehmet BARDHI	857	1,71	1,118
Vasilj SINISTAJ	804	1,65	1,017
Valid N (listwise)	635		

Ukoliko uporedimo aritmetičku sredinu i modus možemo videti sledeće:

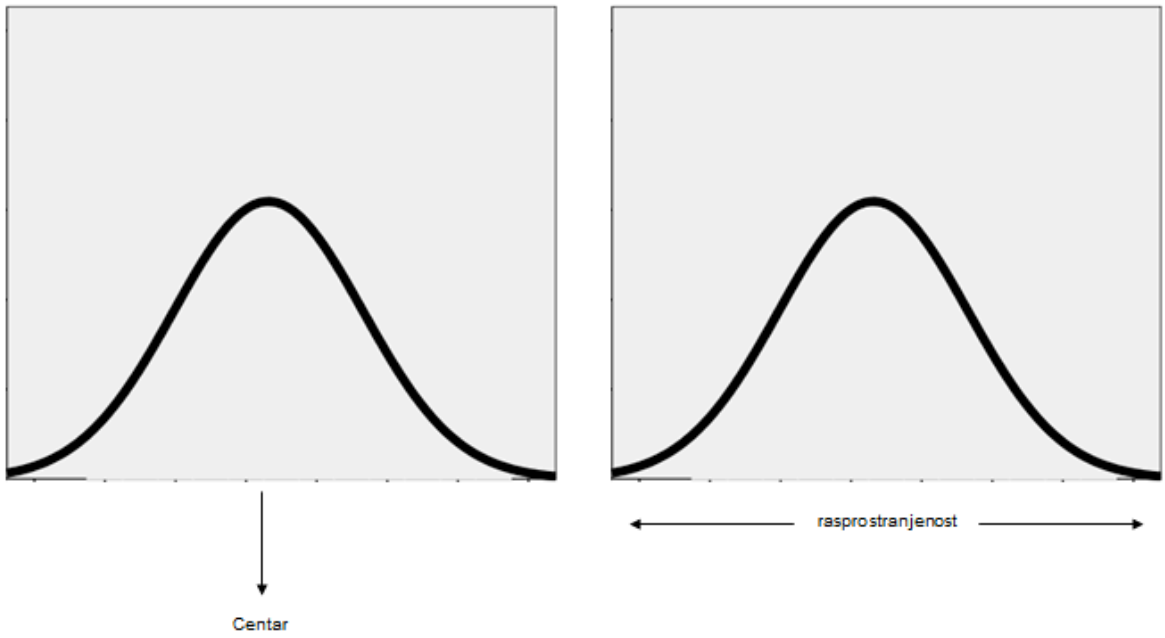
Statistics

	Valid	Mean	Modus
Filip VUJANOVIC	928	2,99	1
Ranko KRIVOKAPIC	919	2,45	1
Zeljko STURANOVIC	915	3,14	4
Gordana DJUROVIC	852	2,86	1
Vujica LAZOVIC	768	2,48	1
Milo DJUKANOVIC	919	3,09	5
Andrija MANDIC	874	2,51	1
Nebojsa MEDOJEVIC	891	3,10	3
Srdjan MILIC	802	2,27	1
Miodrag ZIVKOVIC	883	2,18	1
Predrag POPOVIC	873	2,16	1
Ranko KADIC	843	2,15	1
Zoran ZIZIC	863	2,01	1
Emilo LABUDOVIC	874	2,04	1
Ferhat DINOSA	864	1,88	1
Mehmet BARDHI	857	1,71	1
Vasilj SINISTAJ	804	1,65	1

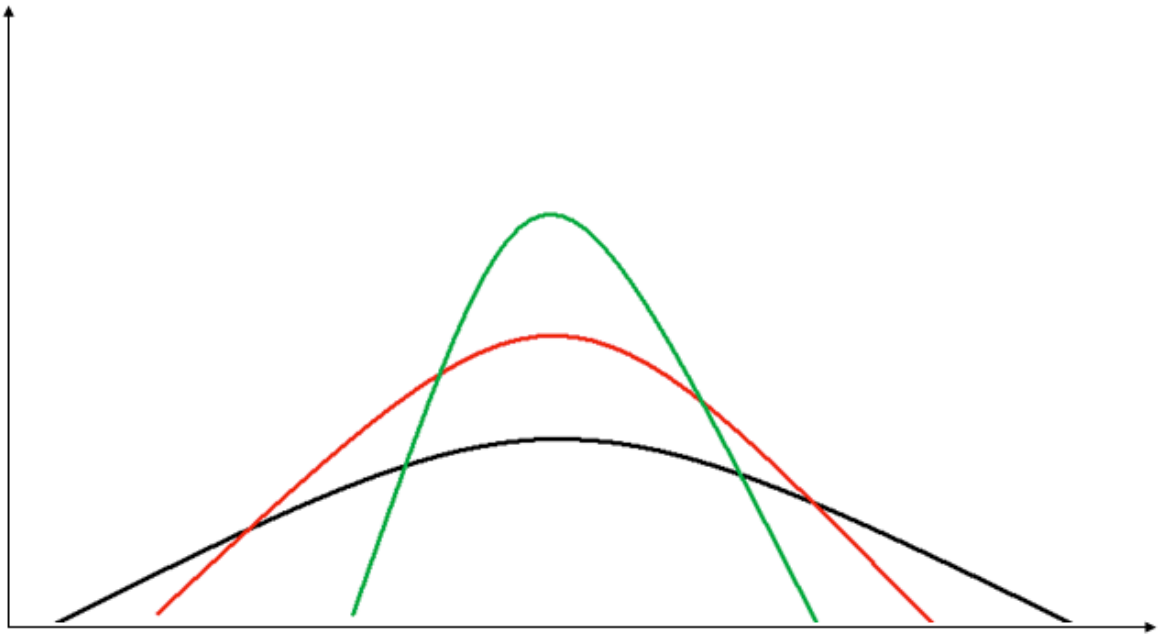
U konkretnoj situaciji, možemo videti da od M, Đukanovića višu prosečnu ocenu imaju Šturanović i Medojević, ali nam modus govori o tome da je Đukanović dobio najveći broj petica. Ovaj podatak je veoma indikativan kada je reč o eventualnom glasanju građana, naime, građani će najverovatnije da glasaju one kojima su dali petice a ne onima kojima su dali trojke i četvorke.

Mere centralne tendencije samo parcijalno opisuju podatke, te su prema tome mere varijablinosti nužne za potpuni opis neke varijable. **Varijabilnost** je sastvani

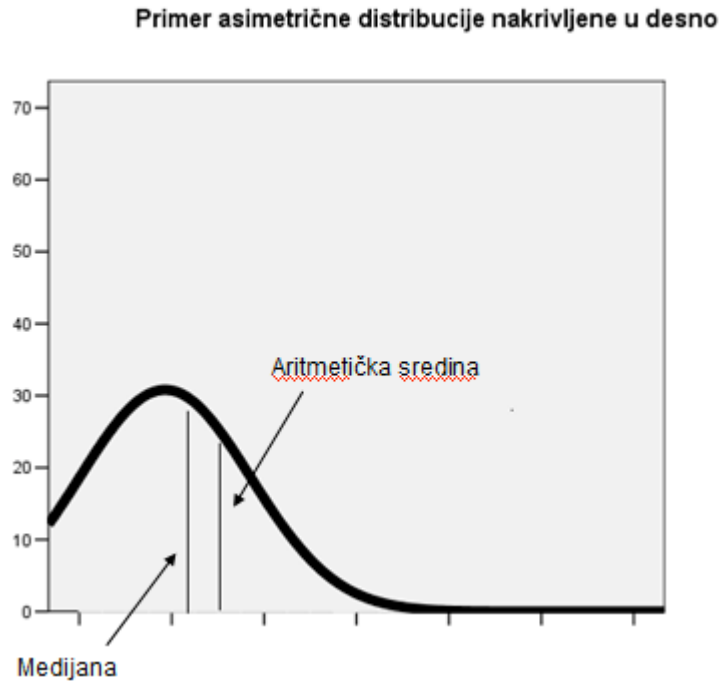
deo distribucije, i njena procena je jednako važna kao i procena centralne tendencije. Drugim rečima, centralna tendencija uz mere varijabilnosti nam pomaže da vizualizujemo oblik jedne distribucije. Svaka distribucija prema ome ima dva aspekta:

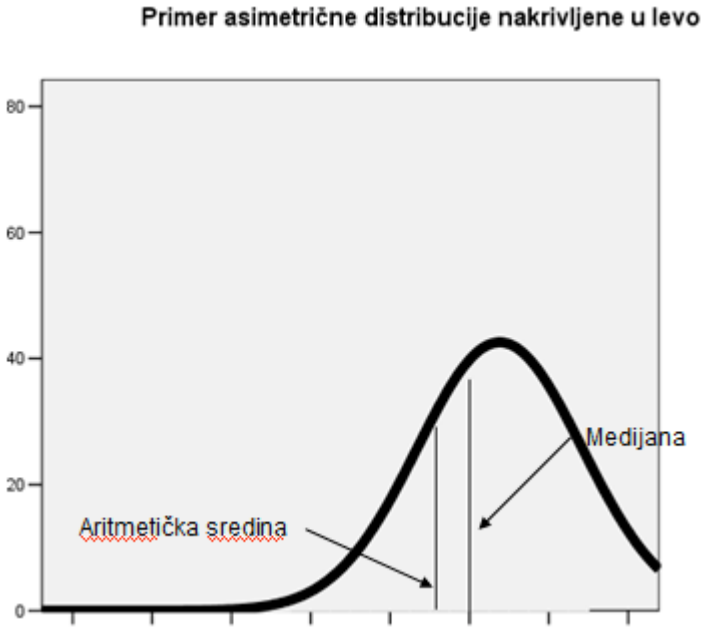


Drugima rečima možemo imati istu aritmetičku sredinu za nekoliko distribucija, a opet da one budu veoma različite zbog varijablinosti, evo prikaza:



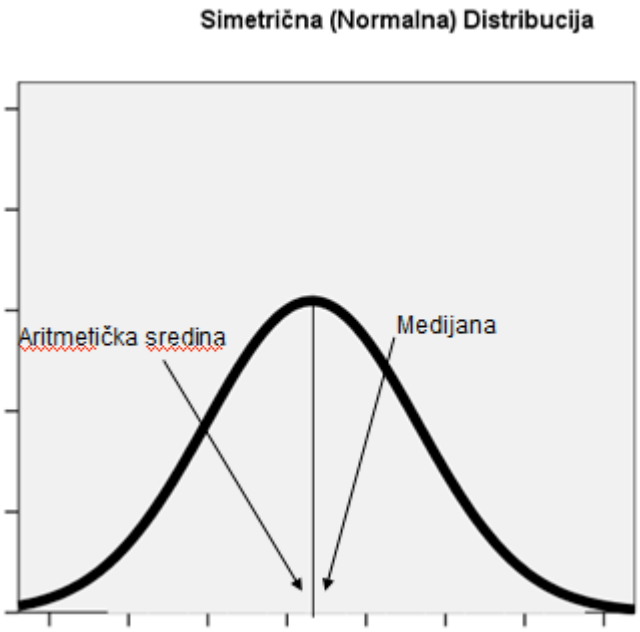
U osnovi svaka distribucija može biti zakrivljena ' u levo' ili 'u desno'. Evo primera:





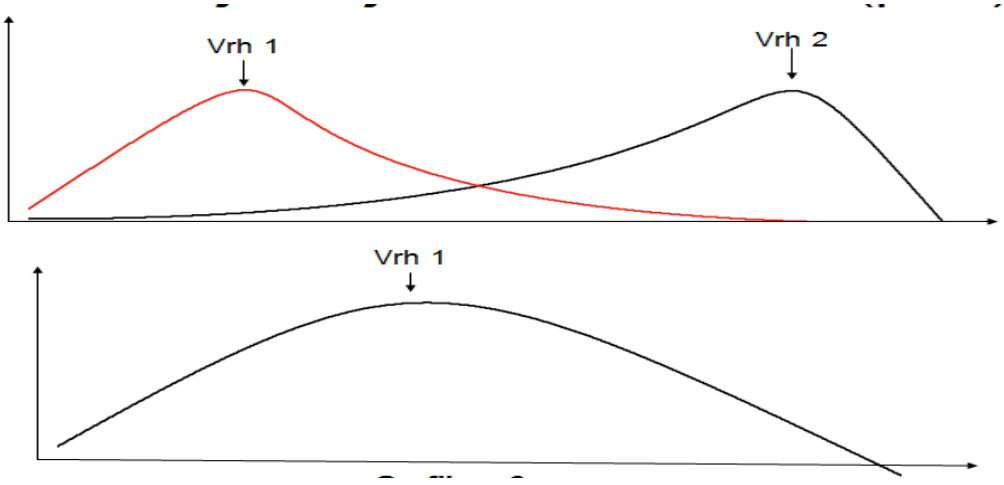
Dakle, ukoliko je distrubicija 'zakrivljena' u desno, vrdsnost aritmetičke sredine biće veća od vrednosti medijane. Obrnuto, ukoliko je distribucija zakrivljena u levo, vrednost medijane biće veća od vrednosti aritmetičke sredine. Konačno, ukoliko je distribucija ravnomerna (kažemo normalna), onda će se u idealnom slučaju poklapati

vrednosti aritmetičke sredine i medijane što se grafički može videti na sledećem primeru:



Normalna distribucija predstavlja onaj tip distribucije u kojem se vrh distribucije poklapa sa aritmetičkom sredinom distribucije. Ovaj tip distribucije ima oblik 'zvona' pa se često upravo ovaj figurativan naziv koristi za normalnu distribuciju. Normalna distribucija, kao koncept, igra veoma važnu ulogu u statistici, naime, veliki broj statističkih testova počiva na pretpostavci da su vrednosti normalno distribuirane.

Nasuprot konceptu normalne distribucije, u statistici često operišemo konceptom **asimetrične distribucije**. Asimetrična distribucija je onaj tip distribucije u kome se aritmetička sredina ne poklapa sa vrhom distribucije. Distribucija može biti asimetrična u dva smera: prvo, ona može biti iskrivljena u desno i ona može biti iskrivljena u levo, kao što smo pokazali na gornjim grafikonima. Najgori mogući slučaj asimetrične distribucije jeste ukoliko je ona nakrivljena istovremeno i u levo i u desno. Za ovaj tip distribucije kažemo da je izrazito asimetrična i ona ima dva pika:



Kada imamo slučaj izrazito asimetrične distribucije, onda nam to govori da mi u okviru populacije imamo dve kategorije (kohorte), koje se suštinski razlikuju po dataj karakteristici te je s stoga uputno da se te dve kategorije različito i tretiraju u daljoj statisičkoj analizi. Usled činjenica da je narušavanje normalne distribucije veoma značajno, u statistici postoji čitav niz parametara i statistika koji imaju za cilj da mere asimetričnost distribucije.

Skewness (asimetrija/iskrivljenost/zakošenost/nagib distribucije) je parametar koji pokazuje da li je distribucija asimetrična ulevo ili udesno. Skewness se može precizno izračunati i u statistici se to naziva treći momenat distribucije. Skewness je jednak 0 ako je distribucija simetrična, a u suprotnom imaće vrednost drugačiju od 0 sa predznakom koji je identičan kao i asimetrija. Ova mera, dakle, može biti 0, ako je distribucija savršeno normalna, i može odstupati od 0, sa pozitivnim (+) i negativnim (-) predznakom. Ako je predznak negativan (-) , to znači da je distribucija nakošena ulevo, tj. u tom slučaju veća je frekvencija natprosečnih vrednosti. Ako je predznak pozitivan (+), onda je distribucija nakošena udesno, i to znači da je frekvencija vrednosti u korist ispodprosečnih vrednosti. Po konvenciji ako je vrednost skewness-a manja od 1, smatra se da distribucija nije asimetrična, i obrnuto, ako je vrednost veća od jedan, onda se ona može tumačiti kao asimetrična

Kurtosis (spljoštenost distribucije) je parametar koji pruža informaciju o rasprostranjenosti distribucije po y osi. Kurtosis zapravo govori o tome u kojoj meri su vrednosti koncentrisane oko aritmetičke sredine. Kurtosis se precizno izračunava kao četvrti momenat distribucije i ima vrednost 3 za normalnu distribuciju. Tačnije, od vrednosti koju smo dobili oduzimamo 3 tako da idalna distribucija ima kurtosis = 0. Spljoštene distribucije imaju negativnu (-) vrednost a zašiljene pozitivnu (+) vrednost kurtosisa.

Na osnovu vizuelnog prikaza, mi smo utvrdili da varijabilnost može da bude drugačija u zavisnosti od distribucije. Obzirom da je varijabilnost veoma važna karakteristika svake distribucije, u statistici postoje numeričke **mere varijabilnosti.** Mere varijablinosti imaju za cilj da numerički precizno ukažu u kojoj meri distribucija vrednosti odstupa od centralne tendencije. **Opseg** (Range) je najjednostavnija mera varijabilnosti i on odgovara razilici između najveće (max) i najmanje (min) izmerene vrednosti u nizu. U našem gornjem primeru u kojima smo dali prikaz dstribucije godina ispitanika Opseg = 90 – 18 = 72. Dakle, najstariji ispitanik je imao 90 godina, najmlađi 18, a range nam govori o tome u kom rasponu se kreće distribucija vrednosti na ovoj varijabli. Treba, međutim, imati u vidu da je opseg jedna od mera varijabilnosti koja je sasvim nedostatna, naime, u okviru opsega (max-min) moguće su sasvim različite disribucije vrednosti.

Standardna devijacija je jedna od ključnih mera varijabilnosti koja ukazuje u kojoj su meri vrednosti udaljene od aritmetičke sredine. Da bi izračunali standardu devijaciju nužno je prvo izračunati **varijansu** a ona pretpostavlja da je n brojeva u datom uzorku jednak sumi kvadrata distance od aritmetičke sredine podeljeno sa ukupnim brojem vrednosti minus 1 (n-1). Varijansa se izračunava po sledećoj formuli:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Npr. ako je niz brojeva 1,2,3,4,5, aritmetička sredina je 3 i onda je varijansa:

$$s^2 = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5-1} = \frac{4+1+0+1+4}{4} = 2.5$$

Na osnovu varijanse se izračunava standardna devijacija, a ona predstavlja pozitivni kvadratni koren varijanse. Evo formule:

$$s = \sqrt{s^2}$$

Prema tome u našem primeru SD je: $s = \sqrt{2.5} = 1.58$

Standardna devijacija uz aritmetičku sredinu nam pruža veoma korisne informacije. Ključna uloga standardne devijacije jeste da se na osnovu nje formiraju intervali poverenja. Uobičajeno je da se vrednost aritmetičke sredine, prikazane uz standardnu devijaciju razumevaju tako što se aritmetičkoj sredini dodaje/oduzima jedna ili dve standardne devijacije po sledećem principu:

± 1 SD $\quad \bar{x} - s, \quad \bar{x} + s$

± 2 SD $\quad \bar{x} - 2s, \quad \bar{x} + 2s$

Dakle, aritmetička sredina starosti naših ispitanika je 44,28, a standardna devijacija je 16,82. Sledi:

± 1 SD

44,28 – 16,82 = **27,46** i 44,28 + 16,82 = **61,1;**

± 2 SD

44,28 – 33,64 = **10,64** i 44,28 + 33,64 = **77,92**;

Promptni zaključak: Distribucija je asimetrična i to udesno, naime, sa +/- dve standardne devijacije mi smo potpuno iscrpeli varijansu u desno (niske vrednosti tj. mlađi), dok još uvek nismo obuhvatili celokupnu varijansu ulevo (visoke vrednosti tj. stariji)

Pojednostavljeno, za početak ćemo definisati empirijsko pravilo za za interpretaciju standardne devijacije. Ukoliko je distribucija normalna važi:

- Oko 68% vrednosti će biti obuhvaćene +/- 1S
- Oko 95% vrednosti će biti obuhvaćene +/- 2S
- Oko 99,7% vrednosti će biti obuhvaćene +/- 3S

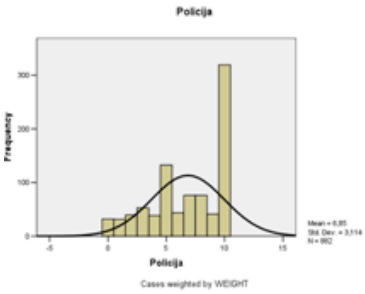
Ovakav pristup se može koristiti na vrlo jednostavan način. Npr. ako su prosečna primanja u Crnoj Gori 300 EUR sa standardnom devijacijom 130 EUR. To znači da oko 68%populacije u Crnoj Gori ima primanja između 170 i 430 EUR i oko 95% populacije ima platu od 40 do 560 EUR (da li je ovo slučaj u Crnoj Gori i ako nije šta iz toga sledi?). Ako prosečan Crnogorac provede 3 sata pored televizora dnevno sa standardnom devijacijom od 1 sat, to znači da oko 68% populacije provodi pored TV-a između 2 i 4 sata i 95% populacije gleda TV između 1 i 5 sati. Ako je prosečna ocena na skali od 1-5 za X političara 3.0 sa standardnom devijacijom 1.5, to znači da ovog političara 68% populacije ocenjuje ocenom od 1.5 do 4.5

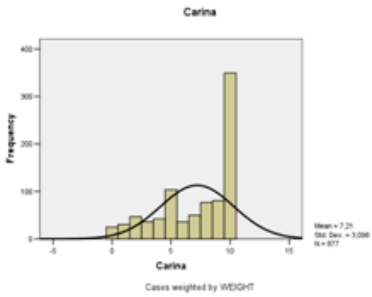
Percentili predstavljaju sumarne kumulativne pokazatelje distribucije iskazane relativnim frekvencijama. Percentili u suštini dele ukupnu kumulativnu distribuciju u jednoj tački, te prema tome se distribucija deli na dva dela. Najčešći u upotrebi je tzv 50-ti percentil, koji ima vrednost medijane. Kada je o njemu reč, ukupna distribucija se deli na način da polovina slučajeva (objekata) ima vrednosti ispod a polovina iznad medijane. U upotrebi su takođe i 25-ti, 75-ti percentil, a nekada se koriste i 10-ti i 90-ti percentil. Upotreba je jednostavna, npr. ako je u našem primeru sa godinama vrednost na 25-om percentilu =30, to znači da 25% populacije jeste mlađe od 30 godina. Budući da je vrednost 75-og percentila 57,6, to znači da je 75% naših ispitanika mlađe od 57,6 dogina. Konsekventno, pošto je vrednost na 90-tom percentilu 68, proističe da je 90% naših ispitanika mlađe od 68 godina. Budući da 25-ti, 50-ti i 75-ti percentil deli populaciju u četiri grupe, njih nazivamo **kvartilima** a razliku između 75-tog i 25-tog, nazivamo **interkvartilni opseg** (IQR – interquartile range).

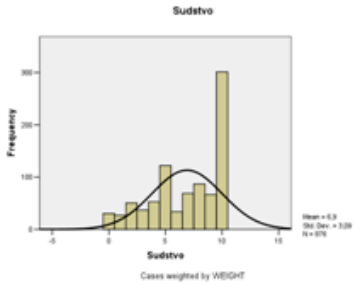
Evo jednog primera merenja percepcije korupcije i distribucije koja je dobijena istraživanjem sa svim statisticima koje smo opisivali:

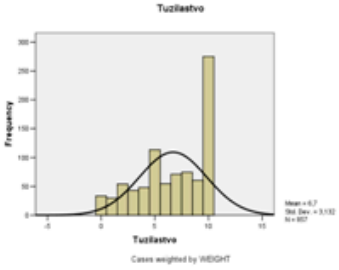
Institucija - oblast / Stav - ocjena	Korupcije											Ne znam, ne mogu da procijenim
	Nema										Ima	
1. Policija	0	1	2	3	4	5	6	7	8	9	10	99
2. Carina	0	1	2	3	4	5	6	7	8	9	10	99
3. Sudstvo	0	1	2	3	4	5	6	7	8	9	10	99
4. Tužilaštvo	0	1	2	3	4	5	6	7	8	9	10	99
5. Zdravstvo	0	1	2	3	4	5	6	7	8	9	10	99
6. Mediji (TV, radio, novine)	0	1	2	3	4	5	6	7	8	9	10	99
7. Sport	0	1	2	3	4	5	6	7	8	9	10	99
8. Osnovno školstvo	0	1	2	3	4	5	6	7	8	9	10	99
9. Srednje školstvo	0	1	2	3	4	5	6	7	8	9	10	99
10. Visoko školstvo (Univerzitet)	0	1	2	3	4	5	6	7	8	9	10	99
11. Državne službe	0	1	2	3	4	5	6	7	8	9	10	99
12. Opštinske službe	0	1	2	3	4	5	6	7	8	9	10	99

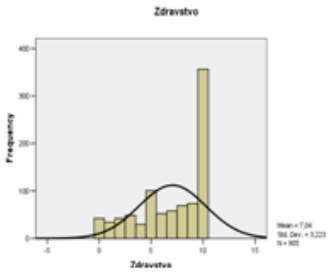
		Policija	Carina	Sudstvo	Tuzilastvo	Zdravstvo	Mediji	Sport
N	Valid	882	877	876	857	905	853	850
	Missing	129	134	135	154	106	158	161
Mean		6,85	7,21	6,90	6,70	7,04	5,88	5,64
Std. Error of Mean		,105	,105	,104	,107	,107	,112	,116
Median		7,00	8,00	8,00	7,00	8,00	6,00	5,00
Mode		10	10	10	10	10	10	10
Std. Deviation		3,114	3,098	3,090	3,132	3,223	3,271	3,389
Variance		9,694	9,597	9,549	9,811	10,389	10,701	11,486
Skewness		-,581	-,783	-,626	-,533	-,752	-,202	-,172
Std. Error of Skewness		,082	,083	,083	,084	,081	,084	,084
Kurtosis		-,853	-,673	-,817	-,922	-,725	-1,187	-1,266
Std. Error of Kurtosis		,164	,165	,165	,167	,162	,167	,168
Range		10	10	10	10	10	10	10
Minimum		0	0	0	0	0	0	0
Maximum		10	10	10	10	10	10	10
Sum		6046	6319	6047	5741	6377	5015	4793
Percentiles	25	5,00	5,00	5,00	5,00	5,00	3,00	3,00
	50	7,00	8,00	8,00	7,00	8,00	6,00	5,00
	75	10,00	10,00	10,00	10,00	10,00	9,00	9,00
	90	10,00	10,00	10,00	10,00	10,00	10,00	10,00

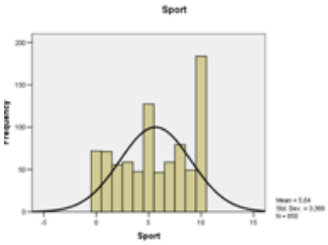


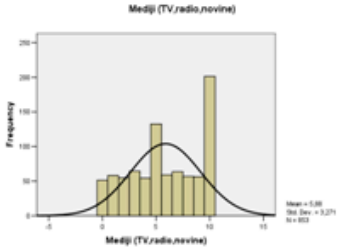










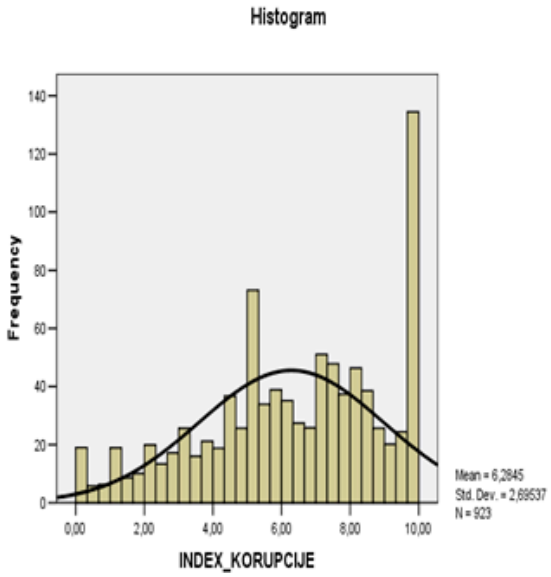


Na osnovu svih ajtema, formiramo jedinstvenu skalu, evo distribucije:

Statistics

INDEX_KORUPCIJE

N	Valid	923
	Missing	88
Mean		6,2845
Std. Error of Mean		,08871
Median		6,5000
Mode		10,00
Std. Deviation		2,69537
Variance		7,265
Skewness		-,405
Std. Error of Skewness		,080
Kurtosis		-,648
Std. Error of Kurtosis		,161
Range		10,00
Minimum		,00
Maximum		10,00
Sum		5801,86
Percentiles	25	4,5833
	50	6,5000
	75	8,4167
	90	10,0000



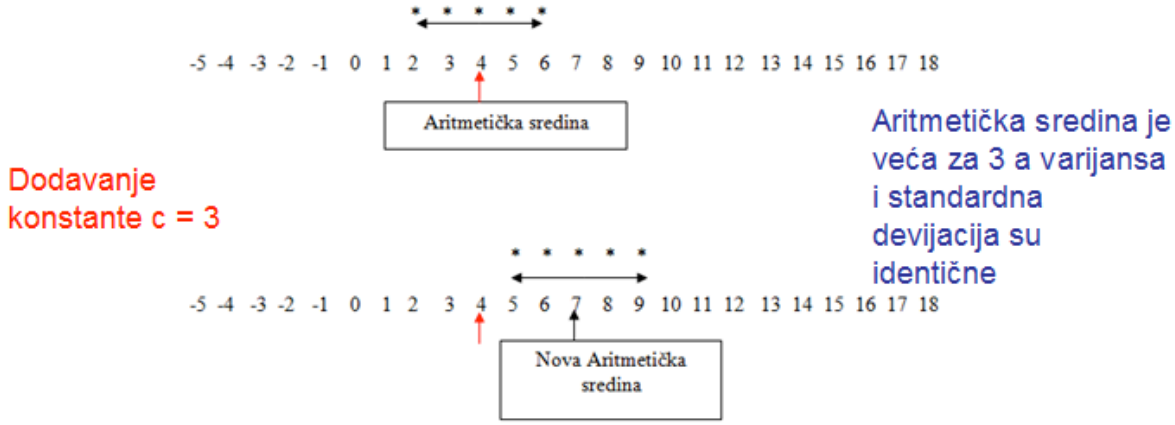
Fomiranje skorova na osovnu većeg broja varijabli i opersanje sa ukupnim skorom kao posebnom varijablom jedna je od veoma čestih i korisnih procedura u društvenim i političkim istraživanjima. U daljem izlaganju imaćemo veliki broj primera koji upravo koričćenjem ove procedure omogućavaju istraživački posao koji kanimo da naučimo na ovom kursu.

Deskriptivna statistika II

U ovom poglavlju naučićemo:

- Da matematički operišemo sa aritmetičkom sredinom i da razumemo šta se dešava kao posledica klasičnih operacija (+,-,* i /)
- Da razumemo potrebu i primenu standardizacije skorova
- Da naučimo kako da tranformišemo varijablu u Z skorove
- Da posredstvom Z skorova razumemo intervale poverenja od 95% i 99%
- Da naučimo da interpretiramo intervale poverenja aritmetičke sredine
- Da naučimo kako da identifikujemo ekstremne vrednosti (outliere)
- Da razumemo alternativne mere centralne tendencije 5% odsečena sredina i M estimator

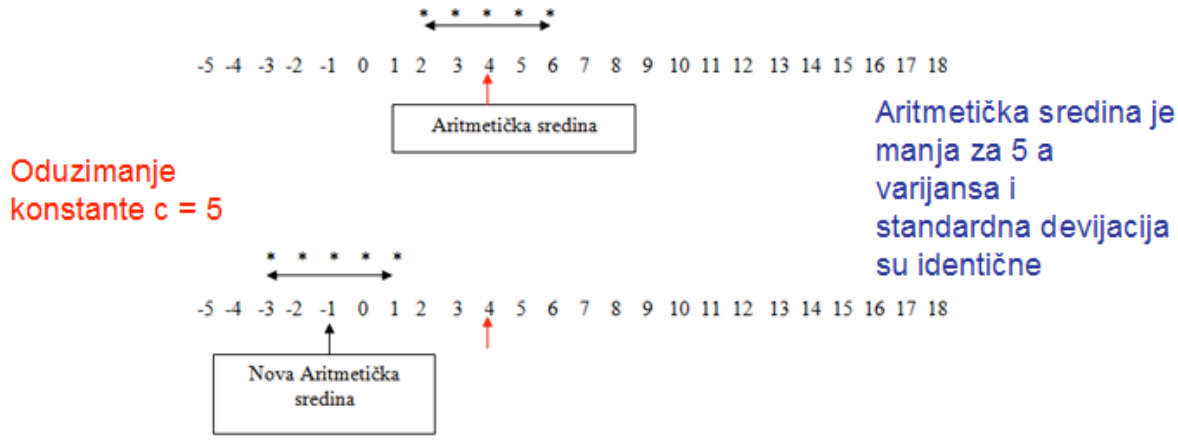
Najpre kada je reč o jednostavnim matematičkim operacijama, pogledajmo šta se dešava kada **dodajemo konstantu**:



Ako na originalne vrednosti jednog seta podataka dodamo konstantu (c) svakoj od vrednosti u nizu, novoformirani skor će imati aritmetičku sredinu koja iznosi *originalna aritmetička sredina + konstanta*

Ako na originalne vrednosti jednog seta podataka dodamo konstantu (c), novoformirani skor će imati istu varijansu i standardnu devijaciju kao što je to bio slučaj sa originalnim setom podataka

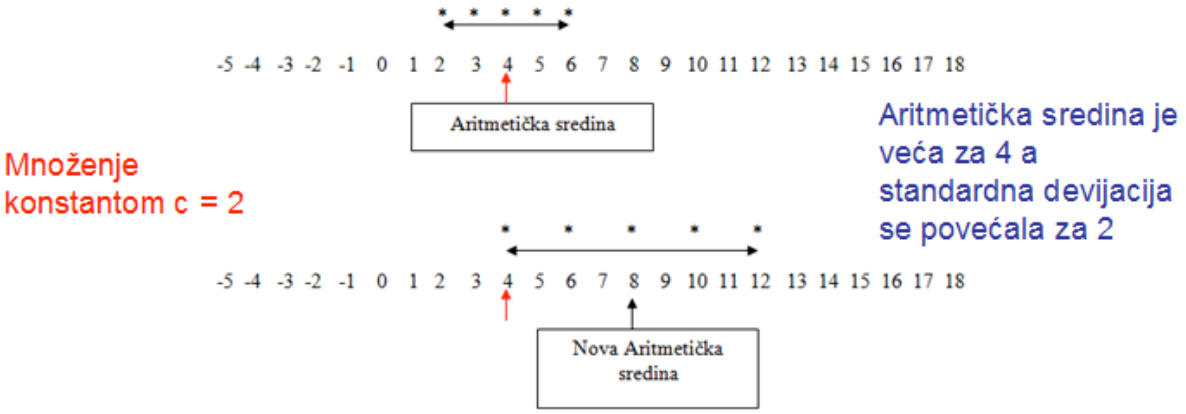
Ukoliko **oduzimamo konstantu** dešava se sledeće:



Ako od originalnih vrednosti jednog seta podataka oduzmemo konstantu (c), od svake vrednosti u niizu, novoformirani skor će imati aritmetičku sredinu koja iznosi *originalna aritmetička sredina – konstanta*

Ako od originalnih vrednosti jednog seta podataka oduzmemo konstantu (c), novoformirani skor će imati istu varijansu i standardnu devijaciju kao što je to bio slučaj sa originalnim setom podataka

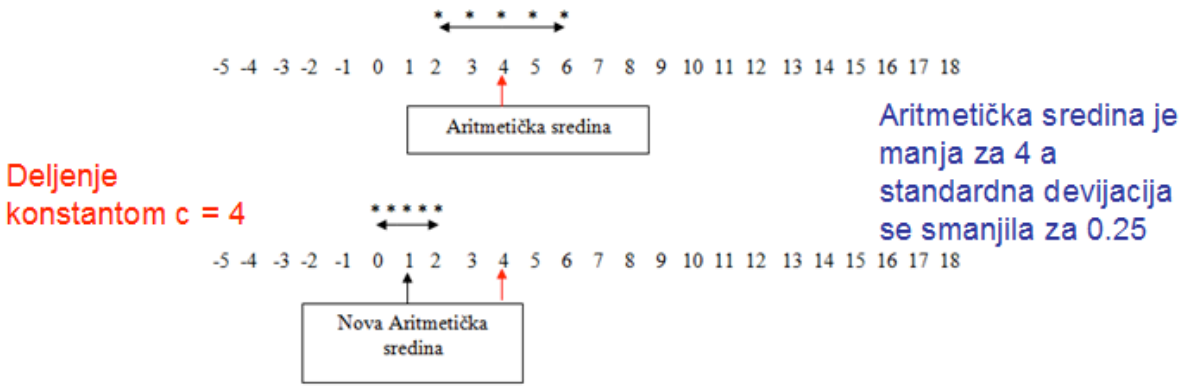
Ukoliko **množimo sa konstantom** dešava se sledeće:



Ako originalne vrednosti jednog seta podataka pomnožimo sa konstantom (c), i to množenjem svake vrednosti u nizu, novoformirani skor će imati aritmetičku sredinu koja iznosi *originalna aritmetička sredina * konstanta*

Ako originalne vrednosti jednog seta podataka pomnožimo konstantom (c), novoformirani skor će imati varijansu puta veću od varijanse originalnog seta i standardnu devijaciju koja je ***c*** puta veća od originalnog seta podataka

Ukoliko **delimo sa konstantom** dešava se sledeće:



Ako originalne vrednosti jednog seta podataka podelimo sa konstantom (c), i to deljenjem svake vrednosti u niizu, novoformirani skor će imati aritmetičku sredinu koja iznosi *originalna aritmetička sredina / konstanta*

Ako originalne vrednosti jednog seta podataka podelimo konstantom (c), novoformirani skor će imati varijansu 1/ puta veću od varijanse originalnog seta i standardnu devijaciju koja je 1/***c*** puta veća od originalnog seta podataka

Čemu služe operacije sa aritmetičkom sredinom? U društvenim istraživanjima uobičajeno da se društvene i političke pojave mere na način formiranja određenih skorova (skala i indexa)
Kada se formiraju skorovi, neretko želimo da optimisujemo skorove sa ciljem da numerički nizovi budu smisleni za interpretaciju (recimo od 1-100 ili od 0 – 10). U ovim postupcima obično smo prinuđeni da promenimo originalan set podataka te prema tome moramo znati šta će se desiti sa aritmetičkom sredinom i konstantom u situacijama da primenimo neke od operacija

Standardizacija skorova

Do sada smo zaključili da svaka vrednost na nekoj skali (npr 115) ima smisla samo ukoliko znamo distribuciju vrednosti na datoj varijabli i ukoliko u odnosu na ovu distribuciju možemo da interpretiramo mesto koje ima ta pojedina vrednost (115). Naime, ukoliko znamo da je aritmetička sredina na toj varijabli 110, onda znamo da je 115 veće od aritmetičke sredine. Ali ni ovaj podatak nam ne govori dovoljno, naime, ova vrednost 115 bitno zavisi od distribucije, tj. postavljamo pitanje da li je ta vrednost unutar jedne, dve ili tri standardne devijacije. Drugim rečima, vrednost od 115 na skali ima potpuno drugačiji smisao u zavisnosti od toga da li je standardna devijacija npr. 5 ili 15 poena.

Kako bi identifikovali relativno mesto opservirane vrednosti na datoj distribuciji, tada nije dovoljno samo da znamo devijaciju, već je potrebno da devijaciju transformišemo u numeričku vrednost koja odgovara standardnoj devijaciji za dati skor. Cilj ovakvog postupka bio bi da lociramo svaku vrednost na način da odredimo koliko je standardnih devijacija ta vrednost udaljena od aritmetičke sredine. Prema tome, ideja standardizacije skorova je veoma jednostavna,

ali važna i upotrebljiva, naime, u okviru svake varijable, na osnovu pretpostavke o normalnoj distribuciji, svaka vrednost će biti transformisana tako da novoformirana vrednost jednostavno numerički izražava koliko je standardnih devijacija udaljena od aritmetičke sredine.

Npr., imamo skor od 450 u distribuciji koja ima aritmetičku sredinu 400 i standardnu devijaciju 25. Budući da skor odstupa od aritmetičke sredine za 50 a standardna devijacija je 25, mi jednostavno znamo da ako podelimo 50 sa 25 (50/25) dobijamo vrednost 2. Ovaj podatak govori o tome da je naš skor u okvirima od 2 standardne devijacije u odnosu na aritmetičku sredinu. Na ovom primeru dat je univerzalan način za formiranje tzv. ***z Skorova***.

Standardizovani *z Skorovi se dakle izračunavaju:*

$$zSkor = \frac{posmatranavrednost - aritmetickasredina}{s\tan dardnadevijacija}$$

Izraženo formulom: $$z = \frac{x_i - \bar{x}}{s}$$

U prethodnom primeru: $$z = \frac{450 - 400}{25} = 2.0$$

U ovoj situaciji nam novoformirana vrednost iskazana preko *z Skora* ukazuje da je originalna vrednost veća za dve standardne devijacije u odnosu na aritmetičku sredinu, što je mnogo informativnije i jednostavnije za interpretaciju. Dalje, u jednom jedinom numeričkom podatku, *z Skoru,* dat je na najjednostavniji način podatak i kolika je aritmetička sredina i kolika je standardna devijacija.

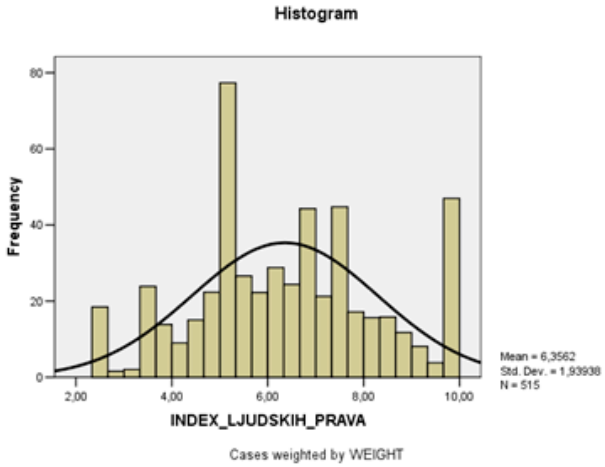
Upotrebljivost *Z skorova* je u tome što mi u statističkoj obradi podataka možemo lako i efikasno da sve vrednosti jedne varijable transformišemo u z skorove. Na ovaj način nova varijabla predstavlja standardizovanu varijablu. Ključno je važno znati da **bez obzira kolika je aritmetička sredina i standardna devijacija originalne varijable, novoformirana standardizovana varijabla koja se bazira na Z skorovima imaće aritmetičku sredinu = 0 i standardnu devijaciju = 1.** Na novoformiranoj varijabli, lako možemo onda analizom distribucije z skorova da na jedan jednostavan način uočimo odstupanja od srednje vrednosti i to u pravim jedinicama odstupanja koje odgovaraju meri standardne devijacije

Evo jednog primera originalne i transformisane varijable:

Originalna varijabla:

INDEX_LJUDSKIH_PRAVA

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2,50	18	1,9	3,6	3,6
	2,78	2	,2	,3	3,9
	3,06	2	,2	,4	4,3
	3,33	7	,8	1,4	5,7
	3,61	17	1,7	3,2	8,9
	3,89	14	1,4	2,7	11,6
	4,17	9	,9	1,7	13,4
	4,44	15	1,5	2,9	16,3
	4,72	22	2,3	4,3	20,6
	5,00	57	5,8	11,0	31,6
	5,28	21	2,1	4,0	35,6
	5,56	27	2,7	5,2	40,8
	5,83	22	2,3	4,3	45,1
	6,11	29	2,9	5,6	50,7
	6,39	24	2,5	4,7	55,4
	6,67	24	2,4	4,6	60,0
	6,94	21	2,1	4,0	64,0
	7,22	21	2,2	4,1	68,2
	7,50	45	4,6	8,7	76,8
	7,78	17	1,7	3,3	80,2
	8,06	16	1,6	3,0	83,2
	8,33	6	,6	1,1	84,3
	8,61	10	1,0	2,0	86,3
	8,89	12	1,2	2,3	88,6
	9,17	8	,8	1,6	90,1
	9,44	4	,4	,7	90,9
	9,72	4	,4	,7	91,6
	10,00	43	4,4	8,4	100,0
	Total	515	52,5	100,0	
Missing	System	465	47,5		
Total		981	100,0		



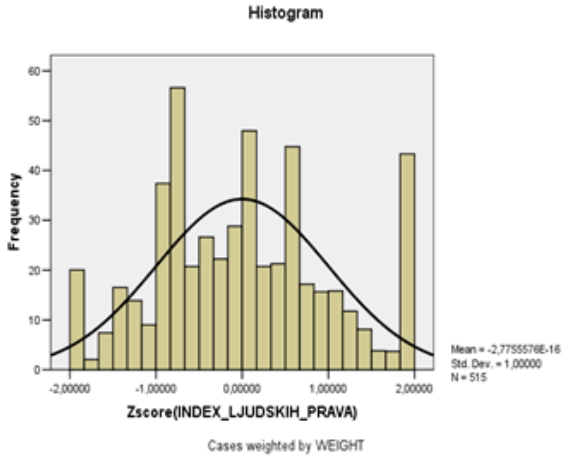
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
INDEX_LJUDSKIH_PRAVA	515	2,50	10,00	6,3562	1,93938
Valid N (listwise)	515				

Transformisana varijabla (z skorovi):

Zscore(INDEX_LJUDSKIH_PRAVA)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1,98835	18	1,9	3,6	3,6
	-1,84512	2	,2	,3	3,9
	-1,70189	2	,2	,4	4,3
	-1,55867	7	,8	1,4	5,7
	-1,41544	17	1,7	3,2	8,9
	-1,27221	14	1,4	2,7	11,6
	-1,12898	9	,9	1,7	13,4
	-,98575	15	1,5	2,9	16,3
	-,84252	22	2,3	4,3	20,6
	-,69929	57	5,8	11,0	31,6
	-,55606	21	2,1	4,0	35,6
	-,41283	27	2,7	5,2	40,8
	-,26960	22	2,3	4,3	45,1
	-,12637	29	2,9	5,6	50,7
	,01686	24	2,5	4,7	55,4
	,16009	24	2,4	4,6	60,0
	,30332	21	2,1	4,0	64,0
	,44655	21	2,2	4,1	68,2
	,58978	45	4,6	8,7	76,8
	,73301	17	1,7	3,3	80,2
	,87624	16	1,6	3,0	83,2
	1,01947	6	,6	1,1	84,3
	1,16270	10	1,0	2,0	86,3
	1,30593	12	1,2	2,3	88,6
	1,44916	8	,8	1,6	90,1
	1,59239	4	,4	,7	90,9
	1,73562	4	,4	,7	91,6
	1,87885	43	4,4	8,4	100,0
	Total	515	52,5	100,0	
Missing	System	465	47,5		
Total		981	100,0		

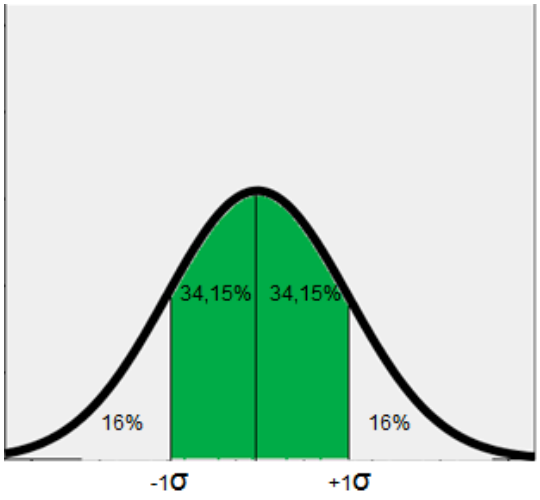


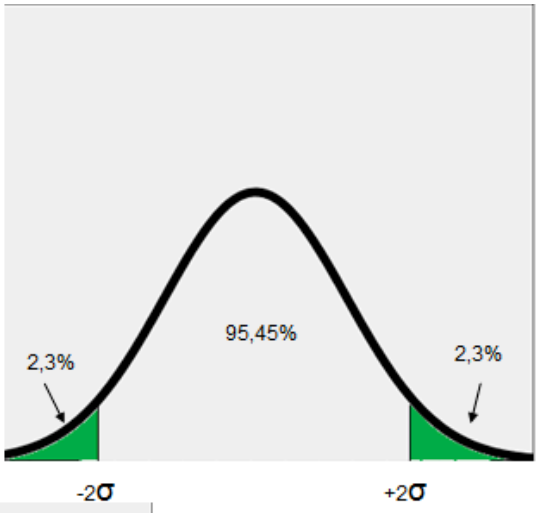
Descriptive Statistics

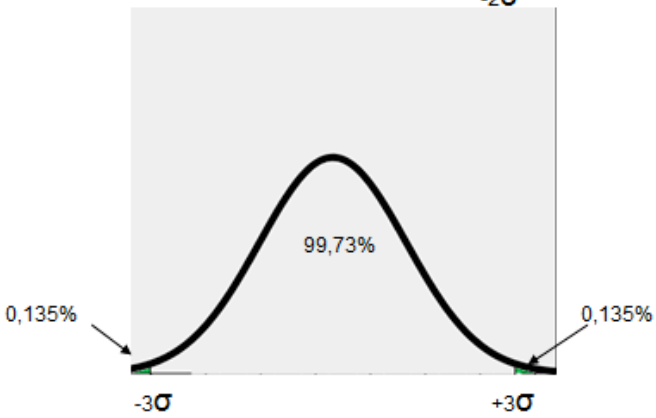
	N	Minimum	Maximum	Mean	Std. Deviation
Zscore(INDEX_LJUDSKIH_PRAVA)	515	-1,98835	1,87885	,0000000	1,00000000
Valid N (listwise)	515				

Zahvaljujući Z skorovima, relativno je jednostavno izračunati koji procenat varijanse je pokriven unutar određenog Z skora (kao što se može videti u Excel tabeli), a

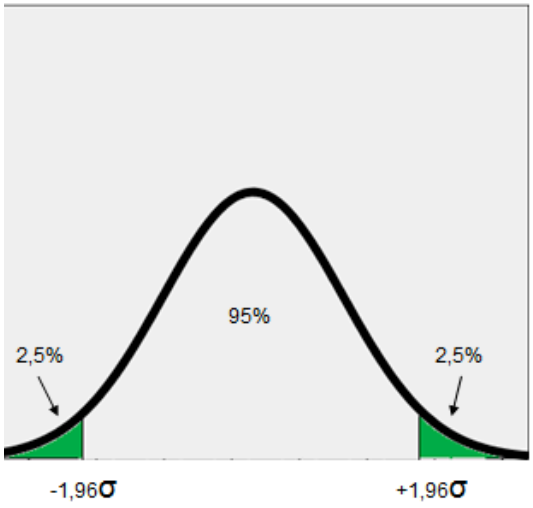
standardni kriterijumi intervala poverenja od 95% i 99% mogu se videti na grafikonima koji slede:

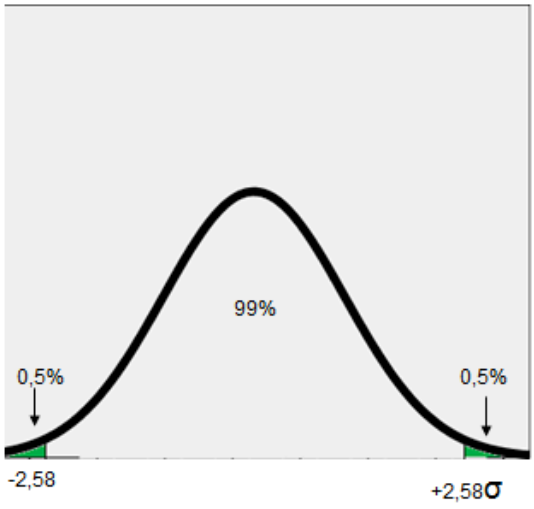






Na osnovu ovoga moguće je relativno jednostavno postavti dva ključna kriterijuma za intervale poverenja 95% i 99%, a na osnovu broja devijacija od aritmetičke sredine i polja koje je ′pokriveno′ u okviru distribucije:





DVA KLJUČNA STANDARDA KOJA ĆE KASNIJE UNIVERZALNO VAŽITI ZA ODREĐIVANJE STATISTIČKE ZNAČAJNOSTI (TZV. **p vrednost (α)**)

sr[illegible]m
[illegible]ce

standardna devijacija. Radi testiranja hipoteza, a ovo će biti predmet u daljem izlaganju, u statistici se koriste dva standarda, 95% i 99% i ovo su prema tome dva uobičajena intervala poverenja u okviru kojih interpretiramo rezultate.

> Na grafikonima uočiti i **zapamtiti** da je 95% interval poverenja +/- 1,96 standardne devijacije, a 99% interval poverenja +/- 2,58 standardne devijacije.

Koliko možemo imati poverenja u aritmetičku sredinu?

Aritmetička sredina je ključna mera centralne tendencije zato što veliki broj statističkih metoda kojima se testiraju hipoteze operiše sa ovim parametrom. No obzirom da je ovaj podatak proizvod procene koji se bazira na uzorku, postavlja se pitanje njegove preciznosti, ili drugim rečima, uzorak po sebi sadrži grešku merenja, jer znamo da je:

μ približno jednako $\bar{x}$

S toga, ključna stvar jeste da na neki validan način procenimo poverenje koje možemo imati u dobijeni podatak. Evo jednog dijaloga koji na mlastičan način ukazuje na problem

Dijalog

Istraživač: Ja sam obavio istraživanje na bazi slučajnog uzorka i na osnovu rezultata sam dobio podatak da je aritmetička sredina ukupnog broja završenih godina školovanja u Crnoj Gori 11,87. Budući da sam očekivao da je ta srednja vrednost manja, mora da je neki problem sa uzorkom

Statističar: Zašto bi problem bio sa uzorkom, je li uzorak bio slučajan ili nije?

Istraživač: Da, bio je slučajan i ukupno je bilo 1000 ispitanika

Statističar: A kolika je standardna devijacija?

Istraživač: 3.083
Statističar: (nekoliko minuta provodi za računarom i zaključuje)...Ne, ne, sve je u redu, podatak koji si dobio je sasvim OK, u čemu je problem?

Istraživač: Pa problem je u tome što ja mislim da je rezultat mog istraživanja proizvod 'loše sreće' u pogledu izbora ispitanika i mislim da kada bi ponovio istraživanje ja ne bih dobio istu vrednost.

Statističar: Vidi, imaš sreće, ja slučajno imam podatke sa popisa o celokupnoj populaciji koji uključuju podatke o broju završenih godina školovanja. Ako želiš mogu da izvučem jedan uzorak od isto tako 1000 ispitanika da proverimo.

Istraživač: Sjajno! Uradi to što pre...

Statističar: Evo odmah, to nije nikakav problem imamo bazu podataka u računaru. Izvukao sam jedan uzorak i dobio sam podatak da je na bazi tog uzorka prosečan broj godina školovanja 11,79, dakle, sve je uredu sa tvojim istraživanjem.

Istraživač: Pa, prosek koji si ti dobio jeste ipak malo manji od onog koji sam ja dobio, biće ipak da sam ja bio loše sreće... Iako je i taj podatak daleko iznad mog očekivanja

Statističar: Ne, ne slažem se da si bio loše sreće evo, napravićemo dvadeset uzoraka pa da proverimo koju ćemo arimtetičku sredinu dobiti:

- Uzorak 2: 11,88 Uzorak 3: 12,01 Uzorak 4: 12,06
- Uzorak 5: 11,92 Uzorak 6: 11,69 Uzorak 7: 11,71
- Uzorak 8: 12,04 Uzorak 9: 11,77 Uzorak 10: 11,99
- Uzorak 11:11,71 Uzorak 12:11,95 Uzorak 13: 12,05
- Uzorak 14:12,00 Uzorak 15:11,90 Uzorak 16: 12,04
- Uzorak 17:11,83 Uzorak 18:11,59 Uzorak 19: 12,01
- Uzorak 20:11,85

Istraživač: Vidi, sve vrednosti koje si dobio su jako blizu, jesi li ti siguran da je sve u redu sa računarom?

Statističar: Naravno da sam siguran, ja ne znam na osnovu kojih informacija si ti bazirao svoja očekivanja, ali koliko vidim od 20 uzoraka, samo jedna vrednost koju sam dobio u uzorku br 18, tačnije da je prosek 11,59, je izvan intervala poverenja koji sam mogao da izračunam na osnovu tvog proseka, dok je prosek svih ostalih uzoraka u okviru intervala poverenja od 95%.

Istraživač: O kakvim to intervalima govoriš?

Statističar: Govorim o intervalu povrenja od 95%, naime to je klasičan standard koji validira dobijene podatke, naročito kada je reč o aritmetičkoj sredini

Istraživač: I kako si to izračunao moliću lepo?

Statističar. Jednostavno, rekao si da si dobio prosek 11,87, da ti je uzorak bio slučajan sa ukupnim brojem od 1000 ispitanika i da je standardna devijacija 3,083

Istraživač: Tačno tako, i šta s tim?

Statističar: Dakle, po tvojim podacima možemo reći da je verovatnoća da je aritmetička sredina koju si dobio rezultat 'loše sreće' jednaka verovatnoći 1: 20

Istraživač: Kako to?

Statističar: Jednostavno, svaka aritmetička sredina po prirodi stvari budući da je rezultat uzorka a ne čitave populacije sadrži standardnu grešku merenja. Ova greška se izračunava tako što se standardna devijacija (3,083) podeli sa kvadratnim korenom ukupnog broja ispitanika:

$$\sigma_{\bar{x}} = \sigma / \sqrt{1000}$$

Statističar: Dakle, kad obavim ovu operaciju dobijam vrednost da je greška aritmetičke sredine 0,098. Na osnovu toga ja znam sa 95% sigurnosti da se prosek ukupnog broja školovanja u Crnoj Gori kreće: 11,87± 1.96*0.098. Tačnije, sa 95% poverenja znam da je tvoja aritmetička sredina između 11,68 i 12,06. Ukoliko pogledaš aritmetičke sredine koje smo dobili na osnovu 20 uzoraka, jasno je da samo jedan uzorak (br 18 gde je aritmetička sredina 11,59) ima aritmetičku sredinu koja nije u okviru ovog intervala, što je potpuno u skladu sa samim intervalom, jer je 1 uzorak od 20 tačno iznosi 5% verovatnoće.

Istraživač: Sad sam zbunjen, šta tačno hoćeš da kažeš?

Statističar: Hoću da kažem da ukoliko biramo 100 uzoraka u 95 od njih naći ćemo da se aritmetička sredina broja završenih godina školovanja kreće u rasponu od 11,68 do 12,06, a u 5 od tih uzoraka možemo naći da to nije tako. Ovo je razlog da govorimo u kategorijama 95% intervala poverenja, i da kažemo da možemo prilično (sa 95% sigurnosti) biti uvereni u podatak.

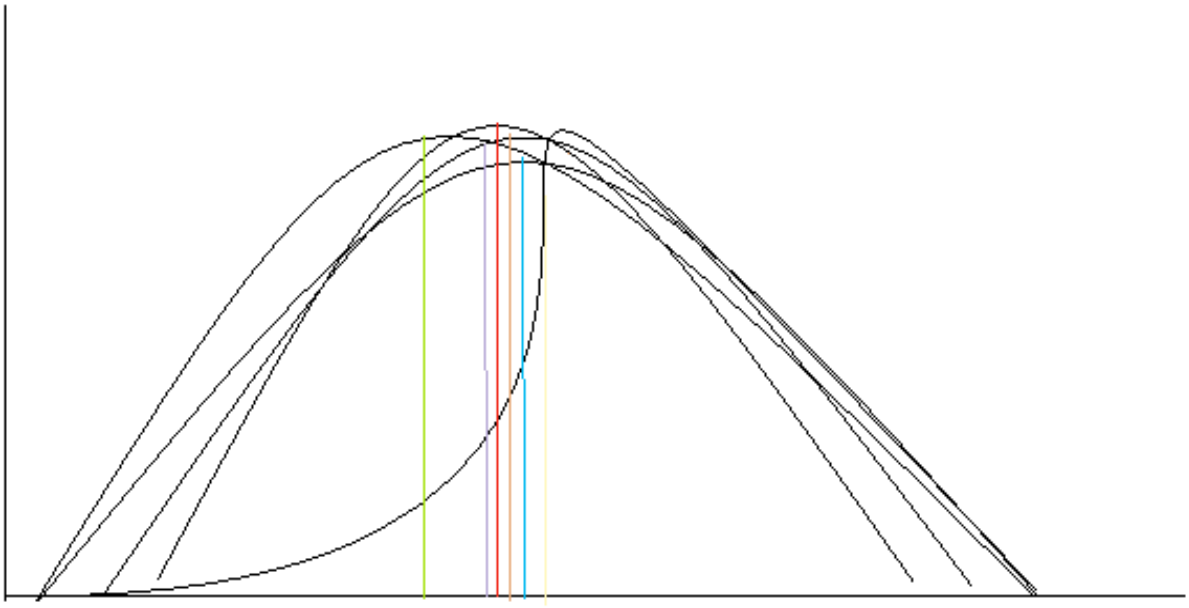
Istraživač: Dobro, ali nikako mi nije jasna matematika koju si izveo za taj interval, tačnije, jasno mi je kako si izračunao standardnu grešku aritmetičke sredine, ali nikako mi nije jasno zašto si tu grešku množio sa 1,96???

Statističar: Jednostavno zato što tako preporučuje centralna granična teorema, naime, ako je distribucija normalna, onda polje koje pokriva 95% vrijanse sa obe strane distribucije ostavlja prostor od po 2,5% na krajevima distribucije a 2,5% polja odgovara vrednosti od 1,96 standardne devijacije. Dakle, 2,5% površine znači da standardnu grešku aritmetičke sredine moramo množiti sa 1.96, a onda dobijenoj vrednosti dodati i oduzeti tih 2,5% sa obe strane

Istraživač: Dobro, dobro, predajem se.... Prihvatama da je podatak koji sam dobio sasvim dobar

Statističar: On je onoliko dobar koliko smo to izrazili 95% intervalom poverenja, ni više ni manje od toga.....

Evo kako izgledaju aritmetičke sredine na većem pbroju uzoraka iste populacije:



Evo kako izgleda primer iz Dijaloga:

N-1000

			Statistic	Std. Error
Ukupan broj zavrsenih godina skolovanja	Mean		11,87	,098
	95% Confidence Interval for Mean	Lower Bound	11,68	
		Upper Bound	12,06	

$\sigma_{\bar{x}} = \sigma / \sqrt{1000}$ $\sigma_{\bar{x}} = 11{,}87/\sqrt{1000} = 0.098$

$95\%CI = \bar{x} \pm 1.96 * \sigma_{\bar{x}}$

(11,87- 1.96*0.098) < 95%CI < (11,87- 1.96*0.098)

95% CI= od 11,68 do 12,06

Ukoliko želimo da izrazimo intervale poverenja od 99% sledi:

$$99\%CI = \bar{x} \pm 2.58 * \sigma_{\bar{x}}$$

$$\sigma_{\bar{x}} = 11,87/\sqrt{1000} = 0.098$$

(11,87- 2.58*0.098) < 95%CI < (11,87- 2.58*0.098)

99% CI= od 11,62 do 12,12

Mean		11,87
95% Confidence Interval for Mean	Lower Bound	11,68
	Upper Bound	12,06

Pored opisanih mera centralne tendencije, u praksi se koriste i nekealternativne mere, koje mogu biti veoma korisne u nekim specifičnim situacijama, naročito onda kada imamo problema sa distribucijom. Evo njpe primera nekoliko statistika koje ćemo elaborirati:

Descriptives

			Statistic	Std. Error
Ukupan broj zavrsenih godina skolovanja	Mean		11,87	,098
	95% Confidence Interval for Mean	Lower Bound	11,68	
		Upper Bound	12,06	
	5% Trimmed Mean		12,08	
	Median		12,00	
	Variance		9,504	
	Std. Deviation		3,083	
	Minimum		0	
	Maximum		22	
	Range		22	
	Interquartile Range		2	
	Skewness		-1,198	,078
	Kurtosis		3,743	,156

Extreme Values

			Case Number	Value
Ukupan broj zavrsen godina skolovanja	Highest	1	727	22
		2	83	21
		3	398	20
		4	616	20
		5	857	20
	Lowest	1	1004	0
		2	956	0
		3	836	0
		4	813	0
		5	776	0[a]

a. Only a partial list of cases with the value 0 are shown in t of lower extremes.

M-Estimators

	Huber's M-Estimator[a]
Ukupan broj zavrsenih godina skolovanja	12,05

a. The weighting constant is 1.339.

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Ukupan broj zavrsenih godina skolovanja	8,00	8,00	11,00	12,00	13,00	16,00	16,00
Tukey's Hinges	Ukupan broj zavrsenih godina skolovanja			11,00	12,00	13,00		

Najpre, identifikovaćemo problem **extremnih vrednosti** (outliers). U svakoj distribuciji koja ima veliki broj vrednosti postoje tzv. ekstremne vrednosti (outlieri). Ekstremne vrednosti su,dakle, one koje su 'neuobičajeno' velike ili 'neuobičajeno'

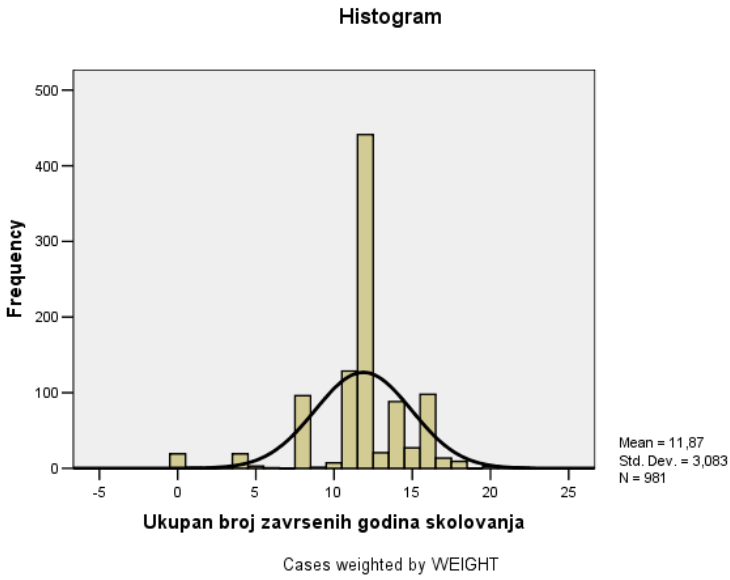
male u odnosu na ostale vrednosti. Budući da su mere centralne tendencije, naročito aritmetička sredina osetljive na extremne vrednosti, njima se u analizi mora posvetiti posebna pažnja. S toga se mi trudimo da identifikujemo extremne vrednosti, pre svega ne bi li na osnovu njihove identifikacije utvrdili da postoji neka greška u merenju. SPSS identifikuje ekstremne vrednosti, i uz to identifikuje slučajeve koji imaju extremne vrednosti i to kako u tabeli tako i koričćenjem tzv box plota:

Extreme Values

		Case Number	Value
Ukupan broj zavrsenih godina skolovanja	Highest 1	727	22
	2	83	21
	3	398	20
	4	616	20
	5	857	20
	Lowest 1	1004	0
	2	956	0
	3	836	0
	4	813	0
	5	776	0[a]

a. Only a partial list of cases with the value 0 are shown in the table of lower extremes.

Ekstremne vrednosti se takođe mogu identifikovati korišćenjem histograma:



Jedna od alternativnih mera centralne tendencije koja je alternativna aritmetičkoj sredini jeste i **5% odsečena sredina**. Ova mera centralne tendencije se formira tako što se od ukupne varijanse 'odseče' (dakle ne kalkulišu se) 5% ekstremnih vrednosti. Dobijeni podatak precizniji je u odnosu na aritmetičku sredina ukoliko ekstremne vrednosti zaista prave problem u distribuciji. U našem primeru

5% odsečena sredina = 12,08 što je znatno više od aritmetičke sredine = 11,87. Ova vrednost 5% odsečene sredine je čak izvan 95% CI, što nam ukazuje da su ekstremno male vrednosti u našoj distribuciji zaista problem u proceni aritmetičke sredine.

M estimatori jesu robusne mere centralne tendencije koje u situaciji kada imamo relativno 'duge krajeve' u distribuciji u kojima su identifikovane ekstremne vrednosti jesu bolje mere centralne tendencije u odnosu na aritmetičku sredinu. M estimatori se izračunavaju na način da se sve vrednosti u distribuciji a koje su osnov za računanje aritmetičke sredine, kalkulišu tako da veći ponder dobijaju one vrednosti koje su blizu aritmetičkoj sredini, a što se neka vrednost više udaljava od aritmetičke sredine, to je njen ponder manji. Na ovaj način, prema tome, srednja vrednost koja se dobije je manje osetljiva na ekstremne vrednosti. Razlika između M estimatora i 5% odsečene sredine je u tome što M estimatori kalkulišu sve vrednosti u distribuciji, samo svaku od tih vrednosti različito ponderišu, dok u 5% odsečena sredina, jednostavno iz kalkulacije izbacuje 5% ekstrema. U svakom slučaju vrednosti ova dva parametara moraju biti relativno slične, u suprotnom imamo ozbiljan problem sa distribucijom. U našem slučaju Huberov M estimator = 12,05 a 5% odsečena sredina = 12,08..... Ovaj podatak nam još jednom ukazuje da smo u našoj distribuciji imali relativno veliki broj malih ekstremnih vrednosti

Testiranje hipoteza korišćenjem z – statistika

Merenje razlika između aritmetičkih sredina veoma je čest postupak u testiranju hipoteza u društvenim istraživanjima. U istraživačkoj praksi, naime, veoma često mi želimo da uporedimo rezultate različitih merenja ne bi li utvrdili da li postoje razlike između dve aritmetičke sredine. Drugim rečima, u praksi nikad

nemamo dve jednake distribucije sa identičnim aritmetičkim sredinama i standardnim devijacijama. To znači da, kada poredimo aritmetičke sredine, mi primećujemo da postoje razlike između njihovih vrednosti. Pitanje je da li ove razlike imamo smatrati **stvarnim** ili su pak razlike rezultat **varijacija** u okvirima intervala poverenja definisnih standardnom greškom merenja aritmetičke sredine

Jedno od najčešćih pitanja koje se postavlja kada se upoređuju dva diskreciona ili kontinuirana niza brojeva jeste: **Da li su razlike između aritmetičkih sredina značajne?**

Evo nekoliko primera:

- Da li je razlika između aritmetičkih sredina koja meri poverenje u policiju na istoj skali od 2,35 i sudstvo 2,22 značajna razlika?
- Da li je razlika u skoru merenja korupcije između zdravstva i sudstva (6,7 i 7,2) značajna razlika?
- Da li su stariji statistički značajno tradicionalniji od mlađih mereno na skali tradicionalizma (3,6 – 3,3)?
- Da li je razlika u ocenama x političkog lidera između severa i centralnog regiona (2,31 naspram 2,15)? statistički značajna razlika?
- Da li zaposljeni muškarci zaista imaju veću platu od zaposljenih žena na istom radnom mestu (235 eur naspram 224 eur)?
- Da li se zaista religioznost povećala u odnosu na pre pet godina obzorom da je skor merenja ove godine veći (67,7 naspram 64,3)?... itd.

Drugim rečima, obzirom da su dobijene aritmetičke sredine rezultat merenja na uzorku, kolika je verovatnoća da su dobijene razlike rezultat 'greške' merenja koja nastaje uzorkovanjem. Naime, mi znamo da svaka aritmetička sredina u stvarnosti može biti malo manja i malo veća, te s toga mi računamo standardnu grešku aritmetičke sredine na osnovu koje formiramo intervale poverenja koji nam govore koje sve vrednosti arimtetičke sredine možemo očekivati u okviru definisanog intervala poverenja. Budući da je isti slučaj na svim varijabalam, otvoreno možemo sumnjati da izmerene razlike između aritmetičkih sredina nisu stvarne, već su samo proizvod varijacija u okviru standardne greške merenja. Da bi smo odgovorili na ovo pitanje, nužno je da obavimo neke operacije, i na osnovu svih dobijenih parametara **testiramo hipoteze**

Statistika zaključivanja

Testiranje hipoteza spada u jednu vrlo važnu oblast koja se zove **statistika zaključivanja.** Osnova statistike zaključivanja leži u činjenici da mi sprovodimo istraživanje na uzorku, a onda koristimo **indukciju** kako bi generalizovali naše zaključke na čitavu populaciju. Statistika zaključivanja je metod posredstvom koga mi oravdavamo generalizaciju, tačnije zaključke do kojih smo došli istraživanjem. Način na koji se u istraživanjima ispituje vrednost generalizacije jeste formulisanje i **testiranje nulte hipoteze (Ho).**

Nulta hupoteza ima krucijalnu funkciju u postupku testiranja hipoteza.

Ona je nulta zato što je definisana negativno. Drugim rečima, nulta hipoteza u svakom konkretnom slučaju tvrdi da između dve grupe ispitanika ne postoje razlike, ili da između dve varijanse ne postoje razlike, ili da između dve aritmetičke sredine ne postoje razlike itd.

Nasuprot nultoj hipotezi (Ho), postupak testiranja zahteva i formulisanje **alternativne hipoteze** (H1)... (mada se u praksi akumulacijom iskustva alternativna hipoteza ne definiše eksplicitno, već se jednostavno podrazumeva). Smisao definisanje alternativne hipoteze jeste njeno formulisanje na način da ona bude suprotna i iscrpna u odnosu na nultu hipotezu. To znači, teorijski, da nulta i alternativna hipoteza moraju biti definisane tako da pokriju svaki mogući ishod testiranja i uz to, da ukoliko je jedna tačna druga nužno ne može biti tačna i obrnuto.

Recimo da želimo da testiramo da je naša populacija u proseku stara 37 godina. U tom slučaju:

(Ho): $\boldsymbol{\mu = 37}$
(H1): $\boldsymbol{\mu \neq 37}$

Dakle, obe hipoteze su definisane na način da su međusobno isključive i da svaki mogući ishod istraživanja pokriva moguću intepretaciju testiranja

U praksi testiranja nulte i alternativne hipoteze, mogući su sledeći slučajevi:

1) (Ho): $\boldsymbol{\mu = a}$
(H1): $\boldsymbol{\mu \neq a}$

2) (Ho): $\boldsymbol{\mu >}$ ili $\boldsymbol{= a}$
(H1): $\boldsymbol{\mu < a}$

3) (Ho): $\boldsymbol{\mu <}$ ili $\boldsymbol{= a}$
(H1): $\boldsymbol{\mu > a}$

Osnovni zadatak testiranja hipoteza u društvenim istraživanjima jeste da se naša potreba da merimo razlike između parametara prevede na jezik nulte i alternativne hipoteze. Ovaj zadatak podrazumeva da ono što jeste istraživačko pitanje prevedemo na jezik **verovatnoće.** Drugim rečima, testiranje nulte hipoteze jeste postupak u kome mi zapravo merimo **veravatnoću nekog ishoda** (iskaza sadržanog u nultoj hipotezi). Na osnovu testiranja mi smo u stanju da kažemo da li je ono što se tvrdi u nultoj hipotezi veoma ili malo verovatno da se ostvari. Ako je verovatnoća ishoda iskaza sadržanog u nultoj hipotezi 'veoma' mala, mi odbacijemo nultu hipotezu, u suprotnom mi ne odbacujemo nultu hipotezu (NAPOMENA: nulta hipoteza samo može biti odbačena ili je ne odbacijemo, izbegava se reći da je potvrđena, iako i ova interpetacija nije netačna)

Pod 'veoma malom' verovatnoćom podrazumevamo statističke standarde koji su poznati kao **statistička značajnost testa**, a standardi su od $p < 0.05$ (95% račun

verovatnoće) i p< 0.01 (99% račun verovatnoće), iako se u svakom pojedinom slučaju može tačno izračunati verovatnoća nekog događaja. To znači da kada na nivou p< 0.01 statističke značajnosti odbacivanja nulte hipoteze da recimo 'ne postoje razlike u stavu prema institucijama između muškaraca i žena', verovatnoća da te razlike ipak ne postoje jesu 1:100. Tačnije, u jednom od 100 uzoraka možemo očekivati suprotan nalaz od našeg, tj., nalaz da zaista ne postoje razlike koje smo izmerili. Isto je i sa i p< 0.05, samo što u ovom slučaju je verovatnoća 5:100

Jedan od načina da testiramo hipoteze jeste korišćenje **z statistika**, a ovo nam je najbliže jer je postupak veoma sličan onom koji važi za transformaciju varijable u z skorove. Prema tome, formula je:

$$Z = \bar{x} - \mu / \sigma_{\bar{x}} \text{ , pri čemu je: } \sigma_{\bar{x}} = \sigma/\sqrt{n}$$

Konsekventno, na uzorku:

$$Z = \bar{x} - \mu / s_{\bar{x}} \text{ , pri čemu je: } s_{\bar{x}} = s/\sqrt{n}$$

Na primer, sproveli smo istraživanje koje meri autoritarnost. Rezultati pokazuju da je prosečna vrednost merenja na datoj skali 20. Nakon nekoliko meseci merili smo autoritarnost identičnim upitnikom na uzorku srednjoškolskih profesora, ne bi li videli da li postoje ili ne postoje razlike u autoritarnosti između čitave populacije i kategorije srednjoškolskih profesora. Na uzorku od 100 profesora srednjih škola, izmerili smo da je prosečna autoritarnost na istoj skali 19,1, sa standardnom devijacijom od 3 indeksna poena. Statistika zaključivanja ne dozvoljava da jednostavno na osnovu ove razlike u aritmetičkim sredinama kažemo da su profesori srednje škole manje autoritarni u odnosu na populaciju. Da bi ovo dokazali koristićemo z statistik. Najpre postavljamo nultu hipotezu koja kaže da **'ne postoji značajna razlika' između dve aritmetičke sredine,** ili prevedeno na jezik istraživanja, nema razlike između populacije i srednjoškolskih profesora u pogledu autoritarnosti.

Ova hipoteza se formalizuje ovako:

(Ho): $\mu = 20$

Naspram ove hipoteze formulisaćemo alternativnu koja izgleda ovako

(H1): $\mu \neq 20$

Na osnovu formula za izračunavanje z statistika, najpre moramo izračunati standardnu grešku aritmetičke sredine, a za ovo, kao što znamo numerator je

standardna devijacija a denominator kvadartni koren ukupnog broja ispitanika (profesora). Dakle:

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

tj.

$$\sigma_{\bar{x}} = 3/\sqrt{100} = 0{,}3$$

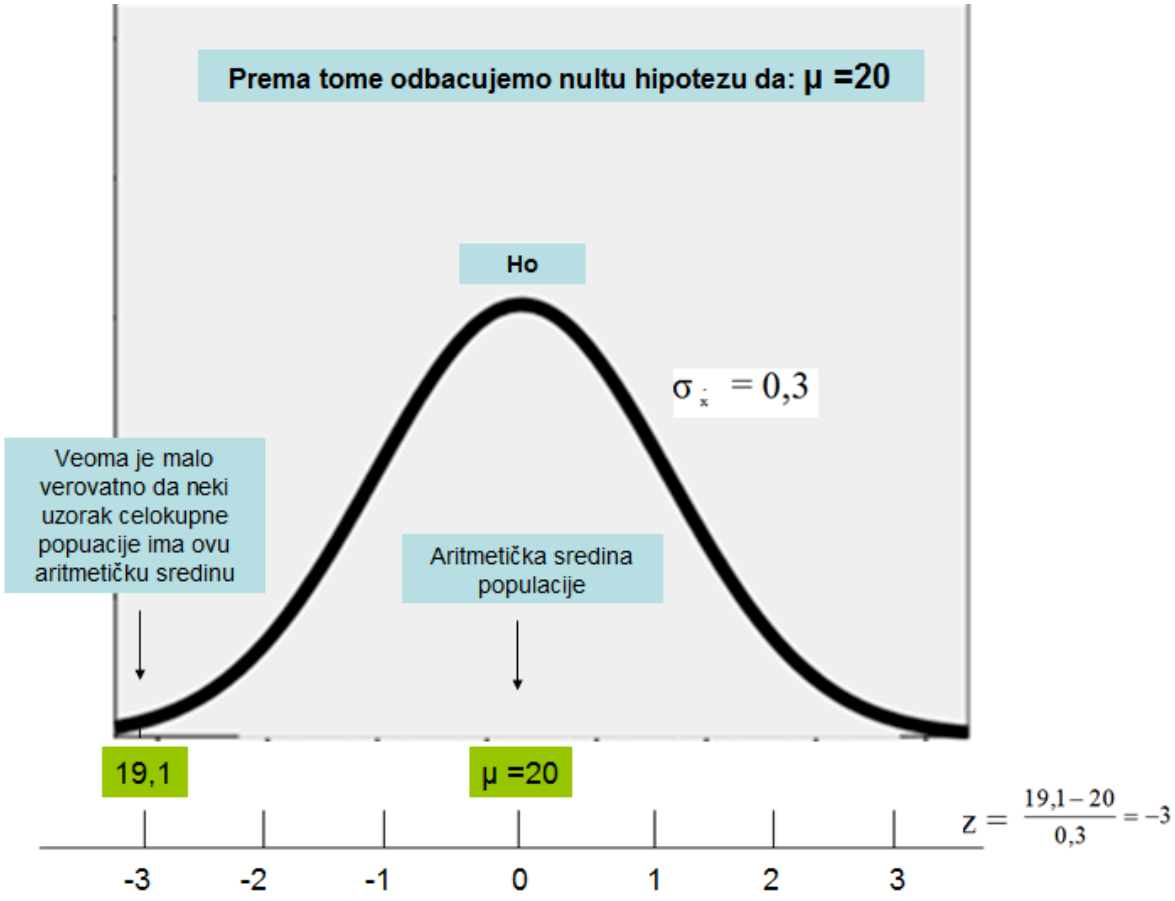
Prema tome, z statistik će biti:

$$Z = \bar{x} - \mu / s_{\bar{x}}$$

tj.

$$\frac{19{,}1 - 20}{0{,}3} = -3$$

Dakle, dobili smo vrednost z statistika: z = - 3. Prema tome, aritmetička sredina uzorka profesora na skali autoritarnosti od 19,1 odstupa od pretpostavljene vrednosti 20 za tri standardne greške merenja! Ključno pitanje sa stanovišta odbacivanja nulte hipoteze je: **Da li je verovatno da smo mi uzorkovanjem 'slučajno' dobili uzorak profesora u kome je aritmetička sredina 19,1 pri čemu znamo da je ova sredina u populaciji 20?** Odgovor je negativan, tačnije, veoma, veoma, veoma malo je verovatno da je naš rezultat proizvod 'loše sreće' uzorkovanja i s toga mi odbacujemo nultu hipotezu da **ne postoje statistički značajne razlike u pogledu autoritarnosti između populacije i profesora srednje škole**. Evo kako to izgleda grafički:



Dijalog

Ministar: pripremili smo nov zakon za obračun korisnika MOP-a. Suština novog zakona je u tome da će neki korisnici dobiti malo manje, neki malo više, ali za državnu kasu biće potpuno svejedno, tačnije država neće izdvajati ni više ni manje nego ranije, jednostavno će sredstva biti pravednije raspodeljena.

Statističar: Hm, jesi li siguran da će u krajnjem ishodu ista suma biti izdvajana iz državne kase na godišnjem nivou.

Ministar: Pa valjda jesam, mi smo tako računali.

Statističar: Dobro, imam predlog, uzećemo uzorak od 100 korisnika, socijalne pomoći, na njihove slučajeve ćemo simulirati primenu novih pravila pa ćemo videti da li će izdvajanja na ovom uzorku biti ista ili drugačija sa stanovišta ukupne sume.

Ministar: Šta ćeš tačno da upoređuješ?

Statističar: Pa, na 100 uzorkovanih ispitanika uporediću prosečnu sumu izdvojenog novca koja se izdvaja po starim pravilima i prosečnu sumu koja bi se izdvajala po novim pravilima

Ministar: Zvuči kao dobra ideja, uradi to što pre!

Statističar: Evo, već sam izvukao slučajan uzorak od 100 korisnika i kada uporedim koliko je prosečno potrebno za njih na osnovu starih pravila obračuna i ovih novih koje mislite da uvedete, imam podatak da niste dobro računali

Ministar: Kako to? Objasni....

Statističar: Pa, jednostavno, Na ovom slučajnom uzorku, primenom novih propisa država je u manjku u proseku -205 evra po korisniku. Obzirom da u Crnoj Gori imamo 10 000 korisnika, znači da državna kasa može da očekuje gubitak oko 200 000 evra na godišnjem nivou.

Ministar: Nemoguće! Nešto nije u redu! Ili uzorak ili taj kompjuter ne računa dobro!
Statističar: Iskreno, malo me brine ovako velika standardna devijacija koja iznosi 725 evra, što je puno obzirom na vrednost aritmetičke sredine, ali još uvek tvrdim da će državna kasa 'najverovatnije' pretrpeti gubitak.

Ministar: Kako to najverovatnije? Šta znači to najverovatnije?
Statističar: To znači da je veoma malo verovatno da kada primenite nova pravila ne pretrpite gubitak koji sam gore opisao.

Ministar: Moraš da me ubediš malo bolje u to....

Statističar: Dobro, evo ovako: Ti tvrdiš da neće biti značajne razlike između starog i novog sistema obračuna sa stanovišta krajnjeg ishoda u smislu ukupne količine para koje država treba da izdvoji?

Ministar: Upravo tako, mi nemamo više para, jednostavno hoćemo pravednije da raspodelimo sredstva

Statističar: Dakle, 'ajde da postavimo tu hipotezu na sledeći način i zovimo to **nultom hipotezom**:

$$\mathbf{H_0: \quad \mu = 0}$$

Statističar: Ja tvrdim da će u kasi biti gubitak, nazovimo ovo **alternativnom hipotezom**:

$$\mathbf{H_1: \quad \mu < 0}$$

Statističar: Sada ćemo testirati nultu hipotezu. Ja sam od 10 000 korisnika izvukao uzorak od 100 korisnika, i izračunao da bi za razliku od prethodnog sistema raspodele primenom novih pravila državna kasa izgubila u proseku 205 evra po korisniku na godišnjem nivou

Ministar: Da, ali ja mislim da je to što si dobio uzorkom od 'samo' 100 ispitanika rezultat 'loše sreće', da si izvukao drugih 100 ispitanika, uvideo bi da nema razlike.

Statističar: U redu, sada ćemo videti da li je moj nalaz rezultat loše sreće. Dakle, hajde prvo da izračunamo standardnu grešku aritmetičke sredine a ona kao što znamo iznosi

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

u našem slučaju:

$$\sigma_{\bar{x}} = 725/\sqrt{100} = 72,5$$

Statističar: Dakle, testiramo **nultu hipotezu:**

$$z = \frac{-205-0}{72,5} = -2,83$$

Ministar: Šta s tim '-2,83'?

Statističar: Pa jednostavno to je na osnovu pravila o z distribuciji – 2,83 standardnih greški udaljeno od aritmetičke sredine, a to je manje od 1% verovatan ishod. Ovo se jasno vidi u tabeli z skorova a još jasnije ako nacrtamo grafikon sa normalnom distribucijom i identifikujemo mesto na kome se nalazi dati z statistik

Ministar: Kakve verovatnoće?

Statističar: verovatnoća da nakon primene novih pravila razlika između starog i novog sistema obračuna iznosi 0! To je bila hipoteza koju smo testirali. Ili drugačije, **verovatnoća da će kao rezultat novih pravila iz državne kase morati da se izdvaja jednako novca kao i ranije je manja od 1: 100. Kada je takva situacija, odbacije se nulta hipoteza, tačnije, odbacujemo na nivou p< 0.01 statističke značajnosti hipotezu da nema razlika između starog i novog sistema**

Ministar: Vidi, ja nisam statističar, ali nije li ta distribucija o kojoj govoriše mera na osnovu koje odbacuješ moju hipotezu?

Statističar: Jeste, zašto?

Ministar: Pa zato što ti kažeš da je rezultat -2,83 u odnosu na normalnu distribuciju, a ova koji si dobio meni baš i ne izgeda 'normalno'

*Statističar:*To jeste tačno, ali mi koristimo histogram verovatnoće za prosek svih uzoraka

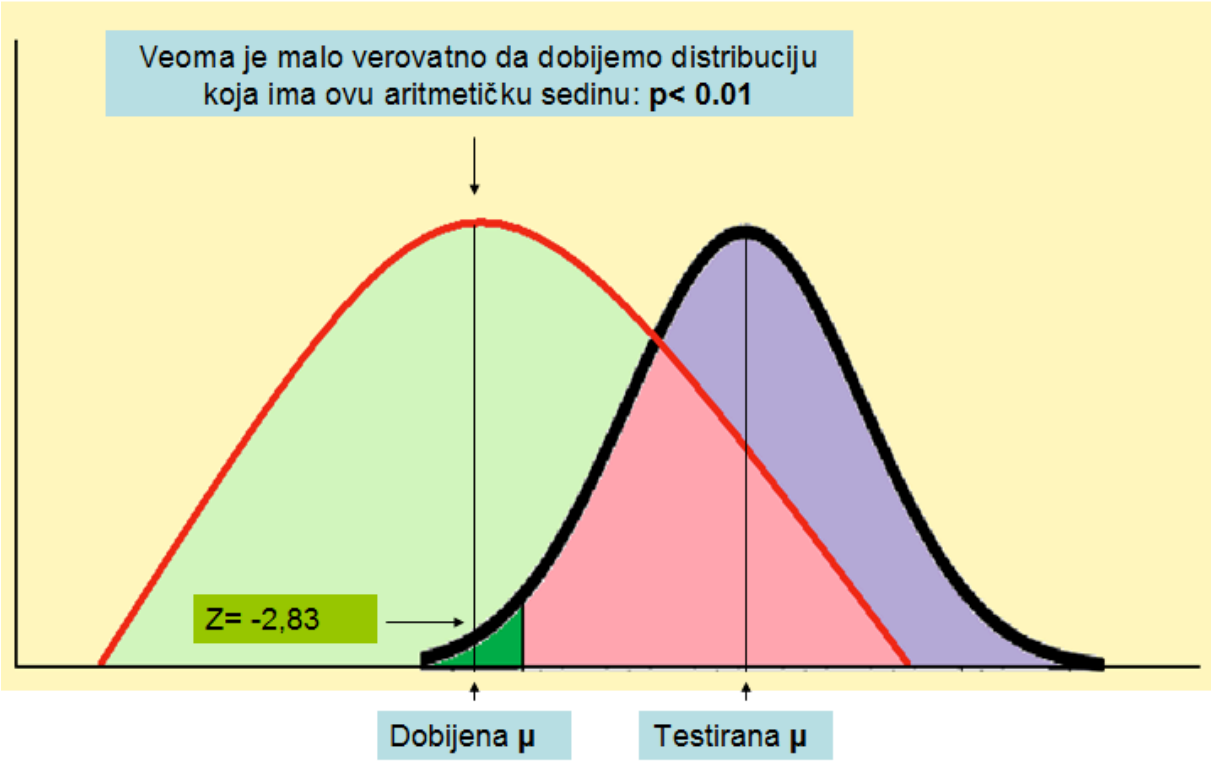
Ministar: Ok, sada razumem, dakle tvoj sud je kombinacija pretpostavki na osnovu zakona verovatnoće i nalaza koji si dobio na osnovu uzorka od 100 korisnika MOP-a

Statističar: Upravo tako! Ti možeš i dalje da insistiraš da je prosek novog i starog sistema na 10 000 korisnika isti, ali u toj situaciji trebaće ti 'čudo' da objasniš zašto se iz kase izdvaja više novca nego ranije, budući da je verovatnoća da se izdvaja isto novca kao i ranije manja od 1:100

Ministar: Možda ipak da se ja držim starog sistema, dakle, šta je tvoj procena da će novi sistem proizvesti

Statističar: Ja smatram da će primena novih pravila dovesti do gubitka od oko 200 eura po korisniku na godišnjem nivou. To možda i nije neka razlika, ali je **stvarna**; ne može se jednostavno odbaciti naš nalaz na uzorku kao rezultat 'loše sreće'

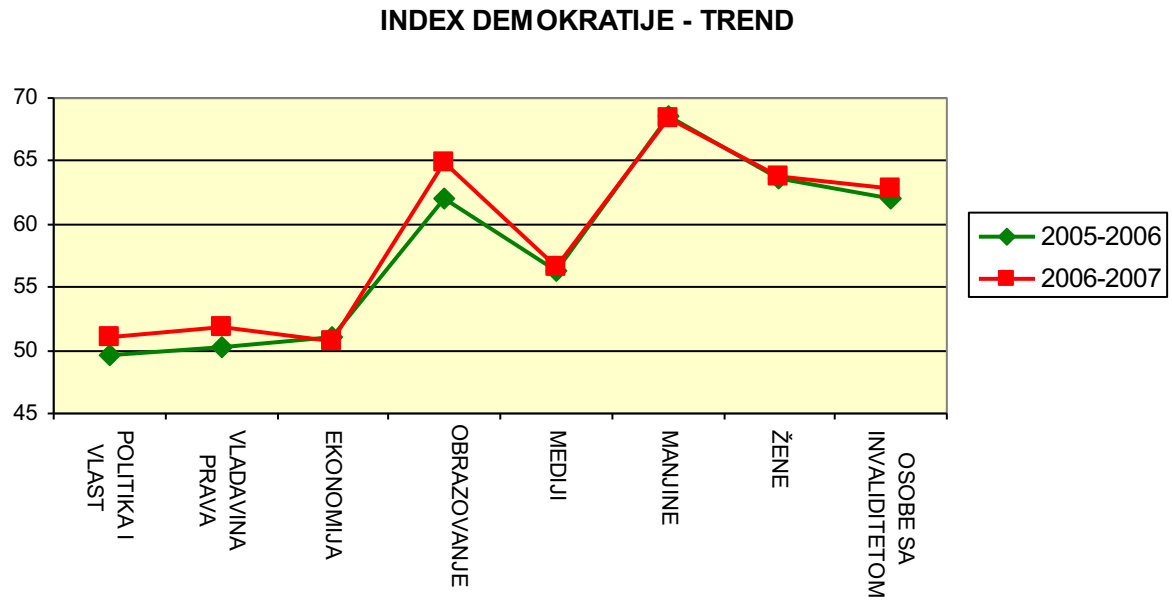
Evo kako ovaj primer izgleda grafički:



Evo još jednog primera iz istraživačke prakse. CEDEM već dve godine za redom sprovodi seriju istraživanja koja za cilj imaju da izmere stanje demokratije. Kao metod koristi se ekstenzivan upitnik sa velikim brojem varijabli (indikatora) a krajnji rezultat jeste formiranje INDEX-a demokratije kao kumulativnog skora. Evo tabelarno i grafički uporednih podataka merenja za 2006 i 2007 godinu:

	2006	2007

POLITIKA I VLAST	49,6	51,0
VLADAVINA PRAVA I ZAKONA	50,2	51,9
EKONOMSKE SLOBODE I EKONOMSKA PARTICIPACIJA	51,0	50,7
DEMOKRATIJA U OBRAZOVANJU	62,1	64,9
DEMOKRATIČNOST MEDIJA	56,3	56,7
POLOŽAJ NACIONALNIH I VJERSKIH MANJINA	68,5	68,4
POLOŽAJ ŽENA	63,7	63,8
POLOŽAJ OSOBA SA INVALIDITETOM	62,0	62,8



Osnovno pitanje u bi bilo, da li su razlike između aritmetičkih sredina koje možemo uočiti značajne? Dakle, vidimo razlike po područjima, ali one nisu naročito izražene. Pitanje je, da li se stanje po područjima zaista poboljšalo? Da li zaista možemo govoriti o tome da je stanje demokratije u Crnoj Gori poboljšano, ili su pak razlike suviše male za ovu tvrdnju? Drugim rečima, ako su razlike značajne, to znači da imamo pozitivan trend, iako je sama razlika mala? Prema tome, ključno je važno da znamo, da li su razlike statistički značajne ili nisu? Za ovo možemo koristiti z statistik. Evo na primeru vladavine prava... Dakle, nulta hipoteza bi glasila da '**ne postoje značajne razlike između izmerenih vrednosti 2006 i 2007 godine'.**

Testiranje hipoteze:

Vladavina prava 2006 = 50,2

Vladavina prava 2007 = 51,9;

standardna greška merenja:

$$\sigma_{\bar{x}} = 0,56$$

Ho: μ = 50,2

$$Z = \bar{x} - \mu / s_{\bar{x}}$$

U našem slučaju:

$$z = \frac{51,9 - 50,2}{0,56} = 3,04$$

Na ovaj način odbacujemo nultu hipotezu na nivou **p< 0.01 statističke značajnosti, jer je veoma malo verovatno da razlike koje smo dobili jesu rezultat greške uzorka.** Drugim rečima, potvrđujemo alternativnu hipotezu i kažemo **da je u Crnoj Gori došlo do napretka u prethodnih godinu dana kada je reč o vladavini prava**.

Evo još jednog primera iz INDEX-a:

Osobe sa invaliditetom 2006 = 62,0
Osobe sa invaliditetom 2007 = 62,8;

Standardna greška merenja:

$$\sigma_{\bar{x}} = 0,59$$

Nulta hipoteza:

Ho: μ = 62,0

Izračunavamo z - statistik

$$Z = \frac{62,8 - 62,0}{0,59} = 1,36$$

Ovaj nalaz nam ukazuje da **ne možimo da odbacimo nultu hipotezu**. Drugim rečima, vrednosti z statistika od 1,36 (pogledaj excel tabelu) je daleko manja nego što je potrebno za bilo koji pristojni nivo statističke značajnosti. Prema tome, naš nalaz kaže da ne možemo da obacimo hipotezu da razlike koje smo dobili mogu biti prosto rezultat greške koja nastaje uzorkovanjem. Drugim rečima, više je nego pristojna verovatnoća da ukoliko bi smo ponovili istraživanje, ne bi pronašli razlike između

aritmetičkih sredina koje smo identifikovali ovim merenjem. Konsekventno, **Odbacujemo alternativnu hipotezu i kažemo da u oblasti zaštita osoba sa invaliditetom u prethodnih godinu dana nije došlo do značajnog napretka**

Šta je prema tome kriterijum da odbacimo nultu hipotezu? Kriterijum je **verovatnoća** da neki uzorak od *n* mogućih uzoraka ima opservirani statistik (u našem slučaju aritmetičku sredinu). Dakle, uz pomoć varijanse, standardne devijacije i standardne greške merenja statistika, upoređuje se opservirani statistik sa testiranim tj. onim koji je definisan nultom hipotezom. Dobijeni rezultat, z statistik ima određenu vrednost koju možemo precizno identifikovati kao meru odstupanja od sredine. Na osnovu uvida u z statistik tabelu i uvida u grafikon, mi zaključujemo **koliko je verovatno** da smo uzorkovanjem ('lošom srećom') dobili distribuciju koja odstupa od očekivane distribucije. **Ako je verovatnoća mala, onda odbacujemo nultu hipotezu.** Kao kriterijum za odbacivanje uzimamo standard p< 0.01 (99% verovatnoće) i p< 0.05 (95% verovatnoće). Ovi kriterijumi su opšte prihvaćeni standardi, a u praksi svaki z statistik ima precizno identifikovanu verovatnoću događanja. Dakle, u krajnjem ishodu rezultat testiranja nulte hipoteze svodi se na određivanje relativnog mesta *z statitika* u odnosu na testirani statistik. U ovom postupku, svaka vrednost *z statistika* ima svoje mesto koje je određeno odstupanjem od testirane vrednosti, i koje se može precizno identifikovati u procentima verovatnoće, kako na osnovu tabele, tako i na osnovu **područja** koje prevazilazi vrednosti z statistika na osnovu pretpostavke o normalnoj distibuciji.

Evo još jednog primera, Recimo npr. da smo izmerili da je aritmetička sredina S.E.I. U Crnoj Gori 2005 iznosila 23,58. Recimo da nas ovoj situaciji interesuje koja je to vrednost z statistika koja je potrebna, kako bi mogli sa izvesnošću da tvrdimo da je došlo do poboljšanja socio-ekonomskog statusa građana u Crnoj Gori u 2007 god. Drugim rečima, ***za koliko je potrebno da aritmetička sredina bude veća od 23.58, pa da kažemo da je došlo do poboljšanja?***

Prema tome:

(Ho): **μ < *23,58*** ili **μ = *23,58***

(H1): **μ > *23,58***

Znamo da je:

$$Z = \bar{x} - \mu / \sigma_{\bar{x}}, \text{ pri čemu je: } \sigma_{\bar{x}} = \sigma/\sqrt{n}$$

Npr., ako z = 1, to znači da da je 1 SD iza **μ =23,58**. Ako je z = 1,5 to znači da je SD iza **μ =23,58.** Koja je vrednost z statistika potrebna kako bi **ušli u region potreban da odbacimo nultu hipotezu na nivou p<** 0.05. Na koji način se definišu područja odbacivanja nulte hipoteze? Evo formalizacije:

(Ho): **μ < *23,58*** ili **μ = *23,58***

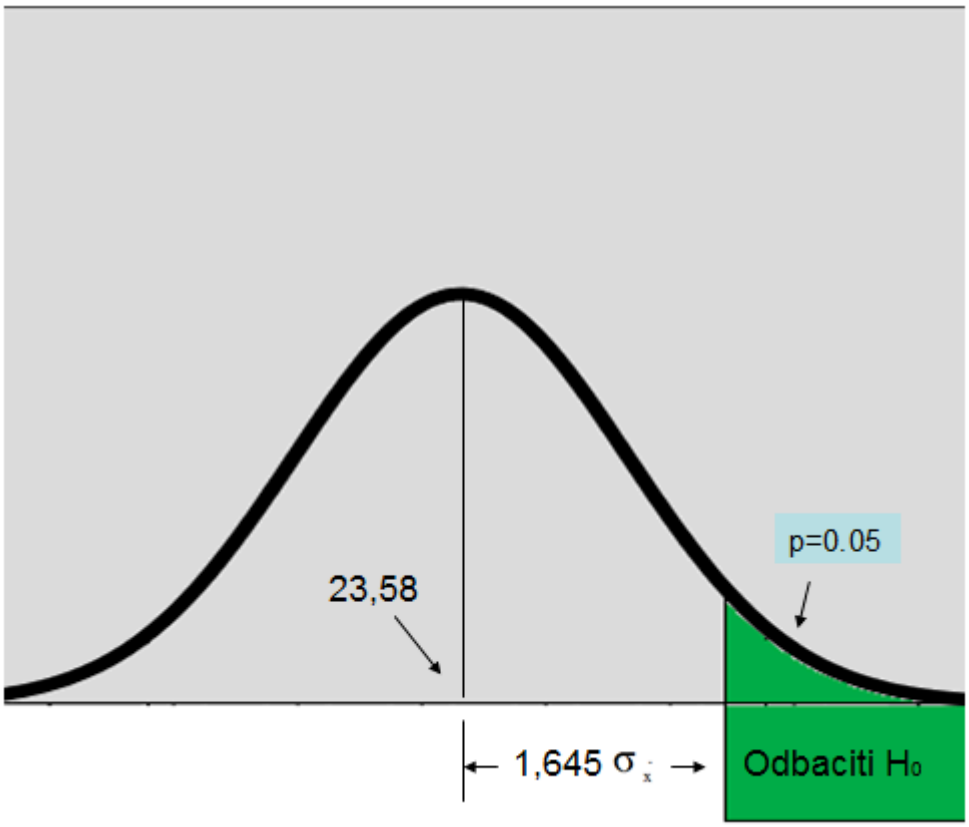
(H1): **μ > *23,58***

$$z = \frac{\bar{x} - 23{,}58}{\sigma_{\bar{x}}}$$

Region odbacivanja: z > 1,645 što odgovara p = 0,05. Dakle, da bi odbacili nultu hipotezu potrebna nam je vrednost: z = 1,645 ili z >1,645? Što je veća vrednost z statistika to imamo više poverenja u našu procenu da odbacimo nultu hipotezu, ali zašto baš granicu predstavlja vrednost z statistika od 1,645? Otkud baš ova vrednost z statistika da je potrebna za odbacivanje nulte hipoteze? Otkud se ova vrednost pojavljuje kao referentna? Zašto ne 1,96, kada smo na osnovu tabele i grafikona utvrdili da je ovo referentna vrednost za 95% interval poverenja z skorova? Ako je vrednost z = 1,645 potrebna za p< 0.05 (95% verovatnoće), koja je vrednost z potrebna za p< 0.01 (99% verovatnoće)? Evo najpre tabelarnog prikaza distribucije z skorova sa referentnim vrednostima:

				Druga decimala za z skorove			
z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949
0.9	.1841	.1814	.1788	.1762	.1726	.1711	.1685
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250

Grafički to izgleda ovako:

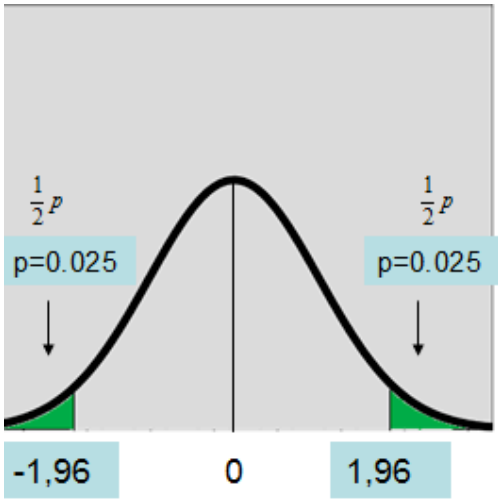


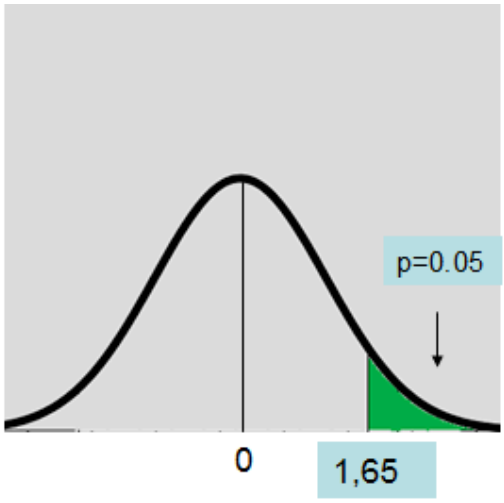
Zašto, prema tome, kriterijum nije z = 1,96 nego je kriterijum z = 1,65?. Zato što nam je za testiranje ove hipoteze potreban **jednostrani** (one-tailed) a ne **dvostrani** (two-tailed) test
Drugim rečima, nulta hipoteza je postavljena na način da ispitujemo samo jedan smer jer je:

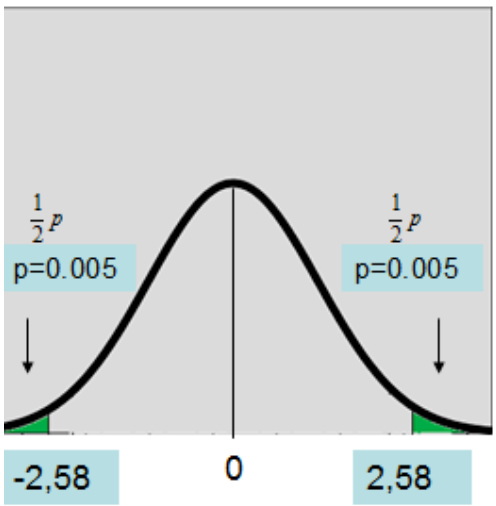
(Ho): $\mu < 23{,}58$ ili $\mu = 23{,}58$

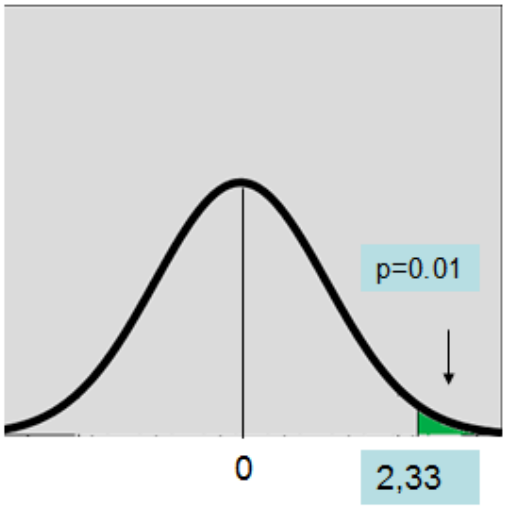
(H1): $\mu > 23{,}58$

Dakle, u ovoj situaciji mi nećemo raspodeliti varijansu na način da ispitujemo regione odbacivanja na oba kraja distribucije, već nas zanima odbacivanje hipoteze samo na jednom kraju distribucije. Dakle, ne koristi se dvostrani **već se koristi jednostrani test**, i konsekventno, na tom kraju mi tražimo vrednost z skora koja na jednoj strani dostiže vrednost 95% varijanse, ili tačnije identifikuje 5 od 100 slučajeva kada je mogući drugačiji ishod od testiranog. Na osnovu tabele z skorova, možemo videti da je u toj situaciji tražena vrednost z skora 1,65 devijacija od testirane vrednosti. Regioni odbacivanja za jednostrane i dvostrane testove statističke značajnosti na 95% i 99% se mogu videti na grafikonima koji slede:









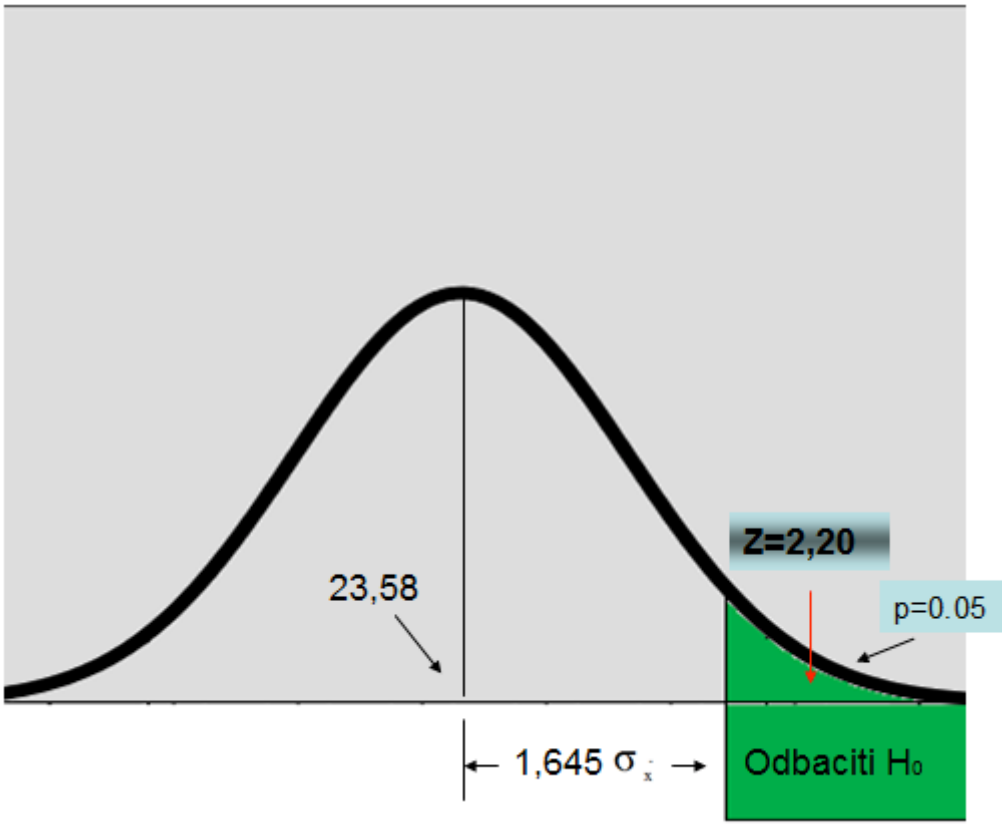
Ako se vratimo na primer S.E.I, dakle, znamo da vrednost z statistika u našem primeru mora biti z= 1,65 ili z > 1,65 kako bi odbacili nultu hipotezu na nivou p< 0.05. Pretpostavimo da smo merenjem S.E.I. u 2007 došli do podatka da je SEI= 24,81 sa standardnom greškom

$$\sigma_{\bar{x}} = 0,56$$

Izračunavamo z statistik:

$$z = \frac{24,81 - 23,58}{0,56} = 2,20$$

U regionu odbacivanja na grafiku se jasno vidi da je dobijena vrednost z=2,20 daleko veća od zahtevane 1.65 i prema tome duboko u regionu odbacivanja nulte hipoteze:



Ako se dobijena vrednost z = 2,20 pronađe u tabeli z skorova, može se videti da je vrednost 0,139. Prevedeno na jezik procenata ovo je 1,39% verovatnoće da je naš nalaz rezultat 'loše sreće' uzorkovanja. Drugim rečima, ako ponovimo merenje100 puta, u manje od dva slučaja (1,39) ćemo pronaći da razlike koje smo pronašli ne postoje. Najpreciznije rečeno, verovatnoća da su pronađene razlike 'slučajnost' iznosi 1,39:100. Ovo je daleko više od zahtevane 5% verovatnoće (5 slučaja u 100). Drugim rečima, veoma je malo verovatno da razlike u SEI 2005 i SEI 2007 ne postoje, i da je dobijeni rezultat proizvod greške uzorka

Zbog ovakvog ishoda, mi odbacujemo nultu hipotezu da **ne postoje razlike između 2005 i 2007 u socioekonomskom statusu merene posredstvom SEI .** Konsekventno , **prihvatamo alternativnu hipotezu** i tvrdimo sa velikom verovatnoćom da su razlike u SEI koje smo pronašli stvarne, i konsekventno tvrdimo da se socioekonomski status u Crnoj Gori poboljšao u 2007 u poređenju sa 2005. godinom

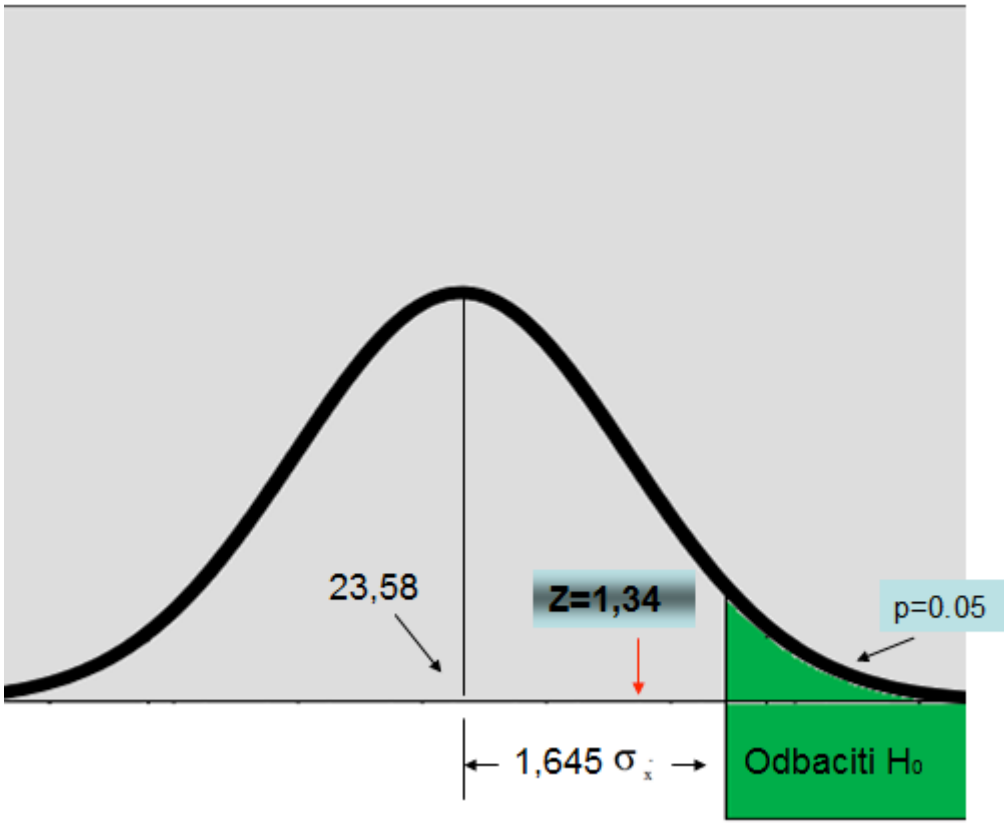
Evo još jednog mogućeg primera, naime, Rekli smo da vrednost z statistika u našem primeru mora biti z= 1,65 ili z > 1,65 kako bi odbacili nultu hipotezu. Pretpostavimo da smo merenjem S.E.I. u 2007 došli do podatka da je SEI= 24,21 sa standardnom greškom:

$\sigma_{\bar{x}} = 0,47$

Izračunavamo z statistik:

$$z = \frac{24,21 - 23,58}{0,47} = 1,34$$

U regionu odbacivanja na grafiku jasno se vidi da je dobijena vrednost z=1,34 manja od zahtevane 1.65 i prema tome nije u regionu odbacivanja nulte hipoteze:



Ako se dobijena vrednost z=1,34 pronađe u tabeli, može se videti da je vrednost 0,09. Prevedeno na jezik procenata ovo je 9% verovatnoće da je naš nalaz rezultat 'loše sreće' uzorkovanja. Drugim rečima, ako ponovimo merenje100 puta, u 9 slučajeva ćemo pronaći da razlike koje smo utvrdili ne postoje. Najpreciznije rečeno, verovatnoća da su pronađene razlike 'slučajnost' iznosi 9:100, a ovo nije malo i ispod je standarda od p< 0.05. Ovo je daleko manje od zahtevane 5% verovatnoće (5 slučaja u 100). Drugim rečima, nije tako 'malo' verovatno da razlike u SEI 2005 i SEI 2007 ne postoje, i razumna je sumnja da je rezultat merenja razlika proizvod greške uzorka. Zbog ovakvog ishoda, mi **ne možemo da odbacujemo nultu hipotezu** da **ne postoje razlike između 2005 i 2007 u socioekonomskom statusu merene posredstvom SEI.** Konsekventno, **odbacujemo alternativnu hipotezu** i tvrdimo da je pristojna verovatnoća da su razlike u SEI koje smo pronašli 'slučajne', i konsekventno ne možemo da tvrdimo da se socioekonomski status u Crnoj Gori poboljšao u 2007 u poređenju sa 2005. godinom

Kako bi smo testirali hipoteze primenom z statistika, potrebno je da preduzmemo **sledećih 10 koraka**:

1. Postaviti problem razlika između statistika na teorijsku ravan

2. Postaviti nultu i alternativnu hipotezu
3. Identifikovati da li je reč o jednostranom ili dvostranom testu
4. Doneti odluku da li se testira hipoteza na $p < 0.05$ ili $p < 0.01$ nivou
5. Izračunati z statistik
6. Pronaći dobijenu vrednost z statistika u tabeli koja definiše područja distribucije
7. Pronaći vrednost iz tabele na grafikonu koji ima osenčeno područje odbacivanja nulte hipoteze
8. Identifikovati da li je dobijena vrednost u području odbacivanja
9. Odbaciti nultu hipotezu ili alternativnu hipotezu
10. Prevesti rezultat testiranja hipoteza na jezik teorije i interpretirati teorijski krajnji nalaz

Testiranje hipoteza korišćenjem t – testa i F - testa

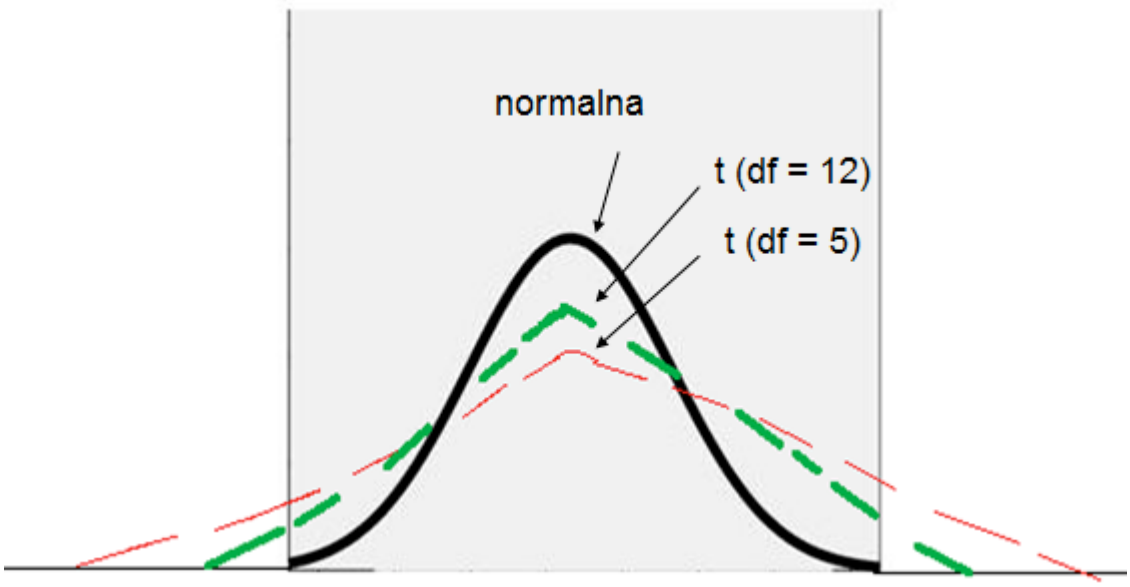
Najpre, treba napomenuti da između T – testa i z statistika postoji veliki broj sličnosti. Te sličnosti su:

- I z statistik i t test postavljaju i testiraju hipoteze
- I z statistik i t test počivaju na proceni relativnog mesta testirane vrednosti u distribuciji svih vrednosti na varijablama
- I z statistik i t test koriste 'zakon verovatnoće' koji se prevodi na jezik statističke značajnosti (p ili α)
- I z statistik i t test upoređuju aritmetičke sredine ne bi li utvrdili da li su razlike koje smo merenjem na uzorku dobili statistički značajne ili nisu.

Sa druge strane između t- testa i z – statistika postoje i određene razlike:

- Z statistik počiva na pretpostavci da je distribucija normalna a t test počiva na **studentovoj** distribuciji (small sample technique)
- Kada koristimo z statistik, mi pretpostavljamo distribuciju aritmetičke sredine sa kojojm se testirana aritmetička sredina upoređuja, a kada koristimo t test, za obe aritmetičke sredine mi znamo tačno i varijansu i standardnu devijaciju i standardnu grešku aritmetičke sredine
- Za razliku od z statistika, u kalkulaciji vrednosti t testa u zavisnosti od broja slučajeva kao osnov za meru statističke značajnosti koristimo broj stepena slobode (df)
- Prema tome, za z statistik i t test koristimo drugačiju tablicu na osnovu koje određujemo statističku značajnost.

Jedna od ključnih pretpostavki na kojoj počiva t – test jeste da distribucija nije identična kao normalna distribucija. Tačnije, studentova distribucija je 'slična' normalnoj distribuciji ali nije identična. Razlog zbog koga je studentova distribucija drugačija jesto što se t test uglavnom oslanja na poređenju aritmetičkih sredina relativno malog broja slučajeva a praksa pokazuje da u toj situaciji imamo drugačiju distribuciju. Karakteristika t distribucije je u tome što njen obliko odgovara normalnoj distribuciji ali sa važnom razlikom, naime, ona je spljoštenija (mezokurtična) i ima duže 'repove' (krajeve). Evo kako to izgleda grafički:



Na grafikonu se može videti poređenje između dve verzije t distribucije. Iz prikaza se može videti da što je veći broj stepeni slobode (*degrees of freedom* - df), to se i t distribucija približava 'normalnoj' distribuciji. Broj stepena slobode (df) je prema tome je prema tome važna i konstitutivna karakteristika same disribucije. Prema tome, distribucija zavisi od broja stepena slobode i u svakom pojedinom slučaju mi moramo statističku značajnost da računamo u odnosu na distribuciju koja je rezultat određenog broja stepena slobode. Broj stepena slobode direktno zavisi od broja opserviranih vrednosti od kojih zavisi standardna greška merenja. Kada je reč o standardnoj greški aritmetičke sredine onda:

df = n-1

Dakle, broj stepeni slobode kada je testirani statistik aritmetička sredina je broj opservacija *minus 1 (napomena: za druge statistike ovaj princip ne važi).* DF je prema tome deskriptivni alat, i on usnovi prikazuje koliko iznosi broj opservacija u setu podataka koji su **slobodni** da variraju kada kalkulišemo željeni statistik. Drugim relima, kada merimo standardnu devijaciju, mi oduzimamo aritmetičku sredinu od svake vrednosti *n*. U ovom postupku, kada oduzmemo pretposlednju vrednost, automatski znamo vrednost finalne devijacije budući da suma svih devijacija mora biti jednaka 0. Prema tome, poslednja devijacija nema slobodu varijacije, samo n-1 može da varira.
T – test se izraunava na sledeći način:

$$t = \frac{x - x}{s_x}$$

Dakle, denominator u formuli izračunavanja t statistika je i sam statistik, što znači da je njegova vrednost podložna fluktuacijama koje su rezultat uzorkovanja. Obzorom da t distribucija počiva na pretpostavci manjeg broja opservacija, sasvim je razumno očekivati spljošteniju distribuciju sa dužim 'krajevima'. Dok je u slučaju normalne distribucije 95% površine unutar +/- 1,96 standardne devijacije, a 99% unutar +/- 2,58 standardne devijacije aritmetičke sredine, ovo nije slučaj kada je reč o t distribuciji. Budući da je t distribucija 'spljoštenija' sa dužim 'krajevima' više od 5% područja biće iza +/- 1,96 standardne devijacije i više od 1% će biti iza +/- 2,58 standardne devijacije. Koliko više, zavisi od konkretne distribucije broja stepeni slobode (df). Što je manji broj stepana slobode, distribucija će biti spljoštenija i 'krajevi' će biti duži. Proističe, da što je manji df mi ćemo morati da idemo dalje od +/- 1,96 standardne devijacije aritmetičke sredine kako bi obuhvatili 95% distribucije i jednako moramo ići dalje od +/- 2,58 standardne devijacije aritmetičke sredine kako bi obuhvatili 99% distribucije.

Isto kao i u slučaju z statistika, i t test koristi tabelu u kojoj za određenu vrednost t testa za dati broj stepena slobode mi možemo odrediti statističku značajnost. Šta se zapravo meri? Isto kao i u slučaju z statistika, mi merimo verovatnoću da je neka distribucija rezultat 'greške' uzorkovanja, dakle, logika je u

oba slučaja identična, samo su kriterijumi u odnosu na različitu distribuciju drugačiji. Konkretno, na osnovu tabele se može videti da je za pokrivanje 95% područja distribucije za df =11 potrebna vrednost t = +/-2,20; dok je za 99% potrebno t= +/-3,11. Međutim, ako je df = 30, onda je za 95% potrebno t=2,04 a za 99% je potrebno t=2,75, što je vrlo blizu z statistik-u (1,96 za 05% i 2,58 za 99%)

Npr., recimo da smo utvrdili da je prosek na skali religioznosti među učenicima četvrtog razreda srednje škole 20 indexnih poena. Pretostavimo da nas interesuje da li je religioznost veća ili manja kod jednog određenog odeljenja u odnosu na čitavu školu. Budući da smo koristili uzorak iz datog odeljenja koje je predmet našeg naše analize, mi imamo samo deset opservacija iz ovog odeljenja. Dakle, t test je jedino rešenje obzirom da se radi o malom broju opservacija. Na uzorku ovog odeljenja od 10 studenata aritmetička sedina je 21,2 a standardna devijacija s = 3,4. Prema tome:

$$t = \frac{\bar{x} - x}{s_{\bar{x}}}$$

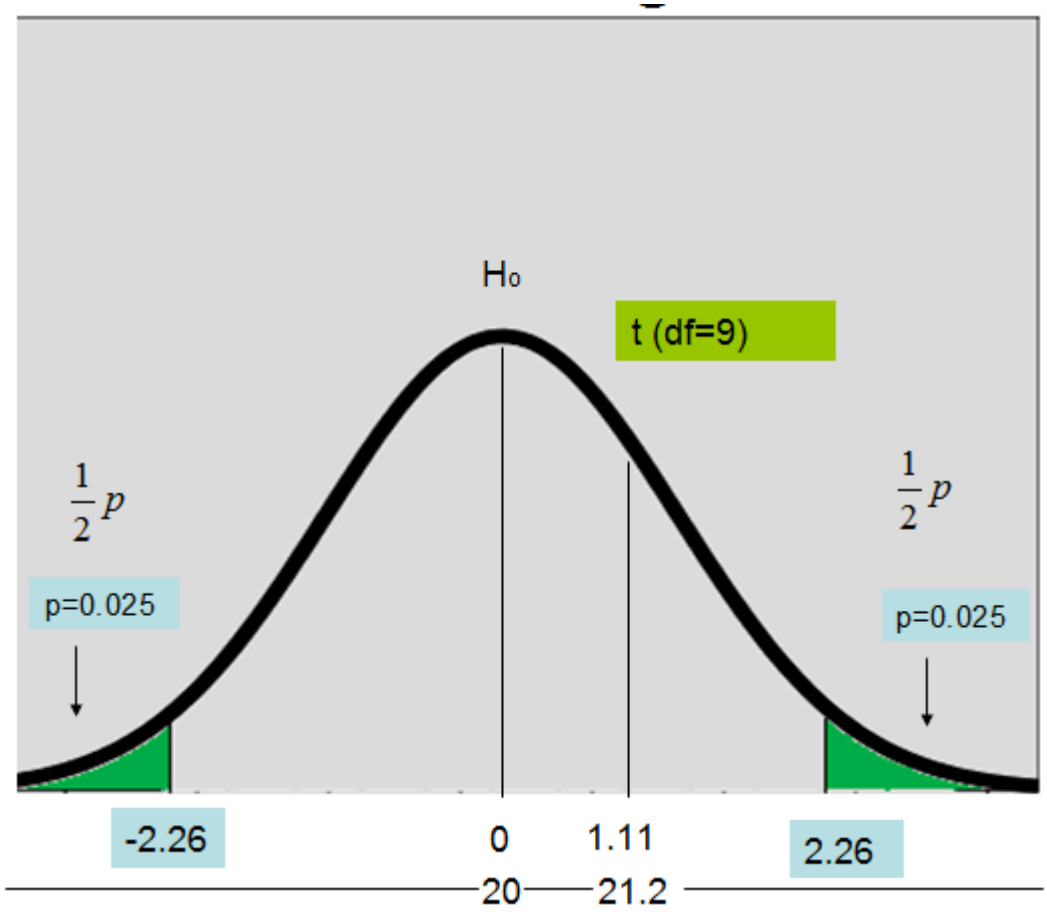
Pri čemu je:

$$s_{\bar{x}} = 3.4/\sqrt{10} = 1.08$$

U načem slučaju:

$$t = \frac{21.2 - 20.0}{1.08} = 1.11 \ldots\ldots Df = 9$$

Za df=9, ako pogledamo tabelu, potrebno je da t bude jednako ili veće od 2,26 kako bi postigli p<0,05. Drugim rečima, t=1.11 je razlika između aritmetičkih sredina koja je pre rezultat 'greške' merenja na osnovu uzorka nego što je je rezultat razlika koje posotoje između jednog i ostalih odeljenja, i prema tome mi odbacujemo nultu hipotezu. Grafički to izgleda ovako:



U situaciji kada na osnovu relativno malog broja slučajeva (recimo manje od 30 – standardni kritetijum za mali uzorak) želimo da uporedimo aritmetičke sredine kako bi testirali hipoteze, koristimo matematičku formulu koja uzima u obzir činjenicu da nam je poznata varijansa za obe distribucije koje su predmet našeg posmatranja. Formula je naizgled složena ali je u biti jednostavna:

$$t = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Primetiti da $n_1 + n_2 - 2$ *predstavlja broj stepena slobode (df)*

Npr., recimo da imamo dve grupe učenika pri čemu su istu materiju ovi studenti savladavali korišćenjem različitih metoda nastave i mi smo im dali isti test na kraju godine ne bi li proverili da li postoji razlika između metoda 1 i metoda 2 nastave. Uzeli smo pet učenika kao reprezentativne za metod 1 i pet učenika za metod 2. Grupa 1 je imala 27 poena na testu a grupa 2 je imala 31 poen. Standardne devijacije:

$s_1 = 9$ dok $s_2 = 12$

Prema tome:

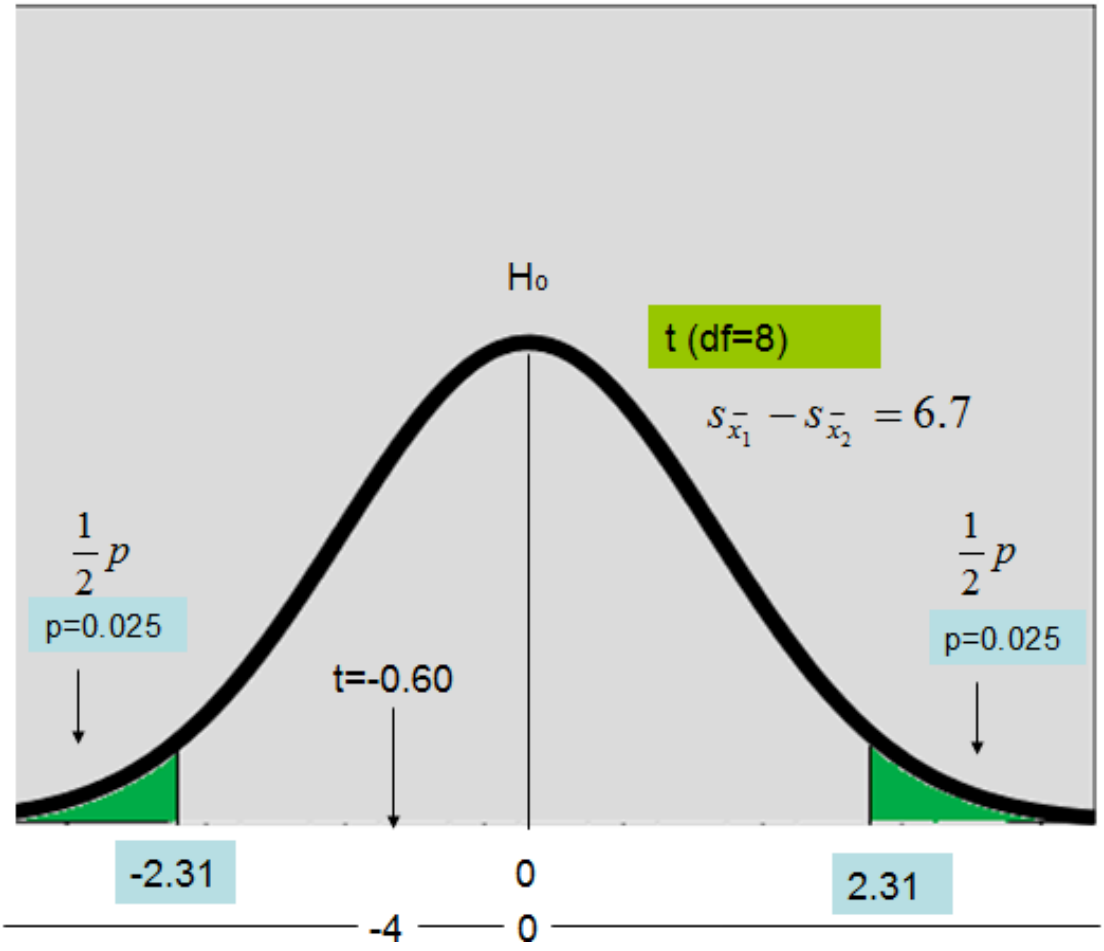
$$t = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Sledi:

$$t = \frac{(27 - 31) - (0)}{\sqrt{\frac{4(9)^2 + 4(12)^2}{5 + 5 - 2}\left(\frac{1}{5} + \frac{1}{5}\right)}} = \frac{-4}{6.7} = -0.60$$

df=5+5-2=8

Grafički:



Evo još jednog primera sa kolokvijuma gde koristimo format tabela u SPSS-u

Group Statistics

	pol	N	Mean	Std. Deviation	Std. Error Mean
suma_poena	F	82	8,6512	3,35024	,36997
	M	25	8,7960	3,78924	,75785

Independent Samples Test

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
suma_poena	Equal variances assumed	-,183	105	,855	-,14478	,78945	-1,71012	1,42056

Ukoliko je reč o uparenom testu, SPSS tabele izgledaju na sledeći način:

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Nacionalno mješoviti brakovi moraju biti nestabilniji nego drugi brakovi. - ?ovjek se moze osje?ati sasvim sigurnim samo kada ħivi u sredini gdje je ve?ina pripadnika njegove nacije.	-,324	1,113	,029	-,382	-,267	-11,056	1440	,000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	99% Confidence Interval of the Difference Lower	Upper			
Pair 1	Nacionalno mješoviti brakovi moraju biti nestabilniji nego drugi brakovi. - ?ovjek se moze osje?ati sasvim sigurnim samo kada ħivi u sredini gdje je ve?ina pripadnika njegove nacije.	-,324	1,113	,029	-,400	-,249	-11,056	1440	,000

Analiza varijanse, poznata kao ANOVA je široko rasprostanjena statistička tehnika koja ima za cilj da uporedi varijansu između dve ili više varijabli. U obradi podataka, ona predstavlja neretko jedan od prvih koraka, budući da svrsishodnost upotrebe ostalih statističkih tehnika neretko zavisi od toga da li je varijansa varijabli koje upoređujemo jednaka ili nije. Za analizu varijanse koristi se F test. F test se jednostavno izračunava kao odnos između dve varijanse:

$$F = \frac{s_1^2}{s_2^2}$$

Pretpostavka, na kojoj počiva F-test (dakle nulta hipoteza) jeste da ne postoje razlike između varijansi seta podataka koje upoređujemo. Dakle, mi pretpostavljamo da:

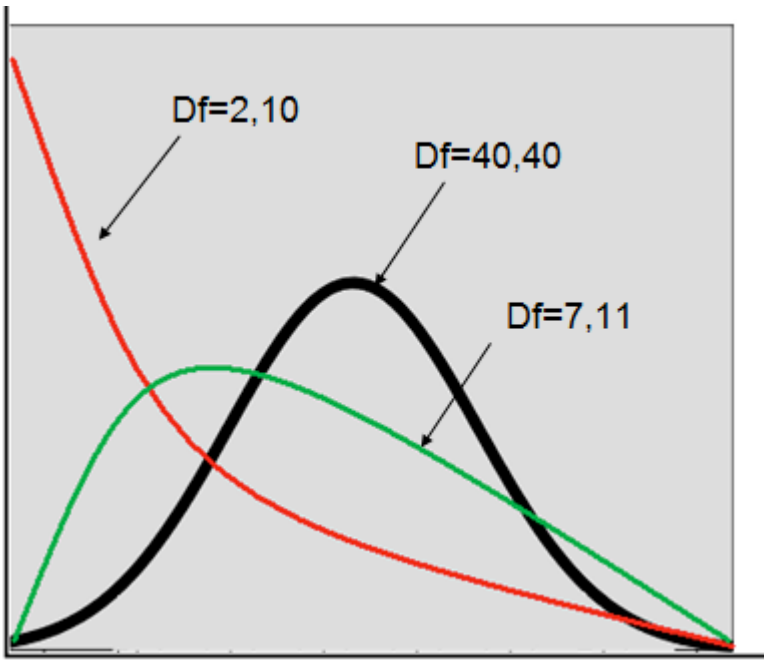
$$E(F) = E(\frac{s_1^2}{s_2^2}) = \frac{E(s_1^2)}{E(s_2^2)} = \frac{\sigma^2}{\sigma^2} = 1$$

Prema tome, budući da dve varijable imaju varijansu s_1^2 i s_2^2, a obe distribucije su proizvod varijacija na osnovu uzorka mi očekujemo da:

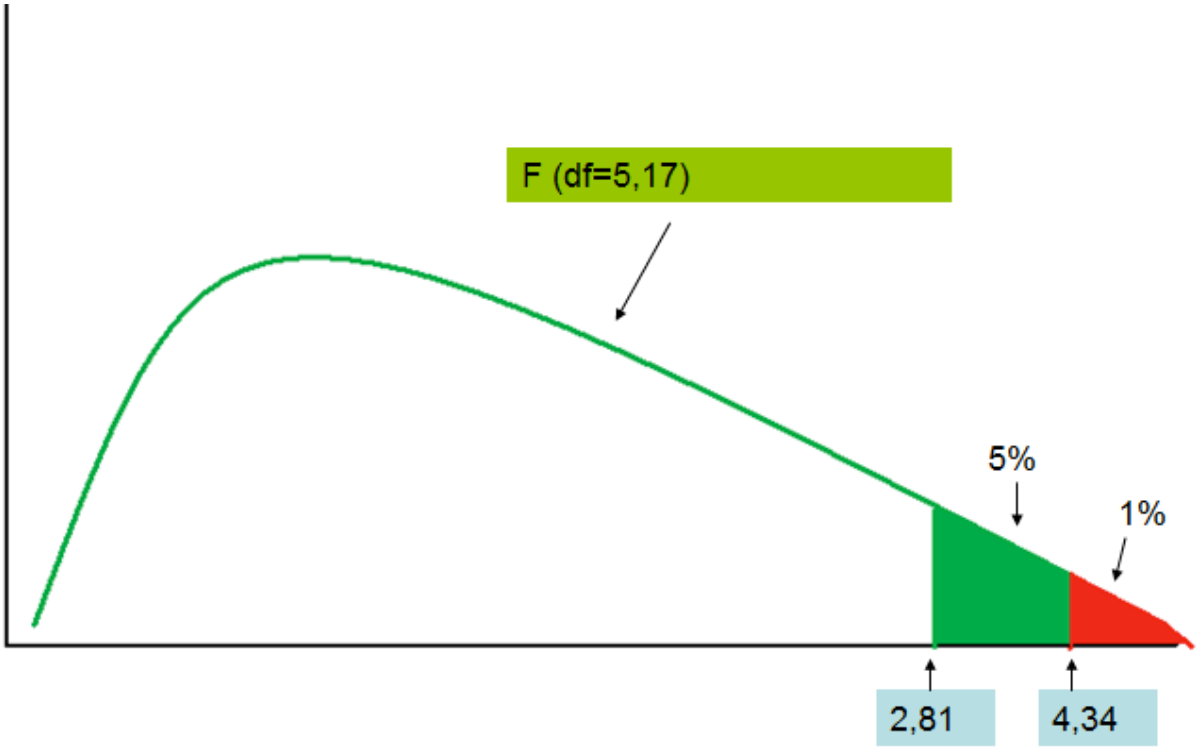
$$F = \frac{s_1^2}{s_2^2} = 1$$

Pošto je vrednost F funkcija dve uzorkovane varijanse, to znači da će vrednost F da zavisi od uzorkovane distribucije na dve varijable. Iako ne znamo tačnu distribuciju dve ili više varijanse (varijable), znamo da će distribucija da zavisi od broja opserviranih vrednosti. To znači, da je oblik distribucije funkcija stepena slobode (df) pri čemu: *df = n-1*. No, budući da mi upoređujemo dve ili više varijanse, to znači za svaku od njih mi posebno obračunavamo *df*. Npr, ako upoređujemo varijansu dve varijable pri čemu je na prvoj n=10 a na drugoj n=15, onda je df=9,14.

Kada je o distribuciji reč, dakle, F distribucija zavisi od df, tačnije od broja slučajeva koji izgrađuju neku distribuciju, to znači da za različiti df mi imamo različitu distribuciju. Drugim rečima, poređenje dve varijanse u svakom pojedinom slučaju sa različitim df ima različitu distribuciju. No, kao i u slučaju normalne i studentove distribucije, i F distribucija ima karakter verovatnoće koji je u statistici dobro poznat. Grafičjki, F distribucija izgleda ovako:



Isto kao i u slučaju t testa i za F test postoje kriterijumi za 95% i 99% pokrivenosti distribucije koji zavise od broja stepena slobode. Dakle, za svaku vrednost F testa, u odnosu na df može se izračunati statistička značajnost testa. Npr. da bi razlike bile statistički značajne između dve varijanse za df=5,17 minimalna porebna vrednost F= 2,81 za p<0,05 i F=4,34 za p<0,01. Evo grafika:



Za izračunavanje F testa koristi se odnos između unutar grupne varijanse (w) i međugrupne varijansa (bg). Ovo je nužno, naime, kada upoređujemo varijanse trebamo imati u vidu da je moguće imati više od dve varijable, pri čemu svaka varijabla ima svoju varijansu. Evo primera:

	Grupa 1	Grupa 2	Grupa 3
	6	18	7
	14	11	11
	19	20	18
	17	23	10
Aritmeticka sredina	14.0	18.0	11.5
Varijansa	32.7	26.0	21.7

$$s_w^2 = \frac{32.7 + 26.0 + 21.7}{3} = 26.7$$

$$s_{bg}^2 = \frac{(14.0-14.5)^2 + (18.0-14.5)^2 + (11.5-14.5)^2}{2} = 10.75$$

Sledi:

$$F = \frac{s_{bg}^2}{s_w^2}$$

Prema tome, u našem slučaju:

$$F = \frac{43.0}{26.8} = 1.60$$

Napomena: Df = 2,9

Na osnovu tabele znamo da je za df=2,9, minimalna kritična vrednost F=4.26 za p<0.05. Šrema tome vrednost koju smo dobili, nije dovoljna da se uđe u zonu odbacivanja nulte hipoteze.

Evo jednog primera u SPSS formatu:

Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Opstanak vlastitog naroda glavni je zadatak svakog pojedinca.	Selo	586	4,16	,724	,030	4,10	4,21	1	5
	Prigradsko naselje	357	4,18	,652	,035	4,11	4,24	1	5
	Grad	528	4,08	,835	,036	4,00	4,15	1	5
	Total	1471	4,13	,750	,020	4,09	4,17	1	5
Svako ima sve ље to mu je potrebno kada je zemlja jaka.	Selo	588	3,97	,834	,034	3,90	4,04	1	5
	Prigradsko naselje	361	3,94	,931	,049	3,84	4,03	1	5
	Grad	535	3,72	1,059	,046	3,63	3,81	1	5
	Total	1484	3,87	,950	,025	3,82	3,92	1	5

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Opstanak vlastitog naroda glavni je zadatak svakog pojedinca.	Between Groups	2,688	2	1,344	2,391	,092
	Within Groups	824,676	1467	,562		
	Total	827,364	1469			
Svako ima sve ље to mu je potrebno kada je zemlja jaka.	Between Groups	18,833	2	9,416	10,565	,000
	Within Groups	1320,022	1481	,891		
	Total	1338,854	1483			

Ili npr.

Descriptives

suma_poena

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Grupa I	19	8,3579	4,27620	,98103	6,2968	10,4190	,85	13,40
Grupa II	12	10,6292	3,58896	1,03604	8,3488	12,9095	3,55	14,75
Grupa III	19	8,5000	3,70278	,84948	6,7153	10,2847	,35	13,40
Grupa IV	15	7,4400	4,03092	1,04078	5,2078	9,6722	1,00	14,05
Grupa V	13	8,8500	2,49266	,69134	7,3437	10,3563	4,25	11,60
Grupa VI	12	8,8583	2,34771	,67773	7,3667	10,3500	5,60	12,40
Grupa VII	15	9,3867	1,76306	,45522	8,4103	10,3630	6,35	12,20
Total	105	8,7771	3,40009	,33182	8,1191	9,4351	,35	14,75

ANOVA

suma_poena

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	78,499	6	13,083	1,141	,345
Within Groups	1123,806	98	11,467		
Total	1202,305	104			

Ispitivanje razlika između aritmetičkih sredina u društvenim i političkim istraživanjima veoma je čest slučaj. U istraživačkoj praksi mi se neretko pitamo da li su razlike koje smo uočili 'stvarne' ili su rezultat 'greške' uzorkovanja. Široka upotrebljivost t testa i F testa pospešuje i trend da se u društvenim istraživanjima svew češće koriste kvazi intervalne skale kao i intervalne skale. Takođe, F test i t test su saastvni deo sofisticiranijih statističkih metoda kao što je npr. regresiona analiza.

Ispitivanje veza između varijabli

- Korelaciona analiza

Veoma često u istraživanju nas zanima **povezanost** ili **asocijacija** između dve varijable. Tačnije, nas zanima da li su vrednosti jedne varijable u asocijaciji (povezane) sa vrednostima druge varijable. Npr. može nas zanimati da li postoji asocijacija između godina ispitanika i autoritarnosti. Ili npr. može nas zanimati da li je stepen obrazovanja povezan sa liberalnom političkom orjentacijom. Ili npr. može nas zanimati da li postoji veza između religioznosti i autoritarnosti itd. Ili, kada je reč o agregatnim podacima, da li su postoji veza između društvenog bruto proizvoda i mortaliteta. Evo najjednostavnijeg primera tabelarno:

	Varijable	
Objekti	x	y
Objekt1	x_1	y_1
Objekt2	x_2	y_2
Objekt3	x_3	y_3
Objekt4	x_4	y_4
...	...	...
...	...	...
...	...	...
...	...	...
Objekt i	x_i	y_i
...	...	...
...	...	...
...	...	...
...	...	...
Objekt n	x_n	y_n

Za razliku od *eksperimentalne asocijacije,* ***korelacija predstavlja onaj tip povezanosti između varijabli u kojem mi nemamo nikakvu kontrolu nad vrednostima varijabli.*** Drugim rečima, kada ispitujemo odnos korelacije između varijabli, svaki objekat može imati drugačije vrednosti na varijablama a da pri tom mi ni na koji način ne utičemo na te vrednosti. Dakle, korelacije se ispituju sa slučajnim varijablama (random variables).

> ***Korelacija predstavlja onaj tip povezanosti između varijabli u kojem mi nemamo nikakvu kontrolu nad vrednostima varijabli***

U praksi društvenih istraživanja, veliki je broj mogućih situacija kada koristimo korelacionu analizu. Npr. povezanost između inteligencije učenika i uspeha u školi. Povezanost između stope kriminaliteta i nezaposlenosti. Povezanosti

između ukupne količine para potrošene na marketing i efekta u pogledu prodaje. Povezanost između političke orijentacije i stava pema institucijama.

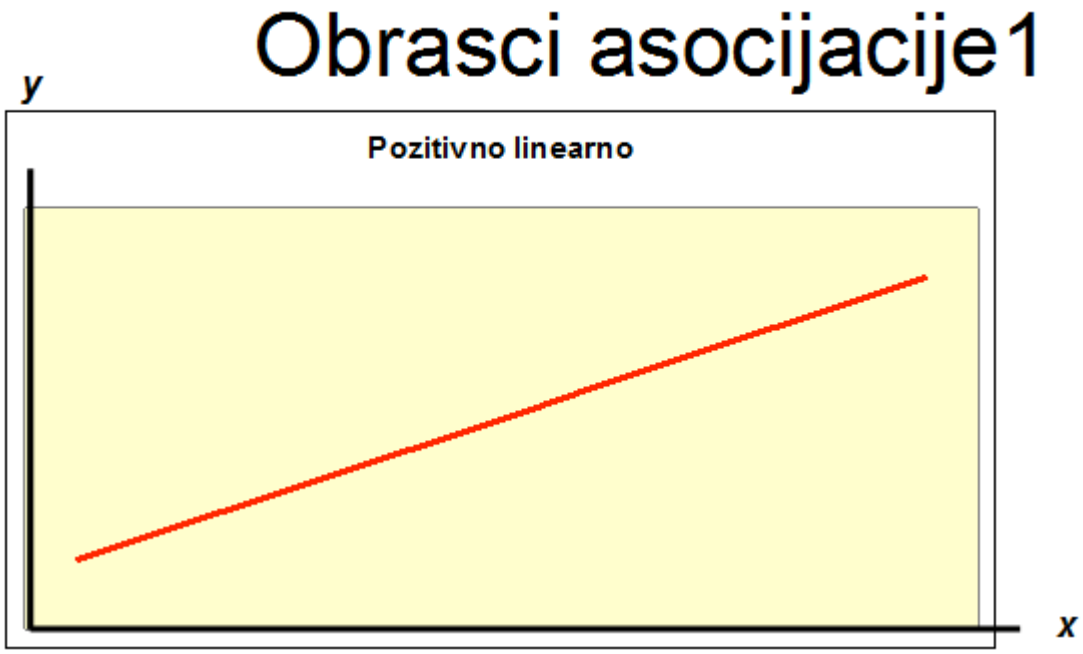
Cilj je, dakle, da koristimo korelacionu analizu kako bi ispitivali povezanost između varijabli. Kada je o korelacijama reč potrebno je da naučimo:

- Kako da izvedemo i interpretiramo korelaciju između varijabli
- Kako da razumemo uslove u kojima se meri korelacija između varijabli
- Kako da interpretiramo koeficijente korelacije
- Kako kako da formiramo i interpretiramo korelacionu matricu
- Kako da formiramo skale i proverimo njihovu pouzdanost (relijabilnost) Cronbach's Alpha koeficijentom

Varijable koje ispitujemo koralacionom analizom moraju da ispunjavaju određene uslove:

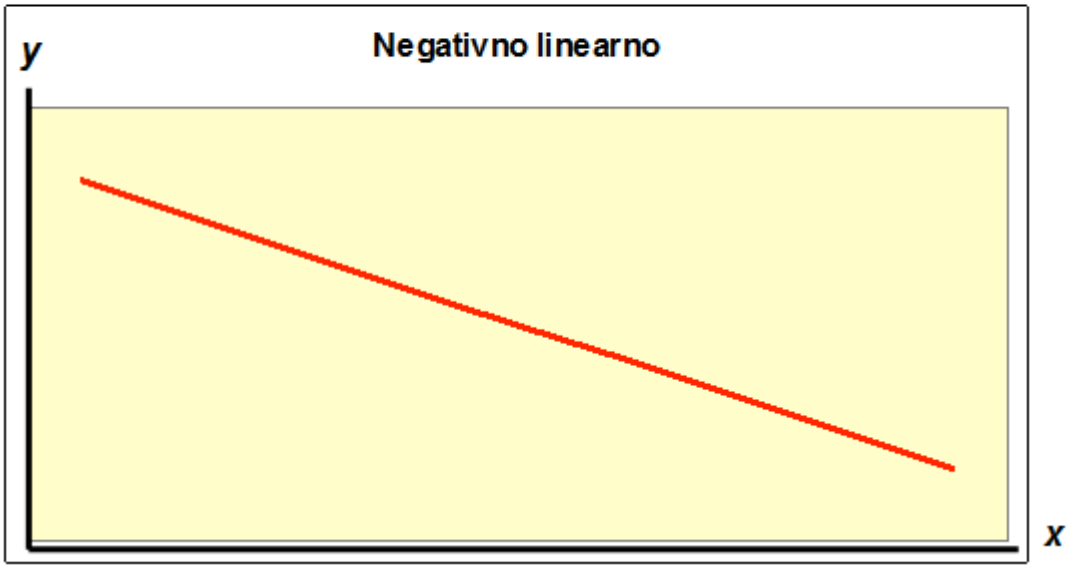
1. One moraju biti slučajne u smislu da mi nismo uticali na distribuciju vrednosti na tim varijablama
2. One moraju imati metrijske karakteristike minimalno ordinalnih skala, pri čemu su intervalne skale pogodnije za korelaciju
3. Nominalne skale (dakle kategorijalne varijable) **nikako, nikada i nigde ne smeju biti involvirane u korelacionu analizu**

Kako bi razumeli korelacionu analizu, najpre pogledajmo u grafikonima koji slede obrasce asocijacije. Najpre, moemo govoriti o **pozitivnoj linearnoj asocijaciji**, a to je kada visoke vrednosti na jednoj varijabli odgovaraju visokim vrednostima na drugoj varijabli i niske vrednosti na jednoj varijabli odgovaraju niskim vrednostima na drugoj varijabli. U praksi se najčešće dešava da ovaj obrazac asocijacije ima neki od oblika zakrivljenosti 'na dole' ili 'na gore' i onda kažemo da je asocijacija linearno konkavna.

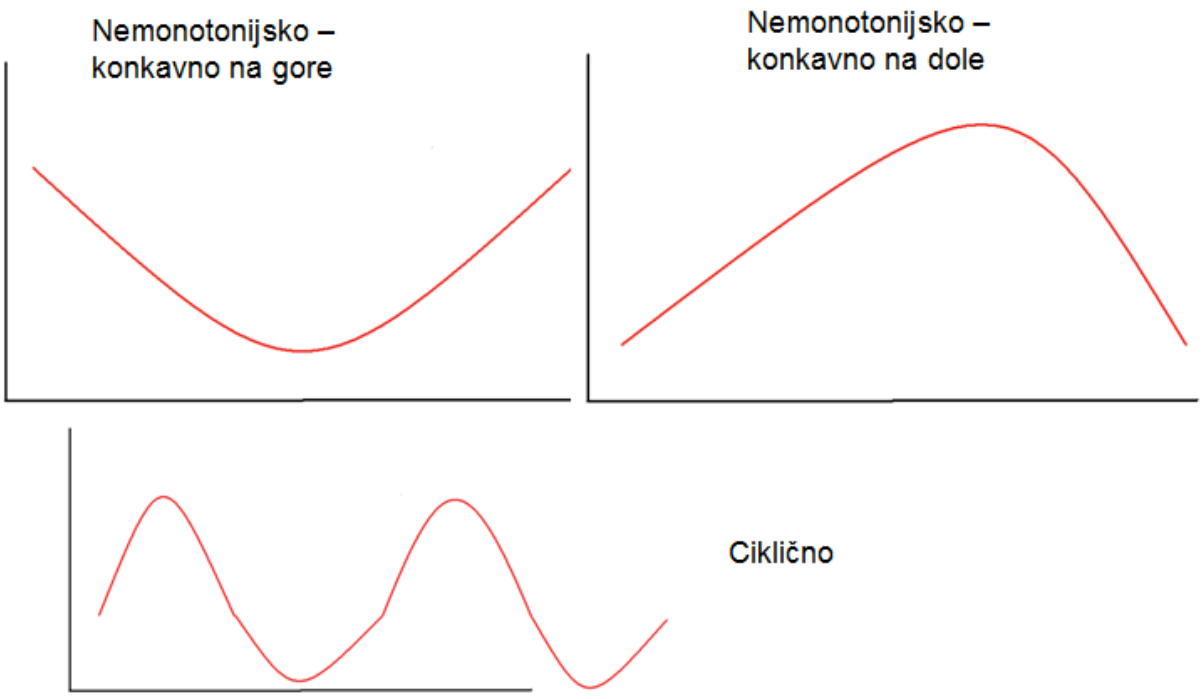


Sa druge strane, moguće je da niske vrednosti na jednoj varijabli odgovaraju visokim vrednostima na drugoj varijabli i visoke vrednosti na jednoj varijabli odgovaraju niskim vrednostima na drugoj varijabli. Onda je reč o **negativnoj linearnoj** povezanosti i ova, takođe, u praksi najčešće jeste konkavno linearna.

Obrasci asocijacije2



Važno, je imati u vidu da je linearna povezanost varijabli, samo jedan od mogućih oblika povezanosti. To znači, da varijable mogu biti povezane, a da ta veza nije linearna. Evo nekoliko klasičnih primera povezanosti između varijabli, a da veza nije linearna:



Suprotno od moguće povezanosti između varijabli mi govorimo o **nepostojanju sistematske asocijacije** između varijabli. Drugim rečima, odsustvo asocijacije znači da visoke ili niske vrednosti na jednoj varijabli jednako verovatno mogu biti 'uparene' sa visokim ili niskim vrednostima druge varijable. Tada kažemo da između dve varijable ne postoji asocijacija, tačnije, kažemo da je odnos ***nezavistan, nepovezan, nekoreliran ili ortogonalan***.

Dakle, korelacija podrazumeva da visoke vrednosti na jednoj varijabli odgovaraju visokim vrednostima na drugoj varijabli ako je korelacija pozitivna (+) ili pak da visoke vrednosti na jednoj varijabli odgovaraju niskim vrednostima na drugoj varijabli ako je korelacija negativna (-). To znači da za svaki pojedini objekt, vrednosti moramo posmatrati u paru, dakle, vrednosti objekta jedan na varijabli x se 'uparuju' sa vrednostima varijable y istog objekta, i ovo se uradi za sve objekte koje imamo u datasetu.

Tzv. 2x2 tabela je dobar način da ispitamo moguću asocijaciju između varijabli koja nas upućuje na moguću korelaciju. Evo primera jedne jednostavne tabele ovog tipa:

PRIHOD	LIBERALNA POLITIČKA ORJENTACIJA	
	Ispod medijane	Iznad medijane
Iznad medijane	(-,+) 10 ispitanika	(+,+) 40 ispitanika
Ispod medijane	(-,-) 40 ispitanika	(+,-) 10 ispitanika

Pretpostavka upotrebe kontingencione tabele jeste, da ukoliko ne postoji asocijacija između varijabli, možemo očekivati jednaku distribuciju ispitanika po ćelijama. U našem slučaju postoji jasna indikacija da se prihod i liberalna politička asocijacija nalaze u korelaciji zato što oni ispitanici koji imaju vrednosti 'iznad medijane' na varijabli 'prihod' u natprosečno većem broju slučajeva imaju vrednosti 'iznad medijane' i na varijabli liberalna politička orjentacija i obrnuto (niske vrednosti na jednoj varijabli odgovaraju niskim vrednostima na drugoj varijabli). Evo mogućih relacija koričćenjem tabele kontingencije koje govore o različitoj asocijaciji između varijabli:

Vrednosti na varijabli *x*	Vrednosti na varijabli *y*	
	Ispod medijane	**Iznad medijane**
Iznad medijane	15 ispitanika	35 ispitanika
Ispod medijane	35 ispitanika	15 ispitanika

POZITIVNA VEZA

Vrednosti na varijabli *x*	Vrednosti na varijabli *y*	
	Ispod medijane	**Iznad medijane**
Iznad medijane	35 ispitanika	15 ispitanika
Ispod medijane	15 ispitanika	35 ispitanika

NEGATIVNA VEZA

Vrednosti na varijabli *x*	Vrednosti na varijabli *y*	
	Ispod medijane	**Iznad medijane**
Iznad medijane	25 ispitanika	25 ispitanika
Ispod medijane	25 ispitanika	25 ispitanika

NEMA POVEZANOSTI

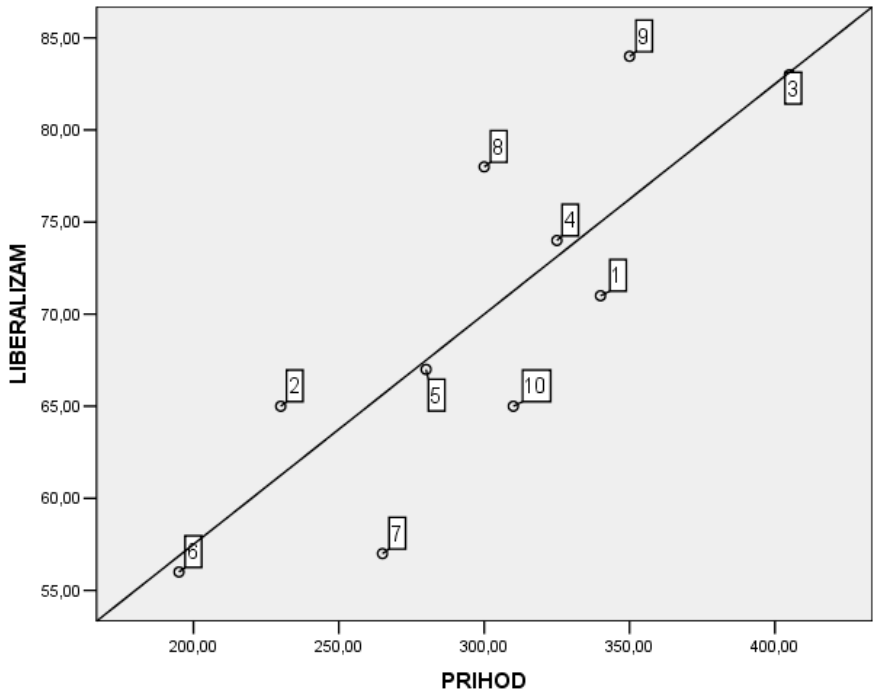
Korišćenje kontingencione tabele jeste prvi, ali svakako nedovoljan korak za identifikaciju korelacije između dve varijable. Naime, informacija o povezanosti dve varijable samo u kategorijama 'iznad' i 'ispod' medijane ne govori nam ništa o magnitudama koje se nalaze iznad i ispod, ili drugim rečima, unutar ove dve kategrije moguće su različite varijacije vrednosti na varijablama. Vrednosti koje su iznad i ispod medijane, drugim rečima mogu biti identične ali isto tako i veoma različite. S toga precizna identifikacija korelacija počiva na poređenju uparenih vrednosti na dve kontinuirane varijable.

Pretpostavimo da želimo da ispitamo korelaciju između prihoda i liberalne političke orijentacije ali koristeći varijable koje imaju kontinuirani niz brojeva. Teorijski, pretpostavka je opravdana, naime, teorija ukazuje da u društvima u kojim postoji tržište i privatna inicijativa, oni koji imaju uspeha na tom tržištu i imaju viši prihod, jesu u većoj meri pristalice liberalne političke orjentacije u odnosu na one koji na tržištu nisu naročito uspešni te imaju niske prihode, tj. pretpostavka je da ovi drugi imaju nizak nivo liberalne političke orjentacije. Tabela sa deset ispitanika bi izgledala ovako:

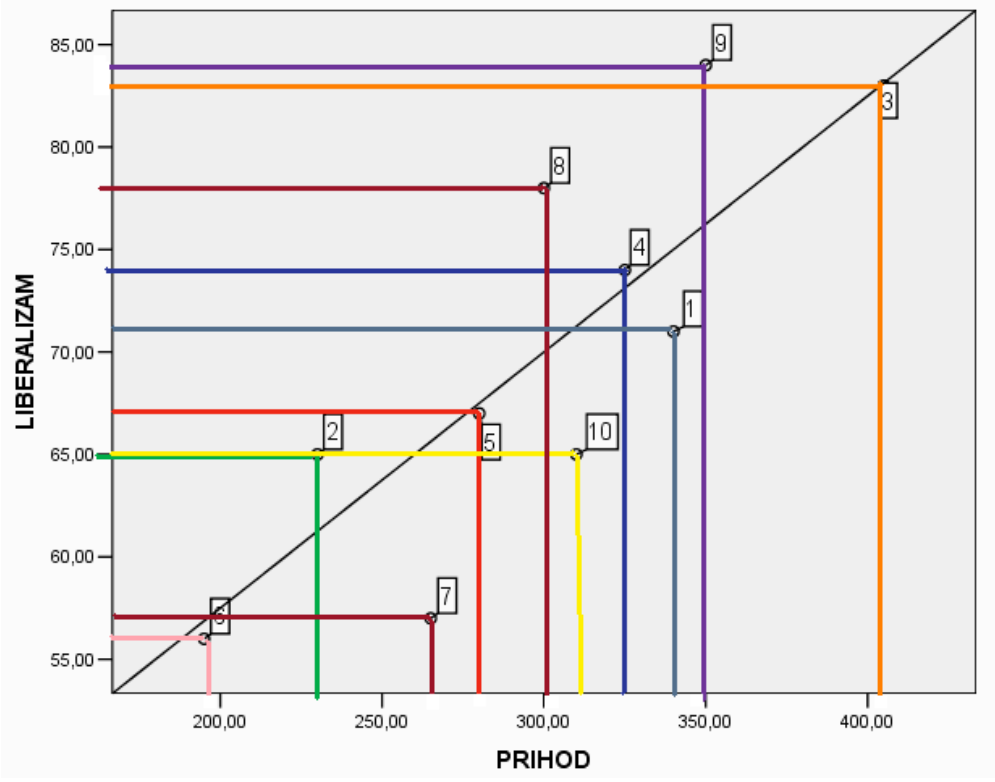
ISPITANICI	**PRIHOD**	**LIBERALIZAM**
1	340	71
2	230	65
3	405	83
4	325	74
5	280	67
6	195	56
7	265	57
8	300	78
9	350	84
10	310	65

Uparene vrednosti na varijablama - prihod i skor na skali liberalne političke orjentacije

Jedan od mogućih načina da utvrdimo linearnu povezanost između varijabli jeste da koristimo grafik '**skater'.** Ovaj tip grafičkog prikazivanja podataka nam omogućuje da uporedimo distribuciju uparenih vrednosti varijabli x i y u dvodimenzionalnom prostoru:



Ukoliko svaku vrednost objekta na varijabli x spojimo sa vrednošću koji taj objekat ima na varijabli y, možemo videti sledeće:



Na osnovu skatera, moguće je vizuelno, dakle, uočiti linerano vezu između dve varijable, u našem slučaju između lineralizma i prihoda. Precizan način za merenje korelacije između dve varijable jeste kalkulisanje tzv. koeficijenta korelacije – r

(ponekad se može naći i oznaka veliko P – ovo nikako i nikada ne sme da se pomeša sa malim p). Obzirom da utvrđivanje korelacije između dve varijable počiva na pretpostavci da postoji linearna povezanost između 'uparenih' vrednosti na dve varijable, matematički kalkulacija izgleda ovako:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Obzirom da je $(x_i - \bar{x})$ devijacija svakog individualnog objekta na varijabli x a $(y_i - \bar{y})$ devijacija svakog objekta na varijabli y pri čemu s_x i s_y predstavljaju standardne devijacije varijabli x i y dok je n broj savakog para opservacija, lako se može zaključiti da su $(x_i - \bar{x})$ i $(y_i - \bar{y})$ standardizovani skorovi na varijablama x i y. Prema tome, konačna kalkulacija za koeficijent korelacije je:

$$r = \frac{\sum z_x z_y}{n-1}$$

Evo najjednostavnijeg prikaza za ovaj kalkulus na našem primeru povezanosti između liberalističke orijentacije i prihoda:

			$z_{x_i} = \frac{x_i - \bar{x}}{s_x}$	$z_{y_i} = \frac{y_i - \bar{y}}{s_y}$	$Z_{x_i} Z_{y_i}$
	PRIHOD	LIBERALIZAM	Z varijabla x	Z varijabla y	Suma devijacija
1	340	71	0,657	0,102	0,0668
2	230	65	-1,149	-0,509	0,5844
3	405	83	1,724	1,322	2,2790
4	325	74	0,410	0,407	0,1670
5	280	67	-0,328	-0,305	0,1002
6	195	56	-1,724	-1,424	2,4543
7	265	57	-0,575	-1,322	0,7597
8	300	78	0,000	0,814	0,0000
9	350	84	0,821	1,424	1,1687
10	310	65	0,164	-0,509	-0,0835
	$\sum x = 3000$	$\sum y = 700$	$\sum z_x = 0$	$\sum z_y = 0$	$\sum z_x z_y = 7,498$
	$\bar{x} = 300$	$\bar{y} = 70.0$	$\bar{z}_x = 0$	$\bar{z}_y = 0$	
	$s_x = 60.92$	$s_y = 9.83$	$s_{zx} = 1$	$s_{zy} = 1$	

$$r = \frac{\sum z_x z_y}{n-1} = \frac{7.498}{9} = .833$$

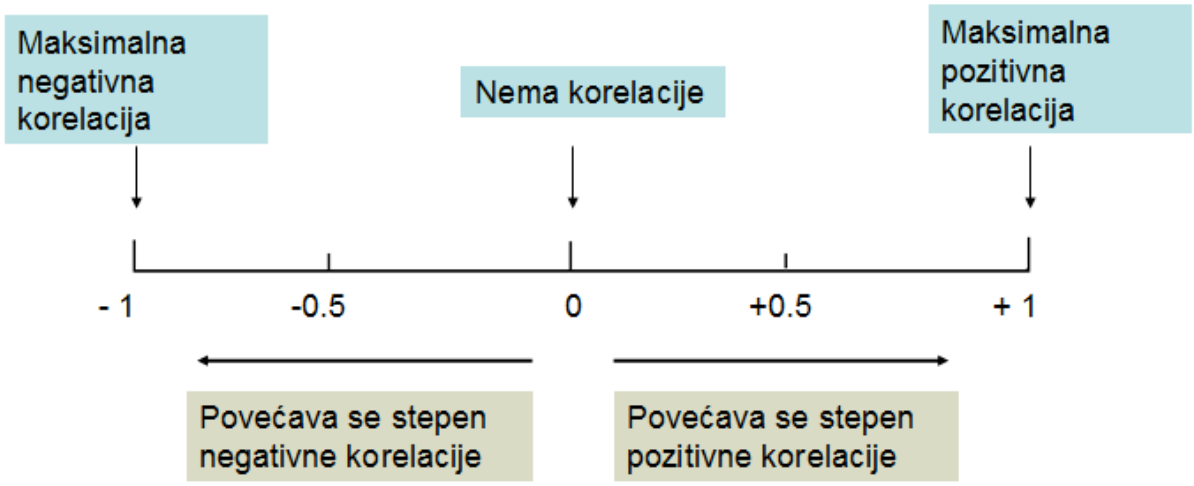
Dakle, kao osnov ispitivanje povezanosti dve varijable mi najpre transformišemo varijable preko z skorova u standardizovane varijable. Na ovaj način, razlike koje postoje u 'sirovim' podacima se transformišu u vrednosti koje izražavaju udaljenost svake vrednosti od aritmetičke sredine na obe varijable. Time se omogućava da se 'upare' vrednosti varijabli koje imaju sasvim različite metrijske karakteristike. Konačno sumiranjem svih standardizovanih skorova na jednoj i drugoj varijabli, te njihovim množenjem i konačno deljenjem sa n-1 mi računamo koeficijent korelacije.

Ukoliko nema linearne povezanosti između uparenih vrednosti z skorova na varijablama onda kažemo da nema korelacije između dve varijable i tom slučaju: r = 0. Ukoliko visoke vrednosti z skorova na jednoj varijabli u paru odgovaraju visokim vrednostima na drugoj varijabli ili ako negativne vrednosti na jednoj varijabli odgovaraju negativnim vrednostima na drugoj varijabli onda je koeficijent korelacije r = (+). Ako visoke vrednosti z skorova na jednoj varijabli odgovaraju niskim uparenim vrednostima z skorova na drugoj varijabli, ili niske vrednosti druge varijable odgovaraju visokim vrednostima prve varijable, onda je r = (-).

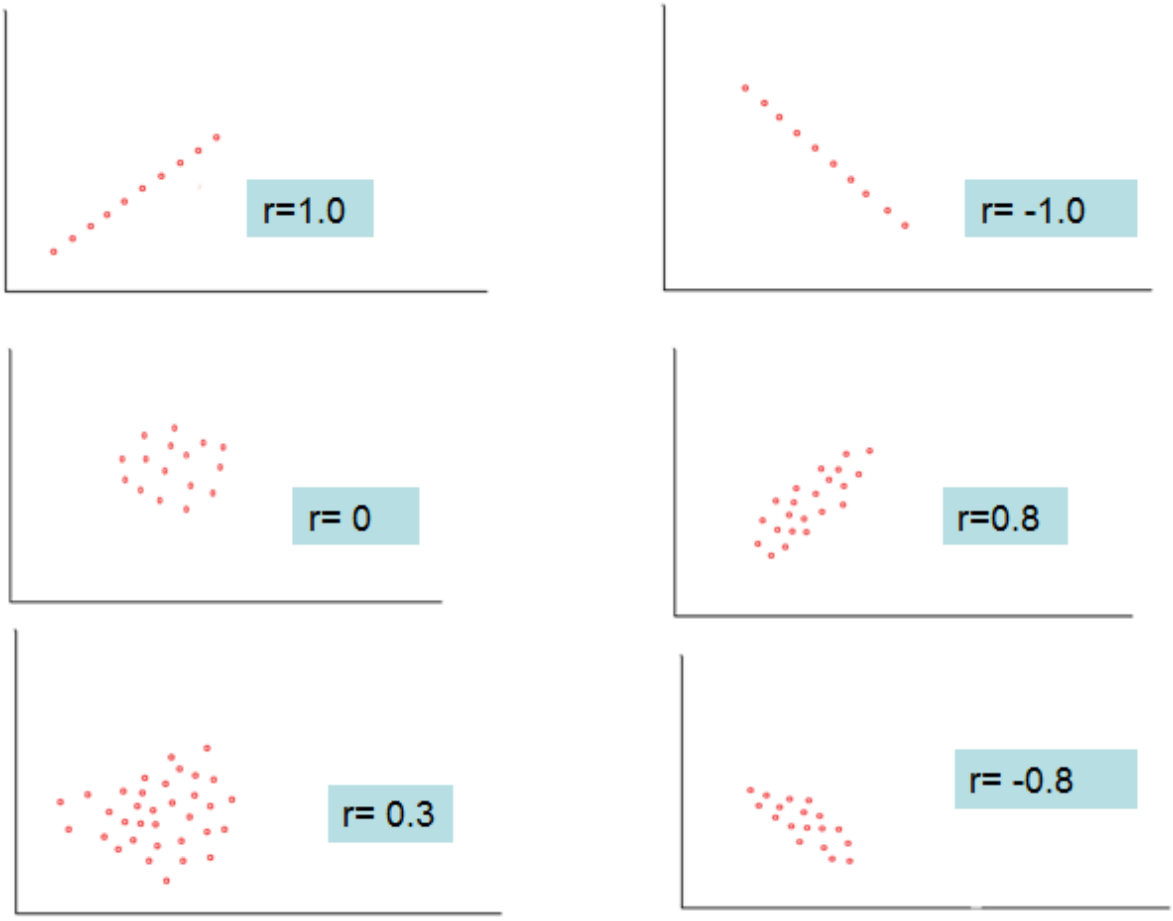
Međutim, baš kao što je slučaj sa svim ostalim statisticima koji su rezultat uzorkovanja, i koeficijent korelacije može biti manji ili veći obzirom na standardnu statističku grešku merenja, tačnije, obzirom na fluktuacije koje se mogu desiti ukoliko bi ponovili merenje na drugim uzorcima. U našem primeru, koeficijent korelacije koji smo dobili r=0,833, nam govori o tome da postoji izrazita pozitivna korelacija između prihoda i liberalne vrednosne orjentacije, ili drugim rečima, veoma,veoma, veoma je

malo verovatno da je dobijeni podatak proizvod 'greške' merenja na osnovu uzorka. Prema tome, i kada je o koeficijentu korelacije reč, važi standard jednostranog i dvostranog testa sa statističkom značajnošću p<0,05 i p<0,01. U našem primeru, ako pogledamo tabelu možemo videti da je za df = n-2, tj df = 10-2, na nivou p<0,01, potreban koeficijent korelacije za jednostrani test potrebno p =/> 0.715 a za dvostran p =/> 0.765. Dakle, dobijeni koeficijent korelacije u našem primeru r= 0,833 je veći od zahtevanih vrednosti, ili drugim rečima, mi tvrdimo da korelacija koju smo identifikovali između prihoda i liberalizna jeste stvarna na nivou 99% poverenja, tj. veoma, veoma, veoma je malo verovatno (1:100) da je korelacije rezultat greške uzorkovanja.

Vrednosti samog koeficijenta korelacije se kreću od -1 do +1. Sa stanovišta statističke značajnosti svaki koeficijent korelacije se može posmatrati u odnosu na dati broj stepena slobode u skladu sa standardima: p<0,01 (99%) i p<0,01 (95%). Ukoliko je test statistički značajan važi sledeći zaključci o povezanosti dve varijable:



Različiti koeficijenti korelacije se mogu grafički videti na sledećim grafikonima:



Evo tabelarnog prikaza najvećeg mogućeg negativnog i pozitivnog koeficijenta korelacije:

var x	var y	z x	z y
1	2	-1,486	-1,486
2	4	-1,156	-1,156
3	6	-0,826	-0,826
4	8	-0,495	-0,495
5	10	-0,165	-0,165
6	12	0,165	0,165
7	14	0,495	0,495
8	16	0,826	0,826
9	18	1,156	1,156
10	20	1,486	1,486

r=1.0

var x	var y	z x	z y
1	10	-1,486	1,486
2	9	-1,156	1,156
3	8	-0,826	0,826
4	7	-0,495	0,495
5	6	-0,165	0,165
6	5	0,165	-0,165
7	4	0,495	-0,495
8	3	0,826	-0,826
9	2	1,156	-1,156
10	1	1,486	-1,486

r= -1.0

Još jednom, važno je imati u vidu da korelacija postoji samo ukoliko je veza između varijabli linearna. Drugim rečima, korelaciona analiza je u stanju da utvrdi povezanost između varijabli samo ukoliko su varijable linearno povezane. U praksi je poznato da varijable mogu biti povezane, a da priroda veze nije linearna kao npr. u slučaju nelinearne i ciklične povezanosti među varijablama. Shodno tome,kada utvrdimo da ne postoji korelacija među varijablama, jedino što možemo tvrditi jeste da varijable nisu u linearnoj vezi a ne možemo tvrditi da između njih ne postoji neki drugi oblik povezanosti.

Postojanje korelacije ne znači nužno da između varijabli postoji kauzalni odnos

Korektna interpretacija korelacija jeste da dve varijable između kojih smo identifikovali korelaciju **kovariraju**. Korelacije se mogu koristiti kao jedan od metoda za utvrđivanje 'mogućeg' odnosa kauzaliteta, ali same korelacije nisu dostatne za utvrđivanje kauzalnih odnosa među varijablama. Utvrđivanje korelacija ima karakter predikcije u smislu da na osnovu identifikovane korelacije mi možemo predvideti odnos između dve varijable.

U društvenim i političkim istraživanjima samo istraživanje je inspirisano pojmovima koje smo identifikovali istraživačkim pitanjem. U procesu operacionalizacije ključnih pojmova, veoma retko je slučaj da su ključni pojmovi operacionalizovani singularnim varijablama. Drugim rečima, za jedan pojam mi koristimo više varijabli. Osnovna pretpostavka je da su sve varijable koje operacionalizuju jedan pojam međusobno povezane, jer ako to nije slučaj, onda smo ili izabrali pogrešne indikatore za neki pojam, ili dimenzije koje smo identifikovali jesu sporne. Utvrđivanje korelacija izeđu varijabli koje operacionalizuju jedan pojam jesu dobar način da se utvrdi povezanost između indikatora koji operacionalizuju dati pojjam. Utvrđivanje međusobnih korelacija između većeg broja varijabli se realizuje posredstvom kalkulacije korelacija između svih varijabli koje su predmet analize

Na ovaj način se dobija korelaciona matrica u kojoj mi možemo da vidimo, precizno, povezanost svih varijabli u modelu. Evo jednog primera korelaione matrice, sa varijablama koje operacionalizuju ocene političara:

Correlations

		Filip VUJANOVIC	Ranko KRIVOKAPIC	Zeljko STURANOVIC	Nebojsa MEDOJEVIC	Milo DJUKANOVIC	Andrija MANDIC	Srdjan MILIC
Filip VUJANOVIC	Pearson Correlation	1	,754**	,799**	-,077**	,784**	-,206**	-,139**
	Sig. (2-tailed)		,000	,000	,026	,000	,000	,000
	N	867	847	840	834	858	813	749
Ranko KRIVOKAPIC	Pearson Correlation	,754**	1	,680**	-,200**	,751**	-,312**	-,190**
	Sig. (2-tailed)	,000		,000	,000	,000	,000	,000
	N	847	860	841	831	846	808	749
Zeljko STURANOVIC	Pearson Correlation	,799**	,680**	1	-,028	,703**	-,209**	-,049
	Sig. (2-tailed)	,000	,000		,416	,000	,000	,184
	N	840	841	855	827	842	808	750
Nebojsa MEDOJEVIC	Pearson Correlation	-,077**	-,200**	-,028	1	-,195**	,498**	,549**
	Sig. (2-tailed)	,026	,000	,416		,000	,000	,000
	N	834	831	827	852	837	814	751
Milo DJUKANOVIC	Pearson Correlation	,784**	,751**	,703**	-,195**	1	-,300**	-,218**
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,000
	N	858	846	842	837	874	815	750
Andrija MANDIC	Pearson Correlation	-,206**	-,312**	-,209**	,498**	-,300**	1	,659**
	Sig. (2-tailed)	,000	,000	,000	,000	,000		,000
	N	813	808	808	814	815	828	746
Srdjan MILIC	Pearson Correlation	-,139**	-,190**	-,049	,549**	-,218**	,659**	1
	Sig. (2-tailed)	,000	,000	,184	,000	,000	,000	
	N	749	749	750	751	750	746	757

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Na ovom primeru može se videti korelacija između ocena za svakog političara koji je bio predmet merenja. Obratiti pažnju da je korelaciona matrica redundandta po svom karakteru, naime, po dijagonali, sve što se nalazi u jednom delu tabele, može se videti i u srugom delu tabele. Evo još jednog primera sa ocenama za institucije:

Correlations

		Skupstinu Crne Gore	Predsjednika Crne Gore	Vladu Crne Gore	Policiju Crne Gore	Sudstvo Crne Gore	Politicke partije u Crnoj Gori	Srpsku pravoslavnu crkvu	Crnogorsku pravoslavnu crkvu
Skupstinu Crne Gore	Pearson Correlation	1	,803**	,821**	,733**	,722**	,643**	-,214**	,535**
	Sig. (2-tailed)		,000	,000	,000	,000	,000	,000	,000
	N	883	871	877	869	859	833	738	689
Predsjednika Crne Gore	Pearson Correlation	,803**	1	,902**	,768**	,733**	,569**	-,273**	,561**
	Sig. (2-tailed)	,000		,000	,000	,000	,000	,000	,000
	N	871	890	881	877	866	839	741	690
Vladu Crne Gore	Pearson Correlation	,821**	,902**	1	,806**	,774**	,593**	-,310**	,577**
	Sig. (2-tailed)	,000	,000		,000	,000	,000	,000	,000
	N	877	881	891	877	868	839	743	692
Policiju Crne Gore	Pearson Correlation	,733**	,768**	,806**	1	,872**	,650**	-,164**	,540**
	Sig. (2-tailed)	,000	,000	,000		,000	,000	,000	,000
	N	869	877	877	893	876	836	744	690
Sudstvo Crne Gore	Pearson Correlation	,722**	,733**	,774**	,872**	1	,665**	-,111**	,539**
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,002	,000
	N	859	866	868	876	880	836	741	686
Politicke partije u Crnoj Gori	Pearson Correlation	,643**	,569**	,593**	,650**	,665**	1	-,032	,417**
	Sig. (2-tailed)	,000	,000	,000	,000	,000		,394	,000
	N	833	839	839	836	836	844	715	668
Srpsku pravoslavnu crkvu	Pearson Correlation	-,214**	-,273**	-,310**	-,164**	-,111**	-,032	1	-,335**
	Sig. (2-tailed)	,000	,000	,000	,000	,002	,394		,000
	N	738	741	743	744	741	715	765	674
Crnogorsku pravoslavnu crkvu	Pearson Correlation	,535**	,561**	,577**	,540**	,539**	,417**	-,335**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	
	N	689	690	692	690	686	668	674	704

**. Correlation is significant at the 0.01 level (2-tailed).

U ovoj korelacionoj matrici se može videti da sa izuzetkom Srpske Pravoslavne Crkve, sve korelaciju su izuzetno visoke, što govori o tome da pozitivno ocenjivanje jedne institucije, korelira sa pozitivnim ocenjivanjem drugihg institucija i obrnuto. Dakle, ovde se može govoriti o tome da građani imaju jedan konzistentan odnos prema svim institucijama. Ovo je jedna od tipičnih situacija kada na osnovu korelacije možemo govoriti o objedinjavanju varijable koje mere isti pojam, u ovoj situaciji to bi bio pojam **poverenje u institucije**. Drugim rečima, ukoliko želimo da merimo poverneje u institucije na jedan sintetičan način, mi bi mogli da koristimo sve gornje varijable (osim SPC) i da formiramo na osnovu njih novu varijablu koja bi merila poverenje u institucije. Ovo je jedna klasična procedura koja se radno naziva, formiranje skala, skorova ili indexa.

Dakle, u situaciji kada više varijabli operacionalizuje jedan pojam mi koristimo korelacionu matricu kako bi potvrdili da su varijable kojima se meri neka pojava međusobno povezane. U ovoj proceduri, cilj je da se ispita **relijabilnost** skale koja meri neki fenomen

Za tu svrhu se koristi **Cronbach's Alpha** koeficijent koji u osnovi predstavlja sumarni pokazatelj svih korelacija varijabli u modelu i koji se izračunava:

$$A=\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k} S_i^2}{S_p^2}\right)$$

Kada je reč o Cronbach's Alpha koeficijentu, za jednu skalu koja se sastoji iz nekoliko ajtema, po konvenciji, standard je da je minimalna vrednost koja opravdava relijabilnost 0.7. Drugim rečima, na osnovu korelacione matrice mi izračunavamo sumarno koeficijent. Cronbach's Alpha i ukoliko je ovaj veći od 0.7, mi kažemo da skala koja bi se sastojala iz varijabli koje smo uključili u merenje jeste **relijabilna**.

Da bi se formimrali skorovi (skale), najpre je potrebna teorijska konceptualizacija pojma. Uzmimo npr. tradicionalizam kao vrednosnu orijentaciju. Tradicionalistička vrednosna orjentacija predstavlja skup stavova/vrednosti koji imaju sledeće dimenzije:

- **Kolektivizam –** usmerenost i vrednovanje vlastitog naroda, nacije kao nosioca socijalno-psihološke sigurnosti za pojedinca
- **Patriotizam –** usmerenost na jaku državu koja pruža zaštitu pojedincu
- **Autoritarnost –** odricanje od individualnosti i identifikacija sa 'vodjom' koji kooptira članove zajednice i pruža zaštitu svakom njenom pripadniku
- **Istoričnost –** Oslanjanje na istorijski karakter zajedništva koji konzervira postojeći socijalni sistem

Evo najjednostavnijeg seta varijabli koje operacionalizuju pojam tradicionalizma sa dobijenim srednjim vrednostima i aritmetuičkom sredinom:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Opstanak vlastitog naroda glavni je zadatak svakog pojedinca.	1471	1	5	4,13	,750
Svako ima sve što mu je potrebno kada je zemlja jaka.	1484	1	5	3,87	,950
Bez vo?e je svaki narod kao ?ovjek bez glave.	1458	1	5	3,80	,983
Zajedni?ko porijeklo pripadnika našeg naroda temelj je našeg povjerenja.	1423	1	5	3,78	,933
Valid N (listwise)	1351				

Ukoliko formiramo korelacionu matricu sa ovim ajtemima možemo videti sledeće:

Inter-Item Correlation Matrix

	Opstanak vlastitog naroda glavni je zadatak svakog pojedinca.	Svako ima sve što mu je potrebno kada je zemlja jaka.	Bez vo?e je svaki narod kao ?ovjek bez glave.	Zajedni?ko porijeklo pripadnika našeg naroda temelj je našeg povjerenja.
Opstanak vlastitog naroda glavni je zadatak svakog pojedinca.	1,000	,406	,369	,394
Svako ima sve što mu je potrebno kada je zemlja jaka.	,406	1,000	,418	,368
Bez vo?e je svaki narod kao ?ovjek bez glave.	,369	,418	1,000	,426
Zajedni?ko porijeklo pripadnika našeg naroda temelj je našeg povjerenja.	,394	,368	,426	1,000

The covariance matrix is calculated and used in the analysis.

Dakle, kada je o Crnoj Gori reč, podaci pokazuju da je tradicionalizam sastavni deo strukture vrednosti u društvu. Na osnovu četiri ajtema koji mere četiri različite dimenzije tradicionalizma, formirali smo skalu koja je relijabilna i koja meri tradicionalizam u Crnogorskom društvu. Sada je potrebno sva četiri ajtema sumirati u zajedničku varijablu. Najjednostavniji način da se to uradi jeste transformacija svih varijabli koje mere tradicionalizam u standardizovane varijable, a onda kreiranje nove varijable (TRADICIONALIZAM) koja predstavlja srednju vrednost svih 4 standardizovanih varijabli. Za ovu svrhu najpre ćemo transformisati originale varijable u z skorove:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Zscore: Opstanak vlastitog naroda glavni je zadatak svakog pojedinca.	1471	-4,17427	1,15730	,0000000	1,00000000
Zscore: Svako ima sve što mu je potrebno kada je zemlja jaka.	1484	-3,02297	1,18739	,0000000	1,00000000
Zscore: Bez vo?e je svaki narod kao ?ovjek bez glave.	1458	-2,85074	1,21684	,0000000	1,00000000
Zscore: Zajedni?ko porijeklo pripadnika našeg naroda temelj je našeg povjerenja.	1423	-2,98258	1,30415	,0000000	1,00000000
Valid N (listwise)	1351				

Onda formiramo jedinstvenu varijablu koja meri tradicionalizam:

Descriptives

			Statistic	Std. Error
Tradicionalizam	Mean		-,0068	,01932
	95% Confidence Interval for Mean	Lower Bound	-,0447	
		Upper Bound	,0311	
	5% Trimmed Mean		,0292	
	Median		,0979	
	Variance		,563	
	Std. Deviation		,75058	
	Minimum		-2,92	
	Maximum		1,24	
	Range		4,16	
	Interquartile Range		,83	
	Skewness		-,649	,063
	Kurtosis		,732	,126

Ovo bi bio jedan od klasičnih načina za formiranje skorova (skala). Postoje i neki drugi načini (recimo faktorski skorovi), koji su specifičniji. Evo još jednog primera kada je reč o liberalističkoj vrednosnoj orijentaciji. Najpre, evo ajtema koji operacionlizuju pojam liberalizma:

Item Statistics

	Mean	Std. Deviation	N
Što vlada manje interveniše u ekonomiji, to bolje za Crnu Goru.	3,02	1,109	1016,95
Vlada ne bi trebalo da pokušava da kontroliše, reguliše ili se na bilo koji drugi na?in miješa u privatne firme.	3,01	1,082	1016,95
Bez privatizacije preduze?a bi bila u još goroj situaciji nego što su sada.	3,21	1,054	1016,95
Sve vrste javnih usluga bi bolje funkcionisale da su privatizovane.	3,18	1,103	1016,95

Evo i korelacione matrice:

Inter-Item Correlation Matrix

	Što vlada manje interveniše u ekonomiji, to bolje za Crnu Goru.	Vlada ne bi trebalo da pokušava da kontroliše, reguliše ili se na bilo koji drugi na?in miješa u privatne firme.	Bez privatizacije preduze?a bi bila u još goroj situaciji nego što su sada.	Sve vrste javnih usluga bi bolje funkcionisale da su privatizovane.
Što vlada manje interveniše u ekonomiji, to bolje za Crnu Goru.	1,000	,196	,019	,081
Vlada ne bi trebalo da pokušava da kontroliše, reguliše ili se na bilo koji drugi na?in miješa u privatne firme.	,196	1,000	,072	,154
Bez privatizacije preduze?a bi bila u još goroj situaciji nego što su sada.	,019	,072	1,000	,571
Sve vrste javnih usluga bi bolje funkcionisale da su privatizovane.	,081	,154	,571	1,000

The covariance matrix is calculated and used in the analysis.

Evo i merenje Cronbach's Alpha koeficijenta:

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,470	,471	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Što vlada manje interveniše u ekonomiji, to bolje za Crnu Goru.	9,40	5,356	,139	,042	,520
Vlada ne bi trebalo da pokušava da kontroliše, reguliše ili se na bilo koji drugi na?in miješa u privatne firme.	9,41	5,126	,204	,058	,459
Bez privatizacije preduze?a bi bila u još goroj situaciji nego što su sada.	9,20	4,651	,338	,327	,333
Sve vrste javnih usluga bi bolje funkcionisale da su privatizovane.	9,24	4,191	,419	,341	,243

Iz ovoga zaključujemo da skala liberalizma koja je operacionalizovana datim ajtemima nije relijabilna za merenje ovog pojma u Crnoj Gori.

Povezanost kategorijalnih varijabli

- Korišenje x^2 statistika za testiranje hipoteza

Neretko u društvenim i političkim istraživanjima mi koristimo kategorijalne varijable a to su one varijable koje operišu sa nominalnim skalama, kao što su npr, varijable: politička afilijacija, religijska denominacija, pol/rod, zanimanje, nacionalna pripadnost itd. Drugim rečima, metrijske karakteristike skala kojima operišu ove varijable ne dozvoljavaju da ispitujemo povezanost između varijabali F testom ili T testom ili korelacionom analizom. U ovim situacijama dizajnirane su posebne statističke tehnike koje nam omogućuju da analiziramo povezanost između kategorijskih varijabli a da pri tom ove tehnike omogućavaju da koristimo testove statističke značajnosti kako bi koristeći račun verovatnoće, ispitali postojanje povezanosti između varijabli o kojima je reč.

Najpre, x^2 se koristi kao neparametrijski test. To znači da se ovaj test koristi tako što se ispituje da li dobijena distribucija odstupa od očekivane distribucije. Upotreba ovakvog načina merenja je jako velika, npr. kada hoćemo da ispitamo da li karaktersistike našeg uzorka po nekoj varijabli odstupaju od distribucije u populaciji

Ovaj statistik upoređuje frekvenciju koju smo dobili istraživanjem sa pretpostavljenom frekvencijom. Terminološki, kalkulacija se zasniva na odnosu između očekivane (expected) i posmatrane (observed) frekvencije. Na ovaj način moguće je odbaciti nultu hipotezu da je distribucija vrednosti 'proporcionala (jednaka)', tačnije, x^2 statistik kalkuliše u kojoj meri posmatrana distribucija vrednosti odstupa od teorijske distribucije. Evo kalkulusa za izračunavanje x^2 statistika:

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Pri čemu:

- O*i* – posmatrana (observed) frekvencija *i* kategorije
- E*i* – očekivana (expected) frekvencije *i* kategorije
- k – ukupan broj kategorija

Evo jednog primera kad akoristimo x^2 statistik kao neparametrijski test. Pretpostavimo da smo u uzorku imali 150 glasača koji bi eventualno na predsedničkim izborima glasali za tri različita kandidata: Filipa Vujanovića, Nebojšu Medojevića i Andriju Mandića. Početna petpostavka, dakle, nulta hipoteza jeste da 150 ispitanika jednako preferira sva tri kandidata. Dakle, pretpostavka sadržana u nultoj hipotezi jeste da je distribucija jednaka: 50 za Filipa, 50 za Meda i 50 za Andriju. Nakon istraživanja smo utvrdili sa je distribucija:43 ispitanika za Filipa, 56

za Meda i 51 za Andriju. Evo kako to izgleda tabelarno, i kakav je kalkulus za izračunavanje x^2 statisika:

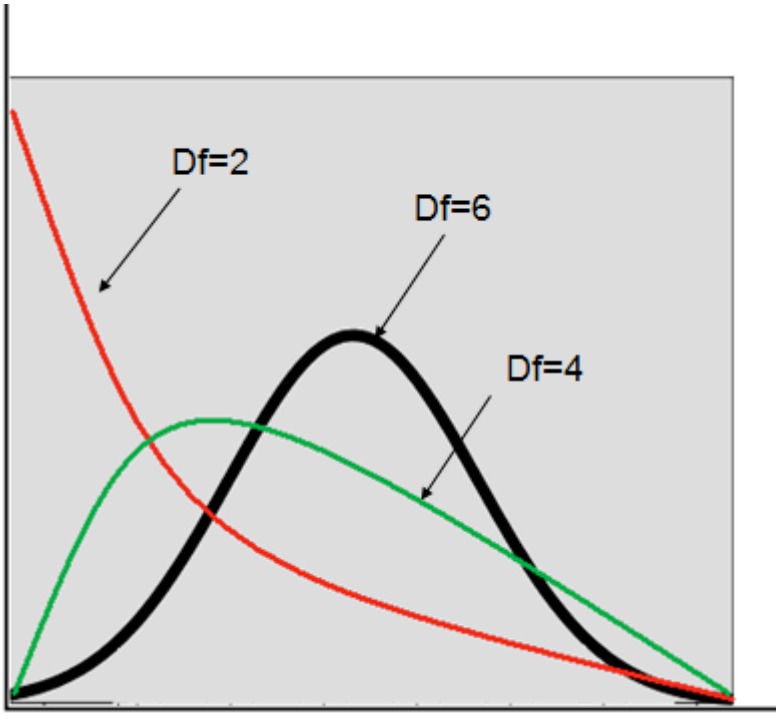
	FILIP	MEDO	ANDRIJA
Očekivana (expected) frekvencija	**50**	**50**	**50**
Posmatrana (observed) frekvencija	**43**	**56**	**51**

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(43-50)^2}{50} + \frac{(56-50)^2}{50} + \frac{(51-50)^2}{50} = 1.72$$

Kao i slučaju F testa i T testa, vrednost x^2 statistika povezana je sa brojem stepena slobode (df). Kada je o x^2 statistiku reč, broj stepena slobode je: broj kategorija minus 1:

df = k -1

U našem slučaju df = 3 – 1 = 2. I jednako, kao što postoji pretpostavka o F distribuciji i T distribuciji, postoji i x^2 distribucija koja varira u odnosu na broj stepena slobode. Evo grafički x^2 distribucija sa različitim brojem stepena slobode:



Statistička značajnost testa, ili drugim rečima testiranje nulte hipoteze, određuje se tako što se vrednost testa uparujuje sa određenim brojem stepeni slobode (df). Na primer za df = 2, kako bi test bio statistički značajan na nivou $p < 0.05$, vrednost namjmanje mora biti x^2 = 5.99. U našem primeru vrednost testa za df = 2, x^2 =1.72, što znači da mi ne možemo da odbacimo nultu hipotezu, ili drugim rečima, ne možemo reći da su razlike koje smo našli statistički značajne, ili pak, ne možemo reći da na osnovu našeg uzorka postoji razlika u verovatnoći da se glasači pre opredele za jednog u odnosu na druga dva kandidata.

Kada imamo multivarijantu matricu (veliki broj varijabli za iste objekte tj. ispitanike), dakle, kada imamo vrednosti na različitim varijablama za iste objekte (ispitanike), mi neretko želimo da utvrdimo postoji li povezanost između varijabli. Tačnije, mi želimo da utvrdimo **jesu li varijabe međusobno povezane ili su pak nezavisne.** Dakle, problem je u osnovi identičan kao u situaciji kada testiramo hipoteze korišćenjem korelacione analize samo je 'priroda' varijabli drugačija. Prema tome, ovog puta se bavimo testiranjem hipoteza koje su postavljene na način da se ispituje povezanost **kategorijalnih** varijabli. Najšira upotreba x^2 testa jeste u situaciji kada mi želimo da utvrdimo povezanost između kategorijalnih varijabli. Za ovu svrhu se koriste tabele kontingencije, ili takozvane unakrsne tabele. Zbog kvaliteta informacija koji se može dobiti uvidom u tabele kontingencije, kao i zbog lakoće testiranje hipoteza, ove tabele se upotrebljavaju jako često. Osnova za korišćenje kontingencionih tabela jeste da se pregleda distribucija vrednosti po kategorijama objekata (ispitanika) pri čemu su i same vrednosti koje se distribuiraju kategorije. Evo jednog primera tavele kontingencije koja dovodi u vezu pol ispitanika i stav glasanj na eventualnom referendumu za NATO:

Da li ste ZA da Crna Gora postane clanica NATO

Count

		Clanica NATO saveza			
		Da	Ne	Nemam odredjeno misljenje	Total
Pol	muski	164	200	94	458
	zenski	149	195	168	512
Total		313	395	262	970

Kontingencione table se koriste za testiranje hipoteza.U ovom situacijama koristi se x^2 test, kako bi se ispitala povezanost između varijabli. Osnovna pretpostavka (nulta hipoteza) na kojoj počiva test jeste da je distribucija po redovima i kolonama jednaka, ili preciznije, test ispituje da li je **distribucija po redovima i kolonama nezavisna.** Drugim rečima, x^2 test se koristi kako bi ispitali da li su razlike između kategorija statistički značajne ili nisu. Kako bi realizovali test potrebno je za svaku ćeliju u tabeli da odredimo koja je očekivana frekvencija budući da se testiranje zasniva na merenju devijacije posmatrane u odnosu na očekivanu frekvenciju. Očekivana frekvencija se određuje tako što se prvo kalkuliše proporcija (ili procenat) a onda se dobijeni podatak ponovo preračunava u fizičke brojeve tako što se izračunava odnos između marginalnih totala i opservirane frekvencije. Recimo da smo merili stavove glasača po pitanju izbora kandidata na predsedničkim izborima. Korišćenjem x^2- ta želeli

smo da utvrdimo da li je jednaka verovatnoća da muškarci i žene glasaju za dva kandidata. Recimo da smo dobili sledeću distribuciju:

	Kandidat A	Kandidat B	TOTAL
Muškarci	300	200	500
Žene	250	150	400
Total	550	350	900

Evo načina za određivanje očekivane frekvencije na osnovu dobijene:

	Kandidat A	Kandidat B	F	p
Muškarci	300	200	500	500/900 = .556
Žene	250	150	400	400/900 = .444
Total	550	350	900	1

	Kandidat A	Kandidat B	F	p
Muškarci	.556x550 =305.8	.556x350= 194.6	500	.556
Žene	.444x550 = 244.2	.444x350 = 155.4	400	.444
Total	550	350	900	1

Prema tome:

	Kandidat A		Kandidat B		F	p
Muškarci	300	305.8	200	194.6	500	.556
Žene	250	244.2	150	155.4	400	.444
Total	550		350		900	1

$$x^2 = \frac{(300-305.8)^2}{305.8} + \frac{(200-194.6)^2}{194.6} + \frac{(250-244.2)^2}{244.2} + \frac{(150-155.4)^2}{155.4}$$

$$x^2 = 0.11 + 0.15 + 0.14 + 0.19 = 0.59 \ldots\ldots df = 1$$

Minimalni vrednost za df=1 je x^2 = 3,84 na nivou p < 0.05. ZAKLJUČUJEMO: ne postoje statistički značajne razlike između muškaraca i žena kada je reč o glasanju za kandidata A i B, ili **ne možemo odbaciti nultu hipotezu da je ravnomerna distribucija po ćelijama**

Broj stepeni slobode se određuje tako što se u kontingencionoj tabeli jednostavno nakon ispunjavanja vrednosti nekih ćelija, nužno znaju vrednosti u ostalim ćelijama. Evo kako to izgleda na primeru 2x2 tabele:

	Kandidat A	Kandidat B	TOTAL
Muškarci	300 (df)		500
Žene			400
Total	550	350	900

Broj stepena slobode je, dakle, veoma važan, jer od ovog podatka zavisi distribucija, a u odnosu na distribuciju mi određujemo statističku značajnost testa. U situaciji kada koristimo kontigencione tabele broj stepeni slobode se određuje: broj redova -1 pomnoženo sa brojem kolona -1, dakle:

df=(r-1)(c-1)

Prema tome kada je reč o 2x2 kontingencionoj tabeli, broj redova je 2-1x broj kolona 2-1, iliti u slučaju 2x2 tabele df=1.

Evo još jednog primera koji govori o mogućoj povezanisti između starosti grašana i glasanja za kandidate na predsedničkim izborima:

Starost	GLASANJE ZA CETIRI KANDIDATA										
	A		B		C		D		TOTAL		
	O	E	O	E	O	E	O	E	O	f	p
18-34 god	92	73.2	39	47.1	49	43.3	53	69.4	233	233	.466
35-54 god	41	60.3	58	38.8	25	35.7	68	57.2	192	192	.384
Preko 55 god	24	23.5	4	15.1	19	14.0	28	22.4	75	75	.150
Total	157		101		123		125		500	100	1

$$x^2 = \frac{(92-73.2)^2}{73.2} + \frac{(39-47.1)^2}{47.1} + \ldots\ldots\ldots\ldots + \frac{(28-22.4)^2}{22.4} = 43.14 \ldots df = 6$$

Za df = 6, na osnovu tabele (distribucije) znamo da je vrednost za p<0,01 potrebno $x^2 = 16{,}81$, prema tome, razlike jesu značajne na nivou 99%.

Dakle, x^2 se koristi za ispitivanje povezanosti između kategorijalnih varijabli Način da to uradimo jeste da varijable koje upoređujemo prevedemo na tabele kontingencije. U tabelama kontingencije se pretpostavlja da je distribucija po redovima i kolonama nezavisna. Drugim rečima, nulta hipoteza je da dve varijable nisu povezane. Testni statistik nam u krajnjem rezultatu, testirajući hipotezu, ukazuje da li je zaista distribucija po redovim i kolonama nezavisna, ili pak, odbacujemo ovu nultu hipotezu i tvrdimo da **ne možemo reći da je distribucija po redovima i kolonama nezavisna,** a ovim konsekventno potvrđujemo alterniativnu hipotezu da su varijable povezane.

Evo nekoliko primera upotrebe x^2 za ispitivanje povezanosti između kategorijalnih varijabli:

Pol * Clanica NATO saveza Crosstabulation

			Clanica NATO saveza			
			Da	Ne	Nemam odredjeno misljenje	Total
Pol	muski	Count	164	200	94	458
		% within Pol	35,8%	43,7%	20,5%	100,0%
	zenski	Count	149	195	168	512
		% within Pol	29,1%	38,1%	32,8%	100,0%
Total		Count	313	395	262	970
		% within Pol	32,3%	40,7%	27,0%	100,0%

$x^2 = 18.74, df = 2$ p < 0,01

age_cat * Clanica NATO saveza Crosstabulation

			Clanica NATO saveza			
			Da	Ne	Nemam odredjeno misljenje	Total
age_cat	18-34	Count	120	117	96	333
		% within age_cat	36,0%	35,1%	28,8%	100,0%
	35-54	Count	118	142	102	362
		% within age_cat	32,6%	39,2%	28,2%	100,0%
	54+	Count	76	135	64	275
		% within age_cat	27,6%	49,1%	23,3%	100,0%
Total		Count	314	394	262	970
		% within age_cat	32,4%	40,6%	27,0%	100,0%

$$x^2 = 12.80, df = 4 \qquad p < 0,05$$

Nacija_1 * Clanica Evropske Unije Crosstabulation

			Clanica Evropske Unije			
			Da	Ne	Nemam odredjeno misljenje	Total
Nacija_1	Crnogorac	Count	365	16	62	443
		% within Nacija_1	82,4%	3,6%	14,0%	100,0%
	Srbin	Count	175	70	89	334
		% within Nacija_1	52,4%	21,0%	26,6%	100,0%
	Bosnjak/Musliman	Count	94	2	14	110
		% within Nacija_1	85,5%	1,8%	12,7%	100,0%
	Albanac	Count	44	0	2	46
		% within Nacija_1	95,7%	,0%	4,3%	100,0%
	Ostalo	Count	30	4	10	44
		% within Nacija_1	68,2%	9,1%	22,7%	100,0%
Total		Count	708	92	177	977
		% within Nacija_1	72,5%	9,4%	18,1%	100,0%

$$x^2 = 129.52, df = 8 \qquad p < 0,01$$

Treba imati u vidu da postoje određena ograničenja kada koristimo kontingencione tabele i x^2 test. Tačnije, postoje empirijske situacije u kojima jeste moguće koristi ovaj test ali rezultati jesu vrlo problematični, Jedna od tipičnih situacija jeste kada se koristi x^2 pri čemu operišemo veoma malim brojem slučajeva (objekata tj. ispitanika). Evo primera:

Obrazovanje * Da li, po Vasem misljenju, drzave bivse Jugoslavije treba u potpunosti da saradjuju sa Haskim tribunalom i da izruce sva lica osumnjicena za ratne zlocine? Crosstabulation

			Da li, po Vasem misljenju, drzave bivse Jugoslavije treba u potpunosti da saradjuju sa Haskim tribunalom i da izruce sva lica osumnjicena za ratne zlocine?			
			Da	Ne	Nemam odredjeno misljenje	Total
Obrazovanje	Bez obrazovanja	Count	4	3	1	8
		% within Obrazovanje	50,0%	37,5%	12,5%	100,0%
	Osnovno obrazovanje	Count	47	47	25	119
		% within Obrazovanje	39,5%	39,5%	21,0%	100,0%
	Srednje	Count	271	181	153	605
		% within Obrazovanje	44,8%	29,9%	25,3%	100,0%
	Vise obrazovanje	Count	68	35	29	132
		% within Obrazovanje	51,5%	26,5%	22,0%	100,0%
	Visoko obrazovanje	Count	75	23	14	112
		% within Obrazovanje	67,0%	20,5%	12,5%	100,0%
Total		Count	465	289	222	976
		% within Obrazovanje	47,6%	29,6%	22,7%	100,0%

U ovoj situaciji broj ispitanika u ćelijama je isuviše mali da bi smo mogli na osnovu x^2 testa doći do bilo kakvog pouzdanog zaključka, i sve tvrdnje o povezanosti između ove dve varijable u takvoj situaciji bile bi veoma problematične.

Takođe, sumnja u iszvesnot konkluzija kada se koristi kontingenciona tabela i x^2 test je otvorena u situaciji kada operišemo sa velikim brojem ćelija. Evo primera:

Pol * Ukoliko bi se naredne nedjelje odrzali parlamentarni izbori kojoj partiji bi dali Vas glas? Crosstabulation

% within Pol

		Pol		
		muski	zenski	Total
Ukoliko bi se naredne nedjelje odrzali parlamentarni izbori kojoj partiji bi dali Vas glas?	DPS / Demokratska partija socijalista	40,3%	46,5%	43,5%
	SDP / Socijaldemokarstka partija	5,1%	4,5%	4,8%
	PZP / Pokret za promjene	18,8%	20,1%	19,5%
	SNS / Srpska narodna stranka	13,1%	12,5%	12,8%
	SNP / Socijalisti?ka narodna partija	7,8%	6,4%	7,1%
	NS / Narodna stranka	,9%	1,1%	1,0%
	DSS /Demokratska srpska stranka	1,2%	,8%	1,0%
	LPCG / Liberalna partija CG	3,3%	1,9%	2,6%
	NSS / Narodna socijalisti?ka stranka	,3%		,1%
	Bošnja?ka stranka	1,5%	,6%	1,0%
	DSCG / Demokratski savez u CG	,3%		,1%
	DUA / Demokratska unija Albanaca	1,5%	1,4%	1,4%
	"Albanska alternativa"	,6%	1,1%	,9%
	Srpski radikali	2,1%	1,1%	1,6%
	Nekoj drugoj	3,3%	1,9%	2,6%
Total		100,0%	100,0%	100,0%

U ovoj situaciji, razlike mogu biti značajne zbog same činjenice da imamo veliki broj ćelija, a ne zato što zaista postoji povezanost između varijabli koje su predmet našeg interesovanja.

Konačno, kada govorimo o samoj upotrebi kontingencionih tabela, valja imati u vidu da se ove tabele mogu koristiti i u deskriptivne svrhe i na drugačiji način. Npr, moguće je koristiti tabele kontingencije na način da se 'ukrste' tri varijable. Evo primera:

Vjeroispovijest * Clanica NATO saveza * Pol Crosstabulation

% within Vjeroispovijest

Pol			Clanica NATO saveza			
			Da	Ne	Nemam odredjeno misljenje	Total
muski	Vjeroispovijest	Pravoslavna	27,2%	52,0%	20,8%	100,0%
		Islamska	65,4%	16,0%	18,5%	100,0%
		Katolicka	64,7%	17,6%	17,6%	100,0%
		Ateista (nije vjernik)	46,2%	38,5%	15,4%	100,0%
	Total		35,9%	44,0%	20,1%	100,0%
zenski	Vjeroispovijest	Pravoslavna	22,8%	45,2%	32,0%	100,0%
		Islamska	52,9%	11,8%	35,3%	100,0%
		Katolicka	48,1%	22,2%	29,6%	100,0%
		Ateista (nije vjernik)	16,7%	16,7%	66,7%	100,0%
	Total		29,1%	38,1%	32,8%	100,0%

U ovoj situaciji posmatramo povezanost između veroispovesti i stava prema NATO posebno za muškarce a poebno za žene. U ovoj situaciji, SPSS obezbeđuje x^2 test, takođe, posebno za muškarce a posebno za žene:

Chi-Square Tests

Pol		Value	df	Asymp. Sig. (2-sided)
muski	Pearson Chi-Square	53,957[a]	6	,000
	Continuity Correction			
	Likelihood Ratio	55,142	6	,000
	Linear-by-Linear Association	15,165	1	,000
	N of Valid Cases	457		
zenski	Pearson Chi-Square	51,182[b]	6	,000
	Continuity Correction			
	Likelihood Ratio	53,353	6	,000
	Linear-by-Linear Association	3,870	1	,049
	N of Valid Cases	512		

a. 3 cells (25,0%) have expected count less than 5. The minimum expected count is 2,62.

b. 3 cells (25,0%) have expected count less than 5. The minimum expected count is 1,75.

Kada se upotrebljavaju kotingencione tabele, treba voditi računa da li su totali dati po redovima ili kolonama. Evo najpre tabele koje daju totale po redovima:

Vjeroispovijest * Clanica NATO saveza Crosstabulation

% within Vjeroispovijest

		Clanica NATO saveza			
		Da	Ne	Nemam odredjeno misljenje	Total
Vjeroispovijest	Pravoslavna	25,1%	48,2%	26,7%	100,0%
	Islamska	58,7%	13,8%	27,5%	100,0%
	Katolicka	57,1%	19,0%	23,8%	100,0%
	Ateista (nije vjernik)	35,0%	30,0%	35,0%	100,0%
Total		32,4%	40,7%	26,9%	100,0%

U ovoj situaciji, dakle, interpretacija bi bila, npr. za pravoslavce: ' Od svih koji su pravoslavne veroispovesti, 25,1% podržava NATO, 48,2% su protiv, dok 26,7% nema stav prema NATO.

Drugi način bio bi da se za iste sve varijable prikaže distribucija totala po kolonama: Evo kako to izgleda:

Vjeroispovijest * Clanica NATO saveza Crosstabulation

% within Clanica NATO saveza

		Clanica NATO saveza			
		Da	Ne	Nemam odredjeno misljenje	Total
Vjeroispovijest	Pravoslavna	59,0%	90,6%	75,9%	76,4%
	Islamska	31,1%	5,8%	17,6%	17,2%
	Katolicka	7,6%	2,0%	3,8%	4,3%
	Ateista (nije vjernik)	2,2%	1,5%	2,7%	2,1%
Total		100,0%	100,0%	100,0%	100,0%

U ovoj situaciji se podaci ne interpetiraju po redovima nego po kolonama npr. za one koi podržavaju NATO, na način: 'Od svih koji podržavaju NATO 59% su pravoslavne veroispovesti, 31,1% islamske, 7,6% katoličke, a 2,2% su ateisti.

Ukoliko želimo potpuniju deskripciju povezanosti između dve varijable SPSS obezbeđuje totale i po redovima i po kolonama. U toj situaciji u zaglavlju je naznačeno u odnosu na šta ke dati procenat. Evo primera:

Vjeroispovijest * Clanica NATO saveza Crosstabulation

			Clanica NATO saveza			
			Da	Ne	Nemam odredjeno misljenje	Total
Vjeroispovijest	Pravoslavna	% within Vjeroispovijest	25,1%	48,2%	26,7%	100,0%
		% within Clanica NATO saveza	59,0%	90,6%	75,9%	76,4%
	Islamska	% within Vjeroispovijest	58,7%	13,8%	27,5%	100,0%
		% within Clanica NATO saveza	31,1%	5,8%	17,6%	17,2%
	Katolicka	% within Vjeroispovijest	57,1%	19,0%	23,8%	100,0%
		% within Clanica NATO saveza	7,6%	2,0%	3,8%	4,3%
	Ateista (nije vjernik)	% within Vjeroispovijest	35,0%	30,0%	35,0%	100,0%
		% within Clanica NATO saveza	2,2%	1,5%	2,7%	2,1%
Total		% within Vjeroispovijest	32,4%	40,7%	26,9%	100,0%
		% within Clanica NATO saveza	100,0%	100,0%	100,0%	100,0%

Dodatno, moguće je da se totali posmatraju ne samo po redovima i kolonama., nego i po ćelijama, dakle, da se za svaku ćeliju odredi koji je to procenat u odnosu na ukupan total. Evo primera:

Vjeroispovijest * Clanica NATO saveza Crosstabulation

			Clanica NATO saveza			
			Da	Ne	Nemam odredjeno misljenje	Total
Vjeroispovijest	Pravoslavna	% within Vjeroispovijest	25,1%	48,2%	26,7%	100,0%
		% within Clanica NATO saveza	59,0%	90,6%	75,9%	76,4%
		% of Total	19,2%	36,9%	20,4%	76,4%
	Islamska	% within Vjeroispovijest	58,7%	13,8%	27,5%	100,0%
		% within Clanica NATO saveza	31,1%	5,8%	17,6%	17,2%
		% of Total	10,1%	2,4%	4,7%	17,2%
	Katolicka	% within Vjeroispovijest	57,1%	19,0%	23,8%	100,0%
		% within Clanica NATO saveza	7,6%	2,0%	3,8%	4,3%
		% of Total	2,5%	,8%	1,0%	4,3%
	Ateista (nije vjernik)	% within Vjeroispovijest	35,0%	30,0%	35,0%	100,0%
		% within Clanica NATO saveza	2,2%	1,5%	2,7%	2,1%
		% of Total	,7%	,6%	,7%	2,1%
Total		% within Vjeroispovijest	32,4%	40,7%	26,9%	100,0%
		% within Clanica NATO saveza	100,0%	100,0%	100,0%	100,0%
		% of Total	32,4%	40,7%	26,9%	100,0%

Takođe, kada analiziramo podatke na način da želimo da utvrdimo statističku značajnot povezanosti između dve varijable, možemo da uporedimo očekivanu i dobijenu frekvenciju. Evo primera:

Vjeroispovijest * Clanica NATO saveza Crosstabulation

			Clanica NATO saveza			
			Da	Ne	Nemam odredjeno misljenje	Total
Vjeroispovijest	Pravoslavna	Count	186	358	198	742
		Expected Count	240,7	301,8	199,4	742,0
	Islamska	Count	98	23	46	167
		Expected Count	54,2	67,9	44,9	167,0
	Katolicka	Count	24	8	10	42
		Expected Count	13,6	17,1	11,3	42,0
	Ateista (nije vjernik)	Count	7	6	7	20
		Expected Count	6,5	8,1	5,4	20,0
Total		Count	315	395	261	971
		Expected Count	315,0	395,0	261,0	971,0

Konačno, jedna potpuna tabela kontingencije može da sadrži sve ove informacije u jednoj jedinoj tabeli:

Vjeroispovijest * Clanica NATO saveza Crosstabulation

			Clanica NATO saveza			
			Da	Ne	Nemam odredjeno misljenje	Total
Vjeroispovijest	Pravoslavna	Count	186	358	198	742
		Expected Count	240,7	301,8	199,4	742,0
		% within Vjeroispovijest	25,1%	48,2%	26,7%	100,0%
		% within Clanica NATO saveza	59,0%	90,6%	75,9%	76,4%
		% of Total	19,2%	36,9%	20,4%	76,4%
	Islamska	Count	98	23	46	167
		Expected Count	54,2	67,9	44,9	167,0
		% within Vjeroispovijest	58,7%	13,8%	27,5%	100,0%
		% within Clanica NATO saveza	31,1%	5,8%	17,6%	17,2%
		% of Total	10,1%	2,4%	4,7%	17,2%
	Katolicka	Count	24	8	10	42
		Expected Count	13,6	17,1	11,3	42,0
		% within Vjeroispovijest	57,1%	19,0%	23,8%	100,0%
		% within Clanica NATO saveza	7,6%	2,0%	3,8%	4,3%
		% of Total	2,5%	,8%	1,0%	4,3%
	Ateista (nije vjernik)	Count	7	6	7	20
		Expected Count	6,5	8,1	5,4	20,0
		% within Vjeroispovijest	35,0%	30,0%	35,0%	100,0%
		% within Clanica NATO saveza	2,2%	1,5%	2,7%	2,1%
		% of Total	,7%	,6%	,7%	2,1%
Total		Count	315	395	261	971
		Expected Count	315,0	395,0	261,0	971,0
		% within Vjeroispovijest	32,4%	40,7%	26,9%	100,0%
		% within Clanica NATO saveza	100,0%	100,0%	100,0%	100,0%
		% of Total	32,4%	40,7%	26,9%	100,0%

Ova tabela zaista daje sve potrebne informacije ali je istovremeno opterećena podacima i osim za svrhe analize u procesu istraživanja, nije uobičajeno da se prikazuje u istraživačkim izveštajima.

Jedna od prednosti korišćenja tabela kontingencije i x^2 testa jeste i mogućnopst korišćenja varijabli koje su različite p svom karakteru. Dakle, nije nužno da se 'ukrštaju' demografske varijable i one koje su predmet našeg interesovanja. Evo jednog primera 'ukrštanja':

Clanica NATO saveza * Navedite TV stanicu u koju imate najvise povjerenja Crosstabulation

% within Clanica NATO saveza

		Navedite TV stanicu u koju imate najvise povjerenja									
		TVCG	ELMAG	Montena	MBC	Pink	In	RTS	TV Atlas	Ostale tv stanice	Total
Clanica NATO saveza	Da	38,8%	,7%	4,0%	,7%	5,8%	39,9%		6,8%	3,2%	100,0%
	Ne	14,0%	27,1%	4,3%	1,0%	20,7%	16,7%	3,3%	6,7%	6,0%	100,0%
	Nemam odredjeno misljenje	24,3%	14,1%	3,8%	1,6%	16,8%	29,7%	1,1%	5,9%	2,7%	100,0%
Total		25,6%	14,3%	4,1%	1,0%	14,3%	28,3%	1,6%	6,6%	4,2%	100,0%

$x^2 = 171{,}68$ za Df = 16 i p < 0,01

Za kraj, nekoliko zaključnih konstatacija. Kontingencione tabele su veoma dobro deskriptivno sredstvo. Kontingencione tabele imaju široku upotrebu, a u slučaju kategorijalnih varijabli neretko su jedino sredstvo analize. Kontingencione tabele se mogu koristiti za analizu povezanosti između varijabli. Kontingencione tabele su obično početni metod analize, koji kasnije, ukoliko želimo veću pouzdanost konkluzija, mora biti testiran rigidnijim tehnikama i metodama. Prema tome, kontingencione tabele nam pre služe da postavimo hipoteze koje ćemo kasnije da proveravamo, zato što imaju veliku moć deskripcije.

Regresiona analiza

Korelaciona analiza samo ispituje kovarijaciju između dve varijable. Korelaciona analiza ništa ne govori o **prirodi veze** između varijabli koje kovariraju. **Regresiona analiza** nam omogućuje da ispitamo **prirodu veze** između dve varijable. Regresiona analiza koristeći određene matematičke formule, nam omogućava da postignemo preciznost i to **u kojoj meri** se vrednosti na jednoj varijabli mogu predvideti distribucijom vrednosti druge varijable.

Regresiona analiza ima za cilj **predviđanje**, dakle, da se distribucija vrednosti jedne varijable predvidi vrednostima druge varijable. Zbog toga se regresiona analiza opisno naziva **prediktorskom analizom.** Obzirom da u regresionoj analizi imamo minimum dve varijable, termini koji se koriste su sledeći:

- **Kriterijumska** varijabla je ona čije vrednosti želimo da predvidimo
- **Prediktorska** varijabla je ona od čija distribucija vrednosti jeste osnov za predikciju vrednosti kriterijumske varijable

Neretko, umesto termina kriterijumska, koristi se termin **zavisna varijabla** a umesto prediktorska, koristi se termin **nezavisna varijabla**.U ovoj situaciji, kažemo da od distribucije vrednosti na nezavisnoj varijabli zavisi distribucija vrednosti na zavisnoj varijabli. Problem sa ovom terminologijom jeste u tome što se na ovaj način prejudicira kauzalitet između varijabli, pri čemu, epistemološki, ova vrsta povezanosti nije nužna na osnovu same činjenice da se na osnovu distribucije vrednosti jedne varijable može predvideti distribucija vrednosti druge varijable. Međutim, teorijski, i ova terminologija je prihvatljiva ukoliko se pod 'zavisna' i 'nezavisna' prosto podrazumeva da u matematičkom modelu postoji odnos zavisnosti numeričkih nizova. Takođe, ukoliko teorijski dizajn istraživanja nedvosmisleno pretpostavlja kauzalitet, onda je ovakva terminologija sasvim opravdana.

U istraživačkoj praksi razlikujemo dva tipa regresija:

- Prvo, to je **jednostavna** (*simple)* regresija, to jest, kada želimo da predvidimo vrednost objekta na kriterijumskoj varijabli na osnovu vrednosti koju taj objekat ima na **jednoj** prediktorskoj varijabli
- Drugo, to je **višestruka** (multiple) regresija, tj. kada želimo da predvidimo vrednost nekog objekta na kriterijumskoj varijabli na osnovu vrednosti koje taj objekat ima na **više** prediktorskih varijabli

Ciljevi regresione analize su prema tome:

1. Da odredimo da li postoji veza između dve varijable
2. Da odredimo prirodu povezanosti, tj. da li se ta veza treba opisati matematičkom formulom
3. Da procenimo stepen preciznosti veze između varijabli
4. U situaciji kada primenjujemo multiple regresiju, da procenimo relativan značaj prediktorskih varijabli u smislu njihovog doprinosa u pogledu predikcije objašnjene varijanse kriterijumske varijable

Evo jednog najjednostavnijeg primera:

	varijabla x	varijabla y
Objekt 1	8	31
Objekt 2	5	22
Objekt 3	11	40
Objekt 4	4	19
Objekt 5	14	49

← Posmatraj vrednosti u paru za svaki objekt

Šta se može najjednostavnije zaključiti iz ove tabele:

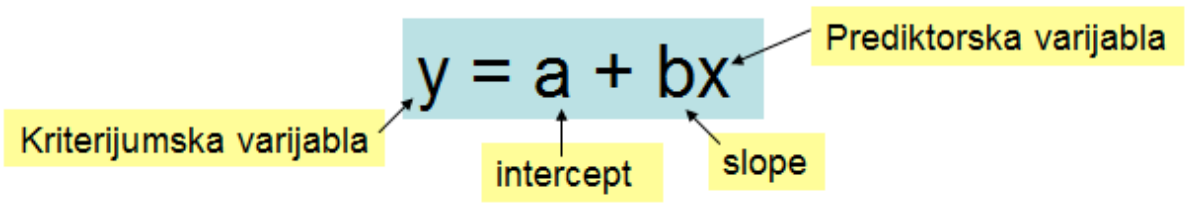
- Vrednosti na varijabli **x** su manje nego vrednosti na varijabli **y**
- Preciznije, vrednosti na varijabli **y** su više nego trostruko veće od vrednosti na varijabli **x**
- Još preciznije, vrednosti na varijabli **y** su **tačno** tri puta veće od vrednosti na varijabli **x** plus konstantna vrednost **7**
- Dakle, y = 3x + 7
- Po konvenciji, konstantnu vrednost pišemo prvo, te prema tome:

$$\mathbf{y = 7 + 3x}$$

Malo podsećanje iz elementarne algebre! y = 7+3x je jednačina koju karakteriše specifična prava linija. Ta prava linija jeste zapravo linearna funkcija koja ima određeni 'nagib' (slope). Formula y = 7+3x se može generalizovati u obliku:

$$\mathbf{y = a + b_x}$$

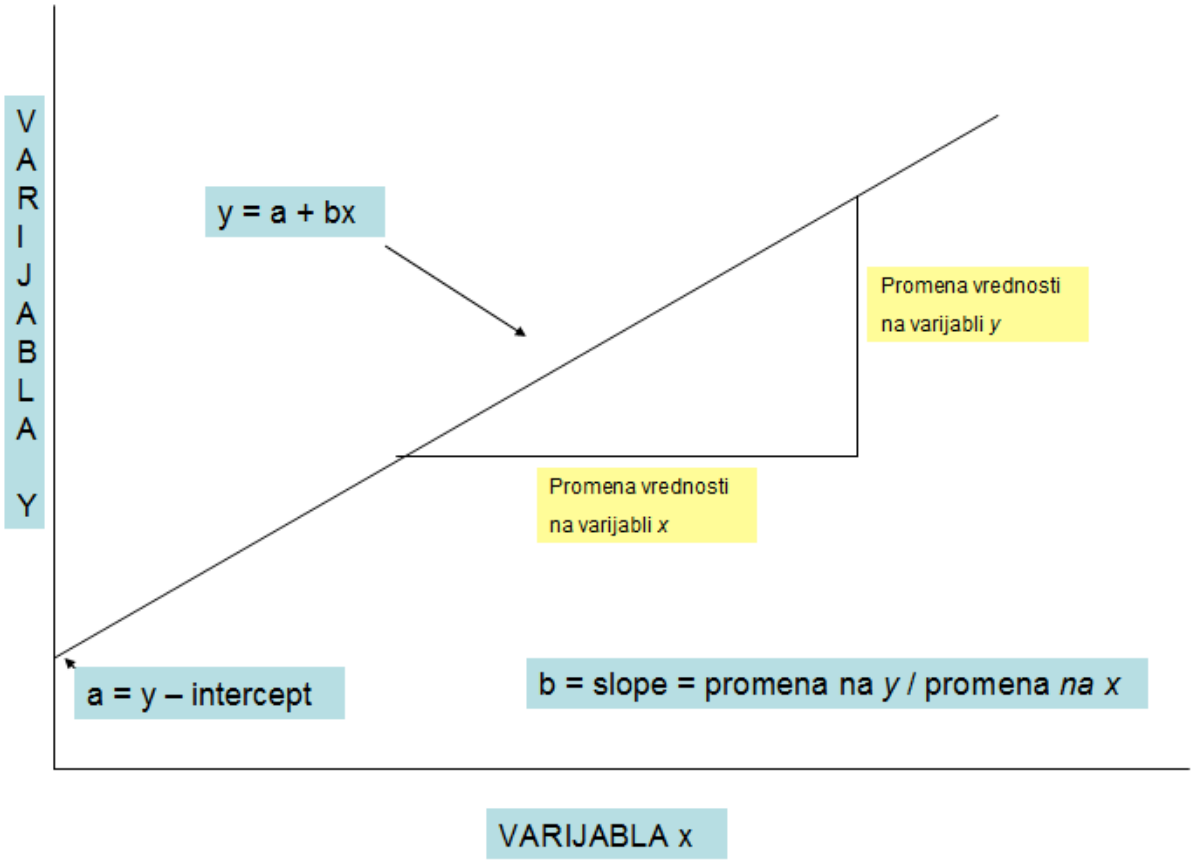
Najpre još jednom da definišemo terminologiju:



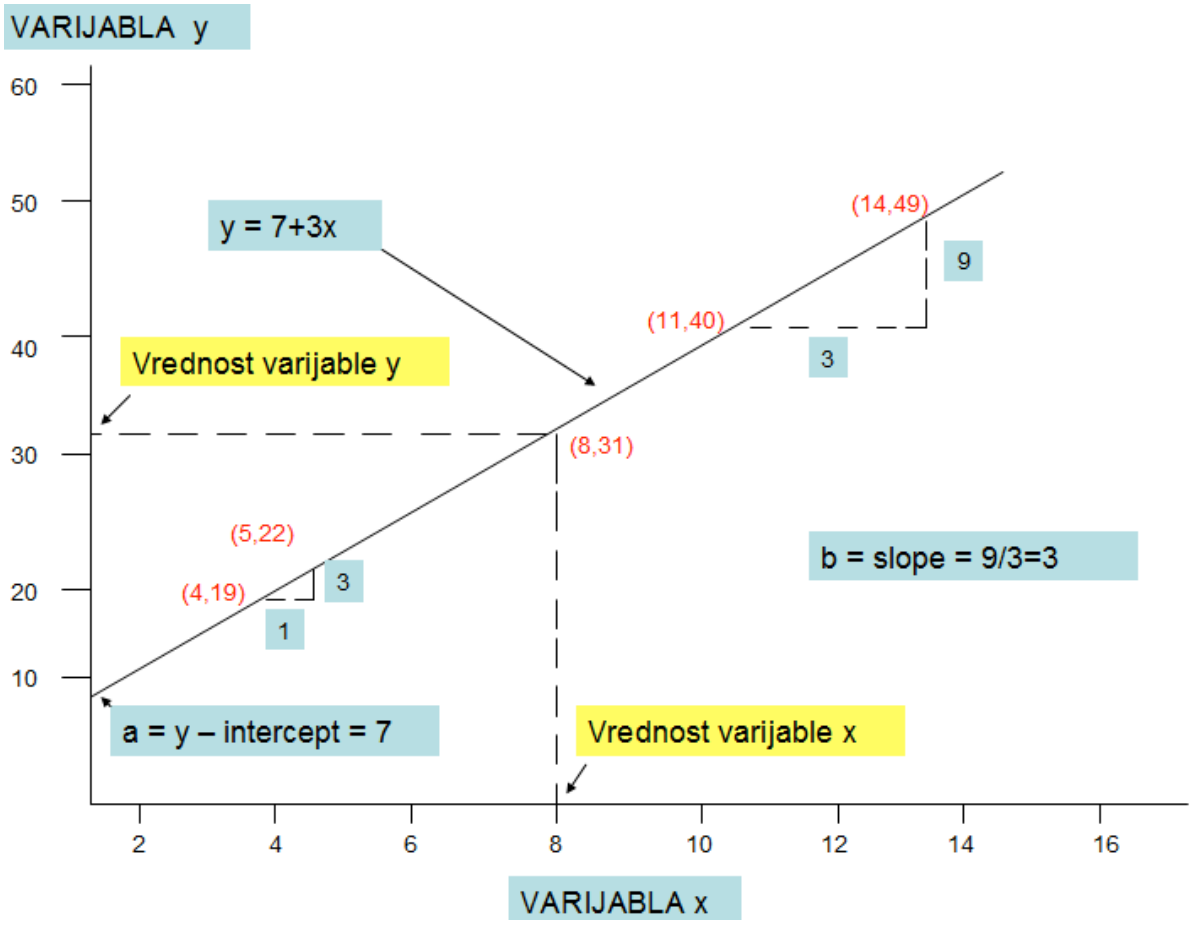
Dakle, vrednost **b** jeste **slope** (nagib linije). Opisno, **b** jeste inklinacija linearne funkcije ili preciznije, **b** je stepen promene vrednosti na varijabli **y** za svaku **jedinicu** promene vrednosti na varijabli **x.** Vrednost **a** je **konstanta** (constant term) i ona predstavlja vrednost na varijabli **y** kada je vrednost na varijabli **x** = 0. Obzirom da

konstanta predstavlja u linearnoj funkciji vrednost **y** u situaciji gde linija presreće y – osu, konstanta se drugačije i u praksi naziva **y- intercept** ili jednostavno **intercept** (presretač).

Grafički se to može prikazati na sledeći način:



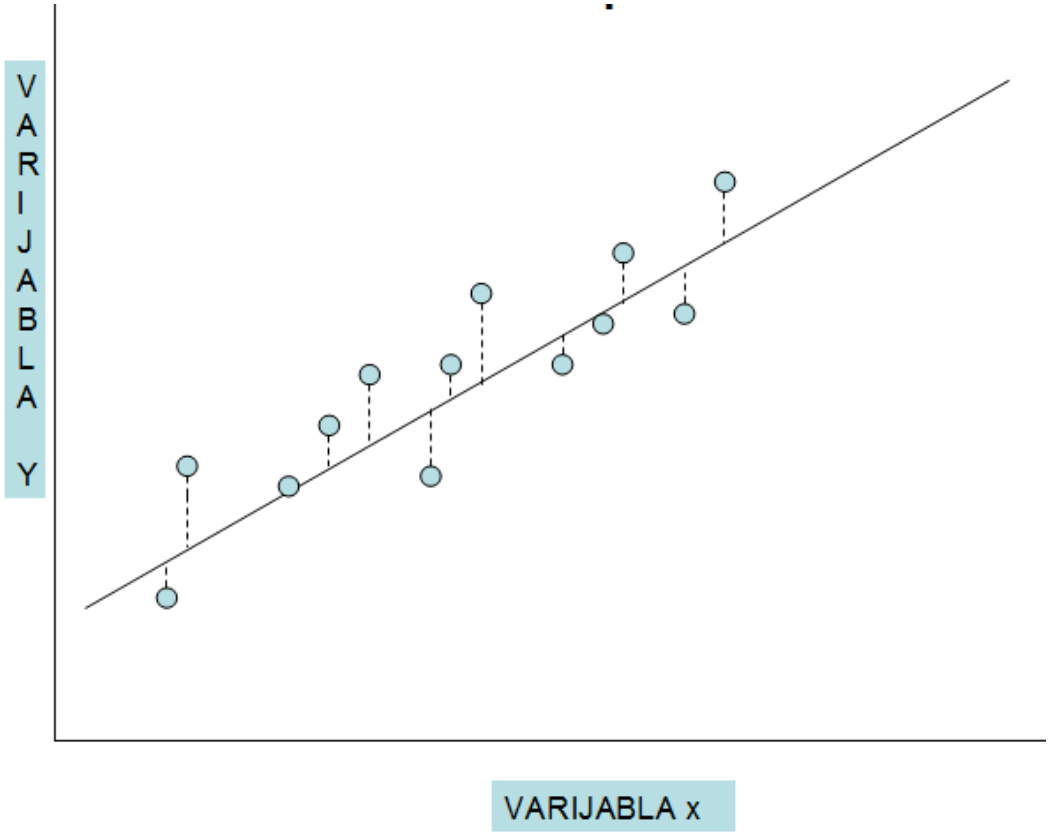
U našem slučaju za dve varijable koje smo prikazali u tabeli, to izlgeda ovako:



Simulacija koju smo prikazali je svakako 'idelano tipska'. U stvarnosti ne postoje tako precizne linearne veze između dve varijable kao u našem primeru: y = 7+3x. U stvarnosti je situacija takva da mi zapravo nalazimo aproksimaciju linearne veze koja više ili manje odudara od idealno tipske predstave, npr.

	varijabla x	varijabla y
Objekt 1	8	31
Objekt 2	5	22
Objekt 3	11	40
Objekt 4	4	19
Objekt 5	14	49
Objekt 6	7	26
Objekt 7	3	18
Objekt 8	6	30
Objekt 9	8	28
Objekt 10	7	30

Dakle, u praksi ne postoje tako 'idealne' linearne veze kao u našem primeru, tačnije, i kada postoje veze one odstupaju manji ili više od idelne linije. To se može videti na sledećem grafikonu:



Koncept **regresione linije** je veoma značajan. Jedno od ključnih pitanja kada je reč o regresionoj analizi jeste, 'gde tačno' povući liniju od svih mogućih linija, kako bi imali liniju koja je 'najbolja' u smislu da su tačke susretanja prediktorske i kriterijumske varijable kumulativno najbliže samoj liniji? Naime, obzirom da postoji bezgraničan broj mogućih linija koji izražavaju **slope** i **intercept**, postavlja se pitanje zašto linija koju povlačimo nije sa većim ili manjim nagibom? Zašto intercept nije više ili niže na y – osi? Ideja je naravno da regresiona linija prođe po 'sredini', ali gde je sredina? Kako određujemo sredinu? Da li imamo objetivan kriterijum da odredimo sredinu?

Recimo da uzmemo da linija koja najbolje 'fituje' treba da prođe kroz $\bar{x}$ i $\bar{y}$ (aritmetičke sredine obe varijable). Ali i u toj situaciji postoji neograničeno veliki broj linija koje možemo povući?! Ovo su, dakle, razlozi zbog kojih nam treba precizno definisana procedura da definišemo najbolju moguću (best fitting) regresionu liniju za:

$$\mathbf{y' = a + b_x}$$

pri čemu je **y′** pretpostavljena vrednost kriterijumske varijable za datu vrednost prediktorske varijable. Koristimo **y′** za predviđenu vrednost kako bi ovu predviđenu vrednost razlikovali od opserviranih vrednosti **y.**

Jedan od standardizovanih i u praksi najkorisnijih načina da se odredi 'linija koja najbolje fituje' jeste kriterijum **sume najmanjih kvadrata** (summ of least square) Na prethodnom grafikonu isprekidanom linijom smo pokazali u kojoj meri svaka tačka susretanja pojedinačno odstupa od regresione linije. Kriterijumom 'least squares' biramo onu liniju, među svim mogućim linijama koja ima najmanju sume kvadratnih devijacija tačaka od linije.
Obzirom da je regresiona linia y'=a+bx, naš zadatak je da identifikujemo vrednosti a i b koje će da minimalizuje:

$$\sum (y_i - y'_i)^2$$

pri čemu je $\mathbf{y}_i$ opservirana vrednost a $\mathbf{y}_i'$ predviđena vrednost

Ovakvim rešenjem mi zapravo tražimo liniju koja izražava najmanju razliku između opserviranih vrednosti i onih koje su rezultat predikcije. Least squared kriterijum je, dakle,sasvim elegantno rešenje za povlačenje regresione linije. Međutim, i dalje ostaje problem da se povuče linija u praksi na osnovu ovog kriterijuma. Drugim rečima, to bi značilo da putem pokušaja i pogrešaka povlačimo veliki broj mogućih linija, izračunamo sumu najmanjih kvadrata devijacija i izaberemo onu liniju koja ima najmanju vrednost sume najmanjih kvadrata. S toga se u praksi moramo osloniti na matematičku formulu koja će nam omogućiti da odredimo liniju koja 'najbolje fituje':

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Po ovoj formuli se mora uraditi sumarizacija za sve *n* parove (x_i, y_i). Vrednost ***a*** tj. y – intercept se može prikazati kao funkcija ***b,*** $\bar{x}$ ***i*** $\bar{y}$ ***:***

$$a = \bar{y} - b\,\bar{x}$$

Evo celokupnog kalkulusa:

r = 0.83

ISPITANICI	PRIHOD (var x)	LIBERALIZAM (var y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	340	71	40	1	40	1600
2	230	65	-70	-5	350	4900
3	405	83	105	13	1365	11025
4	325	74	25	4	100	625
5	280	67	-20	-3	60	400
6	195	56	-105	-14	1470	11025
7	265	57	-35	-13	455	1225
8	300	78	0	8	0	0
9	350	84	50	14	700	2500
10	310	65	10	-5	-50	100
sum x	3000	700			4490	33400
mean	300	70				
std	60,92	9,83				

alternativno

$$b = r\left(\frac{s_y}{s_x}\right)$$

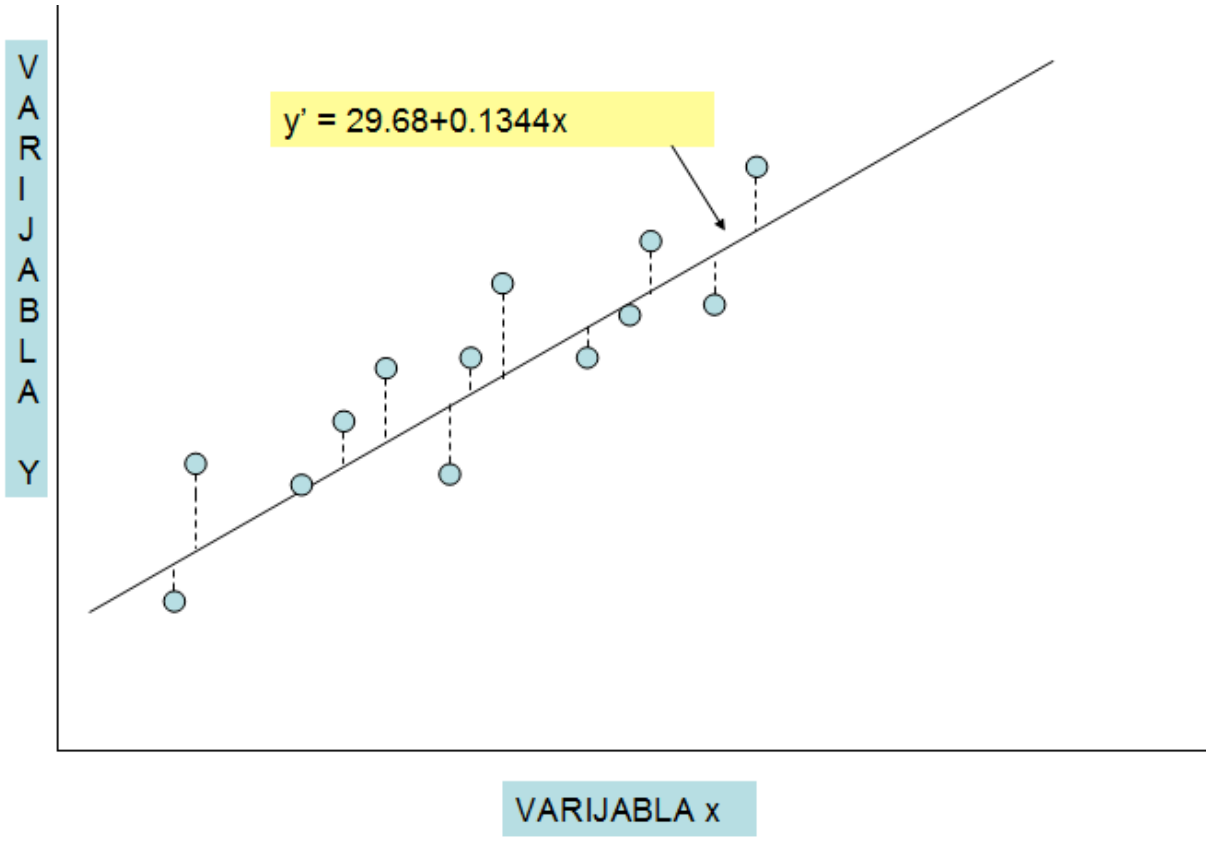
$$b = 0.83\left(\frac{9,83}{60.92}\right) = 0.1344$$

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{4490}{33400} = 0.1344$$

$$a = \bar{y} - b\bar{x} = 70,0 - (0,1344)(300) = 29.68$$

$$y' = a + bx \qquad y' = 29.68 + 0.1344x$$

Grafički se to može najjednostavnije prikazati na sledeći način:



Dakle:

$$y' = a+bx$$

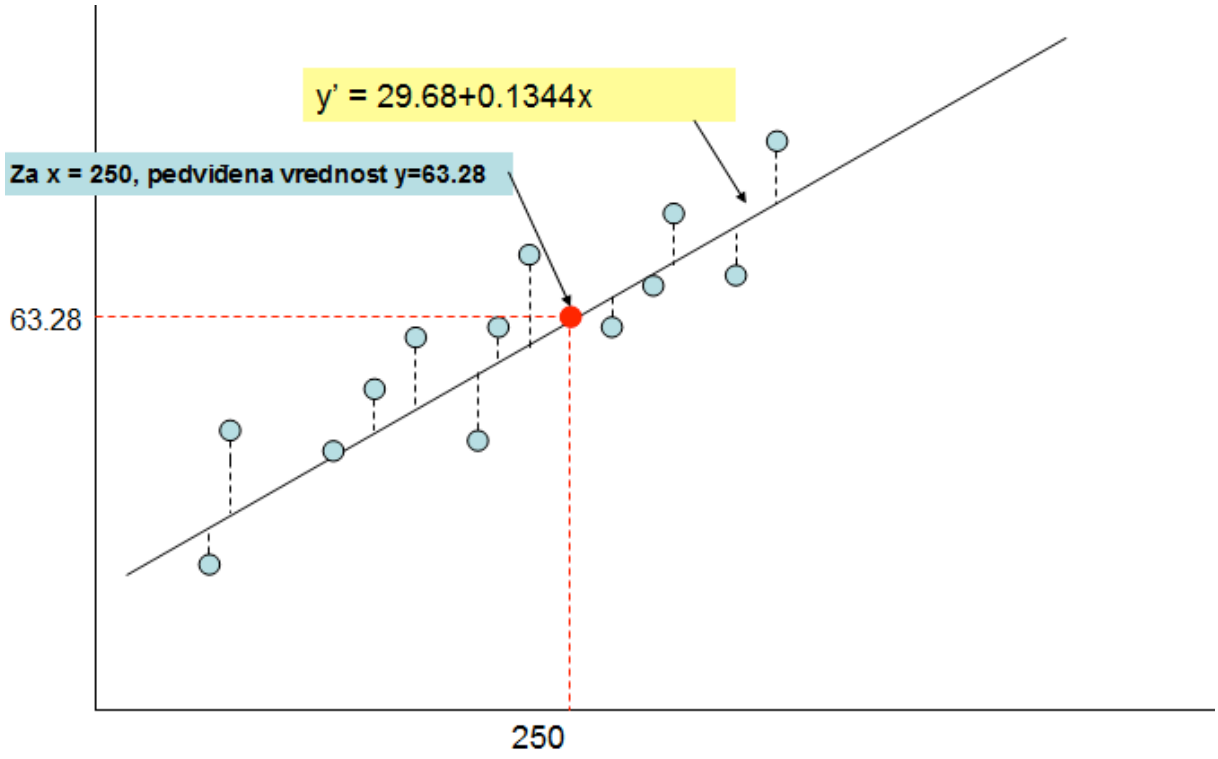
U našem slučaju:

$$y' = 29.68+0.1344x$$

Sledi da ukoliko npr. neki objekt ima vrednost x = 250 onda

$$y'=26.68+0.1344(250) = 63.28$$

Prema tome, ako neki objekat ima prihod od 250 EUR, mi predviđamo da će imati skor od 63.28 na skali liberalizma. Na ovaj način se na osnovu vrednosti prediktorske varijable predviđa vrednost kriterijumske varijable. Evo grafičkog prikaza:



Varijable su u dosadašnjem kalkulusu date u 'sirovom' obliku, tačnije, kalkulus je napravljen na osnovu izvornih vrednosti objekata na autentičnoj skali. Zbog različitih potreba (govorićemo kasnije o ovome), mi neretko imamo potrebu da standardizujemo varijable koje su predmet testiranja hipoteza u regresionoj analizi. U toj situaciji formula:

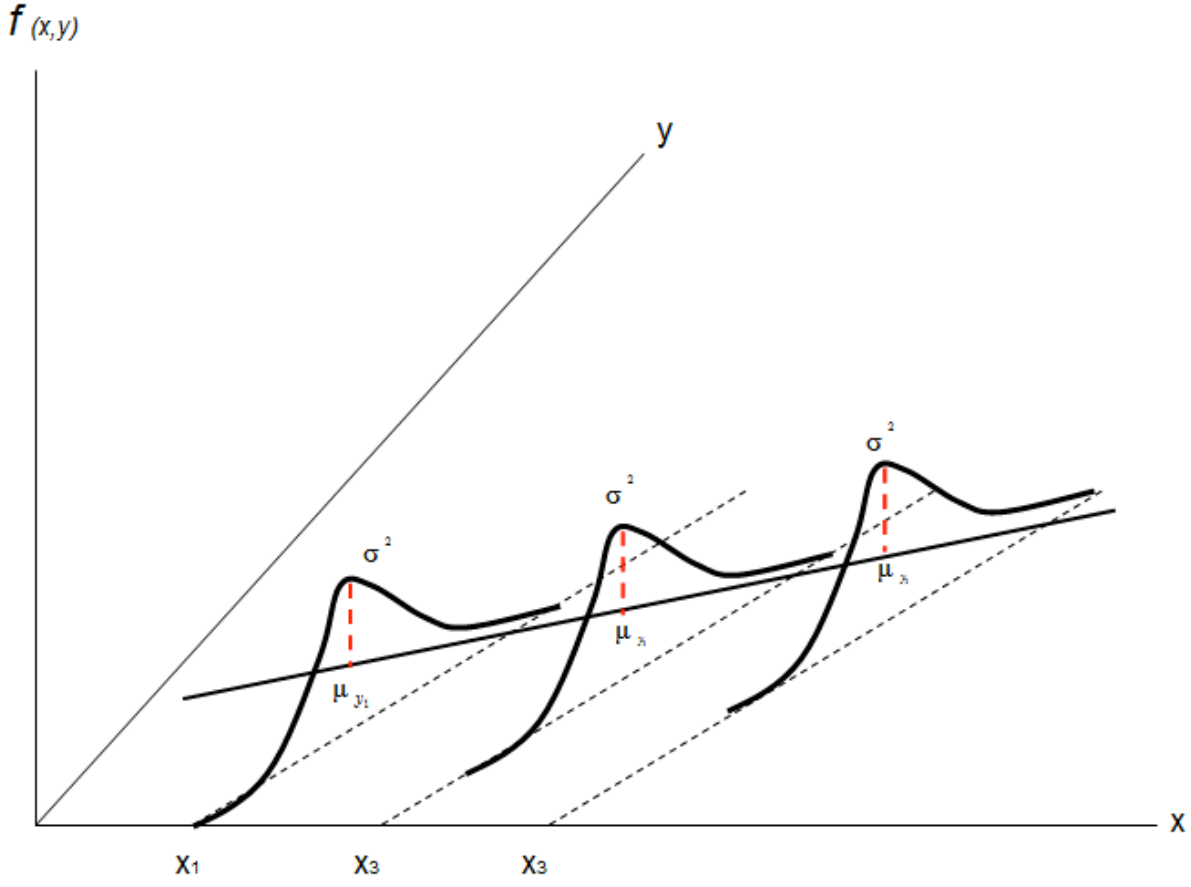
$$\mathbf{y'=a+bx}$$

U standardizovanoj formi je:

zy′=rzx

Dakle, standardizovana predviđena vrednost kriterijumaske varijable odgovara multiplikaciji standardizovane prediktorske varijable i koeficijenta korelacije. Važno je primetiti da u ovoj situaciji nemamo **intercept**, zato što su varijable standardizovane te je aritmetička sredina obe varijable = 0, tj. ako je vrednost prediktorske varijable x=0 onda je i vredost kriterijumske varijable y=0.

Da bi valjano interpretirali regresionu jednačinu u istraživanju, sledeće pretpostavke moraju biti ispunjene. Za svaku vrednost prediktorske varijable **x** postoji verovatna distribucija nezavisnih vrednosti kriterijumske varijable **y**. Varijansa **y** distribucija su 'jednake', uslov koji se zove *homoscedasticity*. Aritmetičke sredine **y** distribucija su na regresionoj liniji $y = \alpha + \beta x$ gde je μ_y aritmetička sredina **y** distribucije za datu vrednost prediktorske varijable **x**, β (beta) je slope same linije dok je α (alpha) y – intercept. To se grafički može prikazati na sledeći način:



Dakle, za svaku vrednost prediktorske varijable **x**, vrednosti kriterijumske varijable **y** variraju 'slučajno' (randomly), oko regresione linije. Konsekventno, svaka pojedinačna opservacija kriterijumske varijable *yi* imaće određenu devijaciju u odnosu na regresionu liniju u samoj polulaciji. Ovu devijaciju tretiramo kao 'grešku' i označavamo sa e_i. Vrednost ei može biti pozitivna ili negativna, obzirom na to da li je pojedinačna opservacija 'ispod' ili 'iznad' regresione linije. Obzirom da e_i svih

opservacija predstavljaju devijacije u odnosu na aritmetičku sredinu **y** distribucije, njihova srednja vrednost jeste = 0.

Prema tome, na osnovu gore iznešenih pretpostavki, svaka pojedina opservacija yi će se izračunavati kao:

$$\mathbf{y_i = \alpha + \beta_x + e_i}$$

To znači, da predviđena vrednost yi predstavlja ima 'stabilan deo' koji se sastoji iz α + βx, 'slučajni deo' koji se sastoji iz e_i, koja je rezultat prirodne varijacije **y** vrednosti regresione linije. Sledi da za svaku vrednost prediktorske varijable **x**, varijansa y_i vrednosti je identična varijansi e_i, i pretpostavka je da je ova varijansa ista bez obzira na vrednosti **x.** Značaj e_i je u tome što on predstavlja primarni izvor 'greške' u procesu predviđanja vrednosti na kriterijumskoj varijabli **y.**

Dakle, ako su pretpostavke na kojima počiva regresioni model ispunjene, možemo biti sigurni da će 'lest squares' metod obezbediti uzorkovanu regresionu liniju y'=a+bx, koja predstavlja procenu 'istinite' ali 'nepoznate regresione linije populacije $\mu_y = \alpha + \beta x$. Međutim, i a i b su rezultat greške merenja na osnovu uzorka, kao i svaki drugi statistik. Prema tome, ključni izvor 'greške' merenja je pokušaj predviđanja pojedinačne vrednosti y jeste 'slučajna' varijansa svakog pojedinog y_i, koja se nalazi 'negde oko' regresione linije tj. e_i za svaku vrednost y_i.

Varijacija y_i vrednosti oko regresione linije može se odrediti matematičkom formulom:

$$s_{yx} = \sqrt{\frac{\sum (y_i - y')^2}{n-2}}$$

Ova formula zapravo meri standardnu devijaciju opserviranih vrednosti **y** i predviđenih vrednosti **y'**. Razlog zašto je n-2 umesto standardnih n -1 jeste usled ćinjenice da imamo dve 'prepreke', tj. slope i y- intercept. Evo celokupnog kalkulusa na našem primeru:

ISPITANICI	PRIHOD (var x)	LIBERALIZAM (var y)	$y' = 29.68 + 0.1344\,x$	$y_i - y'$	$(y_i - y')^2$
1	340	71	75,38	-4,38	19,184
2	230	65	60,59	4,41	19,448
3	405	83	84,11	-1,11	1,232
4	325	74	73,36	0,64	0,410
5	280	67	67,31	-0,31	0,096
6	195	56	55,89	0,11	0,012
7	265	57	65,30	-8,30	68,890
8	300	78	70,00	8,00	64,000
9	350	84	76,72	7,28	52,998
10	310	65	71,34	-6,34	40,196
sum x	3000	700	700,00	0,00	266,466
mean	300	70	= 70	0,00	
std	60,92	9,83	8,19		

$$s_{yx} = \sqrt{\frac{\sum (y_i - y')^2}{n - 2}} \qquad s_{yx} = \sqrt{\frac{266.466}{8}} = 5.77$$

U regresionoj analizi važno nam je da znamo 'meru' ili 'stepen' u kome prediktorska varijabla 'objašnjava' varijansu kriterijumske varjable. Za ovu svrhu najjednostavni način jeste da koristimo koeficijent korelacije **r.** Tačnije, da procenat objašnjene varijanse **y** odgovara kvadratu koeficijenta korelacije **r.** Ova vrednost se drugačije naziva **koeficijent determinacije** i ona se jednostavno prikazuje i računa kao r^2:

$$r^2 = \frac{s_{y'}^{\,2}}{s_y^{\,2}}$$

U našem primeru:

$$r^2 = 0.833^2 = 0,69$$

Ovaj podatak naprosto govori o tome da u našem slučaju 69% varijanse kriterijumske varijable **y** može biti objašnjen prediktorskom varijablom **x**.

Obzirom na značaj koji ima 'slope' (b), veoma je važno za ovu vrednost uraditi test statističke značajnosti. Ukoliko ne postoji 'veza' između varijabli x i y, onda možemo očekivati da će 'slope' regresione jednačine biti =0. Da bi smo testirali

hipotezu Ho: β=0, moramo da izračunamo standardnu grešku uzorkovanja za slope b. Dakle, formula je:

$$s_b = \frac{s_{yx}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

U našem slučaju:

$$s_b \frac{5.77}{33.400} = 0.032$$

Za test statističke značajnosti koristimo t test:

$$t = \frac{b - \mathrm{B}}{s_b}$$ pri čemu je Df = n-2

Dakle, t je u našem slučaju:

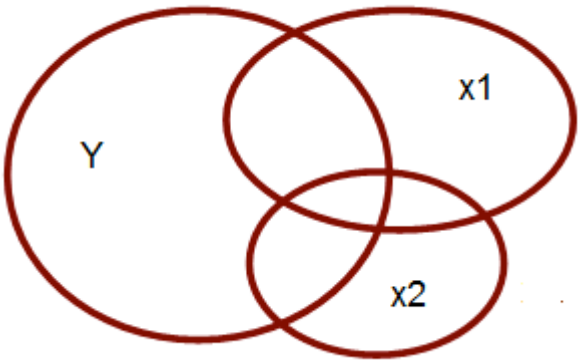
$$t = \frac{0.1344 - 0}{0.032} = 4.20$$

Višestruka regresija

Višestruka (multiple) regresija predstavlja extenziju jednostave regresije. Kalkulus i logika multiple regresije je identična kao kada je reč o jednostavnoj regresiji. Suštinska razlika je u tome, što kada u multiple regresiji, na osnovu distribucije nekoliko prediktorskih varijabli (x1, x2...xn), mi želimo da predvidimo vrednosti na kriterijumskoj varijabli y. Jednačina za multiple regresiju je prema tome:

y′ = a+b1x1+b2x2+..... +bkxk

Za razumevanje linearne regresije možemo koristiti venove dijagrame:



Evo jednog jednostanov primera, naime, na osnovu datih odgovora na dva pitanja, uradićemo priedikciju rezlutata na kolokvijumu studenata:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6,789	1,500		4,526	,000
	Pitanje 1	4,092	,861	,419	4,755	,000
	Pitanje 6	3,666	,560	,577	6,548	,000

a. Dependent Variable: SKOR

b. GRUPA = 1

Model Summary[b,c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,792[a]	,628	,613	3,48280

a. Predictors: (Constant), Pitanje 6, Pitanje 1

b. Dependent Variable: SKOR

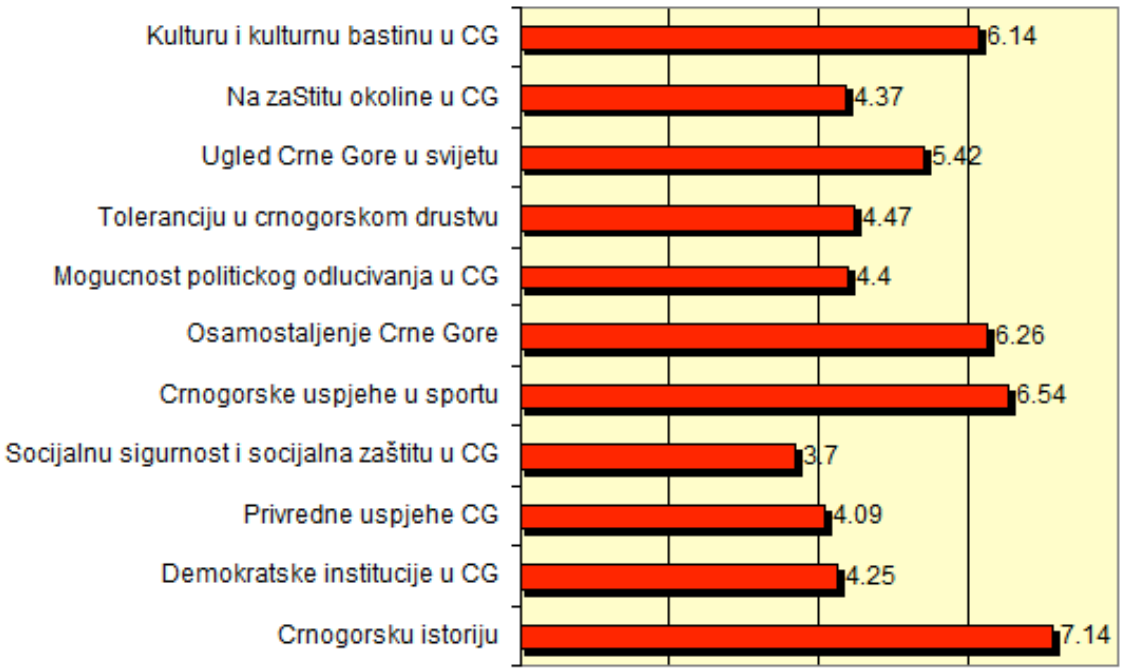
c. GRUPA = 1

U ovoj situaciji, pretpostavimo da je neki student na prvom pitanju osvojio 1,5 poena i na šestom pitanju 2 poena, onda možemo da predvidimo skor koji bi ovaj student imao na čitavom kolokvijumu:

SKOR = 6,79+ (4,09x1,5) + (3,67x2) = 6,79+ 6,14+7,34= 20,27

Evo jednog primera, naime, merili smo koliko su građani ponosni na različite aspekte crnogorskog društva. Evo distribucije ponosa na svakom ajtemu:

Grafikon 7 PONOSAN-A NA: - Srednja vrijednost na ajtemima



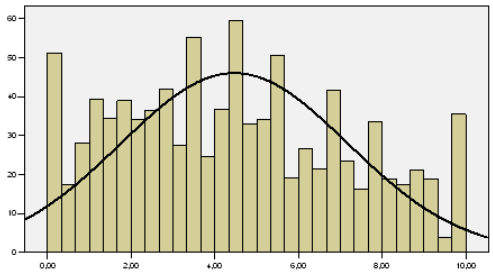
Cronbachc's Alpha = 0,89

Dakle, dobili smo novu varijablu koja meri 'pnos' i koja ima sledeće karakteristike:

Tabela 2 PONOSNI NA CRNU GORU – Kompozitni skor

N	940
Aritmetička sredina	4,4644
Standardna greška aritmetičke sredine	,08861
Medijana	4,3434
Modus	,00
Standardna Devijacija	2,71713
Iskrivljenost distribucije	,254
Spljoštenost distribucije	-,861
Minimum (bez povjerenja)	,00
Maximum (maksimalno povjerenje)	10,00

Grafikon 8 PONOSNI NA CRNU GORU



Evo varijabli i rezultata regresione analize u SPSS formatu:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,128	,484		8,537	,000
	Materjalni_status	,020	,005	,135	4,433	,000
	Crnogorac	,562	,242	,103	2,319	,021
	Srbin	-2,393	,249	-,413	-9,629	,000
	Musliman	,999	,333	,107	2,997	,003
	Sever	,733	,186	,127	3,937	,000
	Primorje	,766	,196	,119	3,918	,000
	Ukupan broj zavrsenih godina skolovanja	-,071	,032	-,065	-2,220	,027
	Zaposlen_stalni_ radni_odnos	,518	,168	,092	3,082	,002
	Penzioner_ka	,704	,236	,087	2,986	,003

a. Dependent Variable: PONOSNI NA CRNU GORU

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,562[a]	,316	,309	2,26498	,316	46,907	9	913	,000	.[b]

a. Predictors: (Constant), Penzioner_ka, Sever, Srbin, Ukupan broj zavrsenih godina skolovanja, Zaposlen_stalni_radni_odnos, Primorje, Musliman, Materjalni_status, Crnogorac

b. Not computed because fractional case weights have been found for the variable specified on the WEIGHT command.

c. Dependent Variable: PONOSNI NA CRNU GORU

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2165,742	9	240,638	46,907	,000[a]
	Residual	4687,919	914	5,130		
	Total	6853,661	923			

a. Predictors: (Constant), Penzioner_ka, Sever, Srbin, Ukupan broj zavrsenih godina skolovanja, Zaposlen_stalni_radni_odnos, Primorje, Musliman, Materjalni_status, Crnogorac

b. Dependent Variable: PONOSNI NA CRNU GORU

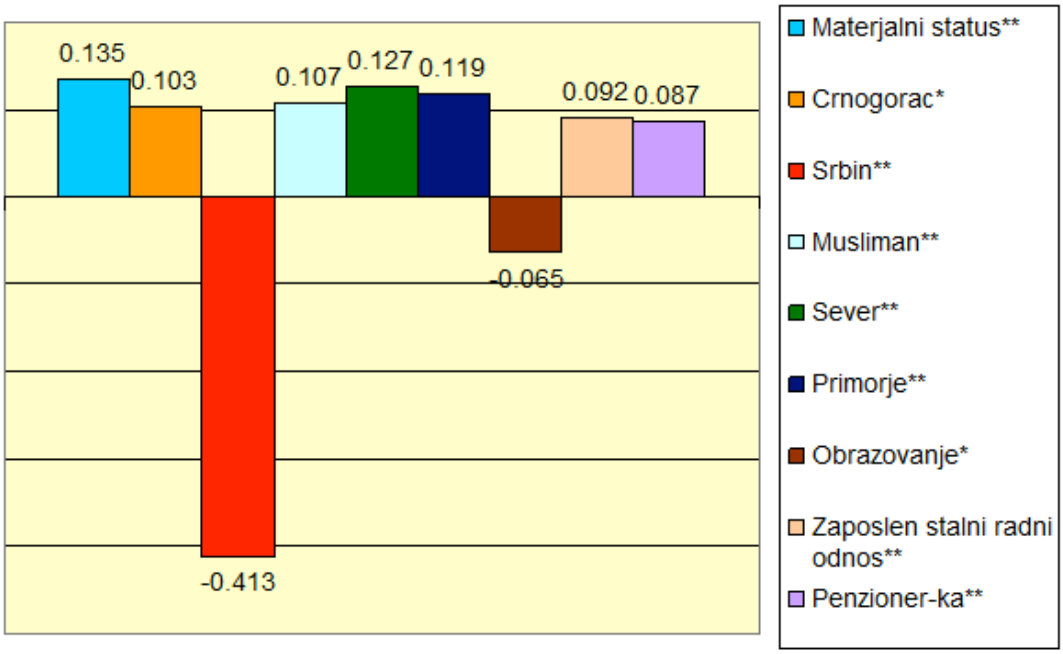
Uobičajen način da se prikazuju rezultati bio bi na našem primeru sledeći:

	B
Konstanta (y-intercept)	4,128**
Materjalni status	,020**
Crnogorac	,562*
Srbin	-2,393**
Musliman	,999**
Sever	,733**
Primorje	,766**
Ukupan broj zavrsenih godina skolovanja	-,071*
Zaposlen stalni radni odnos	,518**
Penzioner/ka	,704**

** p < 0.01
* p < 0.05

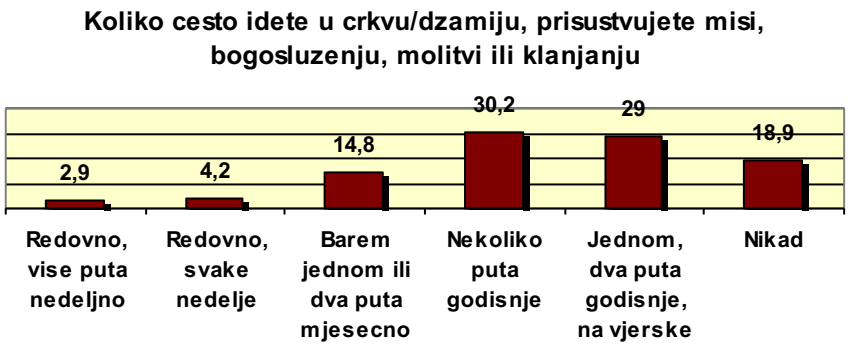
Grafički, to se može prikazati na sledeći način:

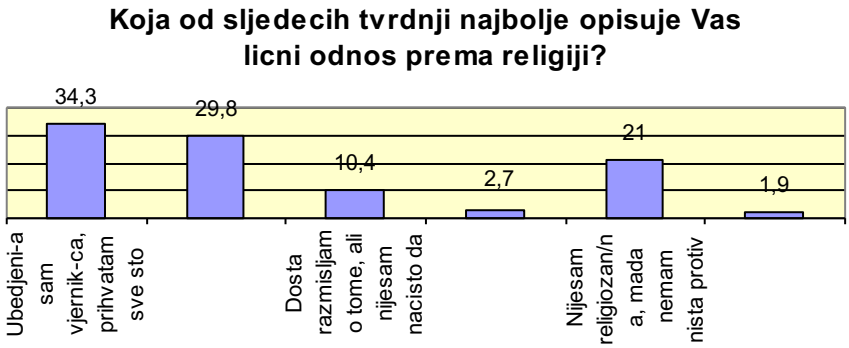
Grafikon 9 OLS – Faktori koji određuju osjećaj ponosa na Crnu Goru

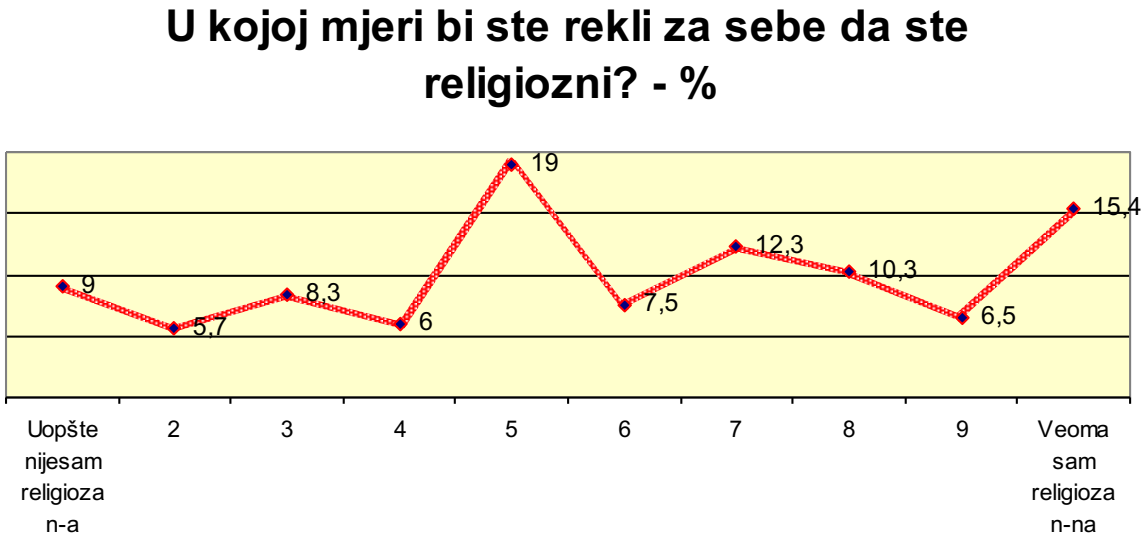


** p < 0,01 * p < 0,05
B (Konstanta) = 4,13 F = 46,9 (df,9) p < 0,01 $R^2 = 0{,}32$

Evo još jednog primera koji ima za cilj da regresionom analizom utvrdi prediktore religioznosti. Najpre, varijable i distribucija vrednosti koje se koriste za merenje religioznosti:



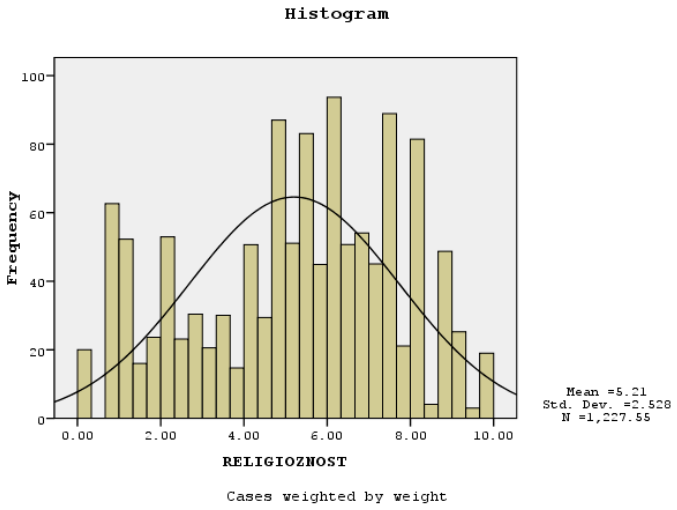




Na osnovu ove tri varijable formirali smo novu koja meri religioznost. Evo distribucije i karakteristika novoformirane varijable:

Descriptives

			Statistic	Std. Error
RELIGIOZNOST	Mean		5,2075	,07215
	95% Confidence Interval for Mean	Lower Bound	5,0659	
		Upper Bound	5,3490	
	5% Trimmed Mean		5,2375	
	Median		5,4815	
	Variance		6,389	
	Std. Deviation		2,52774	
	Minimum		,00	
	Maximum		10,00	
	Range		10,00	
	Interquartile Range		3,85	
	Skewness		-,334	,070
	Kurtosis		-,826	,140



Slede rezlutati regresione analize u SPSS-u:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6,646	,495		13,431	,000
	Godine ispitanika	-,023	,005	-,131	-4,660	,000
	Crnogorac	-,398	,195	-,078	-2,039	,042
	Srbin	1,055	,202	,196	5,232	,000
	Bosnjak	,908	,372	,073	2,443	,015
	Broj clanova domacinstva	,104	,039	,073	2,657	,008
	Ukupan broj zavrsenih godina skolovanja	-,081	,029	-,081	-2,834	,005
	Zaposlen_stalni_radni_odnos	-,481	,147	-,092	-3,280	,001

a. Dependent Variable: RELIGIOZNOST

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	894,407	7	127,772	22,632	,000[a]
	Residual	6677,592	1183	5,646		
	Total	7571,998	1190			

a. Predictors: (Constant), Zaposlen_stalni_radni_odnos, Broj clanova domacinstva, Godine ispitanika, Srbin, Bosnjak, Ukupan broj zavrsenih godina skolovanja, Crnogorac

b. Dependent Variable: RELIGIOZNOST

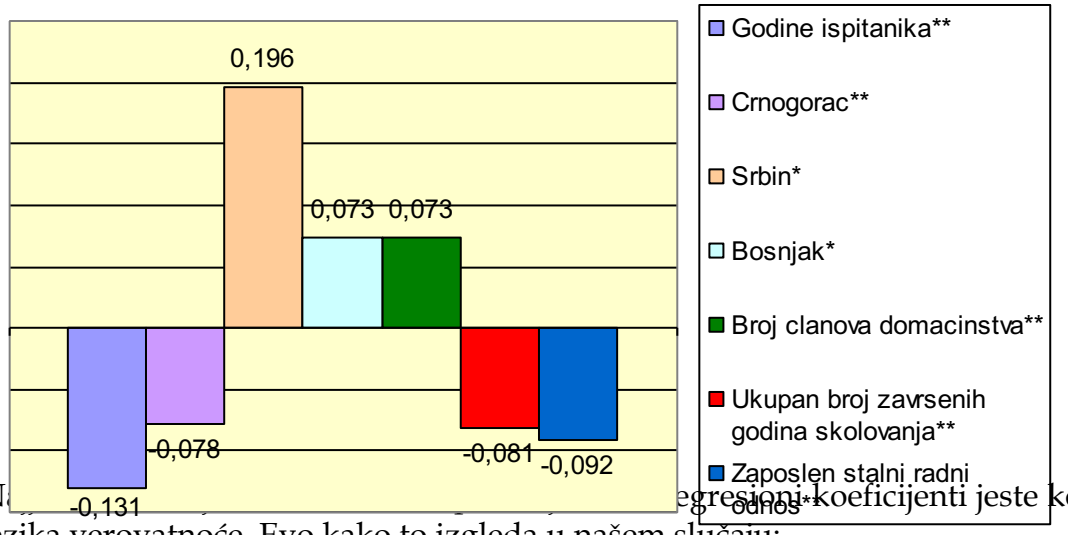
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,344[a]	,118	,113	2,37604

a. Predictors: (Constant), Zaposlen_stalni_radni_odnos, Broj clanova domacinstva, Godine ispitanika, Srbin, Bosnjak, Ukupan broj zavrsenih godina skolovanja, Crnogorac

Grafički se rezultati ove regresione analize mogu prikazati na sledeći način:

Grafikon 45 OLS - Prediktori religioznosti



Na[illegible]egresioni koeficijenti jeste korišćenje jezika verovatnoće. Evo kako to izgleda u našem slučaju:

- Što su ispitanici **stariji** to je i manja vjerovatnoća da će biti religiozni
- Ukoliko se nacionalno izjašnjavaju kao **Crnogorci** veća je vjerovatnoća da neće biti religiozni
- Ukoliko se nacionalno izjašnjavaju kao **Srbi**, veća je vjerovatnoća da će biti religiozni
- Ukoliko se nacionalno izjašnjavaju kao **Bošnjaci**, veća je vjerovatnoća da će biti religiozni
- Što je veći **broj članova domaćinstva**, veća je i vjerovatnoća da će biti religiozni
- Što je veći **stepen obrazovanja**, veća je i vjerovatnoća da neće biti religiozni
- Ukoliko su građani **zaposleni sa stalnim radnim odnosom**, veća je i vjerovatnoća da neće biti religiozni

Regresiona analiza - Formiranje modela

Multiple regresija operiše sa jednom kriterijumskom i većim brojem prediktorskih varijabli. Osnovno pitanje koje se postavlja jeste: 'koje će varijable (od velikog broja varijabli) biti uzete kao mogući prediktori kriterijumske varijable'? Proces u kome se identifikuju prediktorske varijable na osnovu kojih će se procenjivati vrednosti kriterijumske varijable, naziva se 'izgradnaj modela' (model buliding process).

Cilj izgradnje modela jeste da se formira 'dobar' model. Karakteristike 'dobrog' modela su:

- Relativno mali broj varijabli (jednocifren broj)
- Što veći procenat objašnjene varijanse kriterijumske varijable (minimum dvocifren broj)
- Što manji procenat varijanse 'reziduala'
- Veći procenat varijable objašnjene modelom u odnosu na procenat varijanse koji nije objašnjen modelom (minimum dvostruko veći)
- Što manji stepen multikolinearnosti između prediktorskih varijabli
- Što manja razlika između predviđenih i opserviranih vrednosti kriterijumske varijable

Evo jednog primera, najpre, da formiramo skor na osnovu varijabla koje mere poverenje u institucije. Evo distribucije na svim pojedinačnim varijablama:

Statistics

		Skupstinu Crne Gore	Predsjednika Crne Gore	Vladu Crne Gore	Policiju Crne Gore	Sudstvo Crne Gore	Politicke partije u Crnoj Gori
N	Valid	909	920	916	920	894	880
	Missing	104	93	98	93	119	133
Mean		2,71	2,97	2,84	2,70	2,54	2,38
Std. Error of Mean		,040	,045	,045	,043	,041	,037
Median		3,00	3,00	3,00	3,00	3,00	2,00
Std. Deviation		1,212	1,367	1,362	1,294	1,224	1,110
Variance		1,470	1,869	1,855	1,674	1,498	1,233
Skewness		,082	-,070	,012	,149	,281	,306
Std. Error of Skewness		,081	,081	,081	,081	,082	,082
Kurtosis		-1,005	-1,224	-1,251	-1,125	-,944	-,796
Std. Error of Kurtosis		,162	,161	,161	,161	,163	,165
Minimum		1	1	1	1	1	1
Maximum		5	5	5	5	5	5
Percentiles	25	2,00	2,00	2,00	2,00	1,00	1,00
	50	3,00	3,00	3,00	3,00	3,00	2,00
	75	4,00	4,00	4,00	4,00	3,00	3,00

Na osnovu korelacione matrice možemo videti sledeće:

Correlations

		Skupstinu Crne Gore	Predsjednika Crne Gore	Vladu Crne Gore	Policiju Crne Gore	Sudstvo Crne Gore	Politicke partije u Crnoj Gori
Skupstinu Crne Gore	Pearson Correlation	1	,794**	,828**	,658**	,636**	,627**
	Sig. (2-tailed)		,000	,000	,000	,000	,000
	N	909	901	894	888	867	858
Predsjednika Crne Gore	Pearson Correlation	,794**	1	,879**	,627**	,618**	,560**
	Sig. (2-tailed)	,000		,000	,000	,000	,000
	N	901	920	905	899	878	866
Vladu Crne Gore	Pearson Correlation	,828**	,879**	1	,676**	,648**	,567**
	Sig. (2-tailed)	,000	,000		,000	,000	,000
	N	894	905	916	901	875	865
Policiju Crne Gore	Pearson Correlation	,658**	,627**	,676**	1	,775**	,574**
	Sig. (2-tailed)	,000	,000	,000		,000	,000
	N	888	899	901	920	887	866
Sudstvo Crne Gore	Pearson Correlation	,636**	,618**	,648**	,775**	1	,615**
	Sig. (2-tailed)	,000	,000	,000	,000		,000
	N	867	878	875	887	894	853
Politicke partije u Crnoj Gori	Pearson Correlation	,627**	,560**	,567**	,574**	,615**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	
	N	858	866	865	866	853	880

**. Correlation is significant at the 0.01 level (2-tailed).

Reliability Statistics

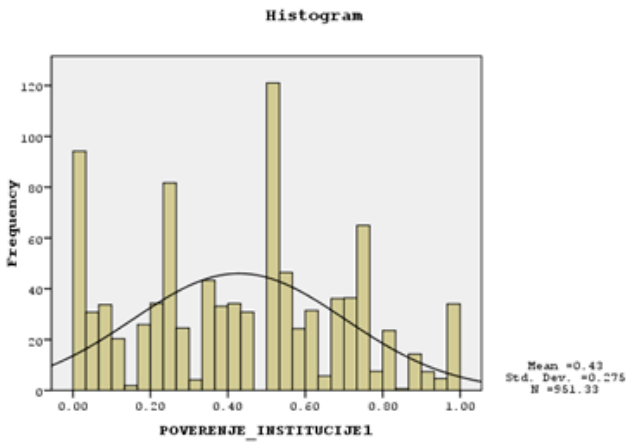
Cronbach's Alpha	N of Items
,930	6

Novoformirana varijabala koja meri povenje u institucije ima sledeće karakteristike:

Statistics

POVERENJE_INSTITUCIJE1

N	Valid	951
	Missing	62
Mean		,4289
Std. Error of Mean		,00891
Median		,4583
Std. Deviation		,27489
Variance		,076
Skewness		,120
Std. Error of Skewness		,079
Kurtosis		-,843
Std. Error of Kurtosis		,158
Minimum		,00
Maximum		1,00
Percentiles	25	,2083
	50	,4583
	75	,6250



U izgradnji modela[1] dakle, mi imamo za cilj da identifikujemo najbolje moguće **prediktorske** varijable, pri čemu je **kriterijumska** varijabla 'poverenje u institucije'. Dakle, ključno je da identifikujemo kriterijumske varijable. Da bi u model uvrstili 'najbolje moguće' kriterijumske varijable, moramo se rukovoditi određenim kriterijumima za izbor varijabli. Postoje tri ključna kriterijuma, **prvo**, moguće je uvrstiti samo one varijable koje imamo u našem datasetu, i ovo je jedno od najozbiljnijih ograničenja. **Drugo**, to je teorijski kriterijum, naime, na osnovu teorijske spekulacije, mi možemo pretpostaviti koje varijable mogu biti najbolji prediktori, i **treće,** modžemo koristiti statističke kriterijume, i kreirati model na osnovu ovih kriterijuma putem pokušaja i pogrešaka.

Vratimo se na naš primer, pretpostavimo da poverenje u institucije najbolje možemo da predvidimo demogrfskim varijablama datim u sledećoj tabeli:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4286	,27511	937
Pol	1,51	,500	937
Starost	43,57	16,106	937
Ukupan broj zavrsenih godina skolovanja	13,80	11,094	937
Prihodi domacinstva - mesecno	10,61	5,290	937

Sledi model koji smo dobili na osnovu ovih varijabli:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,043	4	,261	3,480	,008[a]
	Residual	69,774	932	,075		
	Total	70,817	936			

a. Predictors: (Constant), Prihodi domacinstva - mesecno, Ukupan broj zavrsenih godina skolovanja, Pol, Starost

b. Dependent Variable: POVERENJE_INSTITUCIJE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,300	,047		6,422	,000	,209	,392
	Pol	,036	,018	,066	2,014	,044	,001	,071
	Starost	,001	,001	,085	2,519	,012	,000	,003
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,052	-1,610	,108	-,003	,000
	Prihodi domacinstva - mesecno	,003	,002	,052	1,552	,121	-,001	,006

a. Dependent Variable: POVERENJE_INSTITUCIJE

Na osnovu ovog inicijalnog modela, jednostavno možemo uočiti da je suma kvadrata reyiduala mnogo veća od sume kvadrata reziduala, dakle, model ne objašnjava zadovoljavajući procenat varijanse, ili tačnije, najveći deo varijanse kriterijumske varijable ostaće 'neobjašnjen'. Jednako, primećujemo da od četiri demografske kriterijumske varijable, obrazovanje i prihodi su sasvim jasno neadekvatne usled testa statističke značajnosti, dok je i pol kao varijabla na granici. Prema tome moramo

[1] Svi modeli su dati u autentičnom formatu kako to pruža SPSS

se kretati ka boljem modelu. Ukoliko svim ovim varijablama dodamo nacionalnu pripadnost ispitanika imaćemo sledeći model:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4286	,27511	937
Pol	1,51	,500	937
Starost	43,57	16,106	937
Ukupan broj zavrsenih godina skolovanja	13,80	11,094	937
Prihodi domacinstva - mesecno	10,61	5,290	937
Crnogorac	,4356	,49610	937
Srbin	,3212	,46717	937
Bosnjak_Muslima	,0389	,19357	937
Albanac	,0570	,23195	937

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,501[a]	,251	,245	,23910

a. Predictors: (Constant), Albanac, Prihodi domacinstva - mesecno, Ukupan broj zavrsenih godina skolovanja, Bosnjak_Muslima, Pol, Srbin, Starost, Crnogorac

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17,783	8	2,223	38,881	,000[a]
	Residual	53,034	928	,057		
	Total	70,817	936			

a. Predictors: (Constant), Albanac, Prihodi domacinstva - mesecno, Ukupan broj zavrsenih godina skolovanja, Bosnjak_Muslima, Pol, Srbin, Starost, Crnogorac

b. Dependent Variable: POVERENJE_INSTITUCIJE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,382	,045		8,507	,000	,294	,470
	Pol	,030	,016	,055	1,938	,053	,000	,061
	Starost	,002	,001	,107	3,642	,000	,001	,003
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,058	-2,030	,043	-,003	,000
	Prihodi domacinstva - mesecno	,001	,002	,022	,724	,469	-,002	,004
	Crnogorac	,028	,024	,051	1,188	,235	-,018	,075
	Srbin	-,265	,025	-,451	-10,756	,000	-,314	-,217
	Bosnjak_Muslima	,058	,045	,041	1,300	,194	-,030	,146
	Albanac	-,012	,039	-,010	-,305	,760	-,088	,064

a. Dependent Variable: POVERENJE_INSTITUCIJE

Prvo što se u ovom modelu može uočiti jeste činjenica da istim možemo objasniti 25,1% varijanse kriterijumske varijable (vrednost r^2) što je sasvim zadovoljavajuće. Međutim, i dalje je suma kvadrata reziduala veća od sume kvadrata modela. Ovo je razlog zbog koga je uputno da involviramo još neke moguće prediktorske varijable u model. Budući da je poverenje u institucije u mogućoj vezi sa partijskim preferencijama, uključili smo u model i varijable koje mere partijski preferencijal. Evo modela:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4286	,27511	937
Pol	1,51	,500	937
Starost	43,57	16,106	937
Ukupan broj zavrsenih godina skolovanja	13,80	11,094	937
Prihodi domacinstva - mesecno	10,61	5,290	937
Crnogorac	,4356	,49610	937
Srbin	,3212	,46717	937
Bosnjak_Muslima	,0389	,19357	937
Albanac	,0570	,23195	937
DPS	,3343	,47199	937
SDP	,0323	,17685	937
PZP	,0629	,24298	937
SNS	,1045	,30609	937
SNP	,0834	,27662	937
Ostale_Srpske_partije	,0383	,19196	937
Manjinske_partije	,0174	,13068	937
Apstinenti	,2962	,45684	937

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,656[a]	,431	,421	,20938

a. Predictors: (Constant), Apstinenti, Pol, Crnogorac, Ukupan broj zavrsenih godina skolovanja, Manjinske_ partije, SDP, Starost, Ostale_Srpske_partije, PZP, Prihodi domacinstva - mesecno, Albanac, SNP, Bosnjak_Muslima, SNS, Srbin, DPS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	30,497	16	1,906	43,476	,000[a]
	Residual	40,319	920	,044		
	Total	70,817	936			

a. Predictors: (Constant), Apstinenti, Pol, Crnogorac, Ukupan broj zavrsenih godina skolovanja, Manjinske_partije, SDP, Starost, Ostale_Srpske_partije, PZP, Prihodi domacinstva - mesecno, Albanac, SNP, Bosnjak_Muslima, SNS, Srbin, DPS

b. Dependent Variable: POVERENJE_INSTITUCIJE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,311	,056		5,588	,000	,202	,420
	Pol	,020	,014	,036	1,435	,152	-,007	,047
	Starost	,002	,000	,098	3,723	,000	,001	,003
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,045	-1,781	,075	-,002	,000
	Prihodi domacinstva - mesecno	,001	,001	,020	,776	,438	-,002	,004
	Crnogorac	,011	,021	,020	,539	,590	-,030	,052
	Srbin	-,112	,026	-,191	-4,365	,000	-,163	-,062
	Bosnjak_Muslima	,013	,040	,009	,327	,744	-,066	,092
	Albanac	-,045	,035	-,038	-1,284	,199	-,113	,024
	DPS	,242	,041	,415	5,881	,000	,161	,323
	SDP	,134	,056	,086	2,381	,017	,024	,245
	PZP	-,057	,048	-,050	-1,187	,236	-,151	,037
	SNS	-,103	,047	-,115	-2,189	,029	-,196	-,011
	SNP	-,094	,048	-,094	-1,960	,050	-,187	,000
	Ostale_Srpske_partije	-,065	,054	-,046	-1,209	,227	-,172	,041
	Manjinske_partije	,059	,067	,028	,883	,377	-,072	,190
	Apstinenti	-,030	,041	-,050	-,731	,465	-,111	,051

a. Dependent Variable: POVERENJE_INSTITUCIJE

U ovom modelu odmah možemo uočiti da je procenat varijase kriterijumske varijable objašnjen 43,1%, što je znatno bolje, a i odnos sume kvadrata i reziduala je značajno proporcionalniji. Međutim, na osnovu regresionih koeficijenata varijabli te t- testa, možemo videti da veliki broj varijabli u ovom modelu nije statistički značajan.

Kako prema tome, metodički možemo izgraditi model. Opet, postoje tri uobičajena pristupa:

1. Putem pokušaja i pogrešaka, na način da sami istražujete najbolju moguću kombinaciju od svih varijabli za kloje se pretpostavlja da su prediktori
2. Stepwise: putem kretanja od jedne, pa dodavanjem kriterijumskih varijabli
3. Backward: putem kretanja od svih, a onda izbacivanjem varijabli

U sva tri slučaja kao kriterijum se uzima:

1. Odnos sume kvadrata između modela i reziduala
2. Procenat objašnjene varijanse
3. T-test, i njegova statistička značajnost za procenu statističke značajnosti svake pojedine varijable

Dakle, nastavljamo sa modelom. Najpre ćemo dodati još nekoliko varijabli, i to one koje operacionalizuju odnos ispitanika prema političarima. Evo modela:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4034	,27165	748
Pol	1,48	,500	748
Starost	43,60	15,861	748
Ukupan broj zavrsenih godina skolovanja	13,85	11,246	748
Prihodi domacinstva - mesecno	10,78	5,316	748
Crnogorac	,4349	,49607	748
Srbin	,3404	,47415	748
Bosnjak_Muslima	,0418	,20026	748
Albanac	,0553	,22874	748
DPS	,3305	,47071	748
SDP	,0361	,18674	748
PZP	,0588	,23539	748
SNS	,1195	,32461	748
SNP	,0933	,29103	748
Ostale_Srpske_partije	,0359	,18618	748
Manjinske_partije	,0142	,11827	748
Apstinenti	,2836	,45104	748
Filip VUJANOVIC	2,95	1,475	748
Ranko KRIVOKAPIC	2,27	1,411	748
Milo DJUKANOVIC	2,84	1,622	748
Andrija MANDIC	2,40	1,421	748
Srdjan MILIC	2,64	1,417	748
Nebojsa MEDOJEVIC	2,24	1,229	748

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,813[a]	,662	,651	,16040

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupan broj zavrsenih godina skolovanja, Milo DJUKANOVIC, Starost, Pol, Manjinske_partije, SDP, Ostale_Srpske_ partije, Prihodi domacinstva - mesecno, Albanac, Apstinenti, Bosnjak_Muslima, PZP, SNP, Crnogorac, Andrija MANDIC, Ranko KRIVOKAPIC, SNS, Srdjan MILIC, Filip VUJANOVIC, Srbin, DPS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36,475	22	1,658	64,441	,000[a]
	Residual	18,655	725	,026		
	Total	55,130	747			

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupan broj zavrsenih godina skolovanja, Milo DJUKANOVIC, Starost, Pol, Manjinske_partije, SDP, Ostale_ Srpske_partije, Prihodi domacinstva - mesecno, Albanac, Apstinenti, Bosnjak_ Muslima, PZP, SNP, Crnogorac, Andrija MANDIC, Ranko KRIVOKAPIC, SNS, Srdjan MILIC, Filip VUJANOVIC, Srbin, DPS

b. Dependent Variable: POVERENJE_INSTITUCIJE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,080	,054		-1,468	,143	-,186	,027
	Pol	,007	,012	,013	,611	,542	-,016	,031
	Starost	,001	,000	,058	2,525	,012	,000	,002
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,040	-1,849	,065	-,002	,000
	Prihodi domacinstva - mesecno	,003	,001	,059	2,556	,011	,001	,005
	Crnogorac	-,004	,019	-,008	-,219	,827	-,042	,034
	Srbin	-,021	,024	-,037	-,865	,387	-,068	,027
	Bosnjak_Muslima	,009	,034	,006	,249	,803	-,059	,076
	Albanac	,007	,031	,006	,222	,825	-,054	,068
	DPS	,083	,038	,144	2,160	,031	,008	,158
	SDP	-,017	,049	-,012	-,349	,727	-,114	,080
	PZP	-,001	,045	-,001	-,014	,989	-,088	,087
	SNS	,001	,043	,001	,012	,990	-,084	,085
	SNP	-,024	,044	-,026	-,553	,581	-,111	,062
	Ostale_Srpske_partije	-,021	,049	-,015	-,431	,666	-,118	,075
	Manjinske_partije	,034	,063	,015	,539	,590	-,089	,157
	Apstinenti	,005	,037	,008	,135	,892	-,068	,078
	Filip VUJANOVIC	,055	,008	,296	7,136	,000	,040	,070
	Ranko KRIVOKAPIC	,048	,007	,250	6,742	,000	,034	,062
	Milo DJUKANOVIC	,036	,007	,214	5,094	,000	,022	,050
	Andrija MANDIC	-,004	,007	-,019	-,526	,599	-,017	,010
	Srdjan MILIC	,015	,007	,078	2,070	,039	,001	,029
	Nebojsa MEDOJEVIC	-,005	,006	-,021	-,793	,428	-,016	,007

a. Dependent Variable: POVERENJE_INSTITUCIJE

Ovaj model izgleda kao sasvim solidna početna osnova. Naime, njime možemo objasniti preko 66% kriterijumske varijable, a i suma kvadrata modela je gotovo dvostruko veća od sume kvadrata reziduala. Međutim, analizirajući pojedinačne regresione koeficijente i rezultate t testa, sasvim je jasno da neke varijable ne doprinose modelu. Prema tome, u daljem modeliranju isključićemo u nekoliko iteracija one varijable koje ne daju doprinos predikciji i nisu značajne na osnovu statističkih kriterijuma. Evo identifikovanih varijabli koje ćemo izbaciti iz modela u prvoj iteraciji:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,080	,054		-1,468	,143	-,186	,027
	Pol	,007	,012	,013	,611	,542	-,016	,031
	Starost	,001	,000	,058	2,525	,012	,000	,002
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,040	-1,849	,065	-,002	,000
	Prihodi domacinstva - mesecno	,003	,001	,059	2,556	,011	,001	,005
	Crnogorac	-,004	,019	-,008	-,219	,827	-,042	,034
	Srbin	-,021	,024	-,037	-,865	,387	-,068	,027
	Bosnjak_Muslima	,009	,034	,006	,249	,803	-,059	,076
	Albanac	,007	,031	,006	,222	,825	-,054	,068
	DPS	,083	,038	,144	2,160	,031	,008	,158
	SDP	-,017	,049	-,012	-,349	,727	-,114	,080
	PZP	**-,001**	**,045**	**-,001**	**-,014**	**,989**	**-,088**	**,087**
	SNS	**,001**	**,043**	**,001**	**,012**	**,990**	**-,084**	**,085**
	SNP	-,024	,044	-,026	-,553	,581	-,111	,062
	Ostale_Srpske_partije	-,021	,049	-,015	-,431	,666	-,118	,075
	Manjinske_partije	,034	,063	,015	,539	,590	-,089	,157
	Apstinenti	,005	,037	,008	,135	,892	-,068	,078
	Filip VUJANOVIC	,055	,008	,296	7,136	,000	,040	,070
	Ranko KRIVOKAPIC	,048	,007	,250	6,742	,000	,034	,062
	Milo DJUKANOVIC	,036	,007	,214	5,094	,000	,022	,050
	Andrija MANDIC	-,004	,007	-,019	-,526	,599	-,017	,010
	Srdjan MILIC	,015	,007	,078	2,070	,039	,001	,029
	Nebojsa MEDOJEVIC	-,005	,006	-,021	-,793	,428	-,016	,007

a. Dependent Variable: POVERENJE_INSTITUCIJE

Napomena: vrijable koje izbacujemo iz modela označene su crvenom bojom u tabeli

Dakle, kada je reč o pristalicama PZP-a i pristalicama SNS-a, ove dve varijable imaju jako malu vrednost regresionog koeficijenta u modela kao i jako malu vrednost t testa i nedvosmisleno nemaju statističku značajnost. Zato prvo njih izbacijuemo i ponavljamo regresionu proceduru. Evo rezultata:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4034	,27165	748
Pol	1,48	,500	748
Starost	43,60	15,861	748
Ukupan broj zavrsenih godina skolovanja	13,85	11,246	748
Prihodi domacinstva - mesecno	10,78	5,316	748
Crnogorac	,4349	,49607	748
Srbin	,3404	,47415	748
Bosnjak_Muslima	,0418	,20026	748
Albanac	,0553	,22874	748
DPS	,3305	,47071	748
SDP	,0361	,18674	748
SNP	,0933	,29103	748
Ostale_Srpske_partije	,0359	,18618	748
Manjinske_partije	,0142	,11827	748
Apstinenti	,2836	,45104	748
Filip VUJANOVIC	2,95	1,475	748
Ranko KRIVOKAPIC	2,27	1,411	748
Milo DJUKANOVIC	2,84	1,622	748
Andrija MANDIC	2,40	1,421	748
Srdjan MILIC	2,64	1,417	748
Nebojsa MEDOJEVIC	2,24	1,229	748

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,813[a]	,662	,652	,16018

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupan broj zavrsenih godina skolovanja, Milo DJUKANOVIC, Starost, Pol, Manjinske_partije, SDP, Ostale_Srpske_ partije, Prihodi domacinstva - mesecno, Albanac, Apstinenti, Bosnjak_Muslima, SNP, Crnogorac, Andrija MANDIC, Ranko KRIVOKAPIC, Srdjan MILIC, Srbin, Filip VUJANOVIC, DPS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36,475	20	1,824	71,081	,000[a]
	Residual	18,655	727	,026		
	Total	55,130	747			

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupan broj zavrsenih godina skolovanja, Milo DJUKANOVIC, Starost, Pol, Manjinske_partije, SDP, Ostale_ Srpske_partije, Prihodi domacinstva - mesecno, Albanac, Apstinenti, Bosnjak_ Muslima, SNP, Crnogorac, Andrija MANDIC, Ranko KRIVOKAPIC, Srdjan MIL Srbin, Filip VUJANOVIC, DPS

b. Dependent Variable: POVERENJE_INSTITUCIJE

Šta uočavamo? Jednostavno, u modelu imamo dve varijable manje, a da time nimalo nismo narušili kvalitet modela, tačnije, i suma kvadrata moedela u odnosu na reziduale je ostala gotovo identična, i procenat objašnjene varijanse kriterijumske varijable se nije smanjio. Nastavljamo modeliranje po istom principu:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,080	,044		-1,810	,071	-,166	,007
	Pol	,007	,012	,013	,611	,542	-,016	,031
	Starost	,001	,000	,058	2,538	,011	,000	,002
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,040	-1,855	,064	-,002	,000
	Prihodi domacinstva - mesecno	,003	,001	,059	2,566	,010	,001	,005
	Crnogorac	**-,004**	**,019**	**-,008**	**-,217**	**,828**	**-,042**	**,034**
	Srbin	-,021	,023	-,036	-,890	,374	-,066	,025
	Bosnjak_Muslima	**,009**	**,034**	**,006**	**,251**	**,802**	**-,059**	**,076**
	Albanac	**,007**	**,031**	**,006**	**,225**	**,822**	**-,054**	**,068**
	DPS	,083	,025	,144	3,349	,001	,034	,132
	SDP	-,017	,039	-,012	-,437	,662	-,095	,060
	SNP	-,024	,025	-,026	-,994	,321	-,073	,024
	Ostale_Srpske_partije	-,021	,034	-,015	-,627	,531	-,088	,045
	Manjinske_partije	,034	,055	,015	,609	,543	-,075	,142
	Apstinenti	,005	,019	,008	,262	,793	-,032	,042
	Filip VUJANOVIC	,055	,008	,296	7,194	,000	,040	,069
	Ranko KRIVOKAPIC	,048	,007	,250	6,792	,000	,034	,062
	Milo DJUKANOVIC	,036	,007	,214	5,125	,000	,022	,050
	Andrija MANDIC	-,004	,007	-,019	-,532	,595	-,017	,010
	Srdjan MILIC	,015	,007	,078	2,077	,038	,001	,029
	Nebojsa MEDOJEVIC	-,005	,006	-,022	-,843	,399	-,016	,006

a. Dependent Variable: POVERENJE_INSTITUCIJE

U sledećoj iteraciji, sasvim je opravdano da iz modela izbacimo varijable Crnogorac, Bošljak_Musliman i Albanac. Nakon izbacivanja ovih varijabli ponavljamo proceduru i možemo videti rezultate testiranja:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_INSTITUCIJE	,4034	,27165	748
Pol	1,48	,500	748
Starost	43,60	15,861	748
Ukupan broj zavrsenih godina skolovanja	13,85	11,246	748
Prihodi domacinstva - mesecno	10,78	5,316	748
Srbin	,3404	,47415	748
DPS	,3305	,47071	748
SDP	,0361	,18674	748
SNP	,0933	,29103	748
Ostale_Srpske_partije	,0359	,18618	748
Manjinske_partije	,0142	,11827	748
Apstinenti	,2836	,45104	748
Filip VUJANOVIC	2,95	1,475	748
Ranko KRIVOKAPIC	2,27	1,411	748
Milo DJUKANOVIC	2,84	1,622	748
Andrija MANDIC	2,40	1,421	748
Srdjan MILIC	2,64	1,417	748
Nebojsa MEDOJEVIC	2,24	1,229	748

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,813[a]	,661	,654	,15988

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupan broj zavrsenih godina skolovanja, Milo DJUKANOVIC, Starost, Pol, Manjinske_partije, SDP, Ostale_Srpske_partije, Prihodi domacinstva - mesecno, Apstinenti, SNP, Andrija MANDIC, Srbin, Ranko KRIVOKAPIC, Srdjan MILIC, Filip VUJANOVIC, DPS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36,467	17	2,145	83,918	,000[a]
	Residual	18,663	730	,026		
	Total	55,130	747			

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupan broj zavrsenih godina skolovanja, Milo DJUKANOVIC, Starost, Pol, Manjinske_partije, SDP, Ostale_Srpske_partije, Prihodi domacinstva - mesecno, Apstinenti, SNP, Andrija MAN Srbin, Ranko KRIVOKAPIC, Srdjan MILIC, Filip VUJANOVIC, DPS

b. Dependent Variable: POVERENJE_INSTITUCIJE

Dakle, izbacene su iz modela tri varijable, a sve ključne karakteristike 'dobrog' modela su i dalje očuvane. Prema tome, idemo dalje u redukciju identifikujući sledeće varijable koje ćemo izbaciti iz modela:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,079	,042		-1,898	,058	-,160	,003
	Pol	,007	,012	,013	,603	,547	-,016	,031
	Starost	,001	,000	,058	2,522	,012	,000	,002
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,041	-1,881	,060	-,002	,000
	Prihodi domacinstva - mesecno	,003	,001	,058	2,549	,011	,001	,005
	Srbin	-,018	,018	-,032	-1,036	,301	-,053	,016
	DPS	,083	,025	,143	3,355	,001	,034	,131
	SDP	**-,015**	**,039**	**-,010**	**-,394**	**,694**	**-,091**	**,061**
	SNP	-,024	,025	-,026	-,991	,322	-,073	,024
	Ostale_Srpske_partije	-,022	,034	-,015	-,637	,524	-,088	,045
	Manjinske_partije	,040	,053	,017	,749	,454	-,065	,145
	Apstinenti	**,005**	**,019**	**,008**	**,239**	**,811**	**-,033**	**,042**
	Filip VUJANOVIC	,054	,008	,295	7,196	,000	,040	,069
	Ranko KRIVOKAPIC	,048	,007	,251	6,898	,000	,035	,062
	Milo DJUKANOVIC	,036	,007	,214	5,146	,000	,022	,050
	Andrija MANDIC	-,004	,007	-,020	-,558	,577	-,017	,010
	Srdjan MILIC	,015	,007	,076	2,049	,041	,001	,029
	Nebojsa MEDOJEVIC	-,005	,006	-,022	-,867	,386	-,016	,006

a. Dependent Variable: POVERENJE_INSTITUCIJE

U sledećoj iteraciji izbacujemo varijable SDP i Apstinenti zato što ove dve varijable ne doprinose modelu. Nakon izbacivanja ovih varijabli iz modela moćemo videti sledeće:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4034	,27165	748
Pol	1,48	,500	748
Starost	43,60	15,861	748
Ukupan broj zavrsenih godina skolovanja	13,85	11,246	748
Prihodi domacinstva - mesecno	10,78	5,316	748
Srbin	,3404	,47415	748
DPS	,3305	,47071	748
SNP	,0933	,29103	748
Ostale_Srpske_partije	,0359	,18618	748
Manjinske_partije	,0142	,11827	748
Filip VUJANOVIC	2,95	1,475	748
Ranko KRIVOKAPIC	2,27	1,411	748
Milo DJUKANOVIC	2,84	1,622	748
Andrija MANDIC	2,40	1,421	748
Srdjan MILIC	2,64	1,417	748
Nebojsa MEDOJEVIC	2,24	1,229	748

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,813[a]	,661	,654	,15970

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupa broj zavrsenih godina skolovanja, Milo DJUKANOVIC Starost, Pol, Manjinske_partije, Ostale_Srpske_partij Prihodi domacinstva - mesecno, SNP, Andrija MANDIC, Srbin, DPS, Ranko KRIVOKAPIC, Srdjan MILIC, Filip VUJANOVIC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36,459	15	2,431	95,303	,000[a]
	Residual	18,671	732	,026		
	Total	55,130	747			

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupan broj zavrsenih godina skolovanja, Milo DJUKANOVIC, Starost, Pol, Manjinske_partije, Ostale_Srpske partije, Prihodi domacinstva - mesecno, SNP, Andrija MANDIC, Srbin, DPS, R KRIVOKAPIC, Srdjan MILIC, Filip VUJANOVIC

b. Dependent Variable: POVERENJE_INSTITUCIJE

Sasvim je očito da je izbacivanje ovih varijabli bilo opravdano, naime, iako smo ih izbacili iz modela, model je očuvao ključne karakteristike 'dobrog' modela što se, kao što znamo, može videti iz procenta objašnjene varijanse kriterijumske varijable te odnosa između sume kvadrata modela i reziduala. Prema tome nastavljamo istom procedurom dalje identifikujući sledeće varijable koje ćemo izbaciti iz modela:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,075	,038		-1,973	,049	-,149	,000
	Pol	**,008**	**,012**	**,014**	**,638**	**,524**	**-,016**	**,031**
	Starost	,001	,000	,057	2,508	,012	,000	,002
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,040	-1,866	,062	-,002	,000
	Prihodi domacinstva - mesecno	,003	,001	,059	2,580	,010	,001	,005
	Srbin	-,018	,017	-,032	-1,056	,291	-,053	,016
	DPS	,083	,018	,144	4,652	,000	,048	,118
	SNP	-,026	,023	-,028	-1,097	,273	-,072	,020
	Ostale_Srpske_partije	-,023	,033	-,016	-,711	,477	-,088	,041
	Manjinske_partije	,039	,051	,017	,771	,441	-,061	,139
	Filip VUJANOVIC	,054	,008	,295	7,211	,000	,040	,069
	Ranko KRIVOKAPIC	,047	,007	,246	7,034	,000	,034	,061
	Milo DJUKANOVIC	,036	,007	,213	5,159	,000	,022	,049
	Andrija MANDIC	**-,004**	**,007**	**-,020**	**-,584**	**,560**	**-,017**	**,009**
	Srdjan MILIC	,014	,007	,075	2,018	,044	,000	,028
	Nebojsa MEDOJEVIC	-,005	,006	-,021	-,850	,395	-,016	,006

a. Dependent Variable: POVERENJE_INSTITUCIJE

Na osnovu istih kriterijuma kojima smo se do sada rukovodili, ovog puta smo identifikovali varijablu pol i ocenjivanje Andrije Mandića kao varijable koje ne doprinose modelu te ih izbacijuemo. Evo modela bez ove dve varijable:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_INSTITUCIJE	,4053	,27123	760
Starost	43,70	15,860	760
Ukupan broj zavrsenih godina skolovanja	13,82	11,167	760
Prihodi domacinstva - mesecno	10,80	5,348	760
Srbin	,3377	,47324	760
DPS	,3309	,47086	760
SNP	,0935	,29134	760
Ostale_Srpske_partije	,0380	,19143	760
Manjinske_partije	,0153	,12264	760
Filip VUJANOVIC	2,96	1,473	760
Ranko KRIVOKAPIC	2,28	1,407	760
Milo DJUKANOVIC	2,85	1,617	760
Srdjan MILIC	2,65	1,415	760
Nebojsa MEDOJEVIC	2,25	1,229	760

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,813[a]	,660	,654	,15948

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ostale_Srpske_partije, Starost, Ukupan broj zavrsenih godina skolovanja, Manjinske_partije, Filip VUJANOVIC, Prihodi domacinstva - mesecno, SNP, Srdjan MILIC, DPS, Srbin, Ranko KRIVOKAPIC, Milo DJUKANOVIC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36,886	13	2,837	111,564	,000[a]
	Residual	18,984	746	,025		
	Total	55,870	759			

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ostale_Srpske_partije, Staros Ukupan broj zavrsenih godina skolovanja, Manjinske_partije, Filip VUJANO Prihodi domacinstva - mesecno, SNP, Srdjan MILIC, DPS, Srbin, Ranko KRIVOKAPIC, Milo DJUKANOVIC

b. Dependent Variable: POVERENJE_INSTITUCIJE

Opet jasno i nedvosmisleno možemo uočiti da bez obzira na smanjen broj varijabli, model po svojim ključnim karakteristikama nije gotovo ništa izgubio. Nastavljamo sa identifikacijom varijabli koje ćemo u sledećoj iteraciji izbaciti iz modela:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,066	,033		-2,001	,046	-,130	-,001
	Starost	,001	,000	,058	2,571	,010	,000	,002
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,042	-1,951	,051	-,002	,000
	Prihodi domacinstva - mesecno	,003	,001	,053	2,342	,019	,000	,005
	Srbin	-,023	,017	-,040	-1,364	,173	-,056	,010
	DPS	,079	,018	,137	4,483	,000	,044	,114
	SNP	-,022	,023	-,023	-,944	,346	-,067	,024
	Ostale_Srpske_partije	**-,009**	**,031**	**-,006**	**-,278**	**,781**	**-,070**	**,053**
	Manjinske_partije	**,020**	**,049**	**,009**	**,420**	**,675**	**-,075**	**,116**
	Filip VUJANOVIC	,053	,007	,290	7,247	,000	,039	,068
	Ranko KRIVOKAPIC	,049	,007	,253	7,424	,000	,036	,062
	Milo DJUKANOVIC	,037	,007	,220	5,378	,000	,023	,050
	Srdjan MILIC	,012	,006	,061	2,056	,040	,001	,023
	Nebojsa MEDOJEVIC	-,004	,005	-,018	-,734	,463	-,015	,007

a. Dependent Variable: POVERENJE_INSTITUCIJE

Sledeće varijable koje ćemo neutralisati jesu pristalice ostalih srpskih partija i pristalice manjinskih partija. Dakle, ove dve varijable, na osnovu naših podataka, ne doprinose objašnjenju kriterijumske varijable. Izbacivanjem ovih varijabli možemo videti sledeće:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4053	,27123	760
Starost	43,70	15,860	760
Ukupan broj zavrsenih godina skolovanja	13,82	11,167	760
Prihodi domacinstva - mesecno	10,80	5,348	760
Srbin	,3377	,47324	760
DPS	,3309	,47086	760
SNP	,0935	,29134	760
Filip VUJANOVIC	2,96	1,473	760
Ranko KRIVOKAPIC	2,28	1,407	760
Milo DJUKANOVIC	2,85	1,617	760
Srdjan MILIC	2,65	1,415	760
Nebojsa MEDOJEVIC	2,25	1,229	760

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,812[a]	,660	,655	,15929

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupa broj zavrsenih godina skolovanja, Milo DJUKANOVIC Starost, Prihodi domacinstva - mesecno, SNP, Srbin Srdjan MILIC, DPS, Ranko KRIVOKAPIC, Filip VUJANOVIC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36,879	11	3,353	132,133	,000[a]
	Residual	18,991	748	,025		
	Total	55,870	759			

a. Predictors: (Constant), Nebojsa MEDOJEVIC , Ukupan broj zavrsenih godina skolovanja, Milo DJUKANOVIC, Starost, Prihodi domacinstva - mesecno, SNP, Srbin, Srdjan MILIC, DPS, Ranko KRIVOKAPIC, Filip VUJANOVIC

b. Dependent Variable: POVERENJE_INSTITUCIJE

Dakle, model je jednako dobar iako smo izbacili još dve varijable. Nastavljamo sa identifikacijom varijabli koje ćemo izbaciti iz modela u sledećoj iteraciji:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,064	,033		-1,971	,049	-,128	,000
	Starost	,001	,000	,058	2,565	,010	,000	,002
	Ukupan broj zavrsenih godina skolovanja	-,001	,001	-,042	-1,971	,049	-,002	,000
	Prihodi domacinstva - mesecno	,003	,001	,052	2,315	,021	,000	,005
	Srbin	-,024	,017	-,042	-1,436	,151	-,057	,009
	DPS	,078	,017	,135	4,495	,000	,044	,112
	SNP	**-,021**	**,023**	**-,022**	**-,907**	**,365**	**-,065**	**,024**
	Filip VUJANOVIC	,053	,007	,289	7,241	,000	,039	,068
	Ranko KRIVOKAPIC	,049	,007	,253	7,436	,000	,036	,062
	Milo DJUKANOVIC	,037	,007	,222	5,465	,000	,024	,051
	Srdjan MILIC	,011	,006	,060	2,021	,044	,000	,023
	Nebojsa MEDOJEVIC	**-,004**	**,005**	**-,018**	**-,743**	**,458**	**-,015**	**,007**

a. Dependent Variable: POVERENJE_INSTITUCIJE

Sledeće varijable koje izbacujemo iz modela jesu SNP i ocenjivanje N. Medojevića. Evo modela nakon izbacivanja ovih varijabli:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4076	,27460	778
Starost	43,85	15,927	778
Ukupan broj zavrsenih godina skolovanja	13,79	11,048	778
Prihodi domacinstva - mesecno	10,73	5,333	778
Srbin	,3389	,47366	778
DPS	,3336	,47181	778
Filip VUJANOVIC	2,96	1,480	778
Ranko KRIVOKAPIC	2,29	1,414	778
Milo DJUKANOVIC	2,85	1,620	778
Srdjan MILIC	2,67	1,420	778

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,802[a]	,644	,639	,16488

a. Predictors: (Constant), Srdjan MILIC, Ukupan broj zavrsenih godina skolovanja, Prihodi domacinstva - mesecno, Filip VUJANOVIC, Starost, DPS, Srbin, Ranko KRIVOKAPIC, Milo DJUKANOVIC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37,705	9	4,189	154,101	,000[a]
	Residual	20,876	768	,027		
	Total	58,580	777			

a. Predictors: (Constant), Srdjan MILIC, Ukupan broj zavrsenih godina skolovanja, Prihodi domacinstva - mesecno, Filip VUJANOVIC, Starost, DPS, Srbin, Ranko KRIVOKAPIC, Milo DJUKANOVIC

b. Dependent Variable: POVERENJE_INSTITUCIJE

Rezultati ukazuju da je 'gubitak' modela mali, ali ga moramo identifikovati. Naime, 2% varijanse je manje objašnjeno modelom nakon ove iteracije, a nešto je i lošiji odnos sume kvadrata regresionog modela i reziduala. No, ovaj gubitak nije 'dovoljan' da bi zastali, prema tome nastavljamo sa identifikacijom novih varijabli koje ne doprinose objašnjavanju kriterijumske varijable:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,069	,033		-2,087	,037	-,134	-,004
	Starost	,001	,000	,072	3,165	,002	,000	,002
	Ukupan broj zavrsenih godina skolovanja	**-,001**	**,001**	**-,039**	**-1,819**	**,069**	**-,002**	**,000**
	Prihodi domacinstva - mesecno	,003	,001	,053	2,359	,019	,000	,005
	Srbin	**-,031**	**,017**	**-,054**	**-1,812**	**,070**	**-,065**	**,003**
	DPS	,083	,018	,143	4,713	,000	,049	,118
	Filip VUJANOVIC	,050	,007	,269	6,658	,000	,035	,065
	Ranko KRIVOKAPIC	,049	,007	,253	7,294	,000	,036	,062
	Milo DJUKANOVIC	,037	,007	,216	5,234	,000	,023	,037
	Srdjan MILIC	**,009**	**,005**	**,048**	**1,840**	**,066**	**-,001**	**,009**

a. Dependent Variable: POVERENJE_INSTITUCIJE

Sledeće varijable, na osnovu dobijenih podataka, koje ćemo izbaciti iz modela jesu ukupan broj završenih godina školovanja, Srbin i ocene za S. Milića. Evo modela nakon izbacivanja ove tri varijable:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4207	,27720	840
Starost	43,57	16,035	840
Prihodi domacinstva - mesecno	10,69	5,352	840
DPS	,3433	,47509	840
Filip VUJANOVIC	3,02	1,485	840
Ranko KRIVOKAPIC	2,35	1,433	840
Milo DJUKANOVIC	2,93	1,621	840

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,804[a]	,647	,645	,16524

a. Predictors: (Constant), Milo DJUKANOVIC, Starost, Prihodi domacinstva - mesecno, DPS, Ranko KRIVOKAPIC, Filip VUJANOVIC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41,729	6	6,955	254,715	,000[a]
	Residual	22,748	833	,027		
	Total	64,477	839			

a. Predictors: (Constant), Milo DJUKANOVIC, Starost, Prihodi domacinstva - mesecno, DPS, Ranko KRIVOKAPIC, Filip VUJANOVIC

b. Dependent Variable: POVERENJE_INSTITUCIJE

Iako su tri varijable izbačene, procenat objašnjene varijanse kriterijumske varijable se neznatno povećao, a takođe je i veća vrednost sume kvadrata modela. Dakle, model ima bolje karakteristike nego što je posedovao pre izbacivanja ove tri varijable. Na ovaj način dolazimo do našeg finalnog modela koji izgleda ovako:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,089	,026		-3,386	,001	-,141	-,037
	Starost	,001	,000	,076	3,563	,000	,001	,002
	Prihodi domacinstva - mesecno	,003	,001	,052	2,421	,016	,001	,005
	DPS	,069	,017	,118	4,167	,000	,036	,101
	Filip VUJANOVIC	,057	,007	,307	8,361	,000	,044	,071
	Ranko KRIVOKAPIC	,051	,006	,264	8,140	,000	,039	,063
	Milo DJUKANOVIC	,037	,007	,214	5,562	,000	,024	,050

a. Dependent Variable: POVERENJE_INSTITUCIJE

Zašto je ovo finalni model? Zašto smo prestali sa redukcijom i izbacivanjem varijabli? Koji je kriterijum zapravo da stanemo sa redukcijom u jednom trenutku i prihvatimo model koji se nudi? Ne postoje univerzalni i jednostavni odgovori na ova pitanja. Međutim ima nekoliko smernica koja nam ukazuju kada sa redukcionizmom valja stati. Prvo, tažimo da u modeliu imao što je moguće manje varijabli. Ovo naprosto zato što nam je zgodnije da predviđamo vrednosti kriterijumske varijable na osnovu što manjeg broja prediktorskih varijabli. Sa druge strane, međutim, mi želimo 'dobar' model, što bi značilo da nije dobro smanjivati broj varijabli u situaciji kada model značajno gubi na svom prediktorskom karakteru. Između ova dva kriterijuma treba tražiti rešenje. U konkretnom slučaju, našem primeru modeliranja, šest varijabli je sasvim zadovoljavajuće, imajući u vidu veliki procenat varijanse prediktorske varijable koji možemo objasniti modelom, i imajući u vidu pojedinačne regresione koeficijente te rezultate t testa koji ukazuju da su sve varijable u modelu visoko statistički značajne. Drugim rečima, izbacivanjem neke od ovih varijabli, značajno bi umanjili prediktivnost modela, i ovo se lako može testirati u svakoj konkretnoj istraživačkoj situaciji.

Veoma važno je primetiti da je prvotni model imao 22 varijable a da je objašnjavao 66,2% varijanse kriterijumske varijable. Nakon velikog pbroja iteracija, na osnovu kriterijuma koje smo opisali i objasnili, izbacili smo 16 varijabli te finalni model ima samo šest varijabli. No, bez obzira na ovako opsežnu redukciju varijabli, finalni model objašnjava 64,7% varijanse, što je neznatno smanjenje obzorom na veliki stepen redukcije varijabli.

Uobičajeni način prikazivanja finalnih rezultata regresione analize, bi na našem primeru izgledao ovako:

FAKTORI POVERENJA U INSTITUCIJE - nestandardizovani regresioni koeficijenti -

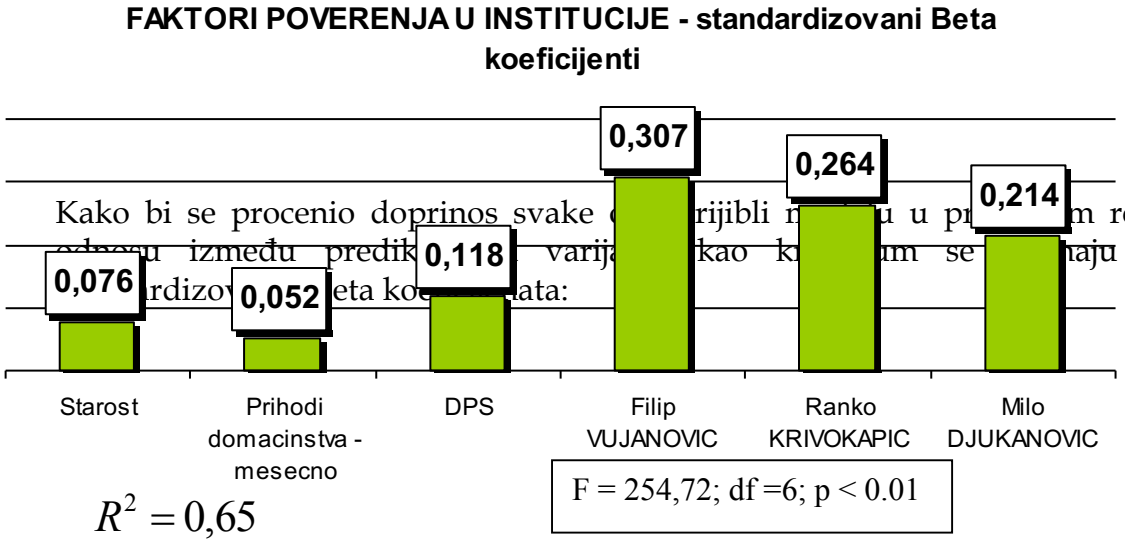
PREDIKTORI	B
B KONSTANTA	-,089**
Starost	,001**
Prihodi domacinstva - mesecno	,003*
DPS	,069**
Filip VUJANOVIC	,057**
Ranko KRIVOKAPIC	,051**
Milo DJUKANOVIC	,037**

** $p < 0.01$
* $p < 0.05$

$R^2 = 0{,}65$

F = 254,72; df =6; p < 0.01

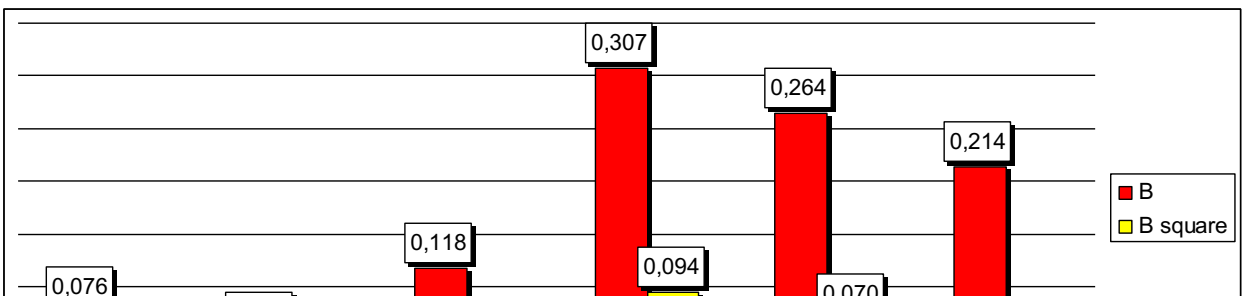
Grafički se to može prikazati na sledeći način:



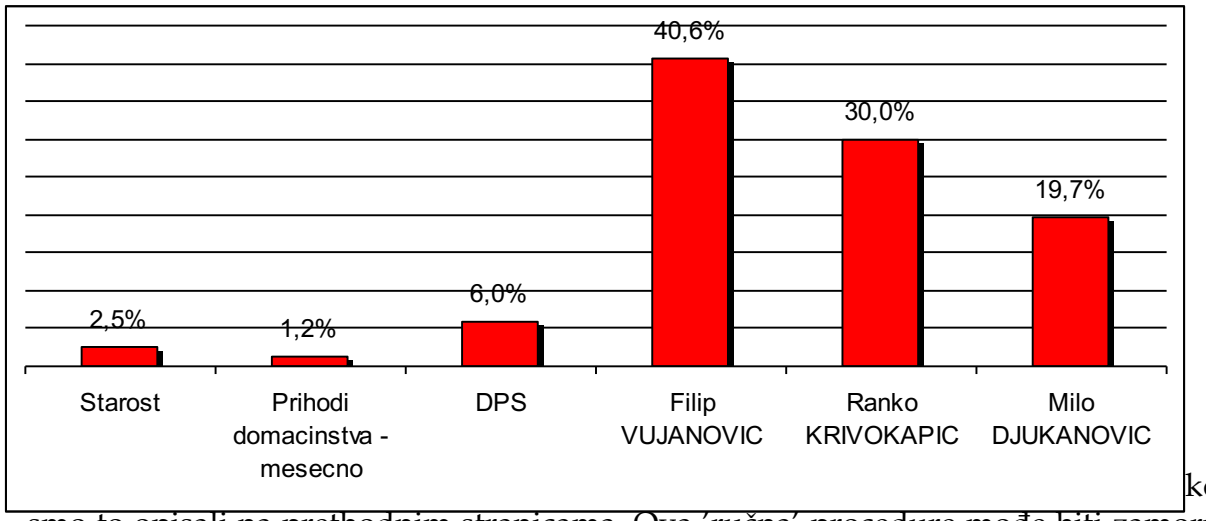
Kako bi se procenio doprinos svake [illegible] u pr[illegible] relativnom odnosu između predik[illegible] varija[illegible] kao k[illegible] se [illegible] kvadrati [illegible]rdizov[illegible]eta koe[illegible]ata:

$R^2 = 0{,}65$

F = 254,72; df =6; p < 0.01



Kvadratne vrednsoti Beta koeficijenata se mogu prikazati preko relativnih frekvencija kako bi se jednostavno ukazalo koliko koja varijabla doprinosi modelu u procentima:



koji smo to opisali na prethodnim stranicama. Ova 'ručna' procedure može biti zamorna i traži da stalno u svakoj novoj iteraciji samostalno izbacujemo varijable iz modela tragajući za 'najboljim' modelom. Ovaj posao, na sreću, može za nas da radi specijalizovani softver, u ovom slučaju SPSS. Dakle, ovaj softver će na osnovu definisanih kriterijuma da izbacuje ili ubacuje varijable, pri čemu smo mi obavezni samo da obezbedimo početni set vrijabli.

Postoje dva načina na koji možemo da oblikujemo model. Prvi podrazumeva da se od svih početnih vaijabli, krene od prve, one koja je statistički najznačajnija, i da se redom dodaju varijable, opet na osnovu testa statističke značajnosti. Procedura staje onog trenutka kada dodavanjem novih varijabli u model, am model ne dobija na ključnim karakteristikama. Ova procedura se zove **stepwise** i zapravo znači da se varijable ubacuju 'korak po korak'. Evo modeliranje stepwise procedurom sa istim setom varijabli kojim smo formirali model n gornjim stranicama:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4055	,27419	763
Pol	1,49	,500	763
Starost	43,68	15,888	763
Ukupan broj zavrsenih godina skolovanja	13,82	11,142	763
Prihodi domacinstva - mesecno	10,73	5,304	763
Crnogorac	,4359	,49620	763
Srbin	,3410	,47436	763
Bosnjak_Muslima	,0410	,19840	763
Albanac	,0569	,23174	763
DPS	,3326	,47146	763
SDP	,0354	,18500	763
PZP	,0577	,23324	763
SNS	,1216	,32705	763
SNP	,0940	,29201	763
Ostale_Srpske_partije	,0369	,18874	763
Manjinske_partije	,0148	,12073	763
Apstinenti	,2781	,44836	763
Filip VUJANOVIC	2,95	1,483	763
Ranko KRIVOKAPIC	2,28	1,418	763
Milo DJUKANOVIC	2,84	1,624	763
Andrija MANDIC	2,41	1,422	763
Srdjan MILIC	2,65	1,421	763

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,005	,015		,362	,717	-,024	,035
	Filip VUJANOVIC	,136	,005	,733	29,734	,000	,127	,145
2	(Constant)	-,012	,014		-,842	,400	-,040	,016
	Filip VUJANOVIC	,088	,006	,477	14,394	,000	,076	,100
	Ranko KRIVOKAPIC	,069	,006	,356	10,751	,000	,056	,081
3	(Constant)	-,020	,014		-1,477	,140	-,047	,007
	Filip VUJANOVIC	,060	,007	,326	8,312	,000	,046	,075
	Ranko KRIVOKAPIC	,054	,007	,277	8,086	,000	,041	,067
	Milo DJUKANOVIC	,044	,007	,262	6,707	,000	,031	,057
4	(Constant)	,005	,015		,370	,711	-,023	,034
	Filip VUJANOVIC	,056	,007	,304	7,807	,000	,042	,070
	Ranko KRIVOKAPIC	,048	,007	,246	7,159	,000	,035	,061
	Milo DJUKANOVIC	,034	,007	,201	4,988	,000	,021	,047
	DPS	,087	,018	,149	4,842	,000	,052	,122
5	(Constant)	-,042	,022		-1,906	,057	-,086	,001
	Filip VUJANOVIC	,055	,007	,295	7,589	,000	,040	,069
	Ranko KRIVOKAPIC	,047	,007	,242	7,076	,000	,034	,060
	Milo DJUKANOVIC	,037	,007	,218	5,358	,000	,023	,050
	DPS	,083	,018	,143	4,672	,000	,048	,118
	Starost	,001	,000	,063	2,848	,005	,000	,002
6	(Constant)	-,084	,028		-3,026	,003	-,138	-,029
	Filip VUJANOVIC	,056	,007	,301	7,750	,000	,042	,070
	Ranko KRIVOKAPIC	,049	,007	,251	7,323	,000	,036	,062
	Milo DJUKANOVIC	,034	,007	,204	4,987	,000	,021	,048
	DPS	,083	,018	,143	4,677	,000	,048	,118
	Starost	,001	,000	,076	3,354	,001	,001	,002
	Prihodi domacinstva - mesecno	,003	,001	,057	2,484	,013	,001	,005

a. Dependent Variable: POVERENJE_INSTITUCIJE

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,733[a]	,537	,537	,18660
2	,774[b]	,599	,597	,17396
3	,788[c]	,621	,620	,16913
4	,795[d]	,632	,630	,16668
5	,798[e]	,636	,634	,16591
6	,800[f]	,639	,636	,16534

a. Predictors: (Constant), Filip VUJANOVIC

b. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC

c. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC, Milo DJUKANOVIC

d. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC, Milo DJUKANOVIC, DPS

e. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC, Milo DJUKANOVIC, DPS, Starost

f. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC, Milo DJUKANOVIC, DPS, Starost, Prihodi domacinstva - mesecno

ANOVA[g]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	30,783	1	30,783	884,102	,000[a]
	Residual	26,490	761	,035		
	Total	57,273	762			
2	Regression	34,281	2	17,140	566,428	,000[b]
	Residual	22,993	760	,030		
	Total	57,273	762			
3	Regression	35,567	3	11,856	414,467	,000[c]
	Residual	21,706	759	,029		
	Total	57,273	762			
4	Regression	36,219	4	9,055	325,904	,000[d]
	Residual	21,055	758	,028		
	Total	57,273	762			
5	Regression	36,442	5	7,288	264,793	,000[e]
	Residual	20,831	757	,028		
	Total	57,273	762			
6	Regression	36,611	6	6,102	223,197	,000[f]
	Residual	20,663	756	,027		
	Total	57,273	762			

a. Predictors: (Constant), Filip VUJANOVIC

b. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC

c. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC, Milo DJUKANOVIC

d. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC, Milo DJUKANOVIC, DPS

e. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC, Milo DJUKANOVIC, DPS, Starost

f. Predictors: (Constant), Filip VUJANOVIC, Ranko KRIVOKAPIC, Milo DJUKANOVIC, DPS, Starost, Prihodi domacinstva - mesecno

g. Dependent Variable: POVERENJE_INSTITUCIJE

Šta možemo primetiti? Najpre, stepwise procedura je na osnovu istih početnih varijabli identifikovali onih šest varijabli koje smo i mi identifikovali postepenim izbacivanjem. Dakle, u krajnjem ishodu rezultat je identičan, samo što je u ovom

slučaju sam softver krenuo od jedne i dodavao po jednu varijabu u svakoj sledećoj iteraciji, dok sa šestim modelom nije došao do šest varijabli koje je uključio u finalni model. U ovom modelovanju, softver je ostavio i jasne tragove koje je varijable ubacivao u kojoj varijaciji, te ključne karakteristike modela u svakoj iteraciji.

Druga procedura je tzv. **backward** procedura, i ova procedura za razliku od stepwise procedure kreće od svih početnih varijabli, i izbacuje varijable u većem broju iteracija, sve dok ne dođe do finalnog modela. Ova procedura je prema tome sličnija onoj koju smo mi sami 'ručno' izveli. Evo finalnog modela backward procedura na našem primeru:

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta	Lower Bound	Upper Bound	B	Std. Error
15	(Constant)	-,071	,029		-2,487	,013	-,128	-,015
	Starost	,001	,000	,076	3,338	,001	,001	,002
	Ukupan broj zavrsenih godina skolovanja	**-,001**	**,001**	**-,038**	**-1,743**	**,082**	**-,002**	**,000**
	Prihodi domacinstva - mesecno	,003	,001	,058	2,545	,011	,001	,005
	DPS	,083	,018	,143	4,695	,000	,049	,118
	Filip VUJANOVIC	,056	,007	,303	7,819	,000	,042	,070
	Ranko KRIVOKAPIC	,048	,007	,248	7,223	,000	,035	,061
	Milo DJUKANOVIC	,034	,007	,204	4,990	,000	,021	,048

$R^2 = 0{,}64$ F = 192.26: df =6: p < 0.01

Ono što možemo videti jeste da na osnovu backward procedure u modelu je ostala jedna varijabla više nego što je to bio slučaj sa našim 'ručnim' modelom i u odnosu na rezultate stepwise procedure. Jednako, ako se pogleda varijabla koja razlikuje ovaj model, reč je o obrazovanju ispitanika izraženo ukupnim brojem završenih godina školovanja. Međutim, primetičemo, da je B koeficijent za ovu varijablu, kao i t test na granici statističke značajnosti, pa se postalja pitanje opravanosti zadržavanja ove varijable u modelu. U praksi, u ovakvim situacijama najčešće je slučaj da određene kohorte ispitanika u okviru ove varijable jesu a određene kohorte nisu značajne. U konkretnoj situaciji, umesto ove varijable uvešćemo dummy varijable obrazovanja po katewgorijama i proverićemo model:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4207	,27720	840
Starost	43,57	16,035	840
Prihodi domacinstva - mesecno	10,69	5,352	840
Bez_i_osnovno_ obrazovanje	,1167	,32120	840
Srednje_obrazovanje	,5592	,49678	840
Vise_obrazovanje	,1571	,36413	840
Visoko_obrazovanje	,1502	,35743	840
Filip VUJANOVIC	3,02	1,485	840
Ranko KRIVOKAPIC	2,35	1,433	840
Milo DJUKANOVIC	2,93	1,621	840
DPS	,3433	,47509	840

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,806[a]	,650	,645	,16505

a. Predictors: (Constant), DPS, Vise_obrazovanje, Starost, Visoko_obrazovanje, Bez_i_osnovno_ obrazovanje, Prihodi domacinstva - mesecno, Ranko KRIVOKAPIC, Filip VUJANOVIC, Milo DJUKANOVIC, Srednje_obrazovanje

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41,890	10	4,189	153,767	,000[a]
	Residual	22,588	829	,027		
	Total	64,477	839			

a. Predictors: (Constant), DPS, Vise_obrazovanje, Starost, Visoko_obrazovanje, Bez_i_osnovno_obrazovanje, Prihodi domacinstva - mesecno, Ranko KRIVOKAPIC, Filip VUJANOVIC, Milo DJUKANOVIC, Srednje_obrazovanje

b. Dependent Variable: POVERENJE_INSTITUCIJE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,138	,050		-2,740	,006	-,236	-,039
	Starost	,001	,000	,062	2,792	,005	,000	,002
	Prihodi domacinstva - mesecno	,003	,001	,065	2,829	,005	,001	,006
	Bez_i_osnovno_ obrazovanje	,089	,048	,104	1,853	,064	-,005	,184
	Srednje_obrazovanje	**,049**	**,045**	**,087**	**1,087**	**,278**	**-,039**	**,136**
	Vise_obrazovanje	**,053**	**,046**	**,069**	**1,142**	**,254**	**-,038**	**,144**
	Visoko_obrazovanje	**,043**	**,047**	**,055**	**,918**	**,359**	**-,049**	**,134**
	Filip VUJANOVIC	,058	,007	,309	8,424	,000	,044	,071
	Ranko KRIVOKAPIC	,050	,006	,259	7,967	,000	,038	,062
	Milo DJUKANOVIC	,037	,007	,215	5,581	,000	,024	,050
	DPS	,068	,017	,117	4,123	,000	,036	,101

a. Dependent Variable: POVERENJE_INSTITUCIJE

Rezultati ukazuju da od četiri kohorte po kategorijama obrazovanja, tri treba izbaciti iz modela i ostaviti samo one varijablu bez obrazovanja i osnovno obrazovanje, a to je dummy varijabla koja dvovalentno razlikue one koji imaju osnovnu školu ili kanje i sve ostale kategorije. Prema tome, naš finalni model će izgledati ovako:

Descriptive Statistics

	Mean	Std. Deviation	N
POVERENJE_ INSTITUCIJE	,4207	,27720	840
Starost	43,57	16,035	840
Prihodi domacinstva - mesecno	10,69	5,352	840
Bez_i_osnovno_ obrazovanje	,1167	,32120	840
Srednje_obrazovanje	,5592	,49678	840
Vise_obrazovanje	,1571	,36413	840
Visoko_obrazovanje	,1502	,35743	840
Filip VUJANOVIC	3,02	1,485	840
Ranko KRIVOKAPIC	2,35	1,433	840
Milo DJUKANOVIC	2,93	1,621	840
DPS	,3433	,47509	840

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,806[a]	,649	,646	,16490

a. Predictors: (Constant), DPS, Prihodi domacinstva - mesecno, Starost, Bez_i_osnovno_obrazovanje, Ranko KRIVOKAPIC, Filip VUJANOVIC, Milo DJUKANOVIC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41,851	7	5,979	219,882	,000[a]
	Residual	22,626	832	,027		
	Total	64,477	839			

a. Predictors: (Constant), DPS, Prihodi domacinstva - mesecno, Starost, Bez_i_ osnovno_obrazovanje, Ranko KRIVOKAPIC, Filip VUJANOVIC, Milo DJUKANOVIC

b. Dependent Variable: POVERENJE_INSTITUCIJE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,090	,026		-3,411	,001	-,141	-,038
	Starost	,001	,000	,063	2,865	,004	,000	,002
	Prihodi domacinstva - mesecno	,003	,001	,062	2,844	,005	,001	,005
	Bez_i_osnovno_ obrazovanje	,041	,019	,047	2,118	,034	,003	,079
	Filip VUJANOVIC	,058	,007	,308	8,419	,000	,044	,071
	Ranko KRIVOKAPIC	,050	,006	,259	7,989	,000	,038	,062
	Milo DJUKANOVIC	,037	,007	,216	5,617	,000	,024	,050
	DPS	,068	,016	,117	4,145	,000	,036	,101

a. Dependent Variable: POVERENJE_INSTITUCIJE

A to ćemo prikazati na sledeći način:

	B
B Konstanta	-,090**
Starost	,001**
Prihodi domacinstva - mesecno	,003**
Bez i osnovno_obrazovanje	,041*
Filip VUJANOVIC	,058**
Ranko KRIVOKAPIC	,050**
Milo DJUKANOVIC	,037**
DPS	,068**

**p < 0.01
*p < 0.05

$R^2 = 0{,}64$ F = 219,882; df =7; p < 0.01

Ako uporedimo model koji smo dobili setepwise procedurom i onaj koji smo dobili sa backward procedurom, možemo videti sledeće:

	B Model 1	B Model 2
B Konstanta	-,089**	-,090**
Starost	,001**	,001**
Prihodi domacinstva - mesecno	,003*	,003**
Bez i osnovno obrazovanje	----------	,041*
Filip VUJANOVIC	,057**	,058**
Ranko KRIVOKAPIC	,051**	,050**
Milo DJUKANOVIC	,037**	,037**
DPS	,069**	,068**
	$R^2 = 0{,}65$	$R^2 = 0{,}64$
	F = 254,7 df 6 p < 0.01	F = 219,9 df 7 p < 0.01

**p < 0.01
*p < 0.05

Dakle, doista, razlike su veoma male i interesanto je da prvi model (onaj koji ne operiše obrazovanjem kao varijablom), objašnjava nešto veći procenat varijanse u odnosu na drugi model, iako ovaj drugi ima jednu varijablu više.

Da zaključimo, proces modeliranja je voma zahtevan i složen. U ovom procesu moramo i teorijski spekulisati i biti veom pažljivi u odnosu na razumevanje statističkih parametara.

U regresionim modelima prediktorske varijable nisu u stanju da objasne 'ukupnu' varijansu kriterijumske varijable. Ostatak, koji prediktorske variable NE objašnjavaju nazivaju se 'reziduali'. Dakle, pod rezidualima se podrazumeva 'ostatak', tačnije, onaj 'deo' varijanse koji prediktorske varijable NE objašnjavaju. Cilj svake regresione analize jeste da što veći procenat varijanse kriterijumske varijable bude objašnjen prediktorskim varijablama a da što manji deo varijanse ostane u kategoriji reziduala. Način na koji se kalkuliše odnos između onog dela varijanse koji objašnjavamo modelom i reziduala, jeste posredstvom analize varijanse odnosa sume kvadrata između 'modela' i 'reziduala'. Evo primera:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41,729	6	6,955	254,715	,000[a]
	Residual	22,748	833	,027		
	Total	64,477	839			

a. Predictors: (Constant), DPS, Prihodi domacinstva - mesecno, Starost, Ranko KRIVOKAPIC, Filip VUJANOVIC, Milo DJUKANOVIC

b. Dependent Variable: POVERENJE_INSTITUCIJE

Dakle, ukupna varijansa se sastoji iz dva dela, 'modela' i 'reziduala'. Ukupna suma kvadrata je raspodeljena na onaj deo varijanse koji je objašnjen modelom i onaj deo koji nije objašnjen modelom (reziduali). Cilj 'dobrog' modela jeste da razlika između sume kvadrata samog modela bude što je moguće veća u odnosu na sumu kvadrata reziduala. Ukoliko je suma kvadrata reziduala 'velika', to znači da je veliki deo varijanse kriterijumske varijable ostao 'neobjašnjen' prediktorskim varijablama. 'Mean square' (srednja vrednost kvadrata) predstavlja 'sumu kvadrata podeljenih sa brojem stepena slobode'. Prema tome, što je veća razlika između 'mean square' modela i reziduala, to model bolje 'fituje'. Sumarno, odnos između 'modela' i 'reziduala' se izračunava F testom, na način da se vrednost 'mean square' modela podeli sa vrednošću 'mean square residuala'

U našoj situaciji F = 6,954838389943/0,02730438227381 = 254,7150973862

Podatak F testa se obavezno daje u prikazu finalnom prikazu modela

Za analizu reziduala postoje i specifični testovi. Jedan od njih je tzv. Durbin – Watsonov test reziduala. Evo primera:

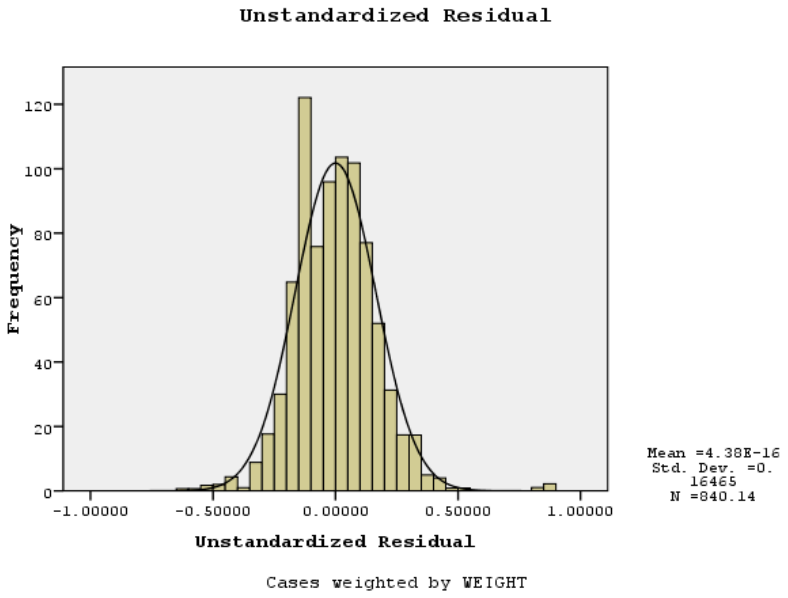
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	,0902	,8419	,4207	,22300	840
Residual	-,61565	,86500	,00000	,16465	840
Std. Predicted Value	-1,482	1,889	,000	1,000	840
Std. Residual	-3,726	5,235	,000	,996	840

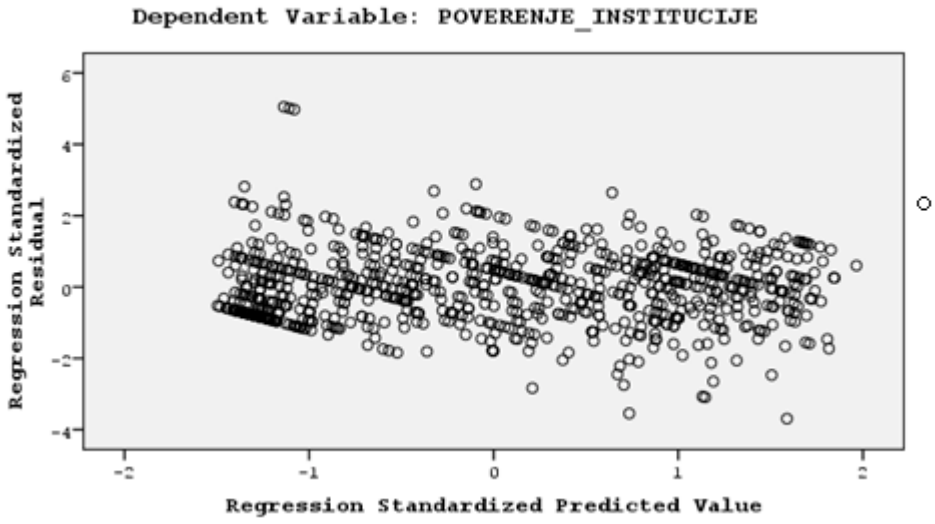
a. Dependent Variable: POVERENJE_INSTITUCIJE

Durbin-Watsonov test prikazuje karakteristike predviđenih vrednosti svih opservacija kako za model tako i za reziduale, i to i za originalne vrednosti i za

standardizovane varijable. Kada je o rezidualima reč, za svaku pojedinu opservaciju, reziduali predstavljaju razliku između opserviranih vrednosti kriterijumske varijable i vrednosti predviđenih modelom. Prema tome, reziduali pokazuju 'stvarnu' grešku modela. **Ukoliko je model odgovarajući, reziduali će biti normalno distribuirani** (za ovo je potrebno da plotujemo reziduale). Evo distribucije reziduala u našem modelu:



Drugi način plotovanja jeste da 'ukrstimo' predikciju i reziduale. U ovoj situaciji smatra se da dobar model 'razbija' reziduale i oni više nisu struktuirani. Evo našeg primera:



Jedan od problem sa kojim se često srećemo u regresionom modelu jeste problem 'multikolinearnosti' ili skraćeno 'kolinearnosti'. Kolinearnost podrazumeva situaciju u kojoj postoji visoka korelacija između prediktorskih varijabli u modelu. Ovo je nepoželjna situacija, i s toga mi se u izgradnji modela trudimo da prediktorske varijable nisu u visokoj korelaciji. To je osnovni razlog za identifikaciju kolinearnosti između prediktorskih varijabli. Jedan od načina za identifikaciju kolinearnosti jeste posredstvom **eigenvalues** (ejgen vrednosti), koja identifikuje koliko ima distinktivnih dimenzija između nezavisnih varijabli. Evo primera:

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	Starost	Prihodi domacinstva - mesecno	Filip VUJANOVIC	Ranko KRIVOKAPIC	Milo DJUKANOVIC	DPS
1	1	5,838	1,000	,00	,00	,00	,00	,00	,00	,01
	2	,622	3,064	,01	,02	,04	,00	,00	,00	,32
	3	,214	5,224	,00	,07	,53	,01	,05	,00	,17
	4	,177	5,739	,01	,25	,02	,02	,10	,04	,42
	5	,073	8,944	,00	,00	,05	,15	,84	,22	,01
	6	,042	11,787	,01	,05	,00	,74	,00	,73	,01
	7	,034	13,117	,98	,60	,35	,07	,00	,00	,06

a. Dependent Variable: POVERENJE_INSTITUCIJE

Ukoliko veći broj prediktorskih varijabli ima eigenvalues blizu vrednosti 0 to znaš da su prediktorske varijable u korelaciji i konsekventno, male promene na prediktorskim varijablama mogu da proizvedu krupne promene na kriterijumskoj varijabli. **Condition INDEX** (Index uslovnosti) jeste kvadratni koren odnosa između najveće eigenvrednosti i prve sukcesivne eigenvrednosti. Po konvenciji, ukoliko je Index veći od 15, to nam ukazuje na mogući problem kolinearnosti, a ukoliko je veći od 30, to sasvim izvesno govori da su prediktorske varijable u visokoj korelaciji.

Jedan od načina za identifikaciju kolinearnosti jeste i korišćenje VIF tj. 'variance inflation factors'. Evo primera:

Coefficients[a]

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)	-,090	,026		-3,411	,001	-,141	-,038		
	Starost	,001	,000	,063	2,865	,004	,000	,002	,862	1,160
	Prihodi domacinstva - mesecno	,003	,001	,062	2,844	,005	,001	,005	,876	1,142
	Bez_i_osnovno_ obrazovanje	,041	,019	,047	2,118	,034	,003	,079	,841	1,189
	Filip VUJANOVIC	,058	,007	,308	8,419	,000	,044	,071	,315	3,179
	Ranko KRIVOKAPIC	,050	,006	,259	7,989	,000	,038	,062	,401	2,493
	Milo DJUKANOVIC	,037	,007	,216	5,617	,000	,024	,050	,285	3,503
	DPS	,068	,016	,117	4,145	,000	,036	,101	,529	1,890

a. Dependent Variable: POVERENJE_INSTITUCIJE

Tolerance se koristi kao mera linearne povezanosti između prediktorskih varijabli. **Tolerancija** predstavlja proporciju varijanse varijable koju ne kalkulišemo u odnosu na ostale prediktorske varijable. Varijable sa malom tolerancijom (recimo, manjom od 0.25) ne doprinose modelu, i treba ih ukloniti. **VIF** (varijance inflation factor) služi za

procenu kolinearnosti između prediktorskih varijabli, i vredsnoti VIF je recipročna vrednostima tolerance. Sa povećavanjem vrednosti VIF smanjuje se i varijansa regresionog koeficijenta
Konsekventno, visoke vrednosti VIF (recimo veće od 10) indiciraju kolinearnost.

Jedan od načina da se proveri validnost modela jeste da se snime (zapomte) vrednosti kriterijimske varijable sa svaku pojedinu opservaciju. Na ovaj način moguće je proveriti i dovesti u odnos vrednosti koje predviđene modelom i opservirane vrednosti. Ovo je, dalje, moguće uraditi na više načina od kojih su četiri najtipičnija:

- Karakteristikama predikcijske varijable
- Prostom korelacijom između kriterijumske i predikcijske varijable (ryy')
- Indexom proporcionalnosti (vrednosti kriterijumske podeliti sa vrednostima predikcijske varijable: y/y')

U našem slučaju:

Statistics

		POVERENJE_INSTITUCIJE	Unstandardized Predicted Value
N	Valid	951	869
	Missing	62	145
Mean		,4289	,4176927
Std. Error of Mean		,00891	,00755457
Median		,4583	,4067653
Mode		,50	,15357
Std. Deviation		,27489	,22265216
Variance		,076	,050
Skewness		,120	,179
Std. Error of Skewness		,079	,083
Kurtosis		-,843	-1,318
Std. Error of Kurtosis		,158	,166
Range		1,00	,75173
Minimum		,00	,09019
Maximum		1,00	,84192
Sum		408,03	362,82039
Percentiles	25	,2083	,1992552
	50	,4583	,4067653
	75	,6250	,6204235

$r_{yy'} = 0{,}804$

$$IP = \frac{\sum y}{\sum y'} = \frac{412{,}87}{367{,}3376} = 1{,}12$$

Dodatno, kontrole modela radi, dobro je identifikovati slučajeve koji spadaju u 'outliere' obzirom na određeni broj standardnih devijacija odstupanja vrednosti opsevacija na y' u odnosu na y. Evo primera sa 3 SD odstuanja:

Casewise Diagnostics[a]

Case Number	Std. Residual	POVERENJE_ INSTITUCIJE	Predicted Value	Residual
89	4,875	1,00	,1944	,80561
170	-3,726	,17	,7823	-,61565
172	-3,115	,17	,6814	-,51470
293	5,187	1,00	,1429	,85711
739	-3,134	,17	,6845	-,51787
746	-3,596	,00	,5943	-,59428
841	5,235	1,00	,1350	,86500
920	3,077	,88	,3666	,50837

a. Dependent Variable: POVERENJE_INSTITUCIJE

Evo odstupanje +/- 2,6 SD:

Casewise Diagnostics[a]

Case Number	Std. Residual	POVERENJE_ INSTITUCIJE	Predicted Value	Residual
89	4,875	1,00	,1944	,80561
170	-3,726	,17	,7823	-,61565
172	-3,115	,17	,6814	-,51470
239	-2,703	,13	,5716	-,44659
293	5,187	1,00	,1429	,85711
329	2,660	,79	,3521	,43960
349	2,640	1,00	,5637	,43628
503	-2,822	,00	,4663	-,46634
739	-3,134	,17	,6845	-,51787
746	-3,596	,00	,5943	-,59428
841	5,235	1,00	,1350	,86500
842	2,776	,58	,1246	,45874
920	3,077	,88	,3666	,50837

a. Dependent Variable: POVERENJE_INSTITUCIJE

Identifikacija slučajeva kod kojih predviđene vrednosti odstupaju od stvarne vrednosti na kriterijumskoj varijabli ukazuju i na slabost modela, i zapravo govore o tome da u nekim slučajevima naša predikcija neće biti valjana. Ovo je sudbina svakog modela, naime, treba imati u vidu da ne postoji idealan model koji će za svaki pojedini slučaj da precizno predvidi vrednost kriterijumske varijable na osnovu prediktorskih varijabli.

Logistička regresija

Logistička regresiona analiza ima istu logiku, smisao i primenu kao i linearna regresija. Dakle, ideja je da se predvidi vrednost kriterijumske varijable na osnovnu nekoliko prediktorskih varijabli. *Napomena*: U SPSS-u, kada se izvodi logistička regresija, kriterijumska varijabla se naziva 'zavisna' a prediktorske varijable se nazivaju 'kovarirajuće'. Osnovna karakteristika logističke regresione analize jeste da

je zavisna varijabla binarna po svom karakteru. Dakle, zavisna varijabla je, ili po prirodi, ili je pripremljena (transformisana) na način da ima samo dve vrednosti, isto kao i svaka druga *dummy* varijabla (1 i 0, ili DA i NE, ili JESTE i NIJE itd.).

Cilj logističke regresije je da na osnovu kriterijumskih varijabli računamo verovatnoću da svaki od slučajeva u našoj datoteci 'uđe' u jednu od dve katagorije (vrednosti) zavisne varijable. Kao krajnji rezultat, mi ćemo utvrditi koje kriterijumske varijable jesu 'značajne' da predvidimo vrednosti 'kriterijumske' varijable, i da prema tome, na osnovu distribucija vrednosti ovih kriterijumskih varijabli, predviđamo vrednosti zavisne varijable. Formula za logističku regresionu analizu je jednostavna i veoma liči na farmulu koju koristimo za linearnu regresiju:

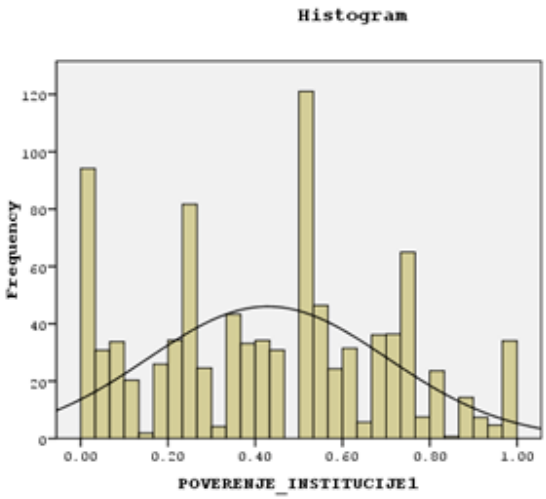
$$\log\left(\frac{Verovatnoca(dogadjanja)}{Verovatnoca(nedogadjanja)}\right) = B_0 + B_1x_1 + B_2x_2 + \ldots\ldots B_nx_n$$

Isto kao kada je reč o linearnoj regresiji, osnovni zadatak logističke regresije jeste izgradnja modela. Cilj dobrog modela jeste isti kao kada je reč o linearnoj regresiji, a to je da se sa što manjim brojem prediktorskih varijabli objasni što veća varijansa kriterijumske varijable. Pri tome postoji čitav niz statističkih procedura i indikatora koji nam mogu omogućiti preciznu procenu precizosti i validnosti modela. Evo primera logističke regresije sa istom kriterijumskom (zavisnom) varijablom, tj, poverenjem u institucije. Evo distribucije:

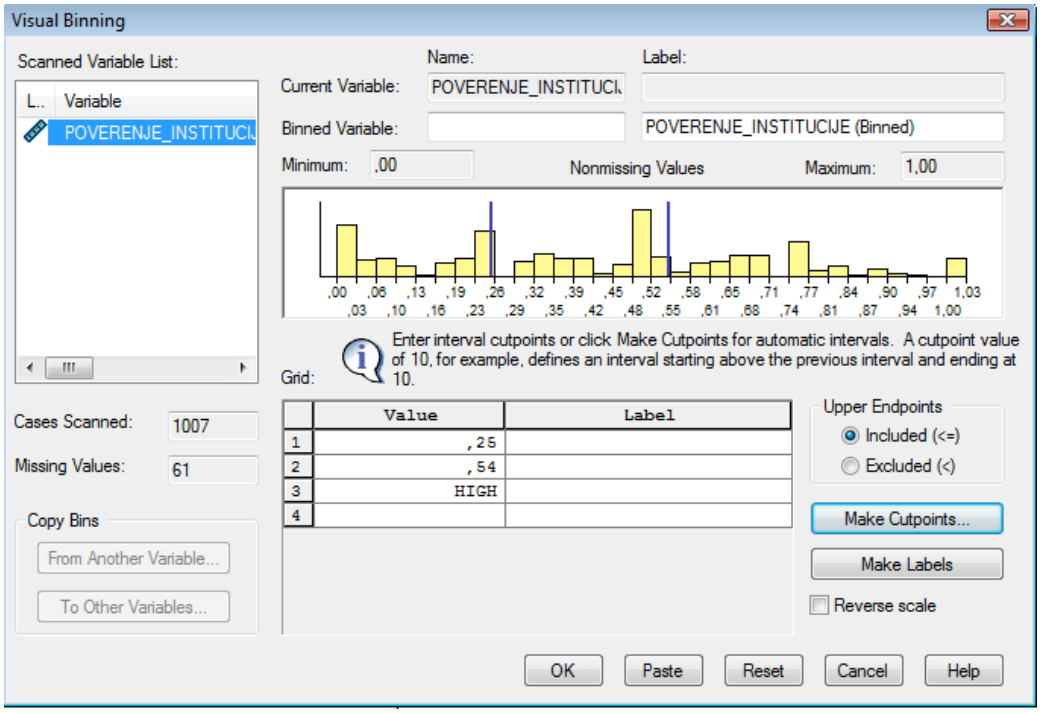
Statistics

POVERENJE_INSTITUCIJE1

N	Valid	951
	Missing	62
Mean		,4289
Std. Error of Mean		,00891
Median		,4583
Std. Deviation		,27489
Variance		,076
Skewness		,120
Std. Error of Skewness		,079
Kurtosis		-,843
Std. Error of Kurtosis		,158
Minimum		,00
Maximum		1,00
Percentiles	25	,2083
	50	,4583
	75	,6250



Kako bi ovo bila zavisna varijabla kojojm je moguće opersiati logističkom regresijom, najpre je nužno da je transformišemo u binarnu varijablu. Za to se možemo poslužiti jednostavnom procedurom u SPSS softweru:



Prema tome, celokupnu varijansu podelićemo najpre na po 33,3% varijanse, a zatim ćemo od ove varijable da kreiramo binarnu, pri ćemu ćemo onu trećinu ispitanika koji se nalaze i trećini sa najvišim vrednostima da identigikujemo kao one koji imaju poverenja, a ostale dve trećine kao one koji nemaju poverenja. Konsekventno, dobili smo dvovalentnu zavisnu varijablu:

POVERENJE_INSTITUCIJE (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NEMA POVERENJE	713	70,4	70,4	70,4
	IMA POVERENJE	300	29,6	29,6	100,0
	Total	1013	100,0	100,0	

Evo regresije sa početnim setom varijabli **forward** procedurom. Ovsa procedura počiva na istom principu kao stepwise procedura kada realizujemo linearnu regresiju. Evo rezultata logističke regresije:

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	v001	,002	1	,966
		P_2	6,679	1	,010
		P_4	,182	1	,670
		P_8	,054	1	,817
		Crnogorac	29,730	1	,000
		Srbin	100,289	1	,000
		Bosnjak_Muslima	6,433	1	,011
		Albanac	3,744	1	,053
		DPS	199,127	1	,000
		SDP	8,447	1	,004
		PZP	3,070	1	,080
		SNS	28,216	1	,000
		SNP	20,962	1	,000
		Ostale_Srpske_partije	8,569	1	,003
		Manjinske_partije	,502	1	,479
		Apstinenti	42,852	1	,000
		P_10_1	223,541	1	,000
		P_10_2	240,098	1	,000
		P_10_3	231,224	1	,000
		P_10_4	189,461	1	,000
		P_10_5	146,322	1	,000
		P_10_6	47,810	1	,000
		P_10_7	29,145	1	,000
		P_10_8	1,359	1	,244
		P_10_9	35,856	1	,000
		P_10_10	14,331	1	,000
		P_10_11	37,063	1	,000
		P_10_12	27,253	1	,000
		P_10_13	35,157	1	,000
		P_10_14	82,158	1	,000
		P_10_15	59,096	1	,000
		P_10_16	62,897	1	,000
		P_10_17	74,644	1	,000
		P_10_18	,181	1	,670
	Overall Statistics		303,377	34	,000

Regresioni koeficijent

Standardna greška

$WALD = \left(\frac{B}{SE}\right)^2$

Expected B predstavlja predviđenu verovatnoću promene na kriterijumskoj varijabli za svaku jedinicu promene na prediktorskoj varijabli

							95,0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	RANKO	1,198	,095	157,521	1	,000	3,315	2,749	3,997
	Constant	-4,161	,302	190,112	1	,000	,016		
Step 2[b]	FILIP	1,028	,140	53,598	1	,000	2,795	2,123	3,680
	RANKO	,668	,108	38,008	1	,000	1,950	1,577	2,411
	Constant	-6,306	,528	142,879	1	,000	,002		
Step 3[c]	DPS	1,241	,277	20,054	1	,000	3,458	2,009	5,951
	FILIP	,872	,147	35,068	1	,000	2,391	1,792	3,190
	RANKO	,559	,113	24,621	1	,000	1,749	1,402	2,181
	Constant	-6,066	,536	128,067	1	,000	,002		
Step 4[d]	DPS	,968	,294	10,844	1	,001	2,633	1,480	4,685
	FILIP	,661	,169	15,325	1	,000	1,937	1,391	2,697
	RANKO	,496	,116	18,281	1	,000	1,643	1,308	2,063
	MILO	,379	,140	7,309	1	,007	1,460	1,110	1,921
	Constant	-6,359	,567	125,668	1	,000	,002		

Model koji smo dobili forward procedurom, na osnovu WALD statistika, je identifikovao 4 varijable u četvrtom koraku, ili tačnije, na osnovu ovog modela mi možemo računati verovatnoću da (ne)će svaki pojedinac imati poverenje u institucije na osnovu njegovog stava prema DPS-u, i ocene koju daje Filipu, Ranku i Milu. Kao i slučaju regresione analize, postavlja se pitanje koliko je model pouzdan, tačnije, sa koliko preciznosti moćemo računati distribuciju na zavisnoj varijabli a na osnovu ove četiri prediktorske varijable. U praksi se koriste dva osnovna statistika. Evo podataka u našem slučaju:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	454,998	,337	,490
2	394,425	,400	,582
3	374,609	,420	,610
4	367,136	,427	,621

← Pseudo R^2

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2,578	3	,461
2	8,498	6	,204
3	3,783	6	,706
4	1,202	6	,977

←

$$X^2_{HL} = \sum_{k=1}^{g} \frac{(O_{1k} - E_{1k})^2}{E_{1k}(1 - \xi_k)}$$

↑

Obrnuta nulta hipoteza

Dakle, najpre, budući da se u slučaju logističke regresije ne može precizno izračunati r^2, u slučaju logističke regresije koristi se tzv. pseudo R^2, koji predstavlja aproksimativnu vednost r^2. SPSS u ovom slučaju obezbeđuje vrednsot pseudo R^2 za dve različite matematičke formule koje se zovu Cox & Snell i Negelkerke (U tabeli se mogu videti značajne razlike pseudo R^2 između ove dve formule. Interpretacija je, međutim, identična kao i kada je slučaj sa regresionom analizom, naime, model je 'bolji' ukoliko se prediktoskim varijablama objašnjava što veći procenat varijanse zavise varijable.

Drugi način da se utvrdi koliko je model adekvatan jeste Hosmer and Lemshow goodnes-of-fit. Na osnovu formule kojom operiše ovaj test, može se videti da je reč o obrnutoj nultoj hipotezi, dakle, interpretacija vrednosti x^2 mora da se uskladi sa ovom činjenicom.

U našem primeru, i na osnovu pseudo R^2 i na osnovu Hosmer and Lemshow testa, moćemo videti da četvrti model koji smo dobili u forward proceduri jeste sasvim zadovoljavajući.

Naravno, kao i u svim drugim prediktorskim analizama, jedno od ključnih pitanja jeste koliko je zasta predikcija precizna. Jedan od najjednostavnijih načina jeste da uporedimo rezultate predikcije (snimljene skorove) sa autentičnom vaijablom. U našem slučaju to izgleda ovako:

Classification Table[a]

Observed			Predicted		
			POVERENJE_INSTITUCIJE (Binned)		
			NEMA POVERENJE	IMA POVERENJE	Percentage Correct
Step 1	POVERENJE_ INSTITUCIJE (Binned)	NEMA POVERENJE	411	32	92,8
		IMA POVERENJE	60	102	62,8
	Overall Percentage				84,7
Step 2	POVERENJE_ INSTITUCIJE (Binned)	NEMA POVERENJE	408	35	92,1
		IMA POVERENJE	52	111	68,1
	Overall Percentage				85,6
Step 3	POVERENJE_ INSTITUCIJE (Binned)	NEMA POVERENJE	403	40	91,0
		IMA POVERENJE	41	121	74,5
	Overall Percentage				86,6
Step 4	POVERENJE_ INSTITUCIJE (Binned)	NEMA POVERENJE	403	40	91,0
		IMA POVERENJE	40	123	75,6
	Overall Percentage				86,9

a. The cut value is ,500

Na osnovu ove tabele možemo jasno videti da u finalnom modelu (Step 4), od svih za koje smo predvideli na osnovu četiri varijable da bi imali poverenje uinstitucije, ova predikcija bi bila tačna u 75,6% slučajeva a za sve za koje bi predvideli da nemaju poverenje, bili bi u pravu za 91% slučajeva, prosečno predviđanje je precizno u 86,9% slučajeva.

Način prikazivanja konačnih rezultata logističke regresije može izgledati ovako:

PREDIKTORI	B koeficijenti
Constanta	−6,36**
DPS	0,97**
FILIP	0,66**
RANKO	0,50**
MILO	0,38**

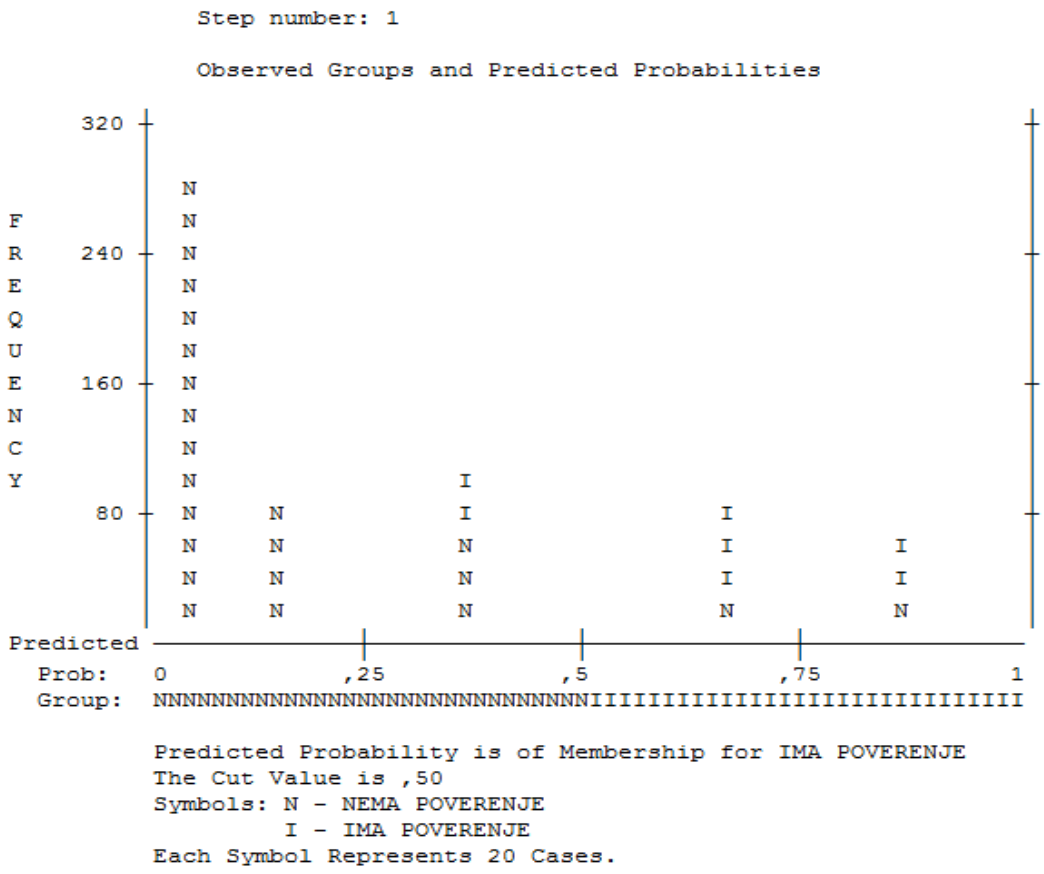
** $p < 0,01$

$$Pseudo - R^2 = 0,43$$

Hosmer and Lemeshow test

$$x^2 = 1,20; df = 6; p = 0,98$$

Jedan od načina testiranje modela jeste koričćenje klasifikacionog grafikona:



Takođe, možemo koristiti i jednostavnu kontingencionu tabelu za analizu opserviranih i predviđenih vrednosti:

POVERENJE_INSTITUCIJE (Binned) * Predicted group Crosstabulation

			Predicted group		
			NEMA POVERENJE	IMA POVERENJE	Total
POVERENJE_ INSTITUCIJE (Binned)	NEMA POVERENJE	Count	548	61	609
		% within POVERENJE_ INSTITUCIJE (Binned)	90,0%	10,0%	100,0%
	IMA POVERENJE	Count	66	206	272
		% within POVERENJE_ INSTITUCIJE (Binned)	24,3%	75,7%	100,0%
Total		Count	614	267	881
		% within POVERENJE_ INSTITUCIJE (Binned)	69,7%	30,3%	100,0%

Kao i u slučaju linearne regresije, SPSS nam omogućava da identifikujemo slučajeve sa lošom predikcijom:

Casewise List[b]

Case	Selected Status[a]	Observed: POVERENJE_INSTITUCIJE (Binned)	Predicted	Predicted Group	Temporary Variable: Resid	Temporary Variable: ZResid
89	S	I**	,008	N	,992	11,151
142	S	N**	,871	I	-,871	-2,600
157	S	I**	,134	N	,866	2,545
170	S	N**	,908	I	-,908	-3,142
211	S	I**	,055	N	,945	4,136
229	S	N**	,908	I	-,908	-3,142
233	S	I**	,082	N	,918	3,340
250	S	N**	,908	I	-,908	-3,142
293	S	I**	,008	N	,992	11,151
301	S	I**	,008	N	,992	11,151
329	S	I**	,068	N	,932	3,717
420	S	I**	,096	N	,904	3,076
438	S	N**	,908	I	-,908	-3,142
504	S	I**	,123	N	,877	2,670
505	S	I**	,134	N	,866	2,545
507	S	I**	,015	N	,985	8,012
559	S	I**	,134	N	,866	2,545
576	S	N**	,908	I	-,908	-3,142
578	S	I**	,079	N	,921	3,423
617	S	I**	,096	N	,904	3,076
638	S	N**	,908	I	-,908	-3,142
668	S	N**	,908	I	-,908	-3,142
768	S	N**	,908	I	-,908	-3,142
817	S	N**	,908	I	-,908	-3,142
830	S	N**	,908	I	-,908	-3,142
841	S	I**	,008	N	,992	11,151
842	S	I**	,008	N	,992	11,151
920	S	I**	,102	N	,898	2,971
957	S	I**	,111	N	,889	2,832
967	S	I**	,079	N	,921	3,423
992	S	N**	,908	I	-,908	-3,142

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Evo jednog primera efektivne primene logistilčke regresione analize u političkim istraživanjima. Godine 2006 održan je referendum o nezavisnosti Crne Gore. Politički, tenzije su postojale a ulog je bio veliki, kako za pristalice samostalne Crne Gore tako i za pristalice zajednice sa Srbijom. Jednako, i odgovornost svih nas koji smo realizovali istraživanja javnog mnjenja sa ciljem da previdimo rezultat na referendumu bio je neuporedivo veći u odnosu na bilo koji drugu istraživačku situaciju. Poslednje istraživanje koje smo obavili 15 dana pred referendum, a koje je realizovano na slučajnom uzorku od 1481 grašana Crne Gore je pokazalo sledeće rezultate:

"@elite li da republika Crna Gora bude nezavisna dr`ava sa punim me|unarodno-pravnim subjektivitetom"?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	DA	671	45,3	45,3	45,3
	NE	519	35,1	35,1	80,4
	Nije odlu~io/la	237	16,0	16,0	96,4
	Ne}e iza}i na referendum	53	3,6	3,6	100,0
	Total	1481	100,0	100,0	

Najveći probel movog podatka bio je visok procenat onih koji nisu odlučili, tačnije, onise nisu mogli jednostavno izbaciti iz distribucije, niti je bilo moguće da se pretpostavi da su oni ravnomerno distribuirani na tri kategorije, a jednako, nije bilo moguće objaviti finalnu predikciju u kojoj bi se reklo da nije poznato za koga će glasati 16% građana. Jedini način bio je da se koristi logistička regresiona analiza kako bi se utvrilo šta će se doista dogoditi sa nepredeljenima na referendumu. Evo rezultata analize koju smo sproveli:

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	glasao_unioniste	421,272	1	,000
		na_zna_koga_bi_glasao	40,155	1	,000
		ne_bi_glasao	39,895	1	,000
		neutralni_cvrstina	139,893	1	,000
		Nije_glasao	80,382	1	,000
		Crnogorac	232,835	1	,000
		Srbin	594,842	1	,000
		Bosnjak_Musliman	57,689	1	,000
		Albanac	75,559	1	,000
	Overall Statistics		980,365	9	,000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13,599	6	,344

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	716,515[a]	,591	,790

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,001.

Classification Table[a]

Observed			Predicted		
			ZA		Percentage Correct
			,00	1,00	
Step 1	ZA	,00	735	75	90,8
		1,00	67	604	90,0
	Overall Percentage				90,4

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	glasao_unioniste	-4,766	,411	134,362	1	,000	,009	,004	,019
	na_zna_koga_bi_glasao	-1,421	,266	28,627	1	,000	,241	,143	,406
	ne_bi_glasao	-1,183	,304	15,136	1	,000	,306	,169	,556
	neutralni_cvrstina	-3,741	,341	120,097	1	,000	,024	,012	,046
	Nije_glasao	-2,053	,216	90,494	1	,000	,128	,084	,196
	Crnogorac	1,240	,465	7,112	1	,008	3,455	1,389	8,594
	Srbin	-1,909	,506	14,233	1	,000	,148	,055	,400
	Bosnjak_Musliman	1,224	,504	5,914	1	,015	3,402	1,268	9,128
	Albanac	2,143	,651	10,838	1	,001	8,524	2,380	30,527
	Constant	1,512	,457	10,967	1	,001	4,535		

a. Variable(s) entered on step 1: glasao_unioniste, na_zna_koga_bi_glasao, ne_bi_glasao, neutralni_cvrstina, Nije_glasao, Crnogorac, Srbin, Bosnjak_Musliman, Albanac.

Na osnovu ovih rezultata, formirana je sledeća predikcija:

"@elite li da republika Crna Gora bude nezavisna dr`ava sa punim me|unarodno-pravnim subjektivitetom"?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	DA	726	49,0	56,3	56,3
	NE	563	38,0	43,7	100,0
	Total	1289	87,0	100,0	
Missing	Ne}e iza}i na referendum	192	13,0		
Total		1481	100,0		

Ova predikcija, kao što je poznato bila je precizna sa intervalom manjim od 1% greške. Na veoma sličan način, urašena je precizna predikcija i predsedničkih izbor akoji su održani 2008 i parlamentarnih koji su održani 2009 godine. Dakle, logistička regresija je nesumnjivo primenljiva u vrlo konkretnim slučajevima kada je reč o političkim istraživanjima.