NAJNOVIJA 2019/2020 god

Skripta iz **statistike** za **usmeni**

autor: **Časlav** Pejdić

(kompletna pitanja za usmeni)

usmeni

1. Pojam, značaj i oblasti statistike

Reč statistika ima 2 značenja. U svakodnevnom životu reč statistika se odnosi na numeričke podatke. Primeri takvih numeričkih podataka su prihod porodice, starost studenta, procenat ubačenih slobodnih bacanja i početna plata pripravnika sa fakultetskom diplomom. Drugo značenje reči statistika odnosi se na statistiku kao naučnu disciplinu. U tom smile statistika se definiše na sledeći način:

Statistika je naučni metod koji se koristi za prikupljanje, prikazivanje, analizu i interpretaciju podataka i donošenje statističkih zaključaka. Statistika ima dva aspekta: teorijski i primenjeni. Teorijska ili matematička statistika bavi se razvojem, izvođenjem i dokazivanjem statističkih teorema, metoda, formula, pravila i zakona. Primenjena statistika podrazumeva primenu tih teorema, metoda, formula, pravila i zakona u rešavanju realnih problema. Primenjena statistika se deli na dve oblasti: deskriptivnu i inferencijalnu statistiku (statističko zaključivanje).

Deskriptivna statistika se sastoji od metoda prikupljanja, sređivanja, prikazivanja i opisivanja podataka pomoću tabela, grafikona i sumarnih pokazatelja.

Inferencijalna statistika obuhvata statističke metode koje primenjujemo da bismo na osnovu rezultata iz dela osnovnog skupa (uzorka) došli do zaključaka o karakteristikama osnovnog skupa. Inače osnovni skup je skup svih elemenata na kojima se pojava statistički posmatra, a deo tog osnovnog skupa koji se sastoji od određenog broja elemenata naziva se uzorkom.

2. Osnovni skup i uzorak (vrste uzoraka)

Osnovni skup ili ciljna populacija se sastoji od svih elemenata ili jedinica posmatranja čije karakteristike ispitujemo – pojedinaca (bića), stvari ili predmeta. Deo osnovnog skupa koji je izabran u svrhe statističke analize naziva se uzorkom. Prikupljanje podataka o elementima osnovnog skupa ili uzorka može se sprovoditi putem ankete. Anketa koja uključuje sve elemente osnovnog skupa naziva se popisom. Ako je anketa sprovedena na uzorku onda se naziva uzoračka anketa. U zavisnosti od toga kako se bira uzorak, on može biti slučajan ili neslučajan. Slučajan uzorak je uzorak u kome elementi prilikom izbora iz osnovnog skupa imaju određenu unapred poznatu, verovatnoću različitu od nule da budu uključeni u uzorak. Kod neslučajnog (namernog) uzorka neki elementi osnovnog skupa nemaju šansu da budu uključeni u uzorak. Dva tipa namernih uzoraka su pogodan uzorak i uzorak zasnovan na subjektivnom sudu istraživača. Kod pogodnog uzorka biraju se najdostupniji elementi osnovnog skupa da bi se brzo došlo do rezultata. Kod uzorka zasnovanog na subjektivnom sudu istraživača, elementi iz osnovnog skupa biraju se na osnovu procene i prethodnog znanja nekog istraživača. Još jedna vrsta uzoraka je kvota uzorak. Kod njega se osnovni skup deli na podskupove na osnovu određenih karakteristika. Zatim iz svakog podskupa biramo poduzorak, tako da svaki poduzorak ima učešće u uzorku proporcionalno učešću tog podskupa u osnovnom skupu.

3. Tehnike izbora slučajnog uzorka

Postoje 4 tehnike za izbor slučajnog uzorka:

1) Prost slučajan uzorak – Ovo je tehnika u kojoj svaki uzorak iste veličine ima istu verovatnoću da bude izabran.

2) Sistematski slučajan uzorak – U sistematskom slučajnom uzorku slučajnim putem izaberemo jedan element iz grupe prvih k elemenata, a zatim u uzorak uključujemo svaki k-ti element, počevši od prvog izabranog elementa koji je uključen u uzorak.

3) Stratifikovan slučajan uzorak – Kod stratifikovanog slučajnog uzorka najpre osnovni skup podelimo na podskupove, koji se zovu stratum. Zatim iz svakog stratum biramo po jedan prost slučajan uzorak. Unija svih prostih slučajnih uzoraka iz svih stratuma predstavlja stratifikovani slučajan uzorak. Veličine uzoraka koji se biraju moraju da budu proporcionalne veličinama stratum iz kojih su izabrani.
4) Klaster uzorak – Kod klaster uzorka osnovni skup se prvo podeli na (geografske) grupe koje se zovu klasteri. Svaki klaster je reprezentativan za osnovni skup. Zatim, slučajno izaberemo klastere. Konačno, iz svakog od izabranih klastera slučajnim putem biramo elemente, koji čine klaster uzorak.

4. Osnovni pojmovi statistike (jedinica posmatranja, promenljiva, opservacija, serija podataka)

Jedinica posmatranja ili element uzorka ili osnovnog skupa jeste određeni subjekat ili objekat o kojem se prikupljaju podaci, odnosno, na kojem se određena pojava statistički posmatra. Promenljiva (obeležje ili varijabla) je osobina (karakteristika) koja se proučava ili istražuje, koja podrazumeva različite vrednosti po jedinicama posmatranja. Za razliku od promenljive, vrednost konstante je uvek fiksna. Vrednost promenljive koja se odnosi na jednu jedinicu posmatranja naziva se opservacijom ili podatkom. Skup podataka koji se odnosi na jednu ili vise promenljivih naziva se serija podataka. Tako na promer, ako posmatramo sve porodice u Srbiji po broju članova domaćinstva, jedinica posmatranja je porodica, promenljiva je broj članova domaćinstva, vrednost promenljive za svaku porodicu je podatak, a svi podaci zajedno su serija podataka.

5. Vrste promenljivih i merne skale

Promenljiva može biti kvantitativna (numerička) ili kvalitativna (atributivna). Kvantitativna promenljiva je promenljiva koja se može brojčano izraziti. Podaci prikupljeni o kvantitativnoj promenljivi nazivaju se kvantitativnim podacima. Kvantitativne promenljive se dele na prekidne i neprekidne. Prekidna promenljiva je promenljiva čiju vrednost možemo da brojimo, što znači da ima samo izolovane vrednosti, najčešće cele brojeve, a ne i međuvrednosti. Neprekidna promenljiva može uzeti bilo koju vrednost u određenom interval ili intervalima.
Promenljiva koja ne može uzeti numeričke vrednosti, ali može da se klasifikuje u dve ili vise kategorija jeste kvalitativna (atributivna) ili kategorijska promenljiva. Podaci prikupljeni o ovakvoj promenljivoj nazivaju se kvalitativnim podacima.
Postoje 4 nivoa merenja i 4 merne skale: nominalna, ordinalna, intervalna i skala odnosa.
Nominalna skala je najnepreciznija. U ovoj skali brojevi se koriste kod pojava koje se mogu klasifikovati samo na određeni broj i tip modaliteta. Tako se klasifikuju na primer pol i bračno stanje (1 neoženjen-neudata, 2 oženjen-udata, 3 udovac-udovica, 4 razveden-razvedena). Dakle broj ne iskazuje kvalitet, već je upotrebljen samo kao simbol. Ordinalna skala svodi merenje modaliteta na njihovo rangiranje po značaju s obzirom na usvojene kriterijume i to brojevima koji označavaju rang, ali ne pokazuju veličinu njihovog razlikovanja. Na primer stepen stručne spreme (1-visoka, 2-viša, 3-srednja i 4-niža).
Intervalna i skala odnosa pokazuju ne samo redosled modaliteta nego i meru njihovog razlikovanja. Pri tome obaveštenje o apsolutnim razlikama omogućuje intervalna, a o relativnim skala odnosa. Intervalnu skalu karakteriše određena jedinica mere, tako da za dva merenja možemo utvrditi apsolutnu razliku izraženu u jedinici mere. Najviši nivo merenja se postiže primenom skale odnosa, koja obezbeđuje značenje bilo kog odnosa merenih objekata, kao što su visina u centimetrima,

telesna masa u kilogramima i slično. Razlika između skale odnosa i intervalne je i u tome što nula kod skale odnosa znači da pojave nema, a kod intervalne da ima. Na promer težina nula znači da težine nema, a temperatura nula ne znači da temperature nema. Zato temperatura i pripada intervalnoj, a težina skali odnosa.

6. Strukturne serije i vremenske serije

Serije podataka se po načinu formiranja i analitičkom sadržaju dele na: strukturne serije i vremenske serije.

Strukturne serije pokazuju raspored statističkog skupa po modalitetima, odnosno po vrednostima obeležja. Sastoje se iz dva reda obaveštenja. U jednom su modaliteti, a u drugom broj jedinica, odnosno frekvencije (učestalost), koje pokazuju koliko se puta pojedini modaliteti javljaju unutar posmatranog statističkog skupa. U zavisnosti od vrste obeležja na osnovu kojega je izvršeno grupisanje podataka, postoje strukturne serije sa atributivnim i strukturne serije sa numeričkim obeležjem. Kod strukturnih serija sa atributivnim obeležjem, vrednosti obeležja se ne mogu izraziti brojčano, a kod strukturnih serija sa numeričkim obeležjem, vrednosti obeležja su brojevi.

Vremenske serije su nizovi statističkih podataka koji pokazuju varijacije pojava tokom vremena. One se kao i strukturne serije prikazuju u dva niza, s tim što se kod ovih serija prvi niz uvek odnosi na vreme (godina, kvartal, mesec,…), a drugi na veličinu (nivo) pojave u posmatranom period. Prema prirodi podataka koje sadrže, vremenske serije se dele na momentne i intervalne. Vremenske serije koje pokazuju veličinu ili nivo pojave u tačno određenim sukcesivnim momentima vremena nazivaju se momentnim serijama. Intervalne vremenske serije pokazuju tok (kretanje) pojave u sukcesivnim vremenski intervalima: izdaci za hranu pod anima, proizvodnja uglja po mesecima …

7. Izvori podataka, popis i uzoračka anketa, slučajne i neslučajne greške

Izvori podataka mogu da se podele u tri kategorije: interni izvori, eksterni izvori, ankete i eksperimenti. Podatke često dobijamo iz internih izvora. Takvi su podaci koje preduzeće poseduje o svojim zaposlenima ili knjigovodstveni podaci preduzeća. Međutim, ako se analiza oslanja na podatke van preduzeća to su eksterni izvori. Podaci dobijeni iz eksternih izvora mogu da budu primarni ili sekundarni. Podaci dobijeni od organizacije koja ih je prikupila nazivaju se primarni podaci. Podaci dobijeni od izvora koji ih nije prikupio su sekundarni podaci. Ponekad potrebne podatke ne možemo dobiti ni iz internih ni iz eksternih izvora, pa se sprovode sopstvene ankete ili eksperimenti. U anketi se podaci prikupljaju od elemenata osnovnog skupa ili uzorka, bez neke kontrole faktora koji mogu da utiču na karakteristike koje nas interesuju ili rezultate ankete. Anketno istraživanje može da bude popis ili uzoračka anketa.

Anketa koja obuhvata svaki element posmatranja osnovnog skupa naziva se popis. U praksi se popis retko koristi, jer je to veoma skupo i dugotrajno istraživanje. Pored toga, u mnogim slučajevima je nemoguće identifikovati svaki element posmatranja ciljne populacije. Obično se za sprovođenje istraživanja izabere deo osnovnog skupa. Taj deo osnovnog skupa naziva se uzorak. Postupak prikupljanja informacija iz dela osnovnog skupa naziva se uzoračka anketa. Anketa može da se sprovede tako što svaki ispitanik odgovara lično: telefonskim putem, putem pošte i intervjuom. Intervju ima prednost zbog velikog odziva i kvaliteta dobijenih odgovora. Međutim, to je najskuplja i najdugotrajnija tehnika. Telefonska anketa takođe daje veliki odziv. Ona je jeftinija i kraće traje od intervjua. Međutim problem sa telefonskom anketom je u tome što mnogi ljudi ne vole da ih neko zove na kućni broj, a osim toga oni koji nemaju telefon izostavljeni su iz nje. Anketa poštom je najjeftiniji metod, ali je odziv obično veoma nizak.

U eksperimentu se podaci prikupljaju od jedinica posmatranja osnovnog skupa ili uzorka, uz kontrolu faktora koji utiču na karakteristike koje nas interesuju ili rezultate eksperimenta. Na primer, ako testiramo novi lek, pacijente ćemo podeliti u dve grupe: kontrolna grupa- članovi grupe ne dobijaju taj lek, eksperimentalna grupa – članovi grupe dobijaju pravi lek.

Anketa koja obuhvata svaki element posmatranja osnovnog skupa naziva se popis. U praksi se popis retko koristi, jer je to veoma skupo i dugotrajno istraživanje. Pored toga, u mnogim slučajevima je nemoguće identifikovati svaki element posmatranja ciljne populacije. Obično se za sprovođenje istraživanja izabere deo osnovnog skupa. Taj deo osnovnog skupa naziva se uzorak. Postupak prikupljanja informacija iz dela osnovnog skupa naziva se uzoračka anketa. Anketa može da se sprovede tako što svaki ispitanik odgovara lično: telefonskim putem, putem pošte i intervjuom. Intervju ima prednost zbog velikog odziva i kvaliteta dobijenih odgovora. Međutim, to je najskuplja i najdugotrajnija tehnika. Telefonska anketa takođe daje veliki odziv. Ona je jeftinija i kraće traje od intervjua. Međutim problem sa telefonskom anketom je u tome što mnogi ljudi ne vole da ih neko zove na kućni broj, a osim toga oni koji nemaju telefon izostavljeni su iz nje. Anketa poštom je najjeftiniji metod, ali je odziv obično veoma nizak.

U eksperimentu se podaci prikupljaju od jedinica posmatranja osnovnog skupa ili uzorka, uz kontrolu faktora koji utiču na karakteristike koje nas interesuju ili rezultate eksperimenta. Na primer, ako testiramo novi lek, pacijente ćemo podeliti u dve grupe: kontrolna grupa- članovi grupe ne dobijaju taj lek, eksperimentalna grupa – članovi grupe dobijaju pravi lek.

Uzorci izabrani iz istog osnovnog skupa se međusobno razlikuju budući da imaju različite elemente, pa prema tome i različite vrednosti. Takođe, rezultati dobijeni na bazi bilo kog uzorka će se razlikovati od rezultata iz osnovnog skupa. Razlika između vrednosti statistike uzorka i parametra osnovnog skupa se naziva slučajnom greškom. Ta razlika predstavlja slučajnu grešku pod uslovom da je reč o slučajnom uzorku i da nije napravljena neka neslučajna greška. Na primeru aritmetičke sredine važi: $slučajna\ greška = \bar{x} - \mu$.

Postoji samo jedna vrsta slučajne greške i ona nastaje usled slučajnog izbora elemenata u uzorak.

Za razliku od slučajnih grešaka, postoje greške koje nastaju iz nekih drugih razloga, kao što su greške prilokom prikupljanja podataka i greške prilikom unosa podataka u tabele i nazivaju se neslučajne (sistematske) greške. Ove greške ne nastaju slučajnim putem, već je njihov uzročnik ljudski faktor.

Glavni razlozi nastanka neslučajnih grešaka su:

1) ako uzorak nije slučajan, razultati na osnovu uzorka mogu se znatno razlikovati od rezultata popisa

2) pitanja mogu biti tako formulisana da nisu razumljiva svim ispitanicima iz uzorka ili iz osnovnog skupa, usled čega se dobijeni odgovori ne mogu smatrati validnim

3) ispitanici mogu namerno da daju netačne informacije kao odgovore na neka osetljiva pitanja

4) anketar može jednostavno napraviti grešku i prilikom evidentiranja podataka ili njihovog unosa u bazu podataka

8. Sređivanje i grafičko prikazivanje kvantitativnih i kvalitativnih podataka

Podaci zapisani redosledom kojim se prikupljaju , pre nego što se urede ili grupišu nazivaju se negrupisani podaci.

Za sređivanje kvalitativnih podataka koristi se raspodela frekvencija, raspodela relativnih frekvencija

i procentualna raspodela. Raspodela frekvencija za kvalitativne podatke sadrži dva niza informacija: kategorije (modalitete) i odgovarajući broj jedinica posmatranja koji pripada svakoj kategoriji (frekvencije). Relativna frekvencija jedne kategorije se dobija kada se njena frekvencija podeli sa sumom svih frekvencija. Prema tome, relativna frekvencija pokazuje koji deo ili procenat sume frekvencija pripada odgovarajućoj kategoriji. Raspodela relativnih frekvencija pokazuje relativnu učestalost svih kategorija. Učešće (procenat) jedne kategorije dobija se kada se njena relativna frekvencija pomnoži sa 100. Procentualna raspodela je raspodela učešća svih kategorija.
Za grafičko prikazivanje kvalitativnih podataka koriste se sledeća dva dijagrama: štapićasti dijagram i strukturni krug (pita).
Da bismo nacrtali štapićasti dijagram, na x-osu nanosimo različite kategorije kvalitativne promenljive. Frekvencije nanosimo na y-osu. Za svaku kategoriju ucrtavamo po jedan stubić čija visina odgovara frekvenciji posmatrane kategorije. Štapićasti dijagram relativnih frekvencija i procentualne raspodele mogu se nacrtati jednostavno tako što ćemo na y-osu umesto frekvencija naneti relativne frekvencije ili učešća.
Strukturni krug (pita) češće se koristi za prikazivanje učešća, iako se može koristiti i za prikazivanje frekvencija ili relativnih frekvencija. Ceo strukturni krug predtavlja veličinu uzorka ili osnovnog skupa. Pita se deli na delove različite veličine, koji predstavljaju učešća različitih kategorija.
I numeričke podatke možemo urediti pomoću raspodele frekvencija, raspodele relativnih frekvencija i procentualne raspodele.
Raspodela frekvencija za numeričke podatke sadrži dva niza informacija. Vrednosti promenljive prikazane grupnim intervalima i njima odgovarajuće brojeve jedinica posmatranja. Podaci prikazani raspodelom frekvencija nazivaju se grupisanim podacima. Granična vrednost grupnog intervala predstavlja sredinu između gornje granice jednog grupnog intervala i donje granice narednog grupnog intervala. Razlika između dve granične vrednosti jednog grupnog intervala. Sredina grupnog intervala dobija se kada se zbir dve granice (ili granične vrednosti) jednog grupnog intervala podeli sa 2.
Prilikom tabelarnog prikazivanja raspodele frekvencija moramo doneti sledeće tri odluke:
1) Koliki će biti broj grupnih interval
2) Kolika će biti širina grupnih interval
3) Kolika će biti donja granica prvog grupnog intervala ili početna tačka
Broj grupnih intervala uglavnom zavisi od broja podataka i najčešće varira između 5 i 20. Često se za određivanje broja interval koristi Sturdžesovo pravilo koje glasi:
$C = 1 + 3{,}3 \cdot logN$, gde C predstavlja broj grupnih intervala, a N ukupan broj podataka.
Grupni intervali mogu biti različite širine, mada se najčešće uzima da je širina jednaka. Najčešće se širina određuje po formuli:
$$aproks.\check{s}irina\ grupnog\ intervala = \frac{najve\acute{c}a\ vrednost - najmanja\ vrednost}{broj\ grupnih\ intervala}.$$
Za početnu tačku može da se uzme bilo koji podesan broj manji ili jednak najmanjoj vrednosti.
Relativna frekvencija grupnog interval se dobija tako što se frekvencija tog interval podeli sa zbirom svih frekvencija.
Učešće se dobija kada relativnu frekvenciju grupnogintervala pomnožimo sa 100.
Za grafičko prikazivanje grupisanih podataka koriste se histogram i poligon.
Kod histograma na x-osu nanosimo grupne interval, a na y-osu frekvencije (ili relativne frekvencije

ili učešća). Zatim za svaki grupni interval ucrtavamo pravougaonik čija visina predstavlja frekvenciju grupnog interval. Pravougaonici se crtaju jedan do drugog bez razmaka. Dobijeni dijagram se naziva histogramom frekvencija ili histogramom relativnih frekvencija ili histogramom učešća.
Da bi se nacrtao poligon frekvencija u dijagram najpre ucrtavamo tačku iznad sredine svakog grupnog interval u visini koja odgovara frekvenciji tog intervala. Na kraju, ucrtavamo tačku za sredinu nepostojećeg interval koji prethodi prvom, kao i tačku za sredinu interval koji sledi iza poslednjeg grupnog interval. Frekvencije tih zamišljenih interval su uvek jednake nuli. Na kraju, spajamo sve ucrtane susedne tačke. Akon a y-osu nanosimo relativne frekvencije onda je to poligon relativnih frekvencija, a ako nanosimo učešća onda je to poligon učešća.

9. Raspodela frekvencija (apsolutnih, relativnih i kumulativnih) i njihovo grafičko prikazivanje

Raspodela frekvencija za numeričke podatke sadrži dva niza informacija. Vrednosti promenljive prikazane grupnim intervalima i njima odgovarajuće brojeve jedinica posmatranja. Podaci prikazani raspodelom frekvencija nazivaju se grupisanim podacima. Granična vrednost grupnog intervala predstavlja sredinu između gornje granice jednog grupnog intervala i donje granice narednog grupnog intervala. Razlika između dve granične vrednosti jednog grupnog intervala. Sredina grupnog intervala dobija se kada se zbir dve granice (ili granične vrednosti) jednog grupnog intervala podeli sa 2.
Prilikom tabelarnog prikazivanja raspodele frekvencija moramo doneti sledeće tri odluke:
1) Koliki će biti broj grupnih interval
2) Kolika će biti širina grupnih interval
3) Kolika će biti donja granica prvog grupnog intervala ili početna tačka
Broj grupnih intervala uglavnom zavisi od broja podataka i najčešće varira između 5 i 20. Često se za određivanje broja interval koristi Sturdžesovo pravilo koje glasi:
$C = 1 + 3{,}3 \cdot logN$, gde C predstavlja broj grupnih intervala, a N ukupan broj podataka.
Grupni intervali mogu biti različite širine, mada se najčešće uzima da je širina jednaka. Najčešće se širina određuje po formuli:
$aproks.\check{s}irina\ grupnog\ intervala = \frac{naje\acute{c}a\ vrednost - najmanja\ vrednost}{broj\ grupnih\ intervala}$.
Za početnu tačku može da se uzme bilo koji podesan broj manji ili jednak najmanjoj vrednosti.
Relativna frekvencija grupnog interval se dobija tako što se frekvencija tog interval podeli sa zbirom svih frekvencija.
Učešće se dobija kada relativnu frekvenciju grupnogintervala pomnožimo sa 100.
Za grafičko prikazivanje grupisanih podataka koriste se histogram i poligon.
Kod histograma na x-osu nanosimo grupne interval, a na y-osu frekvencije (ili relativne frekvencije ili učešća). Zatim za svaki grupni interval ucrtavamo pravougaonik čija visina predstavlja frekvenciju grupnog interval. Pravougaonici se crtaju jedan do drugog bez razmaka. Dobijeni dijagram se naziva histogramom frekvencija ili histogramom relativnih frekvencija ili histogramom učešća.
Da bi se nacrtao poligon frekvencija u dijagram najpre ucrtavamo tačku iznad sredine svakog grupnog interval u visini koja odgovara frekvenciji tog intervala. Na kraju, ucrtavamo tačku za sredinu nepostojećeg interval koji prethodi prvom, kao i tačku za sredinu interval koji sledi iza poslednjeg grupnog interval. Frekvencije tih zamišljenih interval su uvek jednake nuli. Na kraju,

spajamo sve ucrtane susedne tačke. Akon a y-osu nanosimo relativne frekvencije onda je to poligon relativnih frekvencija, a ako nanosimo učešća onda je to poligon učešća.

Raspodela kumulativnih frekvencija prikazuje ukupan broj jedinica posmatranja koje imaju vrednost ispod gornje granice svakog grupnog interval. Raspodela kumulativnih frekvencija se formira samo za kvantitativne podatke.

Kumulativne relativne frekvencije dobijaju se kada se kumulativne frekvencije podele sa zbirom svih frekvencija (ukupnim brojem podataka u skupu ili uzorku). Kumulativna učešća dobijaju se kada se kumulativne relativne frekvencije pomnože sa 100.

Grafički prikaz kumulativnih frekvencija se naziva kumulanta (ogiva). Pri crtanju kumulante, na x-osu se nanose vrednosti promenljive, odnosno gornje granice grupnih interval, a na y-osu se nanose odgovarajuće kumulativne frekvencije ili kumulativne relativne frekvencije ili kumulativna učešća.

10. Deskriptivne mere

Seriju podataka najčešće opisujemo pomoću numeričkih deskriptivnih mera, koje se obično nazivaju karakteristične vrednosti. Dele se na:

1) Mere centralne tendencije
   a) Aritmetička sredina
   b) Medijana
   c) Modus
2) Mere disperzije
   a) Interval varijacije
   b) Varijansa
   c) Standardna devijacija
   d) Koeficijent varijacije
3) Pozicione mere
   a) Kvartili
   b) Interkvartilna razlika
   c) Percentili i rang percentila

Aritmetička sredina: Aritmetička sredina je najčešće korišćena mara centralne tendencije. U slučaju negrupisanih podataka aritmetička sredina se dobija deljenjem zbira svih vrednosti sa brojem vrednosti u toj seriji. Aritmetička sredina izračunata za podatke uzorka se obeležava sa $\bar{x}$, a aritmetička sredina izračunata za podatke osnovnog skupa se obeležava sa μ . Broj vrednosti uzorka se obeležava sa n, a broj vrednosti skupa sa N.

$$\mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{\sum x}{n}$$

Ako su podaci grupisani, tj. dati u vidu tabele frekvencija, mi ne znamo vrednosti pojedinačnih opservacija. Pri izračunavanju aritmetičke sredine grupisanih podataka, prvo odredimo sredinu svakog intervala, a zatim pomnožimo sredine sa frekvencijama odgovarajućih intervala. Suma ovih proizvoda, koja se označava sa $\sum x'f$, predstavlja aproksimaciju sume svih vrednosti, pa se ova suma deli sa ukupnim brojem podataka.

$$\mu = \frac{\sum x'f}{N}$$

$$\bar{x} = \frac{\sum x\prime f}{n}$$

Vrednost aritmetičke sredine skupa je konstanta, dok vrednost aritmetičke sredine uzorka zavisi od izbora uzorka, pa je slučajna promenljiva. Glavna mana aritmetičke sredine je da je ona osetljiva na postojanje ekstremnih vrednosti.

Medijana: Medijana je jednaka vrednosti središnjeg člana serije podataka koji su rangirani u rastućem redosledu. Medijana deli seriju rangiranih podataka na dva jednaka dela. Izračunavanje medijane podrazumeva sledeća dva koraka:

1) Rangiranje podataka od najniže ka najvišoj vrednosti

2) Pronalaženje središnjeg člana. Vrednost ovog člana je jednaka medijani.

Ako je broj podataka u seriji neparan, onda je medijana jednaka vrednosti središnjeg člana rangiranih podataka. Ako je broj podataka paran, onda se ona izračunava kao aritmetička sredina dva središna podatka u seriji. Medijana određuje centar histograma, pri čemu se polovina vrednosti nalazi levo, a polovina desno od nje. Prednost korišćenja medijane kao mere centralne tendencije je u tome što na nju ne utiču ekstremne vrednosti. Dakle, ako serija ima ekstremnih vrednosti, medijana je bolji pokazatelj centralne tendencije od aritmetičke sredine.

Modus: Modus je vrednost koja se javlja sa najvećom frekvencijom u seriji podataka. Najveći nedostatak modusa je u tome što je moguće da jedna serija podataka ili nema ili ima više od jednog modusa. Ukoliko serija ima samo jedan modus onda se zove unimodalna, ako ima dva modusa bimodalna, a ako ima više od dva modusa multimodalna.

Interval varijacije: Interval varijacije je mera disperzije koju je najlakše izračunati. Jenak je razlici između najveće i najmanje vrednosti u seriji podataka. Interval varijacije, kao i aritmetička sredina, ima nedostatak da na njega utiču ekstremne vrednosti. Drugi nedostatak intervala varijacije je što se za njegovo izračunavanje koriste samo dve vrednosti: najveća i najmanja. Sve ostale vrednosti serije se zanemaruju.

Varijansa i standardna devijacija: Standardna devijacija je najčešće korišćena mera disperzije. Vredost standardne devijacije pokazuje koliko blizu su vrednosti serije podataka grupisane oko aritmetičke sredine. Uopšteno, manje vrednosti standardne devijacije ukazuju da su vrednosti te serije raspršene veoma malo oko aritmetičke sredine. Nasuprot tome, veća vrednost standardne devijacije serije podataka ukazuje da su vrednosti te serije raspršene u relativno velikom razmaku oko aritmetičke sredine.

Standardna devijacija se dobija uzimanjem kvadratnog korena varijanse. Varijansa osnovnog skupa se označava sa σ^2, a varijansa uzorka se označava sa s^2. U skladu sa tim, standardna devijacija osnovnog skupa se označava sa σ, a uzorka sa s.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \quad i \quad s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Iz formule se vidi da varijansa skupa pokazuje prosek kvadrata odstupanja vrednosti x od aritmetičke sredine, a varijansa uzorka ne. Kod varijanse uzorka se umesto n koristi n-1. Razlog je što ako bi se koristilo n, varijansa uzorka bi bila pristrasna ocena, a ako se koristi n-1, varijansa uzorka je nepristrasna ocena varijanse skupa. Inače, nepristrasna ocena znači da je u proseku jednaka parametru. Veličine $x-\mu$ i $x-\bar{x}$ se zovu devijacijom vredbnosti x od aritmetičke sredine. Interesantna je osobina da je suma devijacija vrednosti x od aritmetičke uvek jednaka 0, tj. $\sum(x-\mu) = 0 \ i \ \sum(x-\bar{x}) = 0$.

U zadacima se često koriste radne formule za izračunavanje varijanse, jer su brže za izračunavanje:

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad i \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Standardne devijacije skupa i uzorka su kvadratni koren iz varijanse: $= \sqrt{\sigma^2} \ i \ s = \sqrt{s^2}$.

Vrednosti varijanse i standardne devijacije nikada nisu negativne. Standardna devijacija se iskazuje u jedinici

mere originalnih podataka, a varijansa u jedinici mere na kvadrat.
Ukoliko su podaci grupisani koriste se sledeće formule za izračunavanje varijanse:

$$\sigma^2 = \frac{\sum f(x\prime - \mu)^2}{N} \quad i \quad s^2 = \frac{\sum f(x\prime - \bar{x})^2}{n-1}$$

odnosno radne formule:

$$\sigma^2 = \frac{\sum f x^2 - \frac{(\sum f x)^2}{N}}{N} \quad i \quad s^2 = \frac{\sum f x^2 - \frac{(\sum f x)^2}{n}}{n-1} \quad .$$

Koeficijent varijacije: Odnos standardne devijacije i aritmetičke sredine naziva se koeficijent varijacije. Označava se sa CV i pokazuje koliko iznosi standardna devijacija u procentima aritmetičke sredine.

$$CV = \frac{\sigma}{\mu} \cdot 100 \quad i \quad CV = \frac{S}{\bar{x}} \cdot 100$$

Kvartili: Kvartili su deskriptivne mere koje dele seriju podataka, rangiranih po veličini , na četiri jednaka dela. Tri mere će podeliti bilo koju seriju podataka na 4 jednaka dela. Ove tri mere su prvi kvartil (Q_1), drugi kvartil (Q_2) i treći kvartil (Q_3). Drugi kvartil je isto što i medijana. Prvi kvartil je vrednost središnjeg člana među opservacijama koje su manje od medijane, a treći kvartil vrednost središnjeg člana među opservacijama koje su veće od medijane. 25% podataka serije ima vrednost manju, a 75% veću od Q_1. 75% podataka serije ima vrednost manju, a 25% veću od Q_3.
Interkvartilna razlika: Razlika između trećeg i prvog kvartila naziva se interkvartilna razlika. Obeležava se sa IQR.

$$IQR = Q_3 - Q_1$$

Percentili i rang percentila: Percentili su deskriptivne mere koje dele seriju podataka rangiranih po veličini na 100 jednakih delova. Svaka (rangirana) serija podatka ima 99 percentila koji je dele na 100 jednakih delova. K-ti percentil se označava sa P_k, i oko k% vrednosti je manje od P_k, a (100-k)% vrednosti je veće od P_k. Približna vrednost k-tog percentila se određuje na sledeći način:

$$P_k = Vrednost \left(\frac{kn}{100}\right) \check{c}lana\ u\ rangiranoj\ seriji\ podataka$$

Rang percentila za podatak x_ix daje procenat vrednosti u seriji podataka koje su manje od x_i .

$$Rang\ percentila\ za\ x_i = \frac{Broj\ vrednosti\ manjih\ od\ x_i}{Ukupan\ broj\ podataka} \cdot 100 \ .$$

11. Mere centralne tendencije

Mere centralne tendencije su:

a) Aritmetička sredina
b) Medijana
c) Modus

Aritmetička sredina: Aritmetička sredina je najčešće korišćena mara centralne tendencije. U slučaju negrupisanih podataka aritmetička sredina se dobija deljenjem zbira svih vrednosti sa brojem vrednosti u toj seriji. Aritmetička sredina izračunata za podatke uzorka se obeležava sa $\bar{x}$, a aritmetička sredina izračunata za podatke osnovnog skupa se obeležava sa μ . Broj vrednosti uzorka se obeležava sa n, a broj vrednosti skupa sa N.

$$\mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{\sum x}{n}$$

Ako su podaci grupisani, tj. dati u vidu tabele frekvencija, mi ne znamo vrednosti pojedinačnih opservacija. Pri izračunavanju aritmetičke sredine grupisanih podataka, prvo odredimo sredinu svakog intervala, a zatim pomnožimo sredine sa frekvencijama odgovarajućih intervala. Suma ovih proizvoda, koja se označava sa $\sum x'f$, predstavlja aproksimaciju sume svih vrednosti, pa se ova suma deli sa ukupnim brojem podataka.

$$\mu = \frac{\sum x\prime f}{N}$$
$$\bar{x} = \frac{\sum x\prime f}{n}$$

Vrednost aritmetičke sredine skupa je konstanta, dok vrednost aritmetičke sredine uzorka zavisi od izbora uzorka, pa je slučajna promenljiva. Glavna mana aritmetičke sredine je da je ona osetljiva na postojanje ekstremnih vrednosti.

Medijana: Medijana je jednaka vrednosti središnjeg člana serije podataka koji su rangirani u rastućem redosledu. Medijana deli seriju rangiranih podataka na dva jednaka dela. Izračunavanje medijane podrazumeva sledeća dva koraka:

1) Rangiranje podataka od najniže ka najvišoj vrednosti
2) Pronalaženje središnjeg člana. Vrednost ovog člana je jednaka medijani.

Ako je broj podataka u seriji neparan, onda je medijana jednaka vrednosti središnjeg člana rangiranih podataka. Ako je broj podataka paran, onda se ona izračunava kao aritmetička sredina dva središna podatka u seriji. Medijana određuje centar histograma, pri čemu se polovina vrednosti nalazi levo, a polovina desno od nje. Prednost korišćenja medijane kao mere centralne tendencije je u tome što na nju ne utiču ekstremne vrednosti. Dakle, ako serija ima ekstremnih vrednosti, medijana je bolji pokazatelj centralne tendencije od aritmetičke sredine.

Modus: Modus je vrednost koja se javlja sa najvećom frekvencijom u seriji podataka. Najveći nedostatak modusa je u tome što je moguće da jedna serija podataka ili nema ili ima više od jednog modusa. Ukoliko serija ima samo jedan modus onda se zove unimodalna, ako ima dva modusa bimodalna, a ako ima više od dva modusa multimodalna.

12. Ispitivanje asimetrije raspodele (odnos između aritmetičke sredine, medijane i modusa; box plot)

Kriva raspodele frekvencija može biti simetrična ili asimetrična. Postoji više načina za utvrđivanje simetričnosti.

Jedan od načina je pomoću odnosa između srednjih vrednosti. Naime, ako je raspored simetričan, aritmetička sredina, medijana i modus su jednaki. Ako je raspored asimetričan udesno, aritmetička sredina je najveća jer je osetljiva na ekstremne vrednosti koje se nalaze na desnom kraju, modus je najmanji, a medijana je između njih. Ako je raspodela asimetrična ulevo, situacija je obrnuta, tj. aritmetička sredina je najmanja, modus je najveći, a medijana je ponovo između njih.

Drugi način za utvrđivanje asimetričnosti je pomoću koeficijenta asimetrije α_3. Ako je koeficijent asimetrije 0, raspodela je simetrična, ako je pozitivan, raspodela je asimetrična udesno, a ako je negativan raspodela je asimetrična ulevo.

Treći način za ispitivanje asimetrije je pomoću box-plot dijagrama. Box-plot je dijagram koji prikazuje centar, raspršenost i asimetriju serije podataka. Konstruiše se crtanjem pravougaonika pomoću medijane, prvog kvartila, trećeg kvartila i najmanje i najveće vrednosti u seriji podataka između donje i gornje unutrašnje granice.

Box-plot se konstruiše na sledeći način:

Povuče se horizontalna linija i na njoj se označe nivoi podataka, tako da sve vrednosti u seriji budu obuhvaćene. Iznad horizontalne linije se crta pravougaonik, čija je leva strana na poziciji prvog kvartila,

a desna strana na poziciji trećeg kvartila. Unutar pravougaonika se povuče vertikalna linija na nivou medijane. Zatim se povuku dve linije kojima se spajaju najmanja i najveća vrednost između dve unutrašnje granice sa pravougaonikom. Inače donja unutrašnja granica se dobija kada se od prvog kvartila oduzme 1,5·IQR, a gornja unutrašnja granica se dobija kada se na treći kvartil doda 1,5·IQR. 50% vrednosti serije se nalazi unutar pravougaonika, 25% levo od njega, a 25% desno. Takođe 50% podataka je sa leve strane medijane, a 50% sa desne. Vrednosti koje se nalaze izvan dve unutrašnje granice se nazivaju outliers. Oni se dele na dve vrste: umereni i ekstremni. Uvedimo nove pojmove donja spoljašnja granica, koja se dobija kada od prvog kvartila oduzmemo 3·IQR i gornja spoljašnja granica, koja se dobija kada se na treći kvartil sabere 3·IQR. Ako je vrednost iznad unutrašnjih, ali ispod spoljašnjih granica, onda je to umereni outlier, a ako je vrednost iznad spoljašnjih granica onda je to ekstremni outlier.
Ako je serija simetrična, linija medijane se nalazi tačno na sredini pravougaonika.

13. Mere disperzije

<u>Interval varijacije</u>: Interval varijacije je mera disperzije koju je najlakše izračunati. Jenak je razlici između najveće i najmanje vrednosti u seriji podataka. Interval varijacije, kao i aritmetička sredina, ima nedostatak da na njega utiču ekstremne vrednosti. Drugi nedostatak intervala varijacije je što se za njegovo izračunavanje koriste samo dve vrednosti: najveća i najmanja. Sve ostale vrednosti serije se zanemaruju.
<u>Varijansa i standardna devijacija</u>: Standardna devijacija je najčešće korišćena mera disperzije. Vredost standardne devijacije pokazuje koliko blizu su vrednosti serije podataka grupisane oko aritmetičke sredine. Uopšteno, manje vrednosti standardne devijacije ukazuju da su vrednosti te serije raspršene veoma malo oko aritmetičke sredine. Nasuprot tome, veća vrednost standardne devijacije serije podataka ukazuje da su vrednosti te serije raspršene u relativno velikom razmaku oko aritmetičke sredine.
Standardna devijacija se dobija uzimanjem kvadratnog korena varijanse. Varijansa osnovnog skupa se označava sa σ^2, a varijansa uzorka se označava sa s^2. U skladu sa tim, standardna devijacija osnovnog skupa se označava sa σ, a uzorka sa s.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \quad i \quad s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Iz formule se vidi da varijansa skupa pokazuje prosek kvadrata odstupanja vrednosti x od aritmetičke sredine, a varijansa uzorka ne. Kod varijanse uzorka se umesto n koristi n-1. Razlog je što ako bi se koristilo n, varijansa uzorka bi bila pristrasna ocena, a ako se koristi n-1, varijansa uzorka je nepristrasna ocena varijanse skupa. Inače, nepristrasna ocena znači da je u proseku jednaka parametru. Veličine $x-\mu$ i $x-\bar{x}$ se zovu devijacijom vrednosti x od aritmetičke sredine. Interesantna je osobina da je suma devijacija vrednosti x od aritmetičke uvek jednaka 0, tj. $\sum(x-\mu)=0 \; i \; \sum(x-\bar{x})=0$.
U zadacima se često koriste radne formule za izračunavanje varijanse, jer su brže za izračunavanje:

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad i \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Standardne devijacije skupa i uzorka su kvadratni koren iz varijanse: $= \sqrt{\sigma^2} \; i \; s = \sqrt{s^2}$.
Vrednosti varijanse i standardne devijacije nikada nisu negativne. Standardna devijacija se iskazuje u jedinici mere originalnih podataka, a varijansa u jedinici mere na kvadrat.
Ukoliko su podaci grupisani koriste se sledeće formule za izračunavanje varijanse:

$$\sigma^2 = \frac{\sum f(x'-\mu)^2}{N} \quad i \quad s^2 = \frac{\sum f(x'-\bar{x})^2}{n-1}$$

odnosno radne formule:

$$\sigma^2 = \frac{\sum fx^2 - \frac{(\sum fx)^2}{N}}{N} \quad i \quad s^2 = \frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1} \quad .$$

Koeficijent varijacije: Odnos standardne devijacije i aritmetičke sredine naziva se koeficijent varijacije. Označava se sa CV i pokazuje koliko iznosi standardna devijacija u procentima aritmetičke sredine.

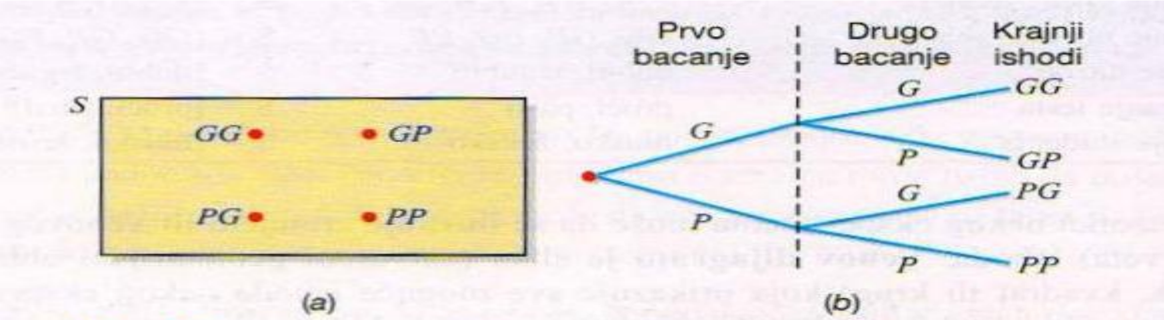
$$CV = \frac{\sigma}{\mu} \cdot 100 \quad i \quad CV = \frac{S}{\bar{x}} \cdot 100$$

14. Eksperiment, ishodi i prostor uzorka

Eksperiment je proces, čiji rezultat izvođenja je jedna i samo jedna od mnogih opservacija. Te opservacije se nazivaju ishodi eksperimenta. Skup svih ishoda jednog eksperimenta se naziva prostor uzorka.

Prostor uzorka nekog eksperimenta može da se ilustruje crtanjem ili Venovog dijagrama ili stabla ishoda. Venov dijagram je slika koja prikazuje sve moguće ishode nekog eksperimenta. Kod stabla ishoda svaki ishod je predstavljen jednom granom.

Na primer, uzmimo eksperiment bacanja novčića dva puta. Sa G ćemo označiti pojavu grba, a sa P pojavu pisma. Mogući ishodi eksperimenta su PP, PG, GP i GG. Ako prostor uzorka obeležimo sa S, možemo napisati: $S = \{PP, PG, GP, GG\}$. Venov dijagram i stablo ishoda imaju sledeći izgled:



Slika 4.2 (*a*) Venov dijagram i (*b*) stablo ishoda za dva bacanja novčića.

15. Koncepti verovatnoće

Tri konceptualna pristupa verovatnoći su:

1) Klasična verovatnoća
2) Koncept verovatnoće kao relativne frekvencije
3) Koncept subjektivne verovatnoće

Klasična verovatnoća: Mnogo puta, različiti ishodi eksperimenta mogu imati istu verovatnoću realizacije. Takvi ishodi se nazivaju jednakoverovatni ishodi. Pravilo klasične verovatnoće se primenjuje da bi se izračunale verovatnoće događaja za neki eksperiment u kome su svi ishodi jednakoverovatni. Po pravilu klasične verovatnoće, verovatnoća elementarnog događaja se dobije kada se 1 podeli sa ukupnim brojem ishoda tog eksperimenta. Za razliku od toga, verovatnoća složenog događaja A se dobija kada se broj povoljnih ishoda za događaj A podeli sa ukupnim brojem ishoda tog eksperimenta. Ako sa E_i označimo elementarni događaj, a sa A složeni događaj, možemo napisati:

$$P(E_i) = \frac{1}{Ukupan\ broj\ ishoda\ eksperimenta}$$

$$P(A) = \frac{Broj\ povoljnih\ ishoda\ za\ A}{Ukupan\ broj\ ishoda\ eksperimenta}$$

Koncept verovatnoće kao relativne frekvencije: Pretpostavimo da želimo da izračunamo sledeće verovatnoće:

1) verovatnoću da će sledeći automobil koji izađe iz fabrike biti „defektan“
2) verovatnoću da slučajno izabrana porodica poseduje kuću
3) verovatnoću da slučajno izabrana žena nikada nije pušila

Ove verovatnoće ne mogu da se izračunaju korišćenjem pravila klasične verovatnoće jer različiti ishodi odgovarajućih eksperimenata nisu podjednako verovatni. U takvim slučajevima, za izračunavanje verovatnoća, koristimo ili postojeće podatke, ili generišemo nove podatke tako što ponavljamo eksperiment veliki broj puta. Relativna frekvencija nekog događaja se koristi kao aproksimacija verovatnoće tog događaja. Ovaj metod dodeljivanja verovatnoće nekom događaju se zove koncept verovatnoće kao relativne frekvencije. Pošto se relativne frekvencije određuju izvođenjem eksperimenta, verovatnoće izračunate korišćenjem relativnih frekvencija mogu da se promene skoro svaki put kada se eksperiment ponovi. Međutim, varijacija će biti mala ako je uzorak veliki. Ako se posmatra ceo skup, relativna frekvencija daje tačnu verovatnoću.

$$P(A) = \frac{f}{n}$$

Sa n je označen ukupan broj ponavljanja eksperimenta, a sa f realizacija događaja A. Relativne frekvencije nisu verovatnoće, već su aproksimacija verovatnoće. Međutim, ako se neki eksperiment ponavlja veliki broj puta, verovatnoća događaja dobijena na osnovu relativne frekvencije teži stvarnoj ili teorijskoj verovatnoći.

Subjektivna verovatnoća: Mnogo puta se srećemo sa eksperimentima koji nemaju ni jednakoverovatne ishode niti mogu da se ponove da bi se generisali podaci. Na primer:

1) verovatnoća da će Milan, koji je student ekonomije, dobiti najbolju ocenu iz statistike
2) verovatnoća da će Crvena Zvezda sledeće godine biti prvak u fudbalu

Za izračunavanje verovatnoće ovih primera ne može da se primeni ni pravilo klasične verovatnoće ni koncept verovatnoće kao relativne frekvencije. Verovatnoća koja se dodeljuje nekom događaju u ovakvim slučajevima se zove subjektivna verovatnoća. Ona se zasniva na sopstvenom mišljenju pojedinca, iskustvu, informacijama i verovanju.

16. Slučajna promenljiva

Slučajna promenljiva je promenljiva čija vrednost je određena ishodom slučajnog eksperimenta. Slučajna promenljiva može da bude diskretna (prekidna) i neprekidna (kontinuirana).

Diskretna slučajna promenljiva je promenljiva koja uzima prebrojivo mnogo vrednosti. Na primer:

1) Broj vozila koja se prodaju kod dilera u toku određenog meseca
2) Broj kuća u jednom bloku
3) Broj riba koje se ulove tokom jednog ribolova

Neprekidna slučajna promenljiva je promenljiva koja može da uzme bilo koju vrednost iz jednog ili više intervala. Pošto je broj vrednosti sadržanih u nekom intervalu beskonačan, mogući broj vrednosti koje može da uzme neprekidna slučajna promenljiva je takođe beskonačan. Na primer:

1) Visina neke osobe
2) Vreme potrebno za izradu testa
3) Cena kuće

17. Raspodela verovatnoća diskretne slučajne promenljive i modeli diskretnih raspodela verovatnoća

Raspodela verovatnoće diskretne slučajne promenljive prikazuje listu svih mogućih vrednosti koje slučajna promenljiva može da uzme i njihovih odgovarajućih verovatnoća. Raspodela verovatnoća diskretne slučajne promenljive ima sledeće dve karakteristike:

1) $0 \leq P(x) \leq 1,$ za svaku vrednost x
2) $\sum P(x) = 1$

Ove dve karakteristike se takođe nazivaju i dva uslova koja raspodela verovatnoća mora da zadovolji. Raspodela verovatnoće diskretne slučajne promenljive može da se prikaže u formi matematičke formule, tabele ili dijagrama. Ukoliko se prikazuje pomoću dijagrama koristi se štapićasti dijagram. Najčešće korišćeni modeli diskrenih raspodela verovatnoće su:

1) Binomna raspodela
2) Hipergeometrijska raspodela
3) Poasonova raspodela

Binomna raspodela: Binomni eksperiment mora da zadovolji sledeća 4 uslova:
1) Postoji n identičnih opita
2) Svaki opit ima samo 2 moguća ishoda
3) Verovatnoće ta dva ishoda ostaju konstantne
4) Opiti su nezavisni

Slučajna promenljiva X koja predstavlja broj uspeha u n opita binomnog eksperimenta naziva se binomna slučajna promenljiva. Raspodela verovatnoće za X u takvim eksperimentima naziva se binomna raspodela verovatnoća.

Hipergeometrijska raspodela: Ako opiti nisu nezavisni, ne možemo primeniti binomnu raspodelu. U takvim slučajevima umesto binomne se koristi hipergeometrijska raspodela. Primer je recimo uzorak izvučen bez ponavljanja iz konačne populacije.

Poasonova raspodela: Za primenu Poasonove raspodele moraju biti ispunjena sledeća tri uslova:
1) X je diskretna slučajna promenljiva
2) Realizacije su slučajne
3) Realizacije su nezavisne

18. Očekivana vrednost i standardna devijacija diskretne slučajne promenljive

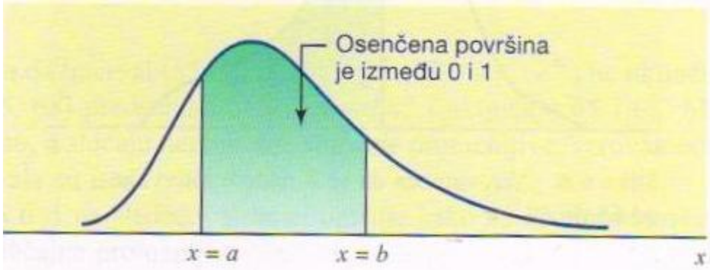
Očekivana vrednost diskretne slučajne promenljive X je vrednost koju očekujemo da se javi u proseku, prilikom izvođenja eksperimenta veliki broj puta. Označava se sa $E(X)$ i izračunava se kao: $(X) = \mu = \sum xP(x)$. Očekivana vrednost slučajne promenljive X predstavlja njenu prosečnu vrednost i takođe se obeležava i sa .

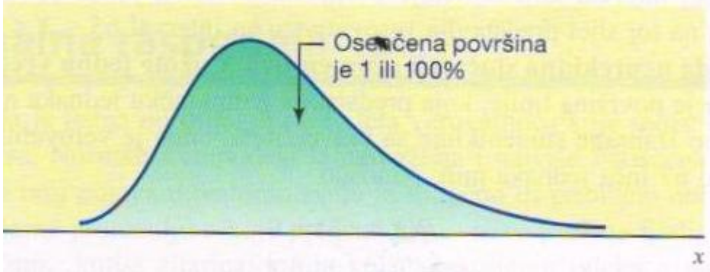
Standardna devijacija diskretne slučajne promenljive X meri raspršenost njene raspodele verovatnoća i izračunava se kao: $\sigma = \sqrt{\sum x^2 P(x) - \mu^2}$. Standardna devijacija diskretne slučajne promenljive nikada nije negativna.

19.Raspodela verovatnoća neprekidne slučajne promenljive

Raspodela verovatnoće neprekidne slučajne promenljive se prikazuje pomoću krive raspodele verovatnoće, koja se još naziva i funkcija gustine verovatnoća.

Raspodela verovatnoća neprekidne slučajne promenljive ima sledeće dve karakteristike:
1) Verovatnoća da X uzme vrednost iz bilo kog intervala se nalazi u intervalu od 0 do 1
2) Ukupna verovatnoća svih (međusobno disjunktnih) intervala u okviru kojih X može da uzme vrednost je 1





Prva karakteristika se odnosi na to da se površina ispod krive raspodele verovatnoća neprekidne slučajne promenljive, između bilo koje dve tačke, nalazi u intervalu od 0 do 1. Druga karakteristika ukazuje na to da je ukupna površina ispod krive raspodele verovatnoća neprekidne slučajne promenljive uvek 1 ili 100%. Verovatnoća da neprekidna slučajna promenljiva X uzme vrednost unutar određenog intervala jednaka je površini ispod krive između dve granice intervala. Za neprekidnu raspodelu verovatnoća, verovatnoća se uvek izračunava za neki interval. Verovatnoća da neprekidna slučajna promenljiva X uzme jednu vrednost je uvek nula.

$$P(X = a) = 0$$

Kao posledica toga kod neprekidnih slučajnih promenljivih uvek važi:

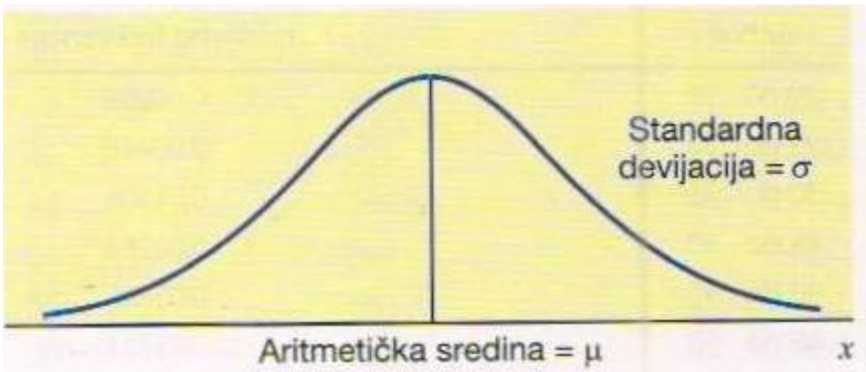
$$P(a \leq X \leq b) = P(a < X < b)$$

Drugim rečima verovatnoća da X uzme vrednost iz intervala od a do b je ista bez obzira da li su vrednosti a i b uključene u interval ili ne.

20. Normalna raspodela, osobine i značaj

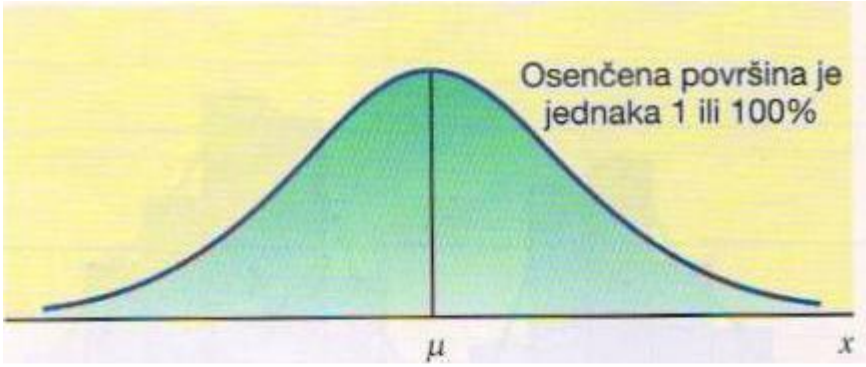
Normalna raspodela je jedna od mnogih raspodela verovatnoća koje može da ima neprekidna slučajna promenljiva. Normalna raspodela je najvažnija i najviše korišćena od svih raspodela verovatnoća. Veliki broj pojava u realnom svetu je ili tačno ili približno normalno raspoređen. Za neprekidne slučajne promenljive koje predstavljaju visine i težine ljudi, razultate na ispitu ili vreme koje je potrebno da se neki posao završi se smatra da imaju (približno) normalnu raspodelu.

Normalna raspodela verovatnoća ili normalna kriva je u obliku zvona (simetrična kriva). Kriva normalnog rasporeda ima sledeći izgled:



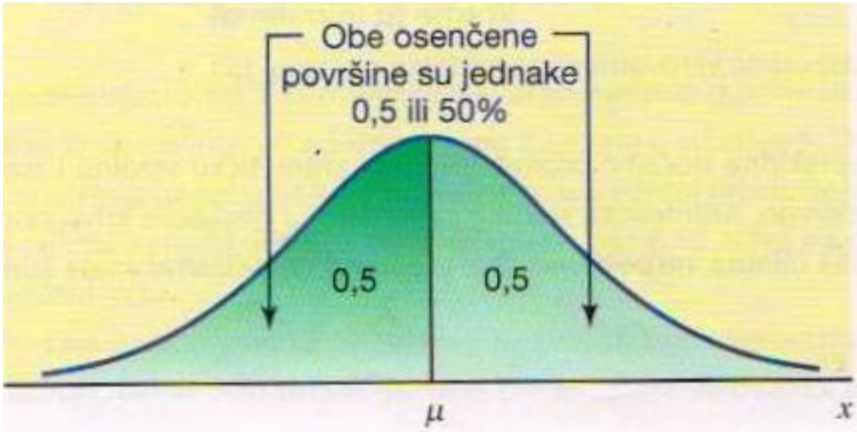
Kriva normalnog raspodela ima sledeće osobine:

1) ukupna površina ispod krive je 1



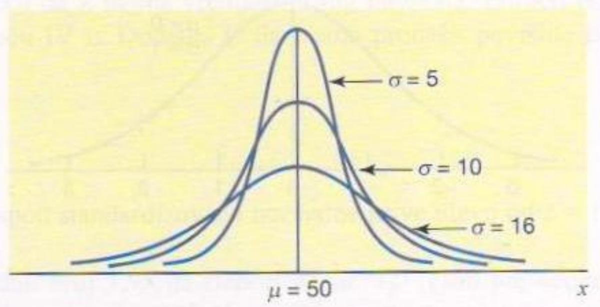
2) kriva je simetrična u odnosu na aritmetičku sredinu

To znači da se 50% ukupne površine ispod krive normalne raspodele nalazi sa leve strane aritmetičke sredine, a 50% sa desne strane aritmetičke sredine.

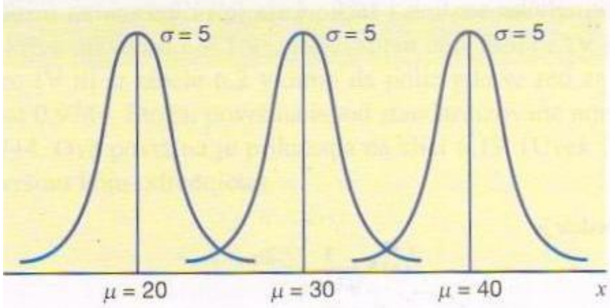


3) dva kraja krive se protežu u beskonačnost, što znači da kriva normalnog raspodela ne dodiruje horizontalnu osu

Aritmetička sredina μ i standardna devijacija σ su parametri normalne raspodele. Kad su date vrednosti ova dva parametra, možemo da odredimo površinu ispod krive normalne raspodele za bilo koji interval. Ne postoji jedinstvena kriva normalne raspodele već familija krivih normalne raspodele, jer svaka različit par vrednosti μ i σ određuje različite normalne raspodele. Vrednost μ određuje centar krive normalne raspodele na horizontalnoj osi, a vrednost σ pokazuje raspršenost krive normalne raspodele.



Slika 6.15 Tri krive normalne raspodele sa istom aritmetičkom sredinom ali različitim standardnim devijacijama.



Slika 6.16 Tri krive normalne raspodele sa različitim aritmetičkim sredinama ali sa istom standardnom devijacijom.

Normalan raspored predstavlja najznačajniji teorijski raspored verovatnoće iz sledećih razloga:

1) Veliki broj pojava ima približno normalan raspored

2) Normalan raspored može poslužiti kao dobra aproksimacija raznih prekidnih rasporeda (npr. Binomnog i Poasonovog) za one vrednosti parametra koje nisu date u tablicama verovatnoće

3) Iz normalnog rasporeda je izveden veliki broj drugih neprekidnih rasporeda, koji takođe imaju značajno mesto u statističkoj analizi, kao što su Studentov raspored, χ^2 raspored i F raspored

4) Veliki broj statističkih problema se rešava ili se može rešavati samo uz pretpostavku da populacija kojoj pripada uzorak ima normalan raspored

21. Empirijsko pravilo za normalnu raspodelu

1) Ukupna površina u opsegu jedne standardne devijacije od aritmetičke sredine je 68,26%. Ova površina se određuje kao razlika između površine ulevo od $z = 1$ i udesno od $z = -1$.

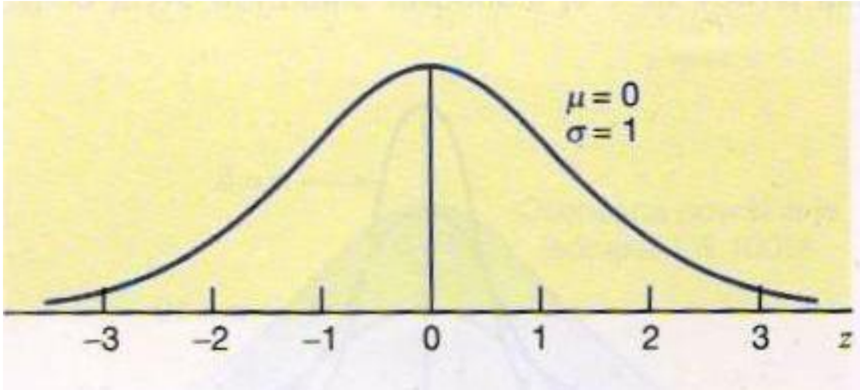
2) Ukupna površina u opsegu dve standardne devijacije od aritmetičke sredine je 95,44%. Ova površina se određuje kao razlika između površine ulevo od $z = 2$ i udesno od $z = -2$.

3) Ukupna površina u opsegu tri standardne devijacije od aritmetičke sredine je 99,74%. Ova površina se određuje kao razlika između površine ulevo od $z = 3$ i udesno od $z = -3$.

Dakle, samo određena vrsta krive u obliku zvona predstavlja normalnu raspodelu. Kriva u obliku zvona kod koje je (oko) 68,26% ukupne površine u opsegu jedne standardne devijacije od aritmetičke sredine, (oko) 95,44% ukupne površine u opsegu dve standardne devijacije od aritmetičke sredine i (oko) 99,74% ukupne površine u opsegu tri standardne devijacije od aritmetičke sredine predstavlja krivu normalne raspodele.

22. Standardizovana normalna raspodela – značaj i primena

Standardizovana normalna raspodela je poseban oblik normalne raspodele. Za standardizovanu normalnu raspodelu, vrednost aritmeti;ke sredine je jednaka nuli, a vrednost standardne devijacije je jednaka 1. Slučajna promenljiva koja ima standardizovanu normalnu raspodelu se obeležava sa Z. Realizovane vrednosti slučajne promenljive koja ima standardizovanu normalnu raspodelu se označavaju sa z i nazivaju se z vrednostima ili z rezultatima. Takođe se nazivaju i standardizovane vrednosti ili standardizovani rezultati.



Na slici z predstavlja rastojanje između aritmetičke sredine i tačke koja je prikazana sa z iskazano u standardnim devijacijama. Tako na primer, tačka čije je z jednako 2, je za dve standardne devijacije veća od aritmetičke sredine, a tačka čije je z jednako -2 je za dve standardne devijacije manja od aritmetičke sredine.

Površina ispod standardizovane normalne krive, između bilo koje dve tačke, može da se tumači kao verovatnoća da Z uzima vrednost iz tog intervala. Ukoliko sa $F(z)$ označimo površinu levo od z, koja može da se nađe u tablici 4, važiće:

1) $P(Z \leq a) = F(a)$
2) $P(Z \geq a) = 1 - F(a)$
3) $P(a \leq Z \leq b) = F(b) - F(a)$

Dakle korišćenjem tablice 4 možemo naći bilo koju površinu ispod standardizovane normalne krive. Međutim, u praktičnim primenama, slučajna promenljiva može da ima normalnu raspodelu sa vrednostima aritmetičke sredine i standardne devijacije koje se razlikuju od 0 i 1. Prvi korak u tom slučaju je da se transformiše data normalna raspodela u standardizovanu normalnu raspodelu. Ovaj postupak se naziva standardizacija normalne raspodele. Realizovane vrednosti slučajne promenljive koja ima normalnu raspodelu (koja nije standardizovana) se obeležavaju sa x. x se u z može transformisati pomoću formule: $z = \frac{x-\mu}{\sigma}$.

Na primer: Ako slučajna prmenljiva X ima normalnu raspodelu sa $\mu = 15$ i $\sigma = 6$, a treba izračunati $P(12 \leq X \leq 19)$, postupak je sledeći:

$$P(12 \leq X \leq 19) = P\left(\frac{12-15}{6} \leq Z \leq \frac{19-15}{6}\right) = P(-0{,}5 \leq Z \leq 0{,}67) =$$
$$F(0{,}67) - F(-0{,}5) = 0{,}7486 - 0{,}3085 = 0{,}4401.$$

Standardizovana normalna raspodela ima razne primene. Neke od njih su:

1) Aproksimacija binomne i Poasonove raspodele
2) ocenjivanje aritmetičke sredine skupa, kada je varijansa skupa poznata
3) testiranje hipoteze o aritmetičkoj sredini skupa kada je varijansa skupa poznata
4) testiranje hipoteze o jednakosti aritmetičkih sredina 2 skupa, kada su varijanse skupova poznate
5) testiranje Spirmanove korelacije ranga, kada je u pitanju veliki uzorak

23. Uzoračka raspodela

Vrednost parametara osnovnog skupa uvek je konstantna, što znači da za bilo koji niz podataka osnovnog skupa postoji samo jedna vrednost aritmetičke sredine . Međutim, isto se ne odnosi i na aritmetičku sredinu uzorka $\bar{X}$. Različiti uzorci iste veličine, izabrani iz istog osnovnog skupa, daju različite vrednosti aritmetičke sredine. Vrednost aritmetičke sredine u konkretnom uzorku će zavisiti od toga koji su elementi izabrani u taj uzorak. Prema tome, aritmetička sredina uzorka $\bar{X}$ je u stvari, slučajna promenljiva. Kao i svaka druga slučajna promenljiva i aritmetička sredina uzorka ima raspodelu verovatnoća, koja se često naziva i uzoračkom raspodelom statistike $\bar{X}$. Raspodela verovatnoća statistike $\bar{X}$ se naziva uzoračkom raspodelom. Ona predstavlja skup parova različitih vrednosti koje može uzeti statistika $\bar{X}$ i odgovarajućih verovatnoća. Raspodela verovatnoća bilo koje statistike uzorka se naziva uzoračkom raspodelom.

24.Raspodela sredina uzoraka

Aritmetička sredina i standardna devijacija koje su izračunae na osnovu uzoračke raspodele promenljive $\bar{X}$ se nazivaju aritmetička sredina i standardna devijacija statistike $\bar{X}$. Aritmetička sredina i standardna devijacija statistike $\bar{X}$ predstavljaju aritmetičku sredinu i standardnu devijaciju aritmetičkih sredina svih mogućih uzoraka iste veličine izabranih iz jednog osnovnog skupa. Standardna devijacija statistike $\bar{X}$ se naziva i standardnom greškom statistike $\bar{X}$. Aritmetička sredina i standardna devijacija statistike $\bar{X}$ se označavaju sa $\mu_{\bar{x}}$ i $\sigma_{\bar{x}}$.

Aritmetička sredina statistike $\bar{X}$ je uvek jednaka aritmetičkoj sredini osnovnog skupa, odnosno $\mu_{\bar{x}} = \mu$. Aritmetička sredina statistike uzorka $\bar{X}$ se naziva i ocena aritmetičke sredine osnovnog skupa . Kada je očekivana vrednost statistike uzorka jednaka vrednosti odgovarajućeg parametra osnovnog skupa, za statistiku uzorka se kaže da predstavlja nepristrasnu ocenu. Pošto je $\mu_{\bar{x}} = \mu$ zaključujemo da je $\bar{X}$ nepristrasna ocena za . Međutim, standardna greška $\sigma_{\bar{x}}$, statistike $\bar{X}$ nije jednaka standardnoj devijaciji σ osnovnog skupa. Naime važi: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Ova formula važi samo ukoliko se iz konačnog osnovnog skupa biraju uzorci sa ponavljanjem ili ako se iz beskonačnog osnovnog skupa biraju uzorci bez ponavljanja. Dakle, formula se koristi samo kad je uzorak dovoljno mali u poređenju sa veličinom skupa. Uzorak se smatra malim u odnosu na osnovni skup onda kada veličina uzorka predstavlja najviše 5% veličine tog skupa, odnosno ako je: $\frac{n}{N} \leq 0{,}05$. Ako ovaj uslov nije zadovoljen, pri izračunavanju koristimo sledeću formulu:

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, gde se faktor $\sqrt{\frac{N-n}{N-1}}$ naziva popravnim faktorom za konačne skupove.

Dakle za standardnu grešku važi:

1) $\sigma_{\bar{x}} < \sigma$, što je i očigledno iz same formule
2) $\sigma_{\bar{x}}$ se smanjuje sa povećanjem veličine uzorka, što je takođe očigledno iz formule

Ako se standardna greška statistike uzorka smanjuje sa povećanjem veličine uzorka, za takvu statistiku kažemo da je konzistentna ocena. Iz formule za $\sigma_{\bar{x}}$, jasno je da povećanjem veličine uzorka se standardna greška smanjuje, što znači da je aritmetička sredina uzorka konzistentna ocena aritmetičke sredine osnovnog skupa.

Kada osnovni skup iz kojeg su izabrani uzorci ima normalnu raspodelu sa aritmetičkom sredinom μ i standardnom devijacijom σ, važi sledeće:

1) aritmetička sredina statistike $\bar{X}$, $\mu_{\bar{x}}$, jednaka je aritmetičkoj sredini osnovnog skupa μ.
2) standardna greška raspodele statistike $\bar{X}$, $\sigma_{\bar{x}}$, jednaka je $\sigma/\sqrt{n}$, pod pretpostavkom da je $n/N \leq 0{,}05$
3) oblik uzoračke raspodele statistike $\bar{X}$ je normalan, bez obzira na veličinu uzorka n.

Dakle, ako osnovni skup iz kojeg potiču izabrani uzorci ima normalnu raspodelu sa aritmetičkom sredinom μ i standardnom devijacijom σ, tada će statistika uzorka $\bar{X}$ takođe imati normalnu raspodelu sa sledećom aritmetičkom sredinom i standardnom greškom, bez obzira na veličinu uzorka:
$\mu_{\bar{x}} = \mu$ i $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, ako je $n/N \leq 0{,}05$.
Međutim, u najvećem broju slučajeva osnovni skupovi iz kojih potiču uzorci nemaju normalnu raspodelu. Tada se oblik uzoračke raspodele statistike $\bar{X}$ određuje na osnovu veoma važne teoreme koja se naziva centralna granična teorema. Prema centralnoj graničnoj teoremi, bez obzira na oblik raspodele osnovnog skupa, aritmetičke sredine $\bar{X}$ velikih uzoraka imaju približno normalnu raspodelu sa aritmetičkom sredinom i standardnom devijacijom:
$\mu_{\bar{x}} = \mu$ i $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, ako je $n/N \leq 0{,}05$.
Uzorci se najčešće smatraju velikim kada je $n \geq 30$.
Dakle, prema cenralnoj graničnoj teoremi:
1) za $n \geq 30$ oblik uzoračke raspodele $\bar{X}$ je približno normalan, bez obzira na oblik raspodele osnovnog skupa
2) aritmetička sredina statistike $\bar{X}$, $\mu_{\bar{x}}$, je jednaka aritmetičkoj sredini skupa, μ .
3) standardna greška, $\sigma_{\bar{x}}$, je jednaka $\sigma/\sqrt{n}$, ako je $n/N \leq 0{,}05$

25. Centralna granična teorema

Prema centralnoj graničnoj teoremi, bez obzira na oblik raspodele osnovnog skupa, aritmetičke sredine $\bar{X}$ velikih uzoraka imaju približno normalnu raspodelu sa aritmetičkom sredinom i standardnom devijacijom:
$\mu_{\bar{x}} = \mu$ i $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, ako je $n/N \leq 0{,}05$.
Uzorci se najčešće smatraju velikim kada je $n \geq 30$.
Dakle, prema cenralnoj graničnoj teoremi:
1) za $n \geq 30$ oblik uzoračke raspodele $\bar{X}$ je približno normalan, bez obzira na oblik raspodele osnovnog skupa
2) aritmetička sredina statistike $\bar{X}$, $\mu_{\bar{x}}$, je jednaka aritmetičkoj sredini skupa, μ .
3) standardna greška, $\sigma_{\bar{x}}$, je jednaka $\sigma/\sqrt{n}$, ako je $n/N \leq 0{,}05$
Prema centralnoj graničnoj teoremi, uzoračka raspodela statistike $\hat{P}$ je približno normalna ukoliko se radi o velikim uzorcima. Kada je reč o proporciji, u praksi se smatra da su uzorci dovoljno veliki ukoliko su ispunjeni sledeći uslovi:
$np > 5 \; i \; nq > 5$.

26. Standardna devijacija skupa, standardna devijacija uzorka i standardna greška aritmetičke sredine

Varijansa i standardna devijacija: Standardna devijacija je najčešće korišćena mera disperzije. Vredost standardne devijacije pokazuje koliko blizu su vrednosti serije podataka grupisane oko aritmetičke sredine. Uopšteno, manje vrednosti standardne devijacije ukazuju da su vrednosti te serije raspršene veoma malo oko aritmetičke sredine. Nasuprot tome, veća vrednost standardne devijacije serije podataka ukazuje da su vrednosti te serije raspršene u relativno velikom razmaku oko aritmetičke sredine.
Standardna devijacija se dobija uzimanjem kvadratnog korena varijanse. Varijansa osnovnog skupa se označava sa σ^2, a varijansa uzorka se označava sa s^2. U skladu sa tim, standardna devijacija osnovnog skupa se označava sa σ, a uzorka sa s.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \quad i \quad s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Iz formule se vidi da varijansa skupa pokazuje prosek kvadrata odstupanja vrednosti x od aritmetičke sredine, a varijansa uzorka ne. Kod varijanse uzorka se umesto n koristi n-1. Razlog je što ako bi se koristilo n, varijansa uzorka bi bila pristrasna ocena, a ako se koristi n-1, varijansa uzorka je nepristrasna ocena varijanse skupa. Inače, nepristrasna ocena znači da je u proseku jednaka parametru. Veličine $x-\mu$ i $x-\bar{x}$ se zovu devijacijom vredbnosti x od aritmetičke sredine. Interesantna je osobina da je suma devijacija vrednosti x od aritmetičke uvek jednaka 0, tj. $\sum(x-\mu)=0 \; i \; \sum(x-\bar{x})=0$.

U zadacima se često koriste radne formule za izračunavanje varijanse, jer su brže za izračunavanje:

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad i \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Standardne devijacije skupa i uzorka su kvadratni koren iz varijanse: $= \sqrt{\sigma^2} \; i \; s = \sqrt{s^2}$.

Vrednosti varijanse i standardne devijacije nikada nisu negativne. Standardna devijacija se iskazuje u jedinici mere originalnih podataka, a varijansa u jedinici mere na kvadrat.

Ukoliko su podaci grupisani koriste se sledeće formule za izračunavanje varijanse:

$$\sigma^2 = \frac{\sum f(x\prime-\mu)^2}{N} \quad i \quad s^2 = \frac{\sum(x\prime-\bar{x})^2}{n-1}$$

odnosno radne formule:

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad i \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} \quad .$$

Međutim, standardna greška $\sigma_{\bar{x}}$, statistike $\bar{X}$ nije jednaka standardnoj devijaciji σ osnovnog skupa. Naime važi: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Ova formula važi samo ukoliko se iz konačnog osnovnog skupa biraju uzorci sa ponavljanjem ili ako se iz beskonačnog osnovnog skupa biraju uzorci bez ponavljanja. Dakle, formula se koristi samo kad je uzorak dovoljno mali u poređenju sa veličinom skupa. Uzorak se smatra malim u odnosu na osnovni skup onda kada veličina uzorka predstavlja najviše 5% veličine tog skupa, odnosno ako je: $\frac{n}{N} \leq 0{,}05$. Ako ovaj uslov nije zadovoljen, pri izračunavanju koristimo sledeću formulu:

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, gde se faktor $\sqrt{\frac{N-n}{N-1}}$ naziva popravnim faktorom za konačne skupove.

Dakle za standardnu grešku važi:

1) $\sigma_{\bar{x}} < \sigma$, što je i očigledno iz same formule

2) $\sigma_{\bar{x}}$ se smanjuje sa povećanjem veličine uzorka, što je takođe očigledno iz formule

Ako se standardna greška statistike uzorka smanjuje sa povećanjem veličine uzorka, za takvu statistiku kažemo da je konzistentna ocena. Iz formule za $\sigma_{\bar{x}}$, jasno je da povećanjem veličine uzorka se standardna greška smanjuje, što znači da je aritmetička sredina uzorka konzistentna ocena aritmetičke sredine osnovnog skupa.

27. Raspodela proporcija uzoraka

Kao i aritmetička sredina $\bar{X}$ i proporcija uzorka $\hat{P}$ je slučajna promenljiva koja ima svoju raspodelu verovatnoća, odnosno uzoračku raspodelu. Raspodela verovatnoća proporcije uzorka $\hat{P}$ se naziva uzoračkom raspodelom proporcije i predstavlja skup parova vrednosti koje može uzeti statistika $\hat{P}$ i odgovarajućih verovatnoća. Aritmetička sredina statistike $\hat{P}$, odnosno aritmetička sredina proporcija uzoraka je uvek jednaka proporciji osnovnog skupa, p , isto kao što je aritmetička sredina aritmetičkih sredina uzoraka $\bar{X}$ uvek jednaka aritmetičkoj sredini osnovnog skupa, μ . Aritmetička sredina proporcija uzoraka se obeležava sa $\mu_{\hat{P}}$ i jednaka je proporciji osnovnog skupa: $\mu_{\hat{P}} = p$. Proporcija uzorka $\hat{P}$, naziva se ocena proporcije osnovnog skupa, p . Standardna devijacija statistike $\hat{P}$, koju obeležavamo sa

$\sigma_{\hat{P}}$ i dobija se pomoću sledeće formule, koja važi samo ako je $\frac{n}{N} \leq 0{,}05$:

$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$.

Međutim, ako je $\frac{n}{N} > 0{,}05$, standardna greška, $\sigma_{\hat{P}}$, se dobija po sledećoj formuli:

$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$.

Izraz $\sqrt{\frac{N-n}{N-1}}$ se naziva popravnim faktorom za konačne skupove.

Ako se standardna greška smanjuje sa povećanjem veličine uzorka, ocena je konzistentna. Iz formule se vidi da je to ovde slučaj, pa je $\hat{P}$ konzistentna ocena proporcije skupa.

Oblik uzoračke raspodele statistike $\hat{P}$ se određuje korišćenjem centralne granične teoreme. Prema centralnoj graničnoj teoremi, uzoračka raspodela statistike $\hat{P}$ je približno normalna ukoliko se radi o velikim uzorcima. Kada je reč o proporciji, u praksi se smatra da su uzorci dovoljno veliki ukoliko su ispunjeni sledeći uslovi:

$np > 5\ i\ nq > 5$.

28.Statističko ocenjivanje

Ocenivanje je postupak kojim se numerička vrednost ili vrednosti dodeljuju parametru osnovnog skupa na osnovu informacija dobijenih iz uzorka. Najčešće se ocenjuju aritmetička sredina skupa, μ , i proporcija u skupu, p , mada mogu da se ocenjuju i drugi parametri kao što su medijana, modus, varijansa i standardna devijacija.

Kada bi mogao da se sprovede popis svaki put kada želimo da nađemo vrednost parametra populacije, ocenjivanje ne bi bilo ni potrebno. Ali, pošto najčešće ne raspolažemo informacijama o celom skupu, već samo o njegovom delu (uzorku), vršimo na osnovu uzorka ocenjivanje parametara osnovnog skupa. Vrednost(i) koja se dodeljuje parametru osnovnog skupa, a koja se bazira na vrednosti statistike uzorka se naziva ocenjena vrednost parametra skupa. Tako na primer, aritmetička sredina uzorka je ocena aritmetičke sredine skupa, a proporcija uzorka je ocena proporcije skupa. Dakle, statistika uzorka koja se koristi za ocenjivanje parametra skupa naziva se ocena. Postupak ocenjivanja uključuje sledeće etape:

1) izbor uzorka
2) prikupljanje neophodnih informacija od jedinica uzorka
3) izračunavanje vrednosti statistike uzorka
4) dodela vrednosti odgovarajućem parametru skupa

Da bi se ocenjivanje moglo sprovoditi uzorak mora da bude prost, slučajan.

29. Parametri skupa i njihove ocene

Ista priča kao i pitanje 28

30. Tačkasta ocena i interval pouzdanosti

Ocenjene vrednosti mogu biti tačkaste ili intervalne.

Ako izaberemo uzorak i izračunamo vrednost statistike uzorka za ovaj uzorak, onda ova vrednost predstavlja tačkastu ocenjenu vrednost odgovarajućeg parametra skupa. Tako na primer, vrednost aritmetičke sredine uzorka, izračunata u nekom uzorku, je tačkasta ocenjena vrednost odgovarajuće aritmetičke sredine skupa. Za svaki uzorak izvučen iz osnovnog skupa se očekuje da da različitu vrednost statistike uzorka. Tako da vrednost koja je dodeljena aritmetičkoj sredini skupa, zasnovana na tačkastoj ocenjenoj vrednosti zavisi od toga koji je od uzoraka izvučen. Kao posledica toga, tačkasta ocenjena vrednost dodeljulje μ vrednost koja se gotovo uvek razlikuje od prave vrednosti aritmetičke sredine skupa.

U slučaju interalnog ocenjivanja, umesto pridruživanja jedne vrednosti parametru osnovnog skupa, konstruiše se interval oko tačkaste ocenjene vrednosti za koji se veruje da sadrži odgovarajući parametar skupa. Tako na primer, ako želimo da ocenimo prosečnu platu u Srbiji. U uzorku smo dobili da prosečna plata iznosi 420 €. Dodajmo i oduzmimo na primer 80 €. Dobijamo interval (340 € , 500 €). Onda tvrdimo da taj interval verovatno sadrži aritmetičku sredinu skupa. Vrednost od 340 € predstavlja donju granicu intervala, a vrednost od 500 €, predstavlja gornju granicu interala. Broj koji se oduzima i dodaje na tačkastu ocenjenu vrednost zove se marginalna greška.

Međutim, postavlja se pitanje, koji broj da oduzmemo i saberemo na tačkastu ocenjenu vrednost da bismo dobili intervalnu ocenjenu vrednost? Odgovor na ovo pitanje zavisi od:

1) standardne devijacije $\sigma_{\bar{x}}$ aritmetičke sredine uzorka $\bar{X}$ (ili standardne greške)

2) nivoa pouzdanosti koji je pripisan intervalu.

Prvo, što je veća standardna devijacija od $\bar{X}$, veći je broj koji oduzimamo i didajemo tačkastoj ocenjenoj vrednosti. Prema tome, očigledno je da ukoliko je raspon vrednosti koje $\bar{X}$ može uzeti veći, to interval formiran oko aritmetičke sredine uzorka, mora biti širi da obuhvati μ .

Drugo, veličina koja je oduzeta i dodata mora biti veća ukoliko želimo interval veće pouzdanosti. Intervalnom ocenjivanju uvek pripisujemo određeno tvrđenje iz teorije verovatnoće, koje se iskazuje preko nivoa pouzdanosti. Interval konstruisan uz određeni nivo pouzdanosti je interval pouzdanosti. Svaki interval se konstruiše uz zadavanje nivoa pouzdanosti i zove se interval pouzdanosti (ili interval poverenja). Interval pouzdanosti je određen na sledeći način:

Tačkasta ocenjena vrednost $\pm$ Marginalna greška

Nivo pouzdanosti koji je pridružen intervalu poverenja pokazuje koliko možemo biti sigurni da ovaj interval sadrži pravu vrednost parametra skupa. Nivo pouzdanosti se označava sa $(1-\alpha)100\%$. Ako nije izražen u procentima, zove se koeficijent pouzdanosti i obeležava se sa $1-\alpha$. Inače α se zove nivo značajnosti. Najčešće se koriste nivoi pouzdanosti od 90%, 95% i 99%.

31. Ocene parametara skupa i njihove osobine

U postupku statističkog ocenjivanja, zaključak o nepoznatoj vrednosti parametra skupa donosimo na osnovu odgovarajuće statistike uzorka: nepoznatu vrednost konstante ocenjujemo na osnovu odgovarajuće slučajne promenljive, odnosno na osnovu njene realizovane vrednosti u uzorku, koja je ponovo konstanta. Ako parametar skupa koji ocenjujemo obeležimo sa θ, statistiku uzorka kojom ocenjujemo taj parametar nazivamo ocenom parametra θ i obeležavamo sa $\hat{\theta}$, a realizovanu vrednost ocene u izabranom uzorku nazivamo ocenjenom vrednošću parametra i obeležavamo je sa $\hat{\theta}'$.

Ocenjena vrednost je jedan broj, tj. predstavlja tačku na numeričkoj skali. Zato $\hat{\theta}'$ nazivamo tačkastom ili brojnom ocenjenom vrednošću, a odgovarajuću ocenu, $\hat{\theta}$, nazivamo tačkastom ocenom. Tačkasta ocena je funkcija predstavljena algebarskim izrazom (formulom), a numerička vrednost te funkcije, za date vrednosti u uzorku, je tačkasta ocenjena vrednost parametra skupa.

Isti parametar možemo da ocenjujemo primenom više različitih ocena. Da bismo se opredelili potrebno je da se upoznamo sa osobinama koje odlikuju ove slučajne promenljive i da izaberemo onu koja ima najviše poželjnih osobina. Osobine ocena su: nepristrasnost, efikasnost, konzistentnost i dovoljnost.

Nepristrasna ocena: Ocena $\hat{\theta}$ je nepristrasna ocena parametra θ, ako je njena očekivana vrednost (aritmetička sredina) jednaka parametru θ:

$$E(\hat{\theta}) = \theta$$

Na primer, aritmeti;ka sredina uyorka je nepristrasna ocena aritmeti;ke sredine skupa, jer je $E(\bar{X}) = \mu$, a isto važi i za medijanu. Varijansa uzorka u imeniocu ima n-1, baš iz razloga da bi bila nepristrasna ocena varijanse skupa.

Efikasna ocena: Jedna nepristrasna ocena je efikasnija od druge, ako ima manju varijansu, odnosno standardnu grešku, za istu veličinu uzorka. Na primer, aritmetička sredina i medijana su obe nepristrasne ocene aritmetičke sredine skupa, ali je aritmetička sredina uzorka efikasnija ocena, jer je njena varijansa manja nego kod medijane. Sam odnos njihovih varijansi je promenljiv i zavisi od rasporeda skupa, ali na primer kod normalnog rasporeda važi $\sigma_{Me}^2 = 1{,}57\sigma_{\bar{x}}^2$. Dakle, aritmetička sredina uzorka je 57% efikasnija od medijane uzorka.

Konzistentna ocena: Ocena je konzistentna ako sa povećanjem veličine uzorka ona teži parametru. Kako n raste, realizovane vrednosti $\hat{\theta}'$ u uzorcima sve se više koncentrišu oko θ, a za $n = \infty$ poklapaju se sa vrednošću parametra.

Dovoljna ocena: Ocena je dovoljna ako koristi sve informacije koje uzorak sadrži o tom parametru. Tako na primer $\bar{X}$ je dovoljna ocena parametra μ , a medijana nije jer ne koristi sve vrednosti iz uzorka već samo središnji član.

32. Ocenjivanje sredine skupa

1) Ako je σ poznato

Postoje tri moguća slučaja:

Slučaj 1: Ako su ispunjena sledeća 3 uslova:

1) standardna devijacija skupa σ je poznata

2) veličina uzorka je mala (tj. $n < 30$)

3) osnovni skup iz koga se uzima uzorak ima normalnu raspodelu

onda koristimo normalnu raspodelu za određivanje intervala poverenja za μ, jer je uzoračka raspodela od $\bar{X}$ normalna sa aritmetičkom sredinom μ i standardnom devijacijom $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, ako je $n/N \leq 0{,}05$.

Slučaj 2: Ako su ispunjena sledeća dva uslova:

1) standardna devijacija skupa σ je poznata

2) veličina uzorka je velika (tj. $n \geq 30$)

opet koristimo normalnu raspodelu za određivanje intervala poverenja za μ , jer prema centralnoj graničnoj teoremi je uzoračka raspodela od $\bar{X}$ normalna sa aritmetičkom sredinom μ i standardnom devijacijom $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, ako je $n/N \leq 0{,}05$.

Slučaj 3: Ako su ispunjena sledeća tri uslova:

1) standardna devijacija skupa σ je poznata

2) veličina uzorka je mala (tj. $n < 30$)

3) osnovni skup iz koga se uzima uzorak nema normalnu raspodelu (ili je njegova raspodela nepoznata)

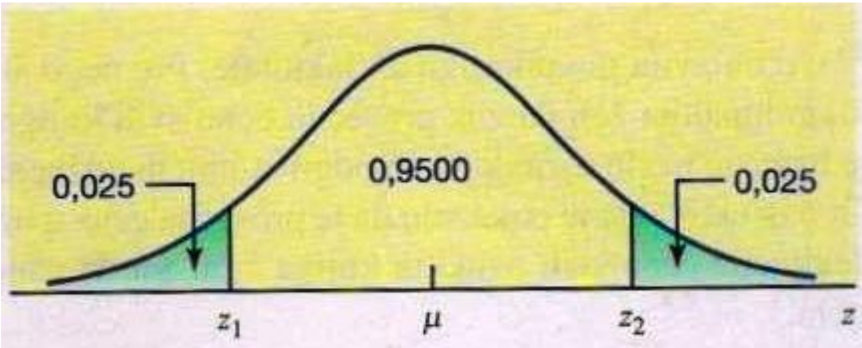
onda koristimo neparametarski metod za određivanje intervala poverenja za μ .

$(1-\alpha)100\%$ interval poverenja za μ u gore pomenutim slučajevima 1 i 2 je:

$$\bar{x} \pm z\sigma_{\bar{x}}$$

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

Vrednost $z\sigma_{\bar{x}}$ u formuli intervala poverenja se zove marginalna greška i označava sa E. Dakle, marginalna greška ocene od μ, koja se obeležava sa E, je vrednost koja je oduzeta i dodata vrednosti $\bar{x}$ kako bi se dobio interval poverenja za μ .
Vrednost z se dobija iz tablice standardizovane normalne raspodele za zadati nivo pouzdanosti. Postupak je sledeći. Ako nam treba na primer 95% interval poverenja za μ . Nivo pouzdanosti od 95% znači da je čitava površina ispod normalne krive za $\bar{X}$ između dve tačke na suprotnim krajevima od μ jednaka 95% ili 0,95. Neka su te dve tačke z_1 i z_2, kao na slici. Ukupna površina između z_1 i z_2 treba da bude $1-\alpha$, u našem slučaju 0,95. Dakle, na krajevima zbir površina treba da bude 0,05, što znači da površina na svakom kraju je 0,025. Da bi površine levo i desno bile po 0,025, ako nađemo u tablici 4 odgovarajuće vrednosti za z, dobićemo da je $z_1 = -1{,}96$ a $z_2 = 1{,}96$. Dakle za nivo pouzdanosti od 95% koristimo vrednost $z = 1{,}96$ u formuli za izračunavanje intervala poverenja.



Inače nivo pouzdanosti od 95% znači, da ako iz osnovnog skupa izvučemo veliki broj uzoraka iste veličine, 95% njih će dati interval poverenja koji će sadržati pravu vrednost za μ .
Širina intervala poverenja zavisi od toga kolika je marginalna greška $z\sigma_{\bar{x}}$, koja zavisi od vrednosti z, σ i n. Međutim, vrednost σ nije pod kontrolom istraživača. Znači, širina intervala zavisi od:
1) vrednosti z, koja je određena nivoom pouzdanosti
2) veličine uzorka n
Povećanje nivoa pouzdanosti povećava i vrednost z, a samim tim i širinu intervala poverenja.
Povećanje veličine uzorka, smanjuje vrednost standardne greške, pa se širina intervala smanjuje.
Tako, ako hoćemo da smanjimo širinu intervala poverenja imamo dve mogućnosti:
1) da smanjimo nivo pouzdanosti
2) povećamo veličinu uzorka

<u>2) Ako σ nije poznato</u>

Postoje 3 slučaja.
Slučaj 1: Ako su ispunjena sledeća tri uslova:
1) standardna devijacija skupa σ nije poznata
2) veličina uzorka je mala ($n < 30$)
3) osnovni skup iz koga se uzima uzorak ima normalnu raspodelu
onda koristimo t raspodelu za određivanje intervala poverenja za μ .
Slučaj 2: Ako su ispunjena sledeća dva uslova:
1) standardna devijacija skupa σ nije poznata
2) veličina uzorka je velika ($n \geq 30$)
onda opet koristimo t raspodelu za određivanje intervala poverenja za μ .
Slučaj 3: Ako su ispunjena sledeća tri uslova:
1) standardna devijacija skupa σ nije poznata

2) veličina uzorka je mala ($n < 30$)
3) osnovni skup iz koga se uzima uzorak nema normalnu raspodelu (ili je njegova raspodela nepoznata) onda koristimo neparametarski metod za određivanje intervala poverenja za μ .
$(1-\alpha)100\%$ interval poverenja za μ u gore pomenutim slučajevima 1 i 2 je:

$$\bar{x} \pm ts_{\bar{x}}$$

gde je $s_{\bar{x}} = s/\sqrt{n}$.
t vrednost je dobijena iz tablice t raspodele za $n-1$ stepeni slobode i za dati nivo pouzdanosti. Ovde je marginalna greška $E = ts_{\bar{x}}$

33. Normalna i studentova raspodela – osobine i primena

ispričati pitanje 22 +
Studentova raspodela je donekle slična normalnoj raspodeli. Kao i normalna raspodela i Studentova t raspodela je simetrična oko aritmetičke sredine i nikada se ne spaja sa horizontalnom osom. Ukupna površina ispod krive t raspodele je 1 ili 100%. Kriva studentove raspodele je više spljoštena od krive standardizovane normalne raspodele, što znači da ima veću disperziju od 1. Međutim, sa povećanjem veličine uzorka, t raspodela se približava standardizovanoj normalnoj raspodeli. Oblik krive t raspodele zavisi od broja stepeni slobode df. Pri ocenjivanju i testiranju aritmetičke sredine broj stepeni slobode je $df = n-1$. Aritmetička sredina Studentove t raspodele je 0, a standardna devijacija je $\sqrt{\frac{df}{df-2}}$ što je uvek veće od 1.
Studentova t raspodela ima razne primene. Neke od njih su:
1) ocenjivanje aritmetičke sredine osnovnog skupa kada σ nije poznata
2) testiranje aritmetičke sredine osnovnog skupa kada σ nije poznata
3) testiranje jednakosti aritmetičkih sredina dva skupa kada devijacije skupova nisu poznate
4) ocenjivanje i testiranje parametara kod regresije i korelacije

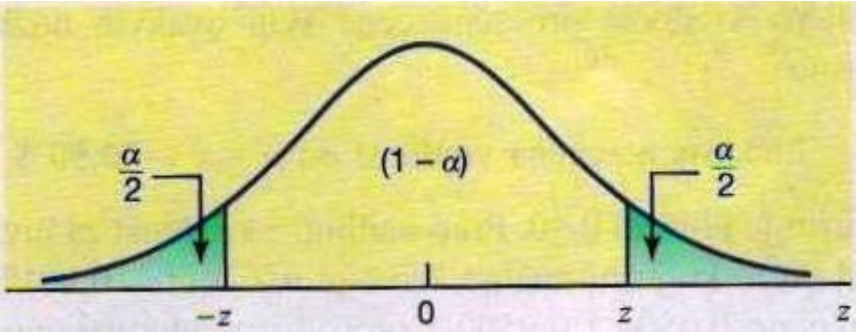
34. Ocenjivanje proporcije skupa

Proporcija skupa p se ocenjuje pomoću proporcije uzorka $\hat{P}$. Za velike uzorke ($n\hat{p} > 5, n\hat{q} > 5$) važi:
1) uzoračka raspodela proporcije uzorka $\hat{P}$ je (približno) normalna.
2) aritmetička sredina uzoračke raspodele $\hat{P}$ (u oznaci $\mu_{\hat{P}}$) je jednaka proporciji skupa p.
3) standardna devijacija uzoračke raspodele proporcije uzorka $\hat{P}$ (u oznaci $\sigma_{\hat{P}}$) je $\sqrt{pq/n}$, pri čemu je q=1-p . Međutim, kada ocenjujemo vrednost proporcije skupa, mi ne znamo ni p ni q, pa ne možemo izračunati $\sigma_{\hat{P}}$. Zbog toga koristimo $s_{\hat{P}}$, koja se naziva ocena za $\sigma_{\hat{P}}$ i izračunava kao: $s_{\hat{P}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$.
Proporcija uzorka, $\hat{P}$, je tačkasta ocena odgovarajuće proporcije skupa p. Da bismo odredili interval poverenja za p, dodajemo proporciji uzorka i oduzimamo od nje broj koji se naziva marginalna greška, E.
$(1-\alpha)100\%$ interval poverenja ya proporciju osnovnog skupa, p, je

$$\hat{p} \pm zs_{\hat{P}}$$

z vrednost koja se ovde koristi dobija se iy tablice standardiyovane normalne raspodele ya dati nivo pouzdanosti, a $s_{\hat{P}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$. Izraz $zs_{\hat{P}}$ se zove marginalna greška, E.

Vrednost z se dobija iz tablice standardizovane normalne raspodele za zadati nivo pouzdanosti. Postupak je sledeći. Ako nam treba na primer 95% interval poverenja za p . Nivo pouzdanosti od 95% znači da je čitava površina ispod normalne krive za $\hat{P}$ između dve tačke na suprotnim krajevima od p jednaka 95% ili 0,95. Neka su te dve tačke z_1 i z_2, kao na slici. Ukupna površina između z_1 i z_2 treba da bude $1 - \alpha$, u našem slučaju 0,95. Dakle, na krajevima zbir površina treba da bude 0,05, što znači da površina na svakom kraju je 0,025. Da bi površine levo i desno bile po 0,025, ako nađemo u tablici 4 odgovarajuće vrednosti za z, dobićemo da je $z_1 = -1{,}96$ a $z_2 = 1{,}96$. Dakle za nivo pouzdanosti od 95% koristimo vrednost $z = 1{,}96$ u formuli za izračunavanje intervala poverenja.



Inače nivo pouzdanosti od 95% znači, da ako iz osnovnog skupa izvučemo veliki broj uzoraka iste veličine, 95% njih će dati interval poverenja koji će sadržati pravu vrednost za p .

Širina intervala poverenja zavisi od toga kolika je marginalna greška $z\sigma_{\bar{x}}$, koja zavisi od vrednosti $z, \hat{p}, \hat{q}$ i n. Međutim, vrednost $\hat{p}, \hat{q}$ nije pod kontrolom istraživača. Znači, širina intervala zavisi od:

1) vrednosti z, koja je određena nivoom pouzdanosti

2) veličine uzorka n

Povećanje nivoa pouzdanosti povećava i vrednost z, a samim tim i širinu intervala poverenja.

Povećanje veličine uzorka, smanjuje vrednost standardne greške, pa se širina intervala smanjuje.

Tako, ako hoćemo da smanjimo širinu intervala poverenja imamo dve mogućnosti:

1) da smanjimo nivo pouzdanosti

2) povećamo veličinu uzorka

35. Preciznost i pouzdanost ocene

Ocena je precizna ukoliko je interval poverenja male širine. Na širinu intervala utičemo promenom:

1) nivoa pouzdanosti

2) veličine uzorka

Povećanje nivoa pouzdanosti utiče na poveća i z i t, a samim tim povećava se i marginalna greška, što znači da se dobija širi interval.

Povećanje veličine uzorka smanjuje standardnu grešku, pa samim tim i marginalnu grešu i širinu intervala poverenja.

Povećati preciznost ocene, tj. smanjiti širinu intervala poverenja možemo na dva načina:

1) smanjenjem nivoa pouzdanosti

2) povećanjem veličine uzorka

Povećanje uzorka je bolje rešenje, ukoliko je to moguće uraditi.

Ukoliko znamo koliko precizan interval želimo, tj. znamo kolika treba da bude marginalna greška E, veličinu uzorka određujemo pomoću formula:

$n = \frac{z^2\sigma^2}{E^2}$, ako ocenjujemo aritmetičku sredinu skupa

$n = \frac{z^2\hat{p}\hat{q}}{E^2}$, ako ocenjujemo proporciju skupa.

36. Testiranje hipoteza

Postupak testiranja hipoteza se sprovodi kada zaključak o parametru osnovnog skupa donosimo na osnovu vrednosti odgovarajuće statistike uzorka. Postoji nulta hipoteza, H_0, i alternativna hipoteza, H_1.

Nulta hipoteza je tvrđenje o nekom parametru osnovnog skupa koji se smatra istinitim sve dok se ne dokaže suprotno. Alternativna hipoteza je tvrđenje o nekom parametru osnovnog skupa koje će biti istinito ako je nulta hipoteza neistinita. Uzmimo primer sa sudskim postupkom. Suđenje počinje pretpostavkom da je nulta hipoteza istinita – odnosno da je osoba nevina. Tužilac sakuplja sve dokaze i prikazuje ih na sudu da bi dokazao da je nulta hipoteza neistinita i da je alternativna hipoteza istinita, odnosno da je osoba kriva.

Na slici ispod koja označava sudski proces, tačka označena sa 0 prikazuje nepostojanje dokaza protiv okrivljene osobe. Što se više udaljavamo udesno po horizontalnoj osi, to su ubedljiviji dokazi da je ta osoba počinila krivično delo. Proizvoljno smo izabrali tačku C na horizontalnoj osi. Pretpostavimo da sudija smatra da je bilo koja količina dokaza desno od tačke C dovoljna, a da bilo koja tačka levo od C nije dovoljna da se osoba proglasi

krivom. Tačka C se u statistici naziva kritična vrednost ili kritična tačka. Ako se količina dokaza koje iznese tužioc nađe levo od tačke C, to znači da nema dovoljno dokaza da se ta osoba proglasi krivom, pa će presuda glasiti da osoba nije kriva. U statistici ova odluka znači da ne odbacujemo H_0. Oblast ulevo od tačke C se naziva oblast neodbacivanja H_0 . Ako se količina dokaza nalazi desno od tačke C, onda postoji dovoljno dokaza da tvrdimo da je osoba kriva i da odbacimo H_0. Oblast udesno od tačke C se naziva oblast odbacivanja H_0.

Međutim, odluka suda ne mora uvek da bude ispravna. Ako je na kraju sudskog procesa osoba proglašena krivom, onda postoje dve mogućnosti:

1) osoba nije počinila krivično delo, ali je proglašena krivom (zbog mogućih lažnih dokaza)

2) osoba je počinila krivično delo i s pravom je proglašena krivom

U prvom slučaju, dud je napravio grešku i osudio nevinu osobu. U statistici ova greška se naziva greška I vrste ili α greška. U drugom slučaju, kada osuđena osoba jeste kriva, sud je doneo ispravnu odluku. Ova dva slučaja su prikazana u drugom redu osenčenog dela tabele.

		Stvarno stanje	
		Osoba nije kriva	Osoba jeste kriva
Odluka suda	Osoba nije kriva	Ispravna odluka	Greška II vrste ili β greška
	Osoba jeste kriva	Greška I vrste ili α greška	Ispravna odluka

Dakle, greška I vrste će se javiti ako je H_0 zapravo istinita, ali smo pogrešno odbacili nultu hipotezu. Vrednost α se naziva nivo značajnosti testa i predstavlja verovatnoću da smo napravili grešku I vrste. Dakle α predstavlja verovatnoću da ćemo odbaciti nultu hipotezu, H_0, kada je ona u stvari istinita.

$$\alpha = P(H_0\ se\ odbacuje\ |H_0\ je\ istinita)$$

Veličina oblasti odbacivanja u statističkom problemu testiranja hipoteze zavisi od vrednosti koju dodeljujemo α.

Pretpostavimo da se desio drugi ishod sudskog procesa, tj. Da okrivljena osoba nije proglašena krivom. Postoje dve mogućnosti:

1) osoba nije počinila krivično delo i nije proglašena krivom

2) osoba je počinila krivično delo, ali zbog nedostatka dokaza nije proglašena krivom

U prvom slučaju odluka suda je ispravna, ali u drugom slučaju sud je napravio grešku jer je oslobodio osobu koja je kriva. U statistici se ovaj tip greške naziva greška II vrste ili β greška. Greška II vrste se javlja kada se neistinita nulta hipoteza ne odbaci. Vrednost β predstavlja verovatnoću javljanja greške II vrste, odnosno:

$$\beta = P(H_0\ se\ ne\ odbacuje\ |H_0\ je\ neistinita)$$

Vrednost $1 - \beta$ se naziva jačina testa i ona predstavlja verovatnoću da se greška II vrste ne javi.

Ove dve vrste grešaka zavise jedna od druge. Ako je veličina uzorka fiksna, onda, pri testiranju jedne hipoteze, vrednosti $\alpha\ i\ \beta$ ne možemo istovremeno da smanjimo; smanjivanjem vrednosti α povećaće se vrednost β i obrnuto.

Tabela 9.2

		Stvarno stanje	
		H_0 je istinita	H_0 nije istinita
Odluka	H_0 se ne odbacuje	Ispravna odluka	Greška II vrste ili β greška
	H_0 se odbacuje	Greška I vrste ili α greška	Ispravna odluka

37. Greške pri testiranju hipoteza i njihove verovatnoće

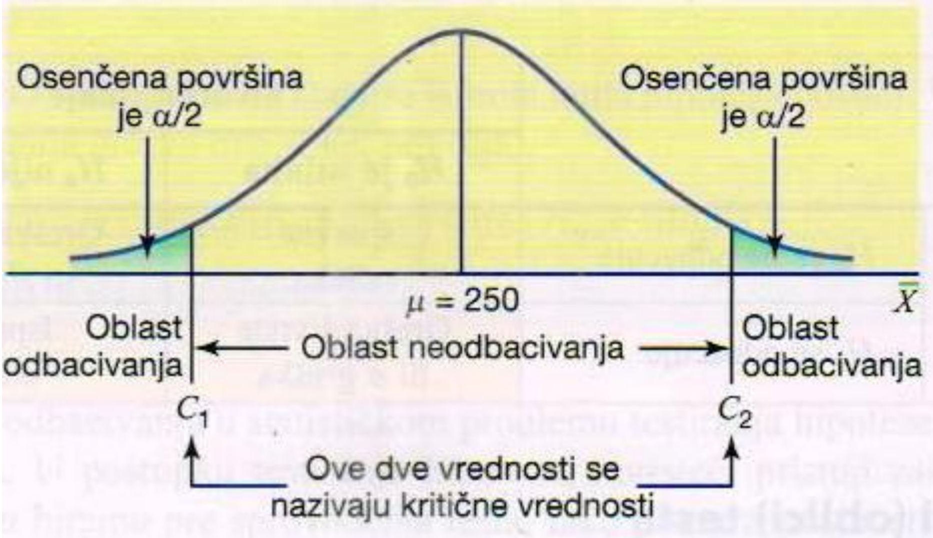
ista priča kao u pitanju 36

38.Oblik (smer) testa

Dvostrani test ima oblast odbacivanja na oba kraja, levostrani test ima oblast odbacivanja na levom kraju, a desnostrani test ima oblast odbacivanja na desnom kraju krive raspodele.

<u>Dvostrani test</u>: Da li je test jednostrani ili dvostrani određuje znak u alternativnoj hipotezi. Ako se u

alternativnoj hipotezi nalazi znak $\neq$, onda je taj test dvostrani. Dvostrani test ima dve oblasti odbacivanja, po jednu na svakom kraju krive raspodele.

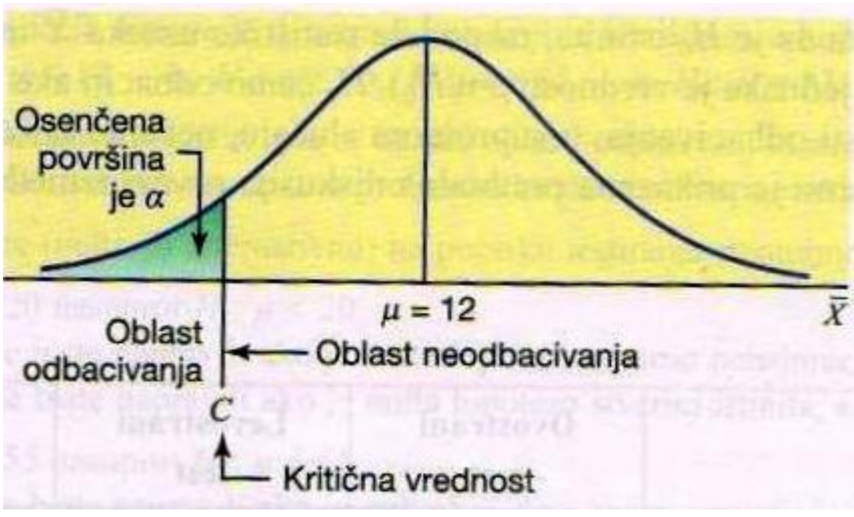


Na slici je prikazana raspodela statistike uzorka $\bar{X}$ pod pretpostavkom da je ona normalna. Ako je H_0 istinita, onda $\bar{X}$ ima normalnu raspodelu sa aritmetičkom sredinom koja je jednaka 250 ($H_0: \mu = 250$). Obe osenčene površine imaju površinu od $\alpha/2$, pa je ukupna osenčena površina jednaka α. Dakle, dvostrani test ima dve kritične vrednosti koje razdvajaju dve oblasti odbacivanja H_0 od oblasti neodbacivanja. H_0 se odbacuje , ako se realizovana vrednost aritmetičke sredine u uzorku, $\bar{x}$, nađe u bilo kojoj od dve oblasti odbacivanja. H_0 nećemo odbaciti ako se vrednost $\bar{x}$ nalazi u oblasti neodbacivanja. Odbacivanjem H_0 tvrdimo da je razlika između hipotetičke vrednosti μ, koja se nalazi u H_0, i vrednosti $\bar{x}$ dobijene iz uzorka suviše velika da bi se javila samo zbog uzoračke greške. Neodbacivanjem H_0 tvrdimo da je razlika između hipotetičke vrednosti μ, koja se nalazi u H_0, i vrednosti $\bar{x}$ dobijene iz uzorka mala i da se mogla javiti samo zbog slučajne greške.

<u>Levostrani test</u>: Pretpostavimo da agencija za zaštitu potrošača želi da ispita da li je prosečna količina soka po limenki manja od 12 unci. Odgovarajuće hipoteze će biti:

$$H_0: \mu = 12 \qquad\qquad H_1: \mu < 12$$

Kada je u alternativnoj hipotezi znak <, kao u ovom slučaju, test je uvek levostran. U levostranom testu oblast odbacivanja se nalazi na levom kraju krive raspodele, a površina koja odgovara oblasti odbacivanja je jednaka α.

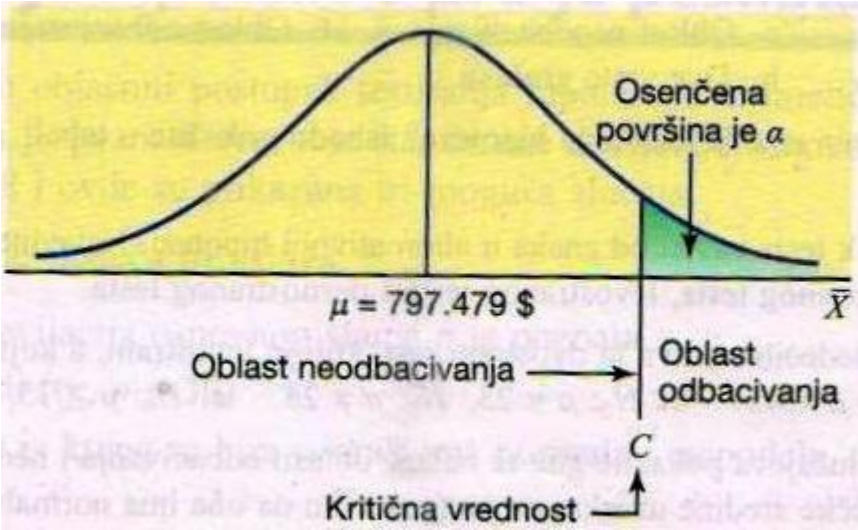


Pod pretpostavkom da je H_0 istinita, raspodela statistike uzorka $\bar{X}$ ima aritmetičku sredinu jednaku 12. Odbacićemo H_0 ako se vrednost $\bar{x}$ dobijena iz uzorka nalazi u oblasti odbacivanja, a u suprotnom nećemo odbaciti H_0.

<u>Desnostrani test</u>: Pretpostavimo da želimo da ispitamo da li je prosečna cena kuća u jednoj imućnoj oblasti veća od 797479$. Odgovarajuće hipoteze će biti:

$$H_0: \mu = 797.479 \qquad H_1: \mu > 797.479$$

Kada je u alternativnoj hipotezi znak >, kao u ovom slučaju, test je uvek desnostran. U desnostranom testu oblast odbacivanja se nalazi na desnom kraju krive raspodele, a površina koja odgovara oblasti odbacivanja je jednaka α.



Pod pretpostavkom da je H_0 istinita, raspodela statistike uzorka $\bar{X}$ ima aritmetičku sredinu jednaku 797.479. Odbacićemo H_0 ako se vrednost $\bar{x}$ dobijena iz uzorka nalazi u oblasti odbacivanja, a u suprotnom nećemo odbaciti H_0.

39. Postupak statističkog testiranja – pristup zasnovan na kritičnoj vrednosti

Ovaj pristup se naziva i tradicionalni ili klasični pristup. U njemu je vrednost nivoa značajnosti α unapred određena. Vrednost α predstavlja ukupnu površinu oblasti odbacivanja. U okviru ove procedure prvo nalazimo kritičnu vrednost z iz tablice za normalnu raspodelu za da ti nivo značajnosti. Zatim izračunavamo realizovanu vrednost statistike testa z za realizovanu vrednost statistike uzorka, $\bar{x}$. Na kraju, realizovanu vrednost statistike z testa poredimo sa kritičnom vrednošću i donosimo odluku. Kod jednostranog testa postoji samo jedna kritična vrednost z, koju određujemo na osnovu vrednosti α. Vrednost α predstavlja površinu na levom ili desnom kraju ispod krive normalne raspodele, u zavisnosti od toga da li je test levostrani ili desnostrani. Međutim, ako je test dvostrani, postoje dve kritične vrednosti z i njih određujemo na osnovu vrednosti ${}^{\alpha}/_{2}$, koja predstavlja površine na oba kraja krive normalne raspodele. Prilikom testiranja hipoteze o μ primenom normalne raspodele, slučajna promenljiva

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}, gde\ je\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

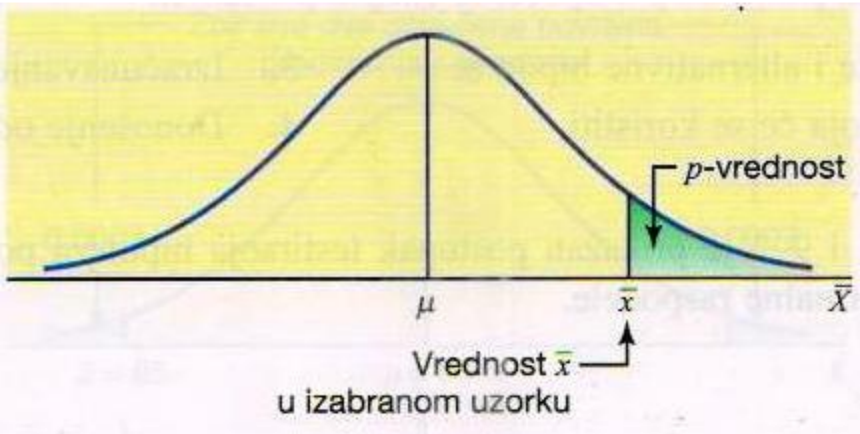
se naziva statistika testa. Statistika testa može da se definiše kao pravilo ili kriterijum koji koristimo pri donošenju odluke da li da odbacimo nultu hipotezu ili da je ne odbacimo.

Postupak testiranja hipoteze primenom pristupa zasnovanog na kritičnoj vrednosti, sprovodimo u 5 etapa:

1) formulisanje nulte i alternativne hipoteze
2) izbor raspodele koja će se koristiti
3) određivanje oblasti odbacivanja i neodbacivanja
4) izračunavanje vrednosti statistike testa
5) donošenje odluke

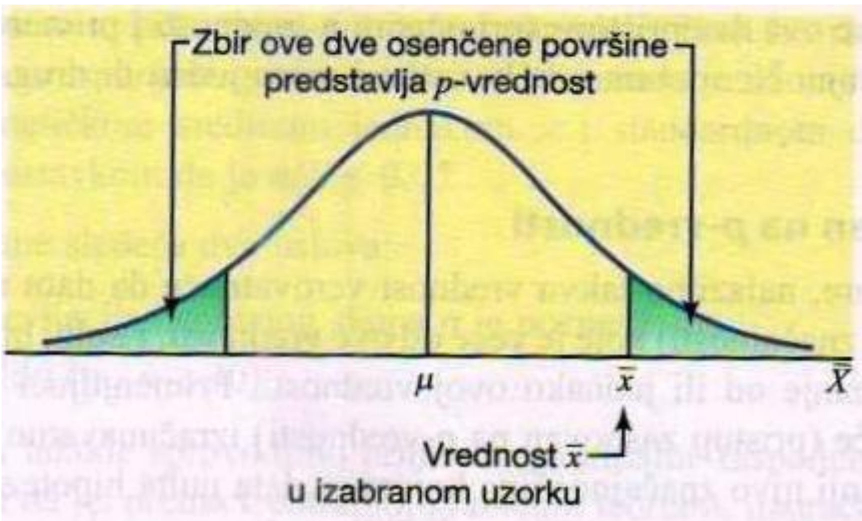
40. Postupak statističkog testiranja – pristup zasnovan na p-vrednosti

Pod pretpostavkom da je nulta hipoteza istinita, p-vrednost može da se definiše kao verovatnoća da statistika uzorka odstupa od hipotetičke vrednosti parametra u smeru alternativne hipoteze, barem toliko koliko i realizovana vrednost statistike uzorka u izabranom uzorku. Nultu hipotezu odbacujemo ako je p-vrednost < α, a nultu hipotezu ne odbacujemo ako je p-vrednost $\geq \alpha$. Kod jednostranog testa p-vrednost je prikazana površinom ispod krive uzoračke raspodele koja se nalazi na kraju raspodele, a omeđena je realizovanom vrednošću statistike uzorka. Na slici je prikazana p-vrednost kod desnostranog testa o μ.



Kod levostranog testa, p-vrednost je prikazana površinom ispod krive uzoračke raspodele, ulevo od vrednosti $\bar{x}$.

Kod dvostranog testa, p-vrednost je verovatnoća prikazana dvostrukom površinom ispod krive uzoračke raspodele koja se nalazi na kraju raspodele, a ograničena je realizovanom vrednošću statistike uzorka.



Da bismo pronašli verovatnoću prikazanu površinom ispod krive normalne raspodele koja je ograničena vrednošću aritmetičke sredine $\bar{x}$, izračunavamo standardizovanu vrednost, z, za realizovanu vrednost $\bar{x}$. Kada o testiranju hipoteze o μ koristimo normalnu raspodelu, onda vrednost z, za vrednost $\bar{x}$ u izabranom uzorku, izračunavamo na sledeći način:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}, gde\ je\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Izračunatu vrednost z, na osnovu $\bar{x}$ iz izabranog uzorka, nazivamo i realizovana vrednost statistike z testa.

Zatim nalazimo verovatnoću prikazanu površinom ispod krive standardizovane normalne raspodele, koja je na kraju krive raspodele ograničena realizovanom vrednošću z. Ova verovatnoća predstavlja p-vrednost, ako je reč o jednostranom testu, ili je jednaka polovini p-vrednosti u slučaju dvostranog testa.

Postupak testiranja primenom pristupa zasnovanog na p-vrednosti sprovodi se u sledeće 4 etape:

1) formulisanje nulte i alternativne hipoteze
2) izbor raspodele koja će se koristiti
3) izračunavanje p-vrednosti
4) donošenje odluke

41. Primena normalne raspodele u ocenjivanju i testiranju aritmetičke sredine skupa – postupak i pretpostavke

Pri ocenjivanju aritmetičke sredina skupa normalna raspodela se koristi u sledeća dva slučaja:

Slučaj 1: Ako su ispunjena sledeća 3 uslova:

1) standardna devijacija skupa σ je poznata
2) veličina uzorka je mala (tj. $n < 30$)
3) osnovni skup iz koga se uzima uzorak ima normalnu raspodelu

onda koristimo normalnu raspodelu za određivanje intervala poverenja za μ, jer je uzoračka raspodela od $\bar{X}$ normalna sa aritmetičkom sredinom μ i standardnom devijacijom $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, ako je $n/N \leq 0{,}05$.

Slučaj 2: Ako su ispunjena sledeća dva uslova:

1) standardna devijacija skupa σ je poznata
2) veličina uzorka je velika (tj. $n \geq 30$)

opet koristimo normalnu raspodelu za određivanje intervala poverenja za μ , jer prema centralnoj graničnoj teoremi je uzoračka raspodela od $\bar{X}$ normalna sa aritmetičkom sredinom μ i standardnom devijacijom $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, ako je $n/N \leq 0{,}05$.

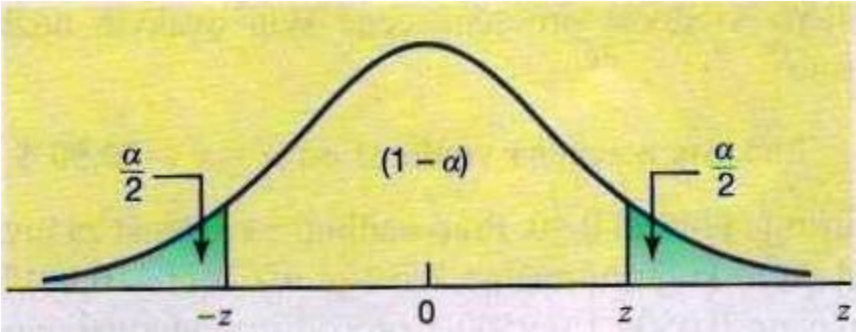
$(1-\alpha)100\%$ interval poverenja za μ u gore pomenutim slučajevima 1 i 2 je:

$$\bar{x} \pm z\sigma_{\bar{x}}$$

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

Vrednost $z\sigma_{\bar{x}}$ u formuli intervala poverenja se zove marginalna greška i označava sa E. Dakle, marginalna greška ocene od μ, koja se obeležava sa E, je vrednost koja je oduzeta i dodata vrednosti $\bar{x}$ kako bi se dobio interval poverenja za μ .

Vrednost z se dobija iz tablice standardizovane normalne raspodele za zadati nivo pouzdanosti. Postupak je sledeći. Ako nam treba na primer 95% interval poverenja za μ . Nivo pouzdanosti od 95% znači da je čitava površina ispod normalne krive za $\bar{X}$ između dve tačke na suprotnim krajevima od μ jednaka 95% ili 0,95. Neka su te dve tačke z_1 i z_2, kao na slici. Ukupna površina između z_1 i z_2 treba da bude $1-\alpha$, u našem slučaju 0,95. Dakle, na krajevima zbir površina treba da bude 0,05, što znači da površina na svakom kraju je 0,025. Da bi površine levo i desno bile po 0,025, ako nađemo u tablici 4 odgovarajuće vrednosti za z, dobićemo da je $z_1 = -1{,}96$ a $z_2 = 1{,}96$. Dakle za nivo pouzdanosti od 95% koristimo vrednost $z = 1{,}96$ u formuli za izračunavanje intervala poverenja.



Inače nivo pouzdanosti od 95% znači, da ako iz osnovnog skupa izvučemo veliki broj uzoraka iste veličine, 95% njih će dati interval poverenja koji će sadržati pravu vrednost za μ .

Pri testiranju hipoteze o aritmetikoj sredini skupa koristi se normalna raspodela u ista dva slučaja kao i kod ocenivanja.

Kada o testiranju hipoteze o μ koristimo normalnu raspodelu, onda vrednost z, za vrednost $\bar{x}$ u izabranom uzorku, izračunavamo na sledeći način:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\ , gde\ je\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Izračunatu vrednost z, na osnovu $\bar{x}$ iz izabranog uzorka, nazivamo i realizovana vrednost statistike z testa.

Zatim u zavisnosti koji pristup radimo, određujemo kritičnu oblast ili tražimo p-vrednost i donosimo zaključak o odbacivanju ili neodbacivanju nulte hipoteze.

42. Primena normalne raspodele u ocenjivanju i testiranju proporcije skupa – postupak i pretpostavke

Proporcija skupa p se ocenjuje pomoću proporcije uzorka $\hat{P}$. Za velike uzorke ($n\hat{p} > 5, n\hat{q} > 5$) važi:

1) uzoračka raspodela proporcije uzorka $\hat{P}$ je (približno) normalna.

2) aritmetička sredina uzoračke raspodele $\hat{P}$ (u oznaci $\mu_{\hat{P}}$) je jednaka proporciji skupa p.

3) standardna devijacija uzoračke raspodele proporcije uzorka $\hat{P}$ (u oznaci $\sigma_{\hat{P}}$) je $\sqrt{pq/n}$, pri čemu je q=1-p . Međutim, kada ocenjujemo vrednost proporcije skupa, mi ne znamo ni p ni q, pa ne možemo izračunati $\sigma_{\hat{P}}$. Zbog toga koristimo $s_{\hat{P}}$, koja se naziva ocena za $\sigma_{\hat{P}}$ i izračunava kao: $s_{\hat{P}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$.

Proporcija uzorka, $\hat{P}$, je tačkasta ocena odgovarajuće proporcije skupa p. Da bismo odredili interval poverenja za p, dodajemo proporciji uzorka i oduzimamo od nje broj koji se naziva marginalna greška, E.

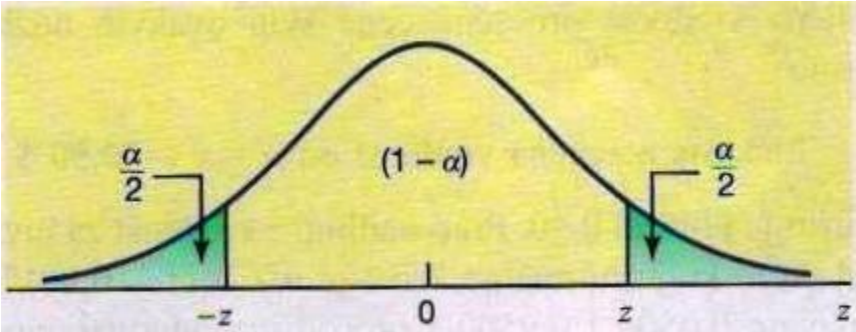
$(1-\alpha)100\%$ interval poverenja ya proporciju osnovnog skupa, p, je

$$\hat{p} \pm z s_{\hat{p}}$$

z vrednost koja se ovde koristi dobija se iy tablice standardiyovane normalne raspodele ya dati nivo pouzdanosti, a $s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$. Izraz $z s_{\hat{p}}$ se zove marginalna greška, E.

Vrednost z se dobija iz tablice standardizovane normalne raspodele za zadati nivo pouzdanosti.

Postupak je sledeći. Ako nam treba na primer 95% interval poverenja za p . Nivo pouzdanosti od 95% znači da je čitava površina ispod normalne krive za $\hat{P}$ između dve tačke na suprotnim krajevima od p jednaka 95% ili 0,95. Neka su te dve tačke z_1 i z_2, kao na slici. Ukupna površina između z_1 i z_2 treba da bude $1-\alpha$, u našem slučaju 0,95. Dakle, na krajevima zbir površina treba da bude 0,05, što znači da površina na svakom kraju je 0,025. Da bi površine levo i desno bile po 0,025, ako nađemo u tablici 4 odgovarajuće vrednosti za z, dobićemo da je $z_1 = -1{,}96$ a $z_2 = 1{,}96$. Dakle za nivo pouzdanosti od 95% koristimo vrednost $z = 1{,}96$ u formuli za izračunavanje intervala poverenja.



Inače nivo pouzdanosti od 95% znači, da ako iz osnovnog skupa izvučemo veliki broj uzoraka iste veličine, 95% njih će dati interval poverenja koji će sadržati pravu vrednost za p .
Normalna raspodela se koristi pri testiranju proporcije skupa, ako je u pitanju veliki uzorak. Uzorak je veliki ako važi $np > 5 \; i \; nq > 5$, gde je p vrednost proporcije iz nulte hipoteze, a $q = 1 - p$. Kada je uzorak veliki proporcija uzorka $\hat{P}$ je približno normalno raspodeljena sa aritmetičkom sredinom jednakom p i standardnom devijacijom $\sqrt{\frac{pq}{n}}$. Statistika testa za proporciju uzorka $\hat{P}$ izračunava se kao:

$Z = \frac{\hat{p}-p}{\sigma_{\hat{P}}}, gde \; je \; \sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$.

Vrednost p je ona koja se koristi u nultoj hipotezi, a $q = 1 - p$.
Vrednost z koja se izračunava za $\hat{p}$ primenom gore navedene jednačine naziva se realizovana vrednost z.
Ukoliko se koristi pristup zasnovan na p-vrednosti, koriste se sledeće 4 etape:
1) formulisanje nulte i alternativne hipoteze
2) izbor raspodele koja će se koristiti
3) izračunavanje p-vrednosti
4) donošenje odluke
p-vrednost se računa kao verovatnoća prikazana površinom ispod krive standardizovane normalne raspodele, koja je na kraju krive raspodele ograničena realizovanom vrednošću z. Ova verovatnoća predstavlja p-vrednost, ako je reč o jednostranom testu, ili je jednaka polovini p-vrednosti u slučaju dvostranog testa.
Ukoliko se koristi pristup zasnovan na kritičnoj vrednosti, koristi se sledećih 5 etapa:
1) formulisanje nulte i alternativne hipoteze
2) izbor raspodele koja će se koristiti
3) određivanje oblasti odbacivanja i neodbacivanja
4) izračunavanje vrednosti statistike testa
5) donošenje odluke

43. Primena studentove t raspodele pri ocenjivanju i testiranju parametara skupa

t raspodela ima sledeće primene:
<u>1) Pri ocenjivanju aritmetičke sredine u sledeća dva slučaja</u>:
Slučaj 1: Ako su ispunjena sledeća tri uslova:
1) standardna devijacija skupa σ nije poznata

2) veličina uzorka je mala ($n < 30$)
3) osnovni skup iz koga se uzima uzorak ima normalnu raspodelu
Slučaj 2: Ako su ispunjena sledeća dva uslova:
1) standardna devijacija skupa σ nije poznata
2) veličina uzorka je velika ($n \geq 30$)
2) Pri testiranju aritmetičke sredine u sledeća dva slučaja:
Slučaj 1: Ako su ispunjena sledeća tri uslova:
1) standardna devijacija skupa σ nije poznata
2) veličina uzorka je mala ($n < 30$)
3) osnovni skup iz koga se uzima uzorak ima normalnu raspodelu
Slučaj 2: Ako su ispunjena sledeća dva uslova:
1) standardna devijacija skupa σ nije poznata
2) veličina uzorka je velika ($n \geq 30$)
3) Pri testiranju hipoteze o jednakosti aritmetičkih sredina 2 skupa ako su ispunjeni sledeći uslovi:
1) dva uzorka su nezavisna
2) standardne devijacije dva osnovna skupa, $\sigma_1\ i\ \sigma_2$, su nepoznate, ali možemo da pretpostavimo da su one jednake $\sigma_1 = \sigma_2$.
3) ispunjen je barem jedan od sledećih uslova
    1) oba uzorka su velika ($n_1 \geq 30\ i\ n_2 \geq 30$)
    2) skupovi iz kojih su izabrani uzorci imaju normalnu raspodelu

44. Testiranje hipoteze o razlici aritmetičkih sredina zasnovano na nezavisnim uzorcima

Često je neophodno uporediti aritmetičke sredine dva osnovna skupa. Dakle treba testirati razliku aritmetičkih sredina $\mu_1 - \mu_2$. Alternativna hipoteza može da glasi da su aritmetičke sredine dva osnovna skupa različite, ili da je aritmetička sredina prvog osnovnog skupa veća od aritmetičke sredine drugog osnovnog skupa, ili da je aritmetička sredina prvog skupa manja od aritmetičke sredine drugog.
1) alternativna hipoteza da se aritmetičke sredine dva osnovna skupa razlikuju može se napisati kao $\mu_1 \neq \mu_2$, a to je isto što i $\mu_1 - \mu_2 \neq 0$.
2) alternativna hipoteza da je aritmetička sredina prvog osnovnog skupa veća od aritmetičke sredine drugog osnovnog skupa može se napisati kao $\mu_1 > \mu_2$, a to je isto što i $\mu_1 - \mu_2 > 0$.
3) alternativna hipoteza da je aritmetička sredina prvog osnovnog skupa manja od aritmetičke sredine drugog osnovnog skupa može se napisati kao $\mu_1 < \mu_2$, a to je isto što i $\mu_1 - \mu_2 < 0$.
Da bi se koristila normalna raspodela moraju biti ispunjeni sledeći uslovi:
1) dva uzorka su nezavisna
2) standardne devijacije dva osnovna skupa, $\sigma_1\ i\ \sigma_2$, su poznate
3) ispunjen je barem jedan od sledećih uslova
    1) oba uzorka su velika ($n_1 \geq 30\ i\ n_2 \geq 30$)
    2) skupovi iz kojih su izabrani uzorci imaju normalnu raspodelu

Kada koristimo normalnu raspodelu, statistika z testa za $\bar{X}_1 - \bar{X}_2$ je data sledećom formulom:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}, gde\ je\ \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ako su zadovoljeni sledeći uslovi:

1) dva uzorka su nezavisna
2) standardne devijacije dva osnovna skupa, σ_1 i σ_2, su nepoznate, ali možemo da pretpostavimo da su one jednake $\sigma_1 = \sigma_2$.
3) ispunjen je barem jedan od sledećih uslova
 1) oba uzorka su velika ($n_1 \geq 30$ i $n_2 \geq 30$)
 2) skupovi iz kojih su izabrani uzorci imaju normalnu raspodelu

onda za testiranje hipoteze o razlici aritmetičkih sredina dva skupa koristimo t raspodelu. Kada su standardne devijacije dva osnovna skupa jednake, umesto σ_1 i σ_2 možemo da koristimo samo σ. Pošto je σ nepoznata, zamenjujemo je njenom tačkastom ocenom S_p koja se zove ponderisana standardna devijacija dva uzorka. Ona se računa po sledećoj formuli:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

gde su n_1 i n_2 veličine dva uzorka, a S_1^2 i S_2^2 varijanse dva uzorka. Ovde je S_p ocena standardne devijacije σ. U ovoj formuli $n_1 - 1$ je broj stepeni slobode za uzorak 1, $n_2 - 1$ je broj stepeni slobode za uzorak 2, dok $n_1 + n_2 - 2$ predstavlja broj stepeni slobode za dva uzorka posmatrana zajedno. Ocena standardne devijacije $\bar{X}_1 - \bar{X}_2$ je: $S_{\bar{X}_1 - \bar{X}_2} = S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Statistika t testa za $\bar{X}_1 - \bar{X}_2$ je data formulom:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}}$$

45. Test prilagođenosti (test oblika raspodele)

Eksperiment koji ima više od dva moguća ishoda (ili kategorije) se naziva multinomni eksperiment i ima četiri osobine. Binomna raspodela je specijalni slučaj multinomne raspodele. Eksperiment sa sledećim osobinama naziva se multinomni eksperiment:
1) sastoji se od n identičnih opita (ponavljanja)
2) svaki opit, rezultira jednim od k mogućih ishoda (ili kategorija), gde je $k > 2$.
3) opiti su nezavisni
4) verovatnoće različitih ishoda ostaju konstantne za svaki opit

Primer multinomnog eksperimenta je bacanje kocke. Ovaj eksperiment se sastoji od više identičnih bacanja kocke (opita): svako bacanje kocke (opit) rezultira jednim od šest mogućih ishoda, nezavisno je od drugih bacanja, a verovatnoće šest ishoda ostaju konstantne za svako bacanje kocke. Frekvencije dobijene u izvođenju eksperimenta nazivaju se ostvarene frekvencije. Testom prilagođenosti testiramo nultu hipotezu da ostvarene frekvencije u eksperimentu slede određenu ili teorijsku raspodelu. Test je nazvan testom prilagođenosti, zato što testiramo hipotezu da se ostvarene frekvencije dobro prilagođavaju određenom modelu. Ostvarene frekvencije se označavaju slovom O. Očekivane frekvencije se označavaju slovom E i predstavlja frekvencije koje očekujemo da dobijemo ako je nulta hipoteza tačna. Očekivana frekvencija za neku kategoriju je:

$$E = np$$

gde je n veličina uzorka, a p verovatnoća da element pripada nekoj kategoriji (modalitetu) ako je nulta hipoteza tačna.

U testu prilagođenosti broj stepeni slobode je

$$df = k - 1$$

gde k označava broj mogućih ishoda (ili kategorija) u eksperimentu.

Statistika testa prilagođenosti obeležava se sa χ^2 i njena vrednost se računa kao:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

gde su

Eksperiment koji ima više od dva moguća ishoda (ili kategorije) se naziva multinomni eksperiment i ima četiri osobine. Binomna raspodela je specijalni slučaj multinomne raspodele. Eksperiment sa sledećim osobinama naziva se multinomni eksperiment:

1) sastoji se od n identičnih opita (ponavljanja)

2) svaki opit, rezultira jednim od k mogućih ishoda (ili kategorija), gde je $k > 2$.

3) opiti su nezavisni

4) verovatnoće različitih ishoda ostaju konstantne za svaki opit

Primer multinomnog eksperimenta je bacanje kocke. Ovaj eksperiment se sastoji od više identičnih bacanja kocke (opita): svako bacanje kocke (opit) rezultira jednim od šest mogućih ishoda, nezavisno je od drugih bacanja, a verovatnoće šest ishoda ostaju konstantne za svako bacanje kocke. Frekvencije dobijene u izvođenju eksperimenta nazivaju se ostvarene frekvencije. Testom prilagođenosti testiramo nultu hipotezu da ostvarene frekvencije u eksperimentu slede određenu ili teorijsku raspodelu. Test je nazvan testom prilagođenosti, zato što testiramo hipotezu da se ostvarene frekvencije dobro prilagođavaju određenom modelu. Ostvarene frekvencije se označavaju slovom O. Očekivane frekvencije se označavaju slovom E i predstavlja frekvencije koje očekujemo da dobijemo ako je nulta hipoteza tačna. Očekivana frekvencija za neku kategoriju je:

$$E = np$$

gde je n veličina uzorka, a p verovatnoća da element pripada nekoj kategoriji (modalitetu) ako je nulta hipoteza tačna.

U testu prilagođenosti broj stepeni slobode je

$$df = k - 1$$

gde k označava broj mogućih ishoda (ili kategorija) u eksperimentu.

Statistika testa prilagođenosti obeležava se sa χ^2 i njena vrednost se računa kao:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

gde su

$$O = ostvarene\ frekvencije\ neke\ kategorije$$

$$E = očekivane\ frekvencije\ neke\ kategorije = np$$

χ^2 test prilagođenosti je uvek desnostrani test.

Da li ćemo odbaciti nultu hipotezu, zavisi od toga koliko se ostvarene i očekivane frekvencije međusobno statistički značajno razlikuju. Da bismo utvrdili koliko su velike razlike između ostvarenih i očekivanih frekvencija, ne možemo posmatrati samo $\sum(O - E)$, zato što će neke od vrednosti $O - E$ biti pozitivne, a neke negativne. Zbog toga, suma razlika biće uvek jednaka nuli. Zato, kvadriramo svaku od vrednosti $O - E$, da bismo dobili $(O - E)^2$, a zatim određujemo njihov značaj u odnosu na recipročan odnos sa očekivanim frekvencijama. Zbir dobijenoh vrednosti daje nam realizovanu vrednost statistike χ^2 testa.

Da bismo primenili test prilagođenosti, veličina uzorka mora biti dovoljno velika tako da očekivana frekvencija za svaku kategoriju bude najmanje 5.

46. Test nezavisnosti

Za svaku jedinicu posmatranja često imamo informacije koje se odnose na više od jedne promenljive. Takve informacije mogu se sumirati i prikazati korišćenjem dvosmerne tabele klasifikacije, koja se takođe naziva i tabela kontingencije ili tabela unakrsne klasifikacije.

Testom nezavisnosti za tabele kontingencije, testiramo nultu hipotezu da dva obeležja (promenljive) posmatranog osnovnog skupa nisu međusobno povezana, to jest, da su nezavisna, nasuprot alternativnoj hipotezi da su obeležja povezana (zavisna). Broj stepeni slobode za test nezavisnosti su:

$$df = (R-1)(K-1)$$

gde su R i K broj redova i kolona, respektivno, u datoj tabeli kontingencije.

Vrednost statistike testa nezavisnosti se računa kao:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

gde su O i E ostvarene i očekivane frekvencije, respektivno, za ćeliju tabele kontingencije.

Nulta hipoteza za test nezavisnosti se uvek formuliše na sledeći način: dva obeležja nisu zavisna.

Alternativna hipoteza se uvek formuliše na sledeći način: dva obeležja su zavisna.

Frekvencije ostvarene sprovođenjem eksperimenta za tabelu kontingencije nazivaju se ostvarene frekvencije. Postupak računanja očekivanih frekvencija za tabelu kontingencije razlikuje se od onog koji smo koristili u testu prilagođenosti. Očekivana frekvencija E za ćeliju tabele kontingencije dobija se korišćenjem formule:

$$E = \frac{(Suma\ reda)(Suma\ kolone)}{Veličina\ uzorka}$$

Kao i kod testa prilagođenosti i test nezavisnosti je uvek desnostrani. Za primenu hi-kvadrat testa nezavisnosti, veličina uzorka mora biti dovoljno velika da bi očekivane frekvencije u svakoj ćeliji bile najmanje 5. Ako očekivana vrednost u ćeliji nije 5, moramo ili da povećamo veličinu uzorka, ili da kombinujemo neke kategorije.

47. Testiranje hipoteze zasnovano na tri i više uzoraka

Ponekad je potrebno testirati jednakost aritmetičkih sredina ili proporcija više od dva skupa.

Za testiranje jednakosti aritmetičkih sredina više od dva skupa koristi se analiza varijanse (ANOVA).

Pretpostavimo da testiramo jednakost aritmetičkih sredina tri skupa. Nulta hipoteza je:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

a alternativna je:

$$H_1: aritmetičke\ sredine\ sva\ tri\ osnovna\ skupa\ nisu\ jednake$$

Jedan od načina za ovo testiranje je da pojedinačno ispitamo jednakost aritmetičkih sredina svaka 2 skupa, pomoću z ili t testa. Međutim, to bi značilo, u ovom primeru, izvršiti tri testiranja hipoteze, dakle testiranje bi dugo trajalo i velika bi bila verovatnoća greške I vrste. Postupak analize varijanse omogućava da istestiramo jednakost aritmetičkih sredina više skupova samo jednim testom.

Opisaćemo postupak jednofaktorske analize varijanse, kojom se porede aritmetičke sredine nekoliko osnovnih skupova. Jednofaktorska se zove jer analiziramo samo jedan faktor ili promenljivu. Tako, na primer, ako testiramo da li se tri grupe studenata koji su učili po tri različita metoda nastave, razlikuju u pogledu prosečnih rezultata ostvarenih na testu, razmatramo samo jedan faktor, a to je uticaj različitih metoda podučavanja na ostvarene rezultate uspeha učenika.

Da bi se primenila jednofaktorska analiza varijanse, potrebno je da budu ispunjeni sledeći uslovi:

1) osnovni skupovi iz kojih se biraju uzorci imaju normalnu raspodelu

2) osnovni skupovi iz kojih se biraju uzorci imaju jednake varijanse

3) uzorci izabrani iz različitih osnovnih skupova su slučajni i nezavisni

Test ANOVA se primenjuje izračunavanjem dve ocene varijanse, σ^2, raspodele osnovnog skupa: varijanse između uzoraka i varijanse unutar uzoraka. Varijansa između uzoraka se naziva i prosek kvadrata odstupanja podataka između uzoraka ili V_A (faktorska varijansa). Varijansa unutar uzoraka se naziva i prosek kvadrata odstupanja podataka unutar uzoraka ili V_R (rezidualna varijansa).

Varijansa između uzoraka, V_A, predstavlja ocenu σ^2, dobijenu na osnovu razlika između aritmetičkih sredina uzoraka uzetih iz različitih osnovnih skupova.

Varijansa unutar uzoraka, V_R, predstavlja ocenu σ^2, dobijenu na osnovu razlika unutar podataka za različite uzorke.

Jednostrani test ANOVA je uvek desnostrani, a oblast odbacivanja je na desnom kraju ispod F raspodele.

Vrednost statistike F testa za test ANOVA se izračunava kao:

$$F = \frac{V_A}{V_R}$$

V_A i V_R se izračunavaju na sledeći način:

$$V_A = \frac{S_A}{k-1} \;\; i \;\; V_R = \frac{S_R}{n-k}$$

gde je $k-1$ broj stepeni slobode brojioca, a $n-k$ broj stepeni slobode imenioca u formuli za izračunavanje vrednosti F statistike.

Suma kvadrata između uzoraka, obeležava se sa S_A i izračunava se na sledeći način:

$$S_A = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right) - \frac{(\sum x)^2}{n}$$

Suma kvadrata unutar uzoraka, obeležava se sa S_R i izračunava se na sledeći način:

$$S_R = \sum x^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right)$$

Ukoliko želimo da istestiramo jednakost proporcija više od dva skupa, koristimo test prilagođenosti.

Uzmimo na primer test jednakosti proporcija pet skupova. Odgovarajuća nulta hipoteza biće:

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5$$

Alternativna hipoteza glasiće:

$$H_1: bar\ dve\ od\ pet\ proporcija\ nisu\ jednake\ sa\ 0{,}20$$

Broj stepeni slobode je $k-1$, u ovom slučaju 4.

Vrednost statistike testa prilagođenosti se računa kao:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

gde su O i E ostvarene i očekivane frekvencije, respektivno.

Kod testiranja jednakosti proporcija, E se dobija kada se veličina uzorka podeli sa k .

48. Testiranje hipoteze o značajnosti razlike sredina više skupova

Za testiranje jednakosti aritmetičkih sredina više od dva skupa koristi se analiza varijanse (ANOVA).

Pretpostavimo da testiramo jednakost aritmetičkih sredina tri skupa. Nulta hipoteza je:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

a alternativna je:

$$H_1: aritmetičke\ sredine\ sva\ tri\ osnovna\ skupa\ nisu\ jednake$$

Jedan od načina za ovo testiranje je da pojedinačno ispitamo jednakost aritmetičkih sredina svaka 2 skupa, pomoću z ili t testa. Međutim, to bi značilo, u ovom primeru, izvršiti tri testiranja hipoteze, dakle testiranje bi dugo trajalo i velika bi bila verovatnoća greške I vrste. Postupak analize varijanse omogućava da istestiramo jednakost aritmetičkih sredina više skupova samo jednim testom.
Opisaćemo postupak jednofaktorske analize varijanse, kojom se porede aritmetičke sredine nekoliko osnovnih skupova. Jednofaktorska se zove jer analiziramo samo jedan faktor ili promenljivu. Tako, na primer, ako testiramo da li se tri grupe studenata koji su učili po tri različita metoda nastave, razlikuju u pogledu prosečnih rezultata ostvarenih na testu, razmatramo samo jedan faktor, a to je uticaj različitih metoda podučavanja na ostvarene rezultate uspeha učenika.
Da bi se primenila jednofaktorska analiza varijanse, potrebno je da budu ispunjeni sledeći uslovi:
1) osnovni skupovi iz kojih se biraju uzorci imaju normalnu raspodelu
2) osnovni skupovi iz kojih se biraju uzorci imaju jednake varijanse
3) uzorci izabrani iz različitih osnovnih skupova su slučajni i nezavisni
Test ANOVA se primenjuje izračunavanjem dve ocene varijanse, σ^2, raspodele osnovnog skupa: varijanse između uzoraka i varijanse unutar uzoraka. Varijansa između uzoraka se naziva i prosek kvadrata odstupanja podataka između uzoraka ili V_A (faktorska varijansa). Varijansa unutar uzoraka se naziva i prosek kvadrata odstupanja podataka unutar uzoraka ili V_R (rezidualna varijansa).
Varijansa između uzoraka, V_A, predstavlja ocenu σ^2, dobijenu na osnovu razlika između aritmetičkih sredina uzoraka uzetih iz različitih osnovnih skupova.
Varijansa unutar uzoraka, V_R, predstavlja ocenu σ^2, dobijenu na osnovu razlika unutar podataka za različite uzorke.
Jednostrani test ANOVA je uvek desnostrani, a oblast odbacivanja je na desnom kraju ispod F raspodele.
Vrednost statistike F testa za test ANOVA se izračunava kao:

$$F = \frac{V_A}{V_R}$$

V_A i V_R se izračunavaju na sledeći način:

$$V_A = \frac{S_A}{k-1} \ \ i \ \ V_R = \frac{S_R}{n-k}$$

gde je $k-1$ broj stepeni slobode brojioca, a $n-k$ broj stepeni slobode imenioca u formuli za izračunavanje vrednosti F statistike.
Suma kvadrata između uzoraka, obeležava se sa S_A i izračunava se na sledeći način:

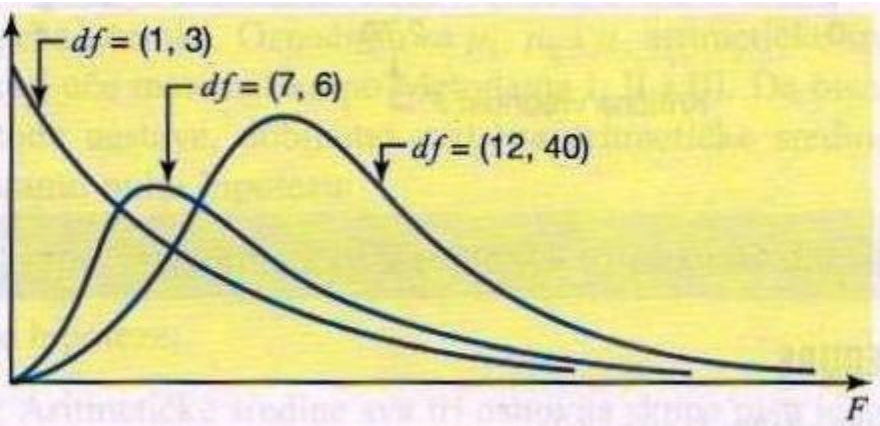
$$S_A = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right) - \frac{(\sum x)^2}{n}$$

Suma kvadrata unutar uzoraka, obeležava se sa S_R i izračunava se na sledeći način:

$$S_R = \sum x^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right)$$

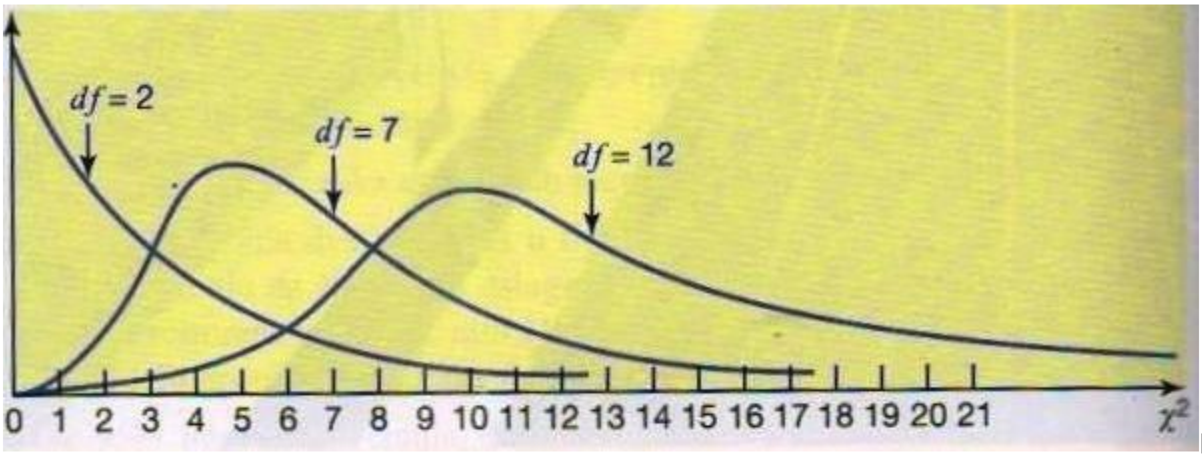
49.F-raspodela i Hi-kvadrat raspodela – osobine i primena

Oblik krive F-raspodele zavisi od broja stepeni slobode. Međutim, F-raspodela ima dva broja stepeni slobode: broj stepeni slobode za brojilac i broj stepeni slobode za imenilac. Ova dva broja su parametri F raspodele. Svaka kombinacija broja stepena slobode za brojilac i imenilac, daje drugačiju krivu F raspodele. Slučajna promenljiva F uzima samo nenegativne vrednosti. F raspodela spada u neprekidne raspodele verovatnoće. Kriva F raspodele je asimetrična udesno, ali se asimetričnost smanjuje sa povećanjem broja stepeni slobode.



F-raspodela se koristi pri testiranju jednakosti aritmetičkih sredina više od dva skupa (ANOVA).

χ^2 raspodela ima samo jedan parametar koji se zove broj stepeni slobode (df). Oblik krive hi-kvadrat raspodele zavisi od broja stepeni slobode. Slučajna promenljiva χ^2 je definisana samo za nenegativne vrednosti. Kriva hi-kvadrat raspodele počinje od nule i prostire se udesno od vertikalne ose. Hi-kvadrat raspodela je asimetrična, ali se ta asimetričnost smanjuje sa povećanjem broja stepeni slobode.



Hi-kvadrat raspodela se koristi kod:

1) testa prilagođenosti
2) testa nezavisnosti
3) ocenjivanja i testiranja nepoznate varijanse skupa

50. Jednofaktorska analiza varijanse

isto kao pitanje 48

51. Funkcionalna i stohastička zavisnost i njihovo prikazivanje

U regresionom modelu, objašnjavajuća promenljiva se obično obeležava sa X, a zavisna promenljiva sa Y. Promenljiva X sa svojim regresionim koeficijentom nalazi se na desnoj strani, a Y na levoj strani nejednakosti. Odsečak se označava sa β_0, a nagib sa β_1. Dakle prost linearni regresion model je oblika:

$$Y = \beta_0 + \beta_1 X$$

β_0 pokazuje kolika je vrednost Y kada je X jednako 0, a β_1 pokazuje za koliko se promeni Y kada se X poveća za jednu jedinicu. Gornji model se naziva deterministički model i pokazuje determinističku (egzaktnu ili funkcionalnu) vezu između promenljivih X i Y. Ovaj model podrazumeva da promenljiva Y egzaktno zavisi od promenljive X i da svakoj vrednosti X odgovara samo jedna vrednost Y. Primeri funkcionalnih veza sreću se recimo u matematici, gde ako je stranica kvadrata 2 cm, obim je 8 cm i ne može da bude ni 9 cm ni 7 cm. Međutim u ekonomiji uglavnom važe stohastičke veze, gde jednoj vrednosti X-a odgovara više različitih vrednosti Y-a. Tako na primer, ako je dohodak raznih porodica isti i iznosi 1000 €, ne znači da sve porodice moraju isto novca trošiti na hranu. Iznos novca koji porodice troše na hranu zavisi i od drugih faktora, kao što su: broj članova porodice, razlika u preferencijama i ukusima, da li imaju sopstvenu proizvodnju hrane … Uticaj tih faktora u prostom linearnom regresionom modelu obuhvatamo dodatnim članom, koji nazivamo slučajnom greškom i najčešće označavamo grčkim slovom ε. Sada je regresioni model sledećeg oblika:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

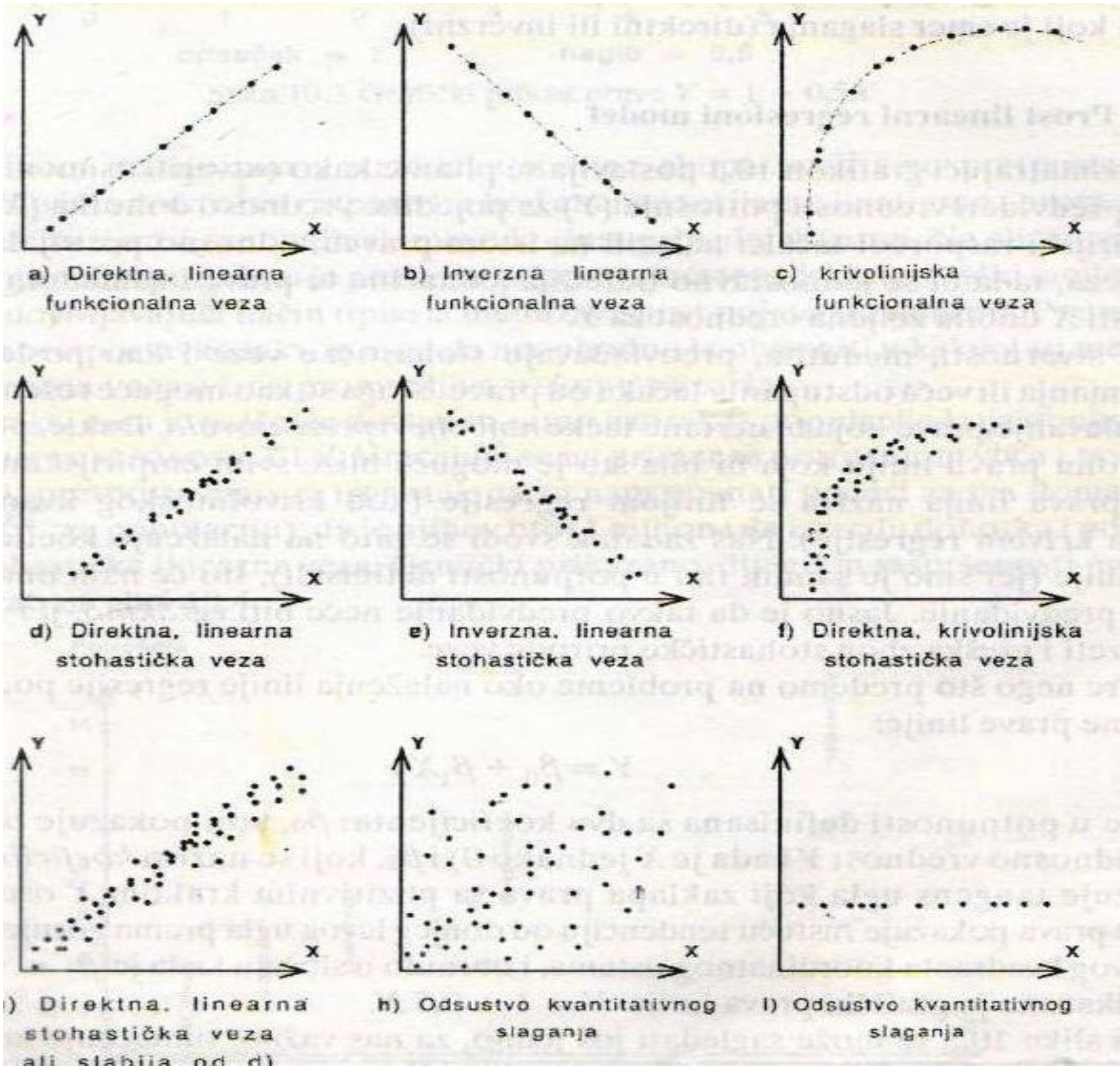
Takav model se naziva stohastički model.

Slučajnom greškom, ε, u modelu su obuhvaćene:

1) Nedostajuće ili izostavljene promenljive. Na primer, već je rečeno da izdaci na hranu zavise osim od dohotka i od mnogih drugih promenljivih

2) Slučajne varijacije. Domaćinstvo, na primer, može da nekog meseca organizuje više zabava, tako da izdaci na hranu budu mnogo viši od uobičajenih, a nekog drugog meseca izdaci budu manji od proseka jer je izdvojen novac za kupovinu nameštaja.

Ukoliko iz skupa izaberemo uzorak i zabeležimo podatke o vrednostima promenljivih X i Y u tom uzorku, a zatim te podatke prikažemo grafički, dobijamo dijagram raspršenosti, koji nam često može ukazati na oblik zavisnosti posmatranih promenljivih. Na slici ispod su prikazane neke karakteristične sizuacije:



52. Utvrđivanje veze između atributivnih obeležja (nominalna i ordinalna skala)

Za utvrđivanje veze između atributivnih obeležja koriste se dva statistička pokazatelja:

1) Spirmenov koeficijent korelacije ranga

2) χ^2 test nezavisnosti

Ukoliko je bar jedno od obeležja na nominalnoj skali, za utvrđivanje veze se koristi χ^2 test nezavisnosti, a ako su oba obeležja na ordinalnoj skali, koristi se Spirmenov koeficijent korelacije ranga.

<u>χ^2 test nezavisnosti:</u>

Za svaku jedinicu posmatranja često imamo informacije koje se odnose na više od jedne promenljive. Takve informacije mogu se sumirati i prikazati korišćenjem dvosmerne tabele klasifikacije, koja se takođe naziva i tabela kontingencije ili tabela unakrsne klasifikacije.

Testom nezavisnosti za tabele kontingencije, testiramo nultu hipotezu da dva obeležja (promenljive) posmatranog osnovnog skupa nisu međusobno povezana, to jest, da su nezavisna, nasuprot alternativnoj hipotezi da su obeležja povezana (zavisna). Broj stepeni slobode za test nezavisnosti su:

$$df = (R-1)(K-1)$$

gde su R i K broj redova i kolona, respektivno, u datoj tabeli kontingencije.

Vrednost statistike testa nezavisnosti se računa kao:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

gde su O i E ostvarene i očekivane frekvencije, respektivno, za ćeliju tabele kontingencije.
Nulta hipoteza za test nezavisnosti se uvek formuliše na sledeći način: dva obeležja nisu zavisna.
Alternativna hipoteza se uvek formuliše na sledeći način: dva obeležja su zavisna.
Frekvencije ostvarene sprovođenjem eksperimenta za tabelu kontingencije nazivaju se ostvarene frekvencije. Postupak računanja očekivanih frekvencija za tabelu kontingencije razlikuje se od onog koji smo koristili u testu prilagođenosti. Očekivana frekvencija E za ćeliju tabele kontingencije dobija se korišćenjem formule:

$$E = \frac{(Suma\ reda)(Suma\ kolone)}{Veličina\ uzorka}$$

Kao i kod testa prilagođenosti i test nezavisnosti je uvek desnostrani. Za primenu hi-kvadrat testa nezavisnosti, veličina uzorka mora biti dovoljno velika da bi očekivane frekvencije u svakoj ćeliji bile najmanje 5. Ako očekivana vrednost u ćeliji nije 5, moramo ili da povećamo veličinu uzorka, ili da kombinujemo neke kategorije. Ukoliko su obeležja zavisna, jačinu te veze možemo utvrditi pomoću koeficijenta kontingencije, C, koji se računa po sledećoj formuli:

$$C = \sqrt{\frac{\chi^2}{n(L-1)}}, \qquad gde\ je\ L = min(R, K)$$

C mo\e da uyme vrednost od 0 do 1.

<u>Spirmenov koeficijent korelacije ranga:</u>

Spirmenov koeficijent korelacije ranga predstavlja koeficijent proste linearne korelacije između rangova. U uzorku se obeležava sa r_S, a u skupu sa ρ_S. Da bi se izračunala vrednost r_S vrši se rangiranje podataka za svaku promenljivu , x i y posebno, a njihovim rangovima se dodeljuju simboli u i v. Razlika između svakog para rangova se obeležava sa $d = u - v$. Zatim se izračunaju kvadrati svake razlike d i saberu se da bi se dobilo $\sum d^2$. Na kraju se izračuna vrednost r_S, po formuli:

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

U sprovođenju testiranja hipoteze o Spirmenovom koeficijentu korelacije ranga ρ_S, koristi se statistika testa r_S, a njegova realizovana vrednost se računa pomoću gornje formule.
r_S može da uzme vrednost između -1 i 1. Ako je $r_S = 0 \Rightarrow$ u uzorku ne postoji monotona veza između X i Y. Ako je $r_S > 0$ u pitanju je direktna monotona veza između X i Y, što znači da sa rastom X raste Y, a ako je Ako je $r_S < 0$ u pitanju je inverzna monotona veza između X i Y, što znači da sa rastom X opada Y.

53. Ciljevi regresione i korelacione analize

Dva osnovna cilja, odnosno dve upotrebe regresionog modela su:
1) Ocenjivanje prosečne vrednosti Y za datu vrednost X. Regresionu pravu uzorka možemo koristiti da bismo ocenili prosečni nivo izdataka za hranu svih domaćinstava osnovnog skupa sa određenim nivoom dohotka
2) Predviđanje pojedinačne vrednosti Y za datu vrednost X. Na osnovu regresione prave uzorka možemo predvideti nivo izdataka za hranu slučajno odabranog domaćinstva sa određenim nivoom dohotka

<u>Upotreba regresionog modela za ocenjivanje prosečne vrednosti Y</u>

Regresioni model osnovnog skupa glasi:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Očekivana vrednost Y za datu vrednost X označavamo sa $\mu_{Y|X}$ i važi:

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

Jedan od ciljeva regresione analiye jeste upravo ocenjivanje $\mu_{Y|X}$ u osnovnom skupu. Vrednost $\hat{Y}$ dobijena zamenom vrednosti X u regresionoj jednačini uzorka, predstavlja tačkastu ocenu prosečne vrednosti promenljive Y u osnovnom skupu, $\mu_{Y|X}$.

Da bismo formirali interval poverenja za $\mu_{Y|X}$, potrebno je da znamo prosečnu vrednost, standardnu devijaciju i oblik uzoračke raspodele tačkaste ocene $\hat{Y}$. Tačkasta ocena $\hat{Y}$ ima normalnu raspodelu sa sredinom $\beta_0 + \beta_1 X$ i standardnom devijacijom:

$$\sigma_{\hat{Y}_p} = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

gde je $\sigma_{\hat{Y}_p}$ standardna greška ocene prosečne vrednosti Y, X_p je vrednost promenljive X za koju vršimo ocenjivanje sredine $\mu_{Y|X}$, a σ_ε predstavlja standardnu devijaciju slučajne greške ε. Budući da je σ_ε nepoznata, ocenjujemo je na osnovu standardne greške regresije, S, a $\sigma_{\hat{Y}_p}$ zamenjujemo njenom ocenom, $S_{\hat{Y}_p}$. Zato formiranje intervala poverenja za $\mu_{Y|X}$ zasnivamo na t raspodeli verovatnoća.

$(1-\alpha)100\%$ interval poverenja za $\mu_{Y|X}$ kada je $X = X_p$, glasi:

$$\hat{Y}_p \pm t S_{\hat{Y}_p}$$

gde je t tablična vrednost za površinu od $\alpha/2$ ispod krive na svakom kraju t raspodele i df=n-2.

Standardna greška ocene prosečne vrednosti zavisne promenljive, $S_{\hat{Y}_p}$, glasi:

$$S_{\hat{Y}_p} = S \sqrt{\frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

<u>Upotreba regresionog modela za predviđanje individualne vrednosti zavisne prom. Y:</u>

Druga važna upotreba regresionog modela je za potrebe predviđanja pojedinačne vrednosti promenljive Y za datu vrednost promenljive X. Predviđena vrednost izdataka pojedinačnog domaćinstva se označava sa Y_p. Kao tačkasta ocena za Y_p koristi se $\hat{Y}$. Interval poverenja za Y_p najčešće se naziva interval predviđanja.

Tačkasta ocena $\hat{Y}$ individualne vrednosti Y_p ima normalnu raspodelu sa sredinom $\beta_0 + \beta_1 X$ i standardnom devijacijom:

$$\sigma_{Y_p} = \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

gde je σ_{Y_p} standardna greška predviđanja individualne vrednosti Y, X_p je vrednost promenljive X za koju vršimo predviđanje individualne vrednosti Y_p, a σ_ε predstavlja standardnu devijaciju slučajne greške ε. Budući da je σ_ε nepoznata, ocenjujemo je na osnovu standardne greške regresije, S, a σ_{Y_p} zamenjujemo njenom ocenom, S_{Y_p}. Zato formiranje intervala poverenja za Y_p zasnivamo na t raspodeli verovatnoća.

$(1-\alpha)100\%$ interval poverenja za Y_p kada je $X = X_p$, glasi:

$$\hat{Y}_p \pm t S_{Y_p}$$

gde je t tablična vrednost za površinu od $\alpha/2$ ispod krive na svakom kraju t raspodele i df=n-2.

Standardna greška predviđanja individualne vrednosti zavisne promenljive, S_{Y_p}, glasi:

$$S_{Y_p} = S\sqrt{1 + \frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

Cilj korelacione analize jeste da se utvrdi da li između varijacija posmatranih pojava postoji kvantitativno slaganje (korelaciona veza) i ako postoji u kom stepenu. Ako se pri tome posmatraju dve pojave govori se o prostoj korelaciji,a prilikom analize više pojava o višestrukoj korelaciji. Koeficijent proste linearne korelacije predstavlja stepen kvantitativnog slaganja između promenljivih. Koeficijent proste linearne korelacije u skupu obeležava se sa ρ, a u uzorku sa r. Koeficijent proste linearne korelacije može uzeti vrednost samo u intervalu između -1 i 1.

$$-1 \leq \rho \leq 1 \quad i \quad -1 \leq r \leq 1$$

Pošto koeficijent korelacije u osnovnom skupu, ρ, nije poznat, linearnu korelacionu analizu sprovodimo na osnovu koeficijenta korelacije uzorka, r.

Ako je $r = 1$, između dve promenljive postoji perfektna pozitivna linearna korelacija. U tom slučaju, sve empirijske tačke na dijagramu raspršenosti se nalaze na rastućoj pravoj. Ako je $r = -1$, između dve promenljive postoji perfektna negativna linearna korelacija. U tom slučaju, sve empirijske tačke na dijagramu raspršenosti se nalaze na opadajućoj pravoj. Ako su empirijske tačke raspršene svuda po dijagramu, tada kažemo da između dve promenljive ne postoji linearna korelacija, odnosno da je koeficijent r približno 0.

U praksi se ne bavimo perfektnom pozitivnom ili negativnom korelacijom, već je u pitanju pozitivna linearna veza ($0 < r < 1$) ili negativna linearna veza ($-1 < r < 0$). U slučaju da je korelacija pozitivna i blizu 1, kažemo da postoji veoma jaka pozitivna linearna korelacija, a ako je bliži nuli slaba pozitivna linearna korelacija. Obrnuto u slučaju da je koeficijent r negativan i bliži -1, kaže se da u uzorku postoji veoma jaka negativna linearna korelacija, a ako je r negativan i bliži nuli, kažemo da u uzorku postoji slaba negativna linearna korelacija.

Koeficijent proste linearne korelacije se računa pomoću formule:

$$r = \frac{SP_{xy}}{\sqrt{SK_{xx}SK_{yy}}}$$

Znak koeficijenta korelacije zavisi od SP_{xy}.

Testiranje koeficijenta linearne korelacije u skupu radi se primenom t raspodele, gde se vrednost t testa računa po formuli:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Broj stepeni slobode je $df = n - 2$.

54. Model proste linearne regresije (pretpostavke i primena)

U regresionom modelu, objašnjavajuća promenljiva se obično obeležava sa X, a zavisna promenljiva sa Y. Promenljiva X sa svojim regresionim koeficijentom nalazi se na desnoj strani, a Y na levoj strani jednakosti. Odsečak se označava sa β_0, a nagib sa β_1. Dakle prost linearni regresion model je oblika:

$$Y = \beta_0 + \beta_1 X$$

β_0 pokazuje kolika je vrednost Y kada je X jednako 0, a β_1 pokazuje za koliko se promeni Y kada se X poveća za jednu jedinicu. Gornji model se naziva deterministički model i pokazuje determinističku (egzaktnu ili funkcionalnu) vezu između promenljivih X i Y. Ovaj model podrazumeva da promenljiva Y egzaktno zavisi od promenljive X i da svakoj vrednosti X odgovara samo jedna vrednost Y. Primeri funkcionalnih veza sreću se recimo u matematici, gde ako je stranica kvadrata 2 cm, obim je 8 cm i ne može da bude ni 9 cm ni 7 cm. Međutim u ekonomiji uglavnom važe stohastičke veze, gde jednoj vrednosti X-a odgovara više različitih vrednosti Y-a. Tako na primer, ako je dohodak raznih porodica isti i iznosi 1000 €, ne znači da sve porodice moraju isto novca trošiti na hranu. Iznos novca koji porodice troše na hranu zavisi i od drugih faktora, kao što su: broj članova porodice, razlika u preferencijama i ukusima, da li imaju sopstvenu proizvodnju hrane … Uticaj tih faktora u prostom linearnom regresionom modelu obuhvatamo dodatnim članom, koji nazivamo slučajnom greškom i najčešće označavamo grčkim slovom ε. Sada je regresioni model sledećeg oblika:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Takav model se naziva stohastički model.

Slučajnom greškom, ε, u modelu su obuhvaćene:

1) Nedostajuće ili izostavljene promenljive. Na primer, već je rečeno da izdaci na hranu zavise osim od dohotka i od mnogih drugih promenljivih

2) Slučajne varijacije. Domaćinstvo, na primer, može da nekog meseca organizuje više zabava, tako da izdaci na hranu budu mnogo viši od uobičajenih, a nekog drugog meseca izdaci budu manji od proseka jer je izdvojen novac za kupovinu nameštaja.

Osnovne pretpostavke regresionog modela su:

1) Linearnost. Između pojedinih vrednosti nezavisne promenljive X i odgovarajućih prosečnih vrednosti Y, postoji linearna veza

2) $E(\varepsilon) = 0$. To znači da je stohastički član (slučajna greška) u proseku jednak nuli.

3) Homoskedastičnost. Ova pretpostavka se tiče opsega odstupanja stohastičkih članova i kaže da svi stohastički članovi imaju jednaka odstupanja,preciznije jednake varijanse:

$$Var(\varepsilon_1) = Var(\varepsilon_2) = \cdots = Var(\varepsilon_N) = \sigma^2$$

4) Nema autokorelacije. To znači da između bilo koja dva stohastička člana ε_i i ε_j ne postoji linearna veza.

5) X nije slučajna promenljiva. Ova pretpostavka ukazuje na to da su vrednosti nezavisne promenljive fiksirane, tj. da ih istraživač unapred mora odabrati pre uzimanja uzorka.

6) ε_i ima normalan raspored. Stoga možemo napisati:

$$\varepsilon_i : N(0, \sigma^2)$$

tj. stohastički član ima normalan raspored sa aritmetičkom sredinom 0 i varijansom σ^2.

Dva osnovna cilja, odnosno dve upotrebe regresionog modela su:

1) Ocenjivanje prosečne vrednosti Y za datu vrednost X. Regresionu pravu uzorka možemo koristiti da bismo ocenili prosečni nivo izdataka za hranu svih domaćinstava osnovnog skupa sa određenim nivoom dohotka

2) Predviđanje pojedinačne vrednosti Y za datu vrednost X. Na osnovu regresione prave uzorka možemo predvideti nivo izdataka za hranu slučajno odabranog domaćinstva sa određenim nivoom dohotka

<u>Upotreba regresionog modela za ocenjivanje prosečne vrednosti Y</u>

Regresioni model osnovnog skupa glasi:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Očekivana vrednost Y za datu vrednost X označavamo sa $\mu_{Y|X}$ i važi:

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

Jedan od ciljeva regresione analiye jeste upravo ocenjivanje $\mu_{Y|X}$ u osnovnom skupu. Vrednost $\hat{Y}$ dobijena zamenom vrednosti X u regresionoj jednačini uzorka, predstavlja tačkastu ocenu prosečne vrednosti promenljive Y u osnovnom skupu, $\mu_{Y|X}$.
Da bismo formirali interval poverenja za $\mu_{Y|X}$, potrebno je da znamo prosečnu vrednost, standardnu devijaciju i oblik uzoračke raspodele tačkaste ocene $\hat{Y}$. Tačkasta ocena $\hat{Y}$ ima normalnu raspodelu sa sredinom $\beta_0 + \beta_1 X$ i standardnom devijacijom:

$$\sigma_{\hat{Y}_p} = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

gde je $\sigma_{\hat{Y}_p}$ standardna greška ocene prosečne vrednosti Y, X_p je vrednost promenljive X za koju vršimo ocenjivanje sredine $\mu_{Y|X}$, a σ_ε predstavlja standardnu devijaciju slučajne greške ε. Budući da je σ_ε nepoznata, ocenjujemo je na osnovu standardne greške regresije, S, a $\sigma_{\hat{Y}_p}$ zamenjujemo njenom ocenom, $S_{\hat{Y}_p}$. Zato formiranje intervala poverenja za $\mu_{Y|X}$ zasnivamo na t raspodeli verovatnoća.
$(1-\alpha)100\%$ interval poverenja za $\mu_{Y|X}$ kada je $X = X_p$, glasi:

$$\hat{Y}_p \pm t S_{\hat{Y}_p}$$

gde je t tablična vrednost za površinu od $\alpha/2$ ispod krive na svakom kraju t raspodele i df=n-2.
Standardna greška ocene prosečne vrednosti zavisne promenljive, $S_{\hat{Y}_p}$, glasi:

$$S_{\hat{Y}_p} = S \sqrt{\frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

<u>Upotreba regresionog modela za predviđanje individualne vrednosti zavisne prom. Y:</u>
Druga važna upotreba regresionog modela je za potrebe predviđanja pojedinačne vrednosti promenljive Y za datu vrednost promenljive X. Predviđena vrednost izdataka pojedinačnog domaćinstva se označava sa Y_p. Kao tačkasta ocena za Y_p koristi se $\hat{Y}$. Interval poverenja za Y_p najčešće se naziva interval predviđanja.
Tačkasta ocena $\hat{Y}$ individualne vrednosti Y_p ima normalnu raspodelu sa sredinom $\beta_0 + \beta_1 X$ i standardnom devijacijom:

$$\sigma_{Y_p} = \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

gde je σ_{Y_p} standardna greška predviđanja individualne vrednosti Y, X_p je vrednost promenljive X za koju vršimo predviđanje individualne vrednosti Y_p, a σ_ε predstavlja standardnu devijaciju slučajne greške ε. Budući da je σ_ε nepoznata, ocenjujemo je na osnovu standardne greške regresije, S, a σ_{Y_p} zamenjujemo njenom ocenom, S_{Y_p}. Zato formiranje intervala poverenja za Y_p zasnivamo na t raspodeli verovatnoća.
$(1-\alpha)100\%$ interval poverenja za Y_p kada je $X = X_p$, glasi:

$$\hat{Y}_p \pm t S_{Y_p}$$

gde je t tablična vrednost za površinu od $\alpha/2$ ispod krive na svakom kraju t raspodele i df=n-2.
Standardna greška predviđanja individualne vrednosti zavisne promenljive, S_{Y_p}, glasi:

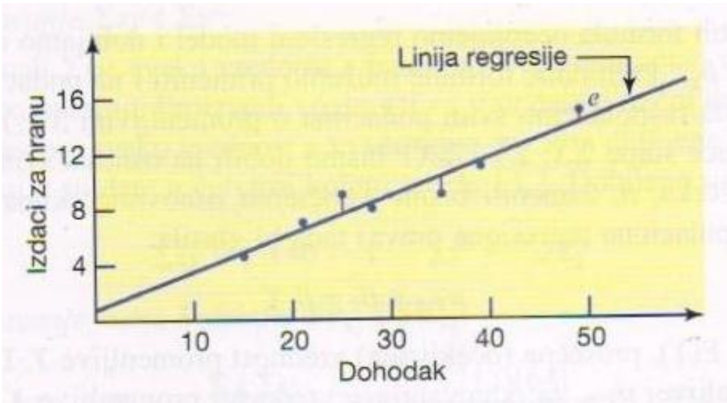
$$S_{Y_p} = S\sqrt{1 + \frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

55. Metod najmanjih kvadrata

Stvarne vrednosti promenljive Y nazivaju se još i empirijske vrednosti. Razlika između stvarne (empirijske) i očekivane (prosečne) vrednosti Y u osnovnom skupu predstavlja slučajnu grešku, ε. Na primer, za posmatrano domaćinstvo, ε predstavlja razliku između iznosa koji je to domaćinstvo proteklog meseca zaista trošilo na hranu i prosečnog iznosa (dobijenog na osnovu regresione prave osnovnog skupa). Budući da regresioni model osnovnog skupa ocenjujemo na regresione prave uzorka, vrednost Y u uzorku za datu vrednost X naziva se ocenjena ili prilagođena vrednost, $\hat{Y}$. Razlika između stvarne i ocenjene vrednosti promenljive Y u uzorku naziva se rezidual i označava se sa e. Rezidual predstavlja ocenu slučajne greške, ε. Dakle, ocenjivanjem modela na osnovu podataka iz izabranog slučajnog uzorka dobijamo:

$$e = ostvareni\ izdaci\ za\ hranu - ocenjeni\ izdaci\ za\ hranu = y - \hat{y}$$

Na slici ispod su prikazani reziduali:



Reziduali su pozitivni ako su stvarne vrednosti izdataka za hranu veće od ocenjenih vrednosti, a negativni ukoliko su stvarne vrednosti izdataka za hranu manje od ocenjenih vrednosti. Suma svih reziduala je uvek jednaka nuli.

$$\sum e = \sum\left(Y - \hat{Y}\right) = 0$$

Pošto je suma reziduala uvek jednaka nuli, njenim minimiziranjem ne možemo dobiti regresionu pravu koja se najbolje prilagođava empirijskim podacima. Umesto toga, minimiziramo sumu kvadrata reziduala (SKR):

$$SKR = \sum e^2 = \sum\left(Y - \hat{Y}\right)^2$$

Primenom metoda najmanjih kvadrata dobijamo vrednost b_0 i b_1 u regresionom modelu uzorka, tako da suma kvadrata reziduala bude minimalna.

Minimiziranjem sume kvadrata reziduala dobijaju se b_0 i b_1, koji predstavljaju ocene regresionih parametara β_0 i β_1, a regresiona prava koja se na osnovu tih ocena dobija naziva se regresiona prava uzorka.

Koeficijenti regresione prave uzorka $\hat{Y} = b_0$+b_1X, odnosno ocene po metodu najmanjih kvadrata glase:

$$b_{\mathbf{1}} = \frac{SP_{xy}}{SK_{xx}} \quad i \quad b_0 = \bar{Y} - b_1\bar{X}$$

gde je $SP_{xy} = \sum XY - \frac{\sum X \sum Y}{n}$ i $SK_{xx} = \sum X^2 - \frac{(\sum X)^2}{n}$

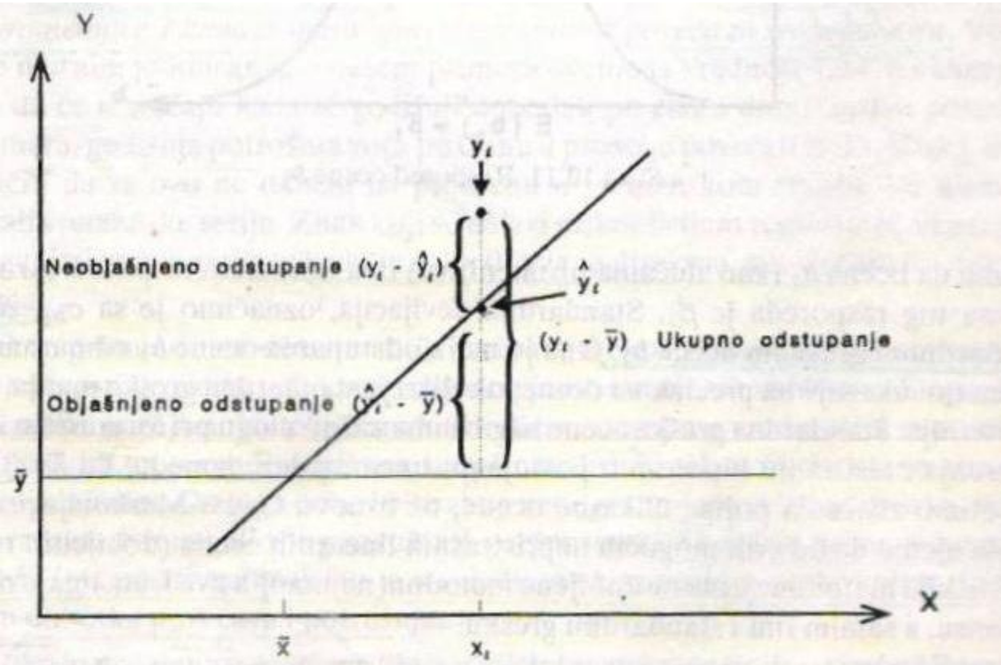
i gde SK i SP označavaju odgovarajuću sumu kvadrata, odnosno sumu proizvoda.

Korišćenjem datih formula ocenjujemo regresioni model i dobijamo ocene po metodu najmnjih kvadrata b_0 i b_1. Prethodne formule možemo primeniti i na podacima osnovnog skupa, pod pretpostavkom da raspolažemo svim podacima o promenljivim X i Y u osnovnom skupu. Pri tome, odgovarajuće sume bismo dobili na osnovu svih podataka osnovnog skupa, a veličinu uzorka, n, zamenili bismo veličinom osnovnog skupa, N. Linija regresije osnovnog skupa tada bi glasila:

$\mu_{Y|X} = \beta_0 + \beta_1$,

gde je $\mu_{Y|X}$, odnosno E(Y), prosečna vrednost promenljive Y.

56. Ukupan, objašnjen i neobjašnje varijabilitet u regresionoj analizi



Na dijagramu smo uzeli proizvoljnu empirijsku vrednost y_i iz uzorka koja odgovara vrednosti nezavisne promenljive x_i. Pošto je aritmetička sredina serije y konkretnog uzorka konstantna, ona ne zavisi od serije x, pa se može ucrtati kao linija paralelna x-osi. Na sličan način kao što smo kod varijanse obeležja pošli od odstupanja pojedinih vrednosti od aritmetičke sredine, tako se i ovde polazi od odstupanja pojedine vrednosti y_i od aritmetičke sredine serije y, $\bar{y}$. Takvo odstupanje naziva se ukupno odstupanje. Budući da posmatramo stohastičku vezu, jedan deo tog odstupanja promenljive Y nije objašnjen promenljivom X, što je na slici dato kao odstupanje $(y_i - \hat{y}_i)$. To odstupanje nazivali smo rezidualom. Kako se ne može pripisati promenljivoj X (već dejstvu slučajne greške, pa tačka nije na pravoj), rezidual smatramo neobjašnjenim odstupanjem. Nasuprot tome, deo odstupanja $(\hat{y}_i - \bar{y})$ je objašnjen regresionom vezom između X i Y i naziva se objašnjenim odstupanjem. Naime, regresionom vezom je objašnjeno da je konkretna vrednost y_i na našem dijagramu veća od aritmetičke sredine $\bar{y}$, zato što je:

a) i odgovarajuća vrednost x_i veća od svoje aritmetičke sredine

b) postoji linearna direktna veza između varijacija promenljivih.

Ukupno odstupanje zavisne promenljive Y stoga možemo tretirati kao zbir objašnjenog i neobjašnjenog odstupanja:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) \quad + \quad (\hat{y}_i - \bar{y})$$

Ukupno odstupanje | Neobjašnjeno odstupanje | Objašnjeno odstupanje

Može se pokazati da će jednakost nastaviti da važi i kada obe strane kvadriramo i sumiramo za sve vrednosti u uzorku. Pošto su sada obuhvaćene sve vrednosti zavisne promenljive u uzorku, kažemo da je ukupan varijabilitet jednak zbiru objašnjenog i neobjašnjenog varijabiliteta:

$$\sum (y_i - \bar{y})^2 = \sum \quad (y_i - \hat{y}_i)^2 \quad + \quad \sum (\hat{y}_i - \bar{y})^2$$

SKU	**SKN**	**SKO**
Ukupna suma kvadrata (Ukupan varijabilitet)	Neobjašnjena suma kvadrata (Neobjašnjen varijabilitet)	Objašnjena suma kvadrata (Objašnjen varijabilitet)

Na taj način, slično kao u analizi varijanse, ukupnu sumu kvadrat razloili smo na dva dela. Objašnjena suma kvadrata često se naziva i regresionom sumom kvadrata, a neobjašnjena suma kvadrata rezidualnom ili sumom kvadrata greške.

57.Ocene regresionih parametara po metodu najmanjih kvadrata – osobine i interpretacija

pitanje 55 +

Ocenjena vrednost b_0 pokazuje odsečak na Y osi u dijagramu raspršenosti i predstavlja prosečnu vrednost y, kada je $x = 0$.

b_1 predstavlja ocenjenu vrednost prosečne promene zavisne promenljive Y kada se nezavisna promenljiva X poveća za svoju jedinicu. Znak koji se nalazi uz koeficijent regresije b_1 ukazuje na smer slaganja između pojava. Kada je $b_1 > 0$ veza je direktna, a u slučaju da je $b_1 < 0$ veza je inverzna.

O kvalitetu ocena dobijenih metodom najmanjih kvadrata u poređenju sa ocenama dobijenim drugim metodima govori Gauss-Markovljeva teorema: „Ako su ispunjene pretpostavke regresionog modela, ocene dobijene metodom najmanjih kvadrata su najbolje (efikasne), nepristrasne linearne ocene.

Na osnovu teoreme sledi da je:

$$E(b_0) = \beta_0$$

i

$$E(b_1) = \beta_1$$

odnosno da su ocene b_0 i b_1 u proseku jednake nepoynatim parametrima β_0 i β_1.

Ocena b_1 kao slučajna promenljiva ima normalan raspored. Aritmetička sredina tog rasporeda je β_1. Standardna devijacija, σ_{b1}, naziva se standardnom greškom ocene b_1. Ona je mera odstupanja ocene b_1 od parametra β_1 i samim tim ukazuje na preciznost ocene, jer ukoliko je standardna greška manja, ocena je kvalitetnija.

Ocene dobijene metodom najmanjih kvadrata imaju najmanju varijansu, a samim tim i standardnu grešku.

58.Koeficijent proste i višestruke determinacije – poređenje

Koeficijent detreminacije prostog linearnog modela , r^2, pokazuje koliko je učešće objašnjenog varijabiliteta (SKO) u ukupnom (SKU), odnosno pokazuje koliki je deo varijacija zavisne promenljive objašnjen prostim regresionim modelom. Koeficijent detreminacije višestrukog regresionog modela se naziva koeficijent višestruke determinacije, označava se sa R^2 i pokazuje koliko je učešće objašnjenog varijabiliteta u ukupnom, odnosno koliki je deo varijacija zavisne promenljive objašnjen višestrukim regresionim modelom. Ovaj koeficijent predstavlja meru validnosti višestrukog regresionog modela, odnosno pokazuje da li izabrane objašnjavajuće promenljive dobro objašnjavaju varijacije zavisne promenljive.

Kao i r^2, vrednost koeficijenta višestruke determinacije, R^2 takođe može uzeti vrednosti u intervalu od 0 do 1, odnosno, $0 \leq R^2 \leq 1$.

Isto kao i u slučaju modela proste linearne regresije, SKU je ukupna suma kvadrata ili ukupni varijabilitet, SKO je suma kvadrata objašnjenog varijabiliteta, dok je SKR(SKN) suma kvadrata reziduala (neobjašnjenog varijabiliteta). SKU je jednak zbiru SKO i SKN. Ove sume se izračunavaju na sledeći način:

$$SKN = \sum e^2 = \sum \left(Y - \hat{Y}\right)^2$$

$$SKU = SK_{yy} = \sum (Y - \bar{Y})^2$$

$$SKO = \sum \left(\hat{Y} - \bar{Y}\right)^2$$

Koeficijent višestruke determinacije predstavlja udeo SKO u SKU:

$$R^2 = \frac{SKO}{SKU}$$

Nedostatak koeficijenta višestruke determinacije R^2 je u tome što se njegova vrednost povećava sa dodavanjem novih promenljivih bez obzira da li one stvarno značajno objašnjavaju varijacije zavisne promenljive. Da bi se eliminisao pomenuti nedostatak koeficijenta R^2, koristi se korigovani koeficijent višestruke determinacije, $\bar{R}^2$. Vrednost $\bar{R}^2$ se računa na sledeći način:

$$\bar{R}^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

Za razliku od R^2 koji nikada ne može biti negativan, $\bar{R}^2$ može uzeti negativne vrednosti.

59. Standardna devijacija slučajne greške i standardna greška regresije

Standardna devijacija slučajne greške, σ_ε, pokazuje koliko je disperzija slučajnih grešaka, odnosno koliko su odstupanja promenljive Y od njenih prosečnih vrednosti na regresionoj pravoj osnovnog skupa. U primeru dohotka i izdataka za hranu, domaćinstva sa istim nivoom dohotka imaju različite izdatke za hranu, odnosno slučajna greška, ε, ima različite vrednosti za posmatrana domaćinstva. Rastojanje svake pojedinačne tačke od tačke na regresionoj pravoj osnovnog skupa predstavlja odgovarajuću vrednost slučajne greške. Standardna devijacija, σ_ε, meri disperziju tih tačaka oko regresione prave osnovnog skupa.

Budući da varijansa, σ_ε^2, i standardna devijacija slučajne greške, σ_ε, nisu poznate, ocenjuju se na osnovu reziduala slučajnog uzorka. Ocena varijanse slučajne greške, S^2, dobijena na osnovu reziduala slučajnog uzorka, naziva se rezidualna varijansa, a ocena standardne devijacije slučajne greške, S, naziva se standardna greška regresije. Standardna greška regresije dobija se na sledeći način:

$$S = \sqrt{\frac{SKR}{n-2}}$$

gde je

$$SKR = \sum \left(Y - \hat{Y}\right)^2$$

Broj stepeni slobode u prostom linearnom regresionom modelu iznosi $n-2$, jer je broj podataka uzorka umanjena za 2 ograničenja.

Za S se češće koristi formula:

$S = \sqrt{\frac{SK_{yy} - b_1 SP_{xy}}{n-2}}$, gde je $SK_{yy} = \sum Y^2 - \frac{(\sum Y)^2}{n}$.

60.Testiranje značajnosti ocena regresionih koeficijenata

Prilikom testiranja hipoteze o regresionom parametru β_1, kod prostog linearnog regresionog modela, nultu hipotezu formulišemo u obliku $H_0: \beta_1 = 0$ (Parametar regresione prave osnovnog skupa jednak je nuli) i ona je ekvivalentna hipotezi da promenljiva X ne utiče na promenljivu Y i da se regresiona prava ne može koristiti u svrhe predviđanja vrednosti Y na osnovu promenljive X. Treba imati u vidu da testiramo samo linearnu zavisnost između dve promenljive.

Da bismo testirali da li X utiče na Y, formulišemo nultu hipotezu u obliku: $H_0: \beta_1 = 0$ (X ne utiče na Y).

Alternativna hipoteza može biti formulisana na jedan od sledeća tri načina:

1) X utiče na Y, odnosno $\beta_1 \neq 0$

2) X pozitivno utiče na Y, odnosno $\beta_1 > 0$

3) X negativno utilče na Y, odnosno $\beta_1 < 0$.

Statistika t testa za testiranje hipoteze o regresionom parametru β_1 glasi:

$$t = \frac{b_1 - \beta_1}{s_{b1}}$$

Vrednost regresionog parametra β_1 u prethodnom izrazu zamenjuje se hipotetičkom vrednošću, koja je definisana nultom hipotezom.

Broj stepeni slobode je: $= n - 2$.

Testiranje hipoteze za bilo koji regresioni parametar β_i u višestrukom regresionom modelu, možemo sprovesti iste procedure koja se koristi prilikom testiranja hipoteze o β_1 u prostom regresionom modelu. Jedina razlika je što je broj stepeni slobode u višestrukoj regresiji jednak $n - k - 1$. Zbog pretpostavke o normalnoj raspodeli slučajnih grešaka, uzoračka raspodela svake ocene b_i je takođe normalna sa srednjom vrednošću β_i i standardnom greškom, σ_{b_i}. Standardna greška σ_{b_i} nije poznata, pa zato koristimo njenu ocenu iz uzorka s_{b_i}, a testiranje hipoteze o parametru β_i zasnivamo na t raspodeli.

Za testiranje hipoteze o regresionom parametru β_i koristi se sledeća statistika testa:

$$t = \frac{b_i - \beta_i}{s_{bi}}$$

Vrednost regresionog parametra β_i u prethodnom izrazu zamenjuje se hipotetičkom vrednošću, koja je definisana nultom hipotezom.

Broj stepeni slobode je: $= n - k - 1$.

61. Primena Studentove t raspodele pri ocenjivanju i testiranju u regresionoj analizi

1) Ocenjivanje regresionog parametra β_1 kod proste linearne regresije

Ocena b_1 predstavlja tačkastu ocenu parametra β_1 regresione prave osnovnog skupa. $(1-\alpha)100\%$ interval poverenja za parametar β_1 glasi:

$b_1 \pm ts_{b_1}$, gde je $s_{b_1} = \frac{s}{\sqrt{SK_{xx}}}$.
Vrednost t statistike dobijamo iz tablice Studentove raspodele za površinu od $\alpha/2$ na oba kraja t raspodele i $n-2$ stepena slobode.

2) Testiranje regresionog parametra β_1 kod proste linearne regresije

Prilikom testiranja hipoteze o regresionom parametru β_1, kod prostog linearnog regresionog modela, nultu hipotezu formulišemo u obliku $H_0: \beta_1 = 0$ (Parametar regresione prave osnovnog skupa jednak je nuli) i ona je ekvivalentna hipotezi da promenljiva X ne utiče na promenljivu Y i da se regresiona prava ne može koristiti u svrhe predviđanja vrednosti Y na osnovu promenljive X. Treba imati u vidu da testiramo samo linearnu zavisnost između dve promenljive.
Da bismo testirali da li X utiče na Y, formulišemo nultu hipotezu u obliku: $H_0: \beta_1 = 0$ (X ne utiče na Y).
Alternativna hipoteza može biti formulisana na jedan od sledeća tri načina:
1) X utiče na Y, odnosno $\beta_1 \neq 0$
2) X pozitivno utiče na Y, odnosno $\beta_1 > 0$
3) X negativno utilče na Y, odnosno $\beta_1 < 0$.
Statistika t testa za testiranje hipoteze o regresionom parametru β_1 glasi:

$$t = \frac{b_1 - \beta_1}{s_{b1}}$$

Vrednost regresionog parametra β_1 u prethodnom izrazu zamenjuje se hipotetičkom vrednošću, koja je definisana nultom hipotezom.
Broj stepeni slobode je: $= n - 2$.

3) Ocenjivanje regresionog parametra β_i kod višestruke linearne regresije

Ocene $b_0, b_1, b_2, b_3, \dots, b_k$ regresionog modela uzorka predstavlja tačkastu ocenu parametara regresionog modela skupa $\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_k$. $(1-\alpha)100\%$ interval poverenja za parametar β_i glasi: $b_i \pm ts_{b_i}$.
Vrednost t statistike dobijamo iz tablice Studentove raspodele za površinu od $\alpha/2$ na oba kraja t raspodele i $n-k-1$ stepena slobode.

4) Testiranje regresionog parametra β_i kod višestruke linearne regresije

Testiranje hipoteze za bilo koji regresioni parametar β_i u višestrukom regresionom modelu, možemo sprovesti iste procedure koja se koristi prilikom testiranja hipoteze o β_1 u prostom regresionom modelu. Jedina razlika je što je broj stepeni slobode u višestrukoj regresiji jednak $n-k-1$. Zbog pretpostavke o normalnoj raspodeli slučajnih grešaka, uzoračka raspodela svake ocene b_i je takođe normalna sa srednjom vrednošću β_i i standardnom greškom, σ_{b_i}. Standardna greška σ_{b_i} nije poznata, pa zato koristimo njenu ocenu iz uzorka s_{b_i}, a testiranje hipoteze o parametru β_i zasnivamo na t raspodeli.
Za testiranje hipoteze o regresionom parametru β_i koristi se sledeća statistika testa:

$$t = \frac{b_i - \beta_i}{s_{bi}}$$

Vrednost regresionog parametra β_i u prethodnom izrazu zamenjuje se hipotetičkom vrednošću, koja je definisana nultom hipotezom.

Broj stepeni slobode je: $= n - k - 1$.

62. Ocenjivanje i predviđanje u regresionoj analizi

Dva osnovna cilja, odnosno dve upotrebe regresionog modela su:

1) Ocenjivanje prosečne vrednosti Y za datu vrednost X. Regresionu pravu uzorka možemo koristiti da bismo ocenili prosečni nivo izdataka za hranu svih domaćinstava osnovnog skupa sa određenim nivoom dohotka

2) Predviđanje pojedinačne vrednosti Y za datu vrednost X. Na osnovu regresione prave uzorka možemo predvideti nivo izdataka za hranu slučajno odabranog domaćinstva sa određenim nivoom dohotka

<u>Upotreba regresionog modela za ocenjivanje prosečne vrednosti Y</u>

Regresioni model osnovnog skupa glasi:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Očekivana vrednost Y za datu vrednost X označavamo sa $\mu_{Y|X}$ i važi:

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

Jedan od ciljeva regresione analiye jeste upravo ocenjivanje $\mu_{Y|X}$ u osnovnom skupu. Vrednost $\hat{Y}$ dobijena zamenom vrednosti X u regresionoj jednačini uzorka, predstavlja tačkastu ocenu prosečne vrednosti promenljive Y u osnovnom skupu, $\mu_{Y|X}$.

Da bismo formirali interval poverenja za $\mu_{Y|X}$, potrebno je da znamo prosečnu vrednost, standardnu devijaciju i oblik uzoračke raspodele tačkaste ocene $\hat{Y}$. Tačkasta ocena $\hat{Y}$ ima normalnu raspodelu sa sredinom $\beta_0 + \beta_1 X$ i standardnom devijacijom:

$$\sigma_{\hat{Y}_p} = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

gde je $\sigma_{\hat{Y}_p}$ standardna greška ocene prosečne vrednosti Y, X_p je vrednost promenljive X za koju vršimo ocenjivanje sredine $\mu_{Y|X}$, a σ_ε predstavlja standardnu devijaciju slučajne greške ε. Budući da je σ_ε nepoznata, ocenjujemo je na osnovu standardne greške regresije, S, a $\sigma_{\hat{Y}_p}$ zamenjujemo njenom ocenom, $S_{\hat{Y}_p}$. Zato formiranje intervala poverenja za $\mu_{Y|X}$ zasnivamo na t raspodeli verovatnoća.

$(1-\alpha)100\%$ interval poverenja za $\mu_{Y|X}$ kada je $X = X_p$, glasi:

$$\hat{Y}_p \pm t S_{\hat{Y}_p}$$

gde je t tablična vrednost za površinu od $\alpha/2$ ispod krive na svakom kraju t raspodele i df=n-2.

Standardna greška ocene prosečne vrednosti zavisne promenljive, $S_{\hat{Y}_p}$, glasi:

$$S_{\hat{Y}_p} = S \sqrt{\frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

<u>Upotreba regresionog modela za predviđanje individualne vrednosti zavisne prom. Y:</u>

Druga važna upotreba regresionog modela je za potrebe predviđanja pojedinačne vrednosti promenljive Y za datu vrednost promenljive X. Predviđena vrednost izdataka pojedinačnog domaćinstva se označava sa Y_p. Kao tačkasta ocena za Y_p koristi se $\hat{Y}$. Interval poverenja za Y_p najčešće se naziva interval predviđanja.

Tačkasta ocena $\hat{Y}$ individualne vrednosti Y_p ima normalnu raspodelu sa sredinom $\beta_0 + \beta_1 X$ i standardnom devijacijom:

$$\sigma_{Y_p} = \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

gde je σ_{Y_p} standardna greška predviđanja individualne vrednosti Y, X_p je vrednost promenljive X za koju vršimo predviđanje individualne vrednosti Y_p, a σ_ε predstavlja standardnu devijaciju slučajne greške ε. Budući da je σ_ε nepoznata, ocenjujemo je na osnovu standardne greške regresije, S, a σ_{Y_p} zamenjujemo njenom ocenom, S_{Y_p}. Zato formiranje intervala poverenja za Y_p zasnivamo na t raspodeli verovatnoća. $(1-\alpha)100\%$ interval poverenja za Y_p kada je $X = X_p$, glasi:

$$\hat{Y}_p \pm t S_{Y_p}$$

gde je t tablična vrednost za površinu od $\alpha/2$ ispod krive na svakom kraju t raspodele i df=n-2. Standardna greška predviđanja individualne vrednosti zavisne promenljive, S_{Y_p}, glasi:

$$S_{Y_p} = S \sqrt{1 + \frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{SK_{XX}}}$$

63. Koeficijent proste korelacije i njegova interpretacija

1) Koeficijent proste linearne korelacije

Cilj korelacione analize jeste da se utvrdi da li između varijacija posmatranih pojava postoji kvantitativno slaganje (korelaciona veza) i ako postoji u kom stepenu. Ako se pri tome posmatraju dve pojave govori se o prostoj korelaciji,a prilikom analize više pojava o višestrukoj korelaciji. Koeficijent proste linearne korelacije predstavlja stepen kvantitativnog slaganja između promenljivih. Koeficijent proste linearne korelacije u skupu obeležava se sa ρ, a u uzorku sa r. Koeficijent proste linearne korelacije može uzeti vrednost samo u intervalu između -1 i 1.

$$-1 \leq \rho \leq 1 \quad i \quad -1 \leq r \leq 1$$

Pošto koeficijent korelacije u osnovnom skupu, ρ, nije poznat, linearnu korelacionu analizu sprovodimo na osnovu koeficijenta korelacije uzorka, r.

Ako je $r = 1$, između dve promenljive postoji perfektna pozitivna linearna korelacija. U tom slučaju, sve empirijske tačke na dijagramu raspršenosti se nalaze na rastućoj pravoj. Ako je $r = -1$, između dve promenljive postoji perfektna negativna linearna korelacija. U tom slučaju, sve empirijske tačke na dijagramu raspršenosti se nalaze na opadajućoj pravoj. Ako su empirijske tačke raspršene svuda po dijagramu, tada kažemo da između dve promenljive ne postoji linearna korelacija, odnosno da je koeficijent r približno 0.

U praksi se ne bavimo perfektnom pozitivnom ili negativnom korelacijom, već je u pitanju pozitivna linearna veza $(0 < r < 1)$ ili negativna linearna veza $(-1 < r < 0)$. U slučaju da je korelacija pozitivna i blizu 1, kažemo da postoji veoma jaka pozitivna linearna korelacija, a ako je bliži nuli slaba pozitivna linearna korelacija. Obrnuto u slučaju da je koeficijent r negativan i bliži -1, kaže se da u uzorku postoji veoma jaka negativna linearna korelacija, a ako je r negativan i bliži nuli, kažemo da u uzorku postoji slaba negativna linearna korelacija.

Koeficijent proste linearne korelacije se računa pomoću formule:

$$r = \frac{SP_{xy}}{\sqrt{SK_{xx}SK_{yy}}}$$

Znak koeficijenta korelacije zavisi od SP_{xy}.

Testiranje koeficijenta linearne korelacije u skupu radi se primenom t raspodele, gde se vrednost t testa računa po formuli:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Broj stepeni slobode je $df = n - 2$.

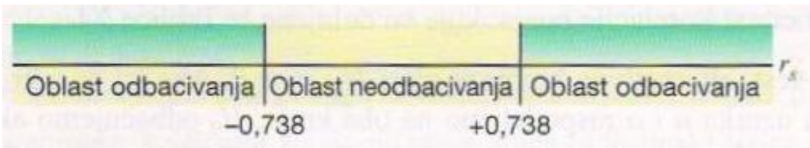
2) Spirmanov koeficijent korelacije ranga

Spirmenov koeficijent korelacije ranga predstavlja koeficijent proste linearne korelacije između rangova. U uzorku se obeležava sa r_S, a u skupu sa ρ_S. Da bi se izračunala vrednost r_S vrši se rangiranje podataka za svaku promenljivu , x i y posebno, a njihovim rangovima se dodeljuju simboli u i v. Razlika između svakog para rangova se obeležava sa $d = u - v$. Zatim se izračunaju kvadrati svake razlike d i saberu se da bi se dobilo $\sum d^2$. Na kraju se izračuna vrednost r_S, po formuli:

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

U sprovođenju testiranja hipoteze o Spirmenovom koeficijentu korelacije ranga ρ_S, koristi se statistika testa r_S, a njegova realizovana vrednost se računa pomoću gornje formule.

r_S može da uzme vrednost između -1 i 1. Ako je $r_S = 0 \Rightarrow$ u uzorku ne postoji monotona veza između X i Y. Ako je $r_S > 0$ u pitanju je direktna monotona veza između X i Y, što znači da sa rastom X raste Y, a ako je Ako je $r_S < 0$ u pitanju je inverzna monotona veza između X i Y, što znači da sa rastom X opada Y.



Kritična vrednost r_S se dobija iz tablice za datu veličinu uzorka i nivo značajnosti. Ako je test dvostrani, koristimo dve kritične vrednosti statistike testa, jednu negativnu i jednu pozitivnu. Međutim, ako je test levostran koristimo samo negativnu vrednost r_S, a ako je desnostrani koristimo samo pozitivnu vrednost r_S.

Nulta hipoteza uvek glasi $H_0: \rho_S = 0$. Realizovana vrednost statistike testa je uvek vrednost r_S koja se izračunava iz podatka iz uzorka. Neka α označava nivo značajnosti, a $-c$ i $+c$ oyna;avaju kriti;ne vrednosti r_S testa za Spirmenov koeficijent korelacije ranga.

1) Za dvostrani test, alternativna hipoteza glasi $H_1: \rho_S \neq 0$. Ako su $\pm c$ kritične vrednosti za datu veličinu uzorka n i α raspodeljeno na oba kraja, H_0 odbacujemo ako je ili $r_S \leq -c$ ili $r_S \geq +c$; to jest, H_0 odbacujemo ako je r_S “isuviše malo” ili “isuviše veliko”.

2) Za desnostrani test, alternativna hipoteza glasi $H_1: \rho_S > 0$. Ako je $+c$ kritična vrednosti za datu veličinu uzorka n i α na jednom kraju, H_0 odbacujemo ako je $r_S \geq +c$; to jest, H_0 odbacujemo ako je r_S “isuviše veliko”.

3) Za levostrani test, alternativna hipoteza glasi $H_1: \rho_S < 0$. Ako je $-c$ kritična vrednosti za datu veličinu uzorka n i α na jednom kraju, H_0 odbacujemo ako je $r_S \leq -c$; to jest, H_0 odbacujemo ako je r_S “isuviše malo”.

64. Korigovani koeficijent višestruke determinacije

isto kao pitanje 58

65. Model višestruke linearne regresije (pretpostavke i primena)

Zavisna promenjiva je determinisana kretanjem jedne ili više objašnjavajućih promenljivih. Kada u regresioni model uključimo dve ili više objašnjavajućih promenljivih, onda je reč o višestrukom regresionom modelu. Model uvek sadrži samo jednu zavisnu promenljivu. Višestruki regresioni model sa zavisnom promenljivom Y i objašnjavajućim promenljivim $X_1, X_2, X_3, \dots, X_k$, dat je u sledećem obliku:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

gde β_0 predstavlja odsečak ili konstantu, $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ su koeficijenti nagiba ili regresioni parametri uz objašnjavajuće promenljive $X_1, X_2, X_3, \dots, X_k$, a ε slučajna greška višestrukog regresionog modela. Model opisuje linearnu zavisnost između zavisne promenljive i k objašnjavajućih promenljivih. Model pretpostavlja da nema interakcije objašnjavajućih promenljivih. Kada X_i ne obuhvata interakciju faktora, već prikazuje uticaj samo jedne objašnjavajuće promenljive , takav model nazivamo višestrukim regresionim modelom prvog reda. Konstanta β_0 predstavlja vrednost Y za vrednosti svih objašnjavajućih promenljivih koje su jednake nuli. Koeficijenti $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ se nazivaju parcijalni regresioni koeficijenti. Na primer β_1 je parcijalni regresioni koeficijent uz promenljivu X_1. Ovaj koeficijent pokazuje prosečnu promenu promenljive Y nastalu usled jedinične promene X_1, pod uslovom da sve ostale objašnjavajuće promenljive ostanu nepromenjene. Slično, β_2 pokazuje prosečnu promenu Y kada se X_2 poveća za jednu svoju jedinicu, a sve ostale jedinice ostanu nepromenjene. Koeficijenti $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ još se nazivaju i stvarni regresioni koeficijenti ili parametri regresionog modela osnovnog skupa.

Pozitivna vrednost koeficijenta β_i ukazuje na pozitivnu zavisnost promenljive Y od promenljive X_i. Negativna vrednost β_i ukazuje na negativnu zavisnost Y od X_i.

Deo $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$ se naziva deterministički deo, a ε je stohastički deo.

U postupku zaključivanja o regresionim parametrima višestrukog regresionog modela koristi se t raspodela, a broj stepeni slobode iznosi: $df = n - k - 1$, gde n predstavlja veličinu uzorka, a k broj objašnjavajućih promenljivih u modelu.

Kada višestruki regresioni model sadrži samo dve objašnjavajuće promenljive $(k = 2)$, model se svodi na:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Regresioni model se ocenjuje na osnovu podataka iz uzorka, a uzoračka regresiona jednačina data je u sledećem obliku:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

Vrednosti $b_0, b_1, b_2, \dots, b_k$ su statistike uzorka koje predstavljaju odgovarajuće tačkaste ocene regresionih parametara osnovnog skupa $\beta_0, \beta_1, \beta_2, \dots, \beta_k$.

Za razliku od modela gde Y predstavlja stvarnu vrednost zavisne promenljive, u regresionom modelu uzorka, $\hat{Y}$ označava ocenjenu vrednost zavisne promenljive. Razlika između Y i $\hat{Y}$ predstavlja rezidual. U slučaju višestrukog regresionog modela suma kvadrata reziduala, SKR (suma kvadrata neobjašnjenog varijabiliteta, SKN) je:

$$SKR = SKN = \sum e^2 = \sum \left(Y - \hat{Y}\right)^2$$

Kao i kod prostog linearnog regresionog modela, ocenjivanje regresione jednačine se sprovodi minimiziranjem sume kvadrata reziduala, odnosno:

$$min \sum \left(Y - \hat{Y}\right)^2$$

a dobijene vrednosti nazivaju se ocenjene vrednosti regresionih parametara po metodu najmanjih kvadrata.

Osnovne pretpostavke višestrukog regresionog modela su:

1) Srednja vrednost slučajne greške ε jednaka je nuli, odnosno $E(\varepsilon) = 0$.

2) Slučajne greške za različite opservacije su međusobno nezavisne (nekorelisane). Pored toga, slučajne greške su normalno raspoređene i imaju konstantnu standardnu devijaciju, σ_ε .

3) Objašnjavajuće promenljive nisu međusobno linearno zavisne. To znači da ne postoji problem multikolinearnosti.

4) Između objašnjavajuće promenljive X_i i slučajne greške ε ne postoji korelaciona veza.

Osnovna primena regresionog modela je za ocenjivanje i predviđanje vrednosti zavisne promenljive.

<u>Upotreba regresionog modela za ocenjivanje prosečne vrednosti Y</u>

Regresioni model osnovnog skupa glasi:

$$Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \varepsilon$$

Očekivana vrednost Y za datu vrednost X_1 i X_2 označavamo sa $\mu_{Y|X}$ i važi:

$$\mu_{Y|X} = \beta_0 + \beta_1 X + \beta_2 X_2$$

Jedan od ciljeva regresione analize jeste upravo ocenjivanje $\mu_{Y|X}$ u osnovnom skupu. Vrednost $\hat{Y}$ dobijena zamenom vrednosti X_1 i X_2 u regresionoj jednačini uzorka, predstavlja tačkastu ocenu prosečne vrednosti promenljive Y u osnovnom skupu, $\mu_{Y|X}$.

Da bismo formirali interval poverenja za $\mu_{Y|X}$, potrebno je da znamo prosečnu vrednost, standardnu devijaciju i oblik uzoračke raspodele tačkaste ocene $\hat{Y}$. Tačkasta ocena $\hat{Y}$ ima normalnu raspodelu sa sredinom $\beta_0 + \beta_1 X + \beta_2 X_2$ i standardnom devijacijom $\sigma_{\hat{Y}_p}$

gde je $\sigma_{\hat{Y}_p}$ standardna greška ocene prosečne vrednosti Y. $\sigma_{\hat{Y}_p}$ zamenjujemo njenom ocenom, $S_{\hat{Y}_p}$. Zato formiranje intervala poverenja za $\mu_{Y|X}$ zasnivamo na t raspodeli verovatnoća.

$(1 - \alpha)100\%$ interval poverenja za $\mu_{Y|X}$ kada je $X_1 = X_{1p}, X_2 = X_{2p}$, glasi:

$$\hat{Y}_p \pm t S_{\hat{Y}_p}$$

gde je t tablična vrednost za površinu od $\alpha/2$ ispod krive na svakom kraju t raspodele i df=n-3.

Standardna greška ocene prosečne vrednosti zavisne promenljive, $S_{\hat{Y}_p}$ se računa pomoću nekog statističkoh softvera.

<u>Upotreba regresionog modela za predviđanje individualne vrednosti zavisne prom. Y:</u>

Druga važna upotreba regresionog modela je za potrebe predviđanja pojedinačne vrednosti promenljive Y za datu vrednost promenljivih X_1 i X_2. Predviđena vrednost izdataka pojedinačnog domaćinstva se označava sa Y_p. Kao tačkasta ocena za Y_p koristi se $\hat{Y}$. Interval poverenja za Y_p najčešće se naziva interval predviđanja.

Tačkasta ocena $\hat{Y}$ individualne vrednosti Y_p ima normalnu raspodelu sa sredinom $\beta_0 + \beta_1 X + \beta_2 X_2$ i standardnom devijacijom σ_{Y_p}, gde je σ_{Y_p} standardna greška predviđanja individualne vrednosti Y.

σ_{Y_p} zamenjujemo njenom ocenom, S_{Y_p}. Zato formiranje intervala poverenja za Y_p zasnivamo na t raspodeli verovatnoća.

$(1-\alpha)100\%$ interval poverenja za Y_p kada je $X_1 = X_{1p}, X_2 = X_{2p}$, glasi:

$$\hat{Y}_p \pm tS_{Y_p}$$

gde je t tablična vrednost za površinu od $\alpha/2$ ispod krive na svakom kraju t raspodele i df=n-3.
Standardna greška predviđanja individualne vrednosti zavisne promenljive, S_{Y_p}, se računa softverski.

66. Statistički testovi u regresionoj i korelacionoj analizi

1) Testiranje regresionog parametra β_1 kod proste linearne regresije

Prilikom testiranja hipoteze o regresionom parametru β_1, kod prostog linearnog regresionog modela, nultu hipotezu formulišemo u obliku $H_0: \beta_1 = 0$ (Parametar regresione prave osnovnog skupa jednak je nuli) i ona je ekvivalentna hipotezi da promenljiva X ne utiče na promenljivu Y i da se regresiona prava ne može koristiti u svrhe predviđanja vrednosti Y na osnovu promenljive X. Treba imati u vidu da testiramo samo linearnu zavisnost između dve promenljive.

Da bismo testirali da li X utiče na Y, formulišemo nultu hipotezu u obliku: $H_0: \beta_1 = 0$ (X ne utiče na Y).

Alternativna hipoteza može biti formulisana na jedan od sledeća tri načina:

1) X utiče na Y, odnosno $\beta_1 \neq 0$

2) X pozitivno utiče na Y, odnosno $\beta_1 > 0$

3) X negativno utilče na Y, odnosno $\beta_1 < 0$.

Statistika t testa za testiranje hipoteze o regresionom parametru β_1 glasi:

$$t = \frac{b_1 - \beta_1}{s_{b1}}$$

Vrednost regresionog parametra β_1 u prethodnom izrazu zamenjuje se hipotetičkom vrednošću, koja je definisana nultom hipotezom.

Broj stepeni slobode je: $= n - 2$.

2) Testiranje regresionog parametra β_i kod višestruke linearne regresije

Testiranje hipoteze za bilo koji regresioni parametar β_i u višestrukom regresionom modelu, možemo sprovesti iste procedure koja se koristi prilikom testiranja hipoteze o β_1 u prostom regresionom modelu. Jedina razlika je što je broj stepeni slobode u višestrukoj regresiji jednak $n - k - 1$. Zbog pretpostavke o normalnoj raspodeli slučajnih grešaka, uzoračka raspodela svake ocene b_i je takođe normalna sa srednjom vrednošću β_i i standardnom greškom, σ_{b_i}. Standardna greška σ_{b_i} nije poznata, pa zato koristimo njenu ocenu iz uzorka s_{b_i}, a testiranje hipoteze o parametru β_i zasnivamo na t raspodeli.

Za testiranje hipoteze o regresionom parametru β_i koristi se sledeća statistika testa:

$$t = \frac{b_i - \beta_i}{s_{bi}}$$

Vrednost regresionog parametra β_i u prethodnom izrazu zamenjuje se hipotetičkom vrednošću, koja je definisana nultom hipotezom.

Broj stepeni slobode je: $= n - k - 1$.

3) Testiranje linearne korelacije u skupu

Testiranje koeficijenta linearne korelacije u skupu radi se primenom t raspodele, gde se vrednost t testa računa po formuli:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Broj stepeni slobode je $df = n - 2$.

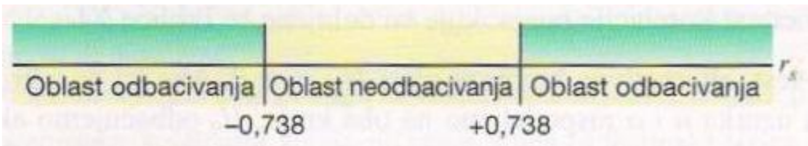
67.Spirmenov koeficijent korelacije ranga

Spirmenov koeficijent korelacije ranga predstavlja koeficijent proste linearne korelacije između rangova. U uzorku se obeležava sa r_S, a u skupu sa ρ_S. Da bi se izračunala vrednost r_S vrši se rangiranje podataka za svaku promenljivu , x i y posebno, a njihovim rangovima se dodeljuju simboli u i v. Razlika između svakog para rangova se obeležava sa $d = u - v$. Zatim se izračunaju kvadrati svake razlike d i saberu se da bi se dobilo $\sum d^2$. Na kraju se izračuna vrednost r_S, po formuli:

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

U sprovođenju testiranja hipoteze o Spirmenovom koeficijentu korelacije ranga ρ_S, koristi se statistika testa r_S, a njegova realizovana vrednost se računa pomoću gornje formule.

r_S može da uzme vrednost između -1 i 1. Ako je $r_S = 0 \Rightarrow$ u uzorku ne postoji monotona veza između X i Y. Ako je $r_S > 0$ u pitanju je direktna monotona veza između X i Y, što znači da sa rastom X raste Y, a ako je Ako je $r_S < 0$ u pitanju je inverzna monotona veza između X i Y, što znači da sa rastom X opada Y.



Kritična vrednost r_S se dobija iz tablice za datu veličinu uzorka i nivo značajnosti. Ako je test dvostrani, koristimo dve kritične vrednosti statistike testa, jednu negativnu i jednu pozitivnu. Međutim, ako je test levostran koristimo samo negativnu vrednost r_S, a ako je desnostrani koristimo samo pozitivnu vrednost r_S.

Nulta hipoteza uvek glasi $H_0: \rho_S = 0$. Realizovana vrednost statistike testa je uvek vrednost r_S koja se izračunava iz podatka iz uzorka. Neka α označava nivo značajnosti, a $-c$ i $+c$ oyna;avaju kriti;ne vrednosti r_S testa za Spirmenov koeficijent korelacije ranga.

1) Za dvostrani test, alternativna hipoteza glasi $H_1: \rho_S \neq 0$. Ako su $\pm c$ kritične vrednosti za datu veličinu uzorka n i α raspodeljeno na oba kraja, H_0 odbacujemo ako je ili $r_S \leq -c$ ili $r_S \geq +c$; to jest, H_0 odbacujemo ako je r_S "isuviše malo" ili "isuviše veliko".

2) Za desnostrani test, alternativna hipoteza glasi $H_1: \rho_S > 0$. Ako je $+c$ kritična vrednosti za datu veličinu uzorka n i α na jednom kraju, H_0 odbacujemo ako je $r_S \geq +c$; to jest, H_0 odbacujemo ako je r_S "isuviše veliko".

3) Za levostrani test, alternativna hipoteza glasi $H_1: \rho_S < 0$. Ako je $-c$ kritična vrednosti za datu veličinu uzorka n i α na jednom kraju, H_0 odbacujemo ako je $r_S \leq -c$; to jest, H_0 odbacujemo ako je r_S "isuviše malo".

68. Ocenjivanje i testiranje značajnosti koeficijenta proste korelacije

Isto kao 63

69. Parametarski i neparametarski pokazatelj korelacije (pretpostavke i primena)

isto kao 63

70. Ekstrapolacija i njena ograničenja u regresionoj analizi

Regresiona prava ocenjena na osnovu uzorka odnosi se na opseg vrednosti X koje su obuhvaćene uzorkom. Ako ocenjivanje ili predviđanje sprovodimo za vrednosti X van opsega uzorka, tada govorimo o ekstrapolaciji. U praksi se ekstrapolacija sprovodi za vrednosti bliže domenu uzoračkih podataka. Međutim, sa udaljavanjem vrednosti X za koju vršimo ocenjivanje i predviđanje, od domena uzoračkih podataka, zaključke treba donositi sa većim stepenom opreznosti.

Slično, ako regresioni model ocenjujemo na osnovu podataka vremenskih serija, predviđena vrednost zavisne promenljive za perio van uzoračkog perioda na bazi kojeg je ocenjen model, takođe se mora interpretirati sa određenim oprezom. Kada se regresioni model koristi za ekstrapolaciju, pretpostavljamo da ista linearna veza između dve promenljive važi i za vrednosti X van opsega uzoračkih podataka. Međutim, moguće je da veza između promenljivih za vrednosti izvan uzoračkih podataka ne bude linearna. Pored toga, čak i u slučaju da je veza linearna, dodavanje novih podataka zahteva ocenjivanje nove regresione prave.

71. Individualni i grupni indeksi

Varijacije vremenskih serija se mogu posmatrati kao apsolutne I relativne promene u zavisnosti od jedinica mera kojima se iskazuju. Apsolutne varijacije izračunavamo kao razlike između nivoa pojave u dva uzastopna ili neka druga vremenska trenutka ili intervala. Izražavamo ih u jedinicama mere u kojima je izražena i sama pojava. Mada pružaju korisne informacije o dinamici pojave, apsolutni pokazatelji nisu pogodni za uporednu analizu varijacija različitih pojava tokom vremena. Zato se u analizi varijacija vremenskih serija češće koriste relativni pokazatelji ili indeksi.

Indeksni brojevi su relativni brojevi koji predstavljaju odnos nivoa pojave u tekućem periodu i njenog nivoa u baznom periodu.

Tekući period je vremenski period u kome se posmatra nivo pojave, a period sa kojim se taj nivo poredi nazivamo baznim periodom. Količnik nivoa pojave u tekućem i baznom periodu se množi sa 100 i kao takav pokazuje procentualnu promenu nivoa pojave.

Postoji više podela indeksa s obzirom na različite kriterijume. U zavisnosti od toga da li pokazuju relativne promene jedne ili grupe srodnih pojava, indeksi mogu biti individualni ili grupni. S obzirom na to da li je prilikom njihovog izračunavanja bazzni period promenljiv ili ne, razlikujemo bazne indekse (indekse sa stalnom bazom) i lančane indekse (indekse sa promenljivom bazom). Prema ekonomskoj veličini na koju se odnose, indekse delimo na indekse cena, indekse fizičkog obima (količina) i indekse vrednosti.

Indeks iznad 100 ukazuje na rast nivoa pojave u tekućem periodu u odnosu na njen nivo u baznom periodu i to za onoliko procenata za koliko je indeks veći od 100. Obrnuto, indeks ispod 100 ukazuje da je reč o smanjenju nivoa pojave za onoliko procenata za koliko je indeks manji od 100. Ako je indeks jednak 100, zaključujemo da je nivo pojave u tekućem periodu nepromenjen u poređenju sa njenim nivoom u baznom periodu.

72.Indeksni brojevi i indeksni poeni

Varijacije vremenskih serija se mogu posmatrati kao apsolutne I relativne promene u zavisnosti od jedinica mera kojima se iskazuju. Apsolutne varijacije izračunavamo kao razlike između nivoa pojave u dva uzastopna ili neka druga vremenska trenutka ili intervala. Izražavamo ih u jedinicama mere u kojima je izražena i sama pojava. Mada pružaju korisne informacije o dinamici pojave, apsolutni pokazatelji nisu pogodni za uporednu analizu varijacija različitih pojava tokom vremena. Zato se u

analizi varijacija vremenskih serija češće koriste relativni pokazatelji ili indeksi.
Indeksni brojevi su relativni brojevi koji predstavljaju odnos nivoa pojave u tekućem periodu i njenog nivoa u baznom periodu.
Tekući period je vremenski period u kome se posmatra nivo pojave, a period sa kojim se taj nivo poredi nazivamo baznim periodom. Količnik nivoa pojave u tekućem i baznom periodu se množi sa 100 i kao takav pokazuje procentualnu promenu nivoa pojave.
Razlika između dva indeksna broja se izražavaju indeksnim poenima. Ovi poeni ne pokazuju relativnu, već apsolutnu razliku između nivoa pojave u dva perioda. Drugim rečima, na osnovu indeksnog poena možemo da izračunamo apsolutnu promenu nivoa pojave između dva perioda.

73. Indeksi sa stalnom i promenljivom bazom i veza između njih

Individualni indeksi pokazuju relativne promene samo jedne pojave u posmatranom (tekućem) u odnosu na izabrani bazni period. Individualni indeksi se dele na bazne i lančane.
Bazni indeksi (indeksi sa stalnom bazom) dobijaju se tako što podatke vremenske serije u svakom posmatranom period stavimo u odnos sa podatkom u periodu koji smo izabrali kao bazu koju tokom vremena ne menjamo.

$$I_t = \frac{Y_t}{Y_0} \cdot 100$$

Bazni indeks se kao i svi indeksi tumači u odnosu na 100. Ako je bazni indeks veći od 100 to pokazuje da je pojava u tekućem periodu toliko procenata koliko je indeks veći od 100, veća nego što je u baznom periodu. Ako je bazni indeks manji od 100 to pokazuje da je pojava u tekućem periodu toliko procenata koliko je indeks manji od 100, manja nego što je u baznom periodu. Ako je bazni indeks 100 to znači da je nivo pojave isti u tekućem i baznom periodu.
Individualne indekse možemo izračunati i kao odnos nivoa pojave u tekućem i prethodnom periodu. Takvi indeksi se zovu lančanim indeksima (indeksima sa promenljivom bazom):

$$L_t = \frac{Y_t}{Y_{t-1}} \cdot 100$$

Lančani indeksi izražavaju tempo promene nivoa pojave u sukcesivnim vremenski periodima: indeks iznad 100 ukazuje na rast, a ispod 100 na pad nivoa pojave u tekućem u odnosu na prethodni period.
a) Preračunavanje baznih indeksa sa jedne na drugu bazu (promena baza)
Često su nam u statističkim publikacijama dostupne samo serije indeksa sa stalnom bazom u jednoj godini, dok je za potrebe naše empirijske analize potrebno da relativne promene budu izražene u odnosu na neku drugu baznu godinu. Tada je potrebno izvršiti preračunavanje baznih indeksa sa jedne na drugu baznu godinu koja nas interesuje. Postupak se sastoji u sledećem: sve bazne indekse sa bazom u staroj godini stavljamo u odnos sa baznim indeksom u godini koja je izabrana kao nova baza. Na primer:

$$I_{2001(2002=100)} = \frac{I_{2001(2001=100)}}{I_{2002(2001=100)}} \cdot 100$$

b) Preračunavanje baznih na lančane indekse
Na osnovu baznih indeksa vremenske serije takođe možemo izračunati i lančane indekse. Postupak se sastoji u deljenju baznog indeksa u tekućem period sa baznim indeksom u prethodnom periodu i množenjem tako dobijenog količnika sa 100. Na taj način se potire stalna baza, a nova baza postaje prethodni period. Na primer:

$$L_{2002} = \frac{I_{2002(2001=100)}}{I_{2001(2001=100)}} \cdot 100$$

c) Preračunavanje lančanih u bazne

Polazimo os nivoa pojave u period koji smo izabrali kao bazni i za koji je bazni indeks jednak 100. Postupak preračunavanja se razlikuje u zavisnosti od toga da li je reč o periodima koji prethode ili o periodima koji slede izabrani bazni period.

Za periode koji prethode izabranoj bazi, bazni indeks I_t delimo lančanim indeksom u istom period L_t i količnik množimo sa 100:

$$I_{t-1} = \frac{I_t}{L_t} \cdot 100$$

gde se t-1 odnosi na periode koji prethode izabranom baznom period.

Na osnovu prethodnog izraza možemo doći i do formule za izračunavanje baznog indeksa za periode koji slede izabrani period:

$$I_t = \frac{L_t I_{t-1}}{100}$$

gde se oznaka t odnosi na periode koji slede izabrani bazni period.

74. Problem izbora baze i interpretacija indeksnih brojeva

Kod izračunavanja i primene indeksnih brojeva najznačajnije pitanje predstavlja izbor baze koja služi kao osnova za upoređivanje. Bazni period može biti bilo koji podatak u vremenskoj seriji. Za bazni period treba uzeti onaj član vremenske serije koji svojom veličinom reprezentuje nivo posmatrane pojave. Po pravilu to je onaj vremenski period ili moment u kome posmatrana pojava pokazuje bar približno prosečni nivo u svom razvoju. U poljoprivredi, na primer, za baznu godinu ne treba uzeti ni sušnu ni rodnu godinu. Isto tako, ne treba uzeti godinu ili period u kome su bile izizetne prilike (ratovi, revolucije, velike poplave itd)

Značaj pravila izbora baze je u tome što podatak koji se uzima za bazu opredeljuje veličinu izračunatih indeksa. Ako je veličina posmatrane pojave u baznom periodu suviše mala ili suviše velika, indeksi će pokazivati deformisanu sliku njenog razvoja – nerealno veliki porast ili opadanje.

Izbor stalne baze ne može biti konačan. Tako, na primer, izmene u strukturi skupova, koje se iskazuju grupnim indeksima, zahtevaju promenu baznog perioda, pogotovo kada se ponderacioni faktori određuju prema relativnom značaju sastavnih serija u baznom periodu. Relativni značaj pojedinih serija može tokom vremena bitno da se izmeni. Struktura posmatranog skupa će se po pravilu više menjati ukoliko je bazni period vremenski udaljeniji od posmatranog perioda.

Zbog toga se kao baza uzima period koji nie mnogo udaljen od posmatranog. Kad u strukturi nastanu veće i značajnije promene bazni period se mora menjati – približava se posmatranom.

Poželjno je da se u baznom periodu ne ispoljava naglašena fluktuacija cena, niti veliki disparitet cena. Ona ne traba da se nađe ni blizu najviše ili najniže tačke ciklusa cena.

Da bi se indeksi pravilno interpretirali i koristili, neophodno je voditi računa o njihovom osnovnom svojstvu – da pokazuju samo relativne promene, koje ne daju nikakvu informaciju o veličini same pojave. Jenakost indeksnih brojeva znači samo podjednak relativni porast (pad), a ne i isti nivo tih pojava. Na primer ako je indeks troškova života u Srbiji isti kao u Nemačkoj to ne znači da su troškovi života isti, već da je rast cena na malo isti u odnosu na bazni period.

75. Ponderisanje indeksnih brojeva

Korišćenjem proste aritmetičke sredine svim sastavnim serijama se daje podjednak značaj, iako ga one u stvari nemaju. U praksi je najčešće slučaj da sastavne serije agregata za koji računamo grupni indeks nemaju isti značaj. Neponderisani grupni indeks količine izražava relativne promene u količinama proizvedene robe, ali ne vodi računa o tome da ista količina proizvoda koji čine grupu, ne mora imati istu vrednost, pa samim tim ni isti značaj u posmatranoj grupi. Takođe, kada ne bismo ponderisali grupni indeks cena u trgovini na malo, to bi značilo da veća cena po jedinici proizvoda dobija veći značaj. Drugim rečima, skuplji proizvodi imali bi veći uticaj na opšti nivo cena iako upravo ti proizvodi, po pravilu, imaju manji obim prometa u trgovini. Prilikom konstrukcije grupnih indeksa se, dakle, mora voditi računa o tome koliko se puta cena po jedinici svakog proizvoda javlja, odnosno mora se pridati veći značaj (ponder) cenama onih proizvoda koje imaju veći uticaj na opšti nivo cena, a to su proizvodi sa većim učešćem u prometu.

Cilj ponderisanja grupnih indeksa je da istakne relativni značaj svake sastavne serije koju uključujemo u grupni indeks. Na taj način se povećava reprezentativnost grupnih indeksa kao indikatora relativnih varijacija više vremenskih serija. U praksi se najćešće primenjuju dva postupka ponderisanja grupnih indeksa u zavisnosti od toga koji je metod njihovog konstruisanja korišćen:

1) Ponderisanje indeksa dobijenih po metodu prosečnih odnosa

Ako prilikom konstrukcije grupnih indeksa koristimo metod prosečnih odnosa, onda se postupak ponderisanja sprovodi po pravilima izračunavanja ponderisanih srednjih vrednosti. Svaki individualni indeks se množi svojim ponderom, a suma ovih proizvoda se deli sumom pondera:

$$\bar{I}_{(t)} = \frac{\sum I_{i(t)}\alpha_i}{\sum \alpha_i}$$

gde su $I_{i(t)}$ individualni indeksi (količine ili cena). Kao ponderacioni faktor, α_i, najčešće se koristi vrednost iz baznog perioda $(\alpha = p_0 q_0)$, pa se dobija:

$$\bar{I}_{(t)} = \frac{\sum I_{i(t)} p_{i0} q_{i0}}{\sum p_{i0} q_{i0}}$$

2) Ponderisanje indeksa dobijenih po metodu agregata

U postupku ponderisanja grupnih indeksa cena dobijenih po metodu agregata kao pondere koristimo količine proizvoda, a u slučaju grupnih indeksa količina ponderi su cene. U zavisnosti od toga da li se ponderi određuju prema strukturi sastavnih serija u baznom ili tekućem periodu razlikujemo Laspejresov metod i Pašeov metod ponderisanja.

Laspejresov metod ponderisanja kao pondere koristi cene ili količine iz baznog perioda. To znači da u konstrukciji agregatnog indeksa cena, postupak ponderisanja po Laspejresovom metodu se sastoji u množenju cene svakog od n proizvoda u tekućem i u baznom periodu količinom istog proizvoda iz baznog perioda. Dakle količine se uzimaju nepromenjene, tako da će dobijeni rezultat pokazati samo relativnu promenu cena.

Slično, pri računanju agregatnog indeksa količine, ponderisanje sprovodimo tako što količinu svakog od n proizvoda u tekućem i baznom periodu množimo odgovarajućom cenom iz baznog perioda. Dakle cene se uzimaju nepromenjene, tako da će dobijeni rezultat pokazati samo relativnu promenu količina.

$$I_p = \frac{\sum p_t q_0}{\sum p_0 q_0} \cdot 100 \qquad I_q = \frac{\sum q_t p_0}{\sum q_0 p_0} \cdot 100$$

Po Pašeovom metodu ponderisanja, ponderi se određuju prema strukturi sastavnih serija u tekućem periodu. Kao ponderi se koriste cene ili količine iz tekućeg perioda. To znači da u konstrukciji agregatnog indeksa cena, postupak ponderisanja po Pašeovom metodu se sastoji u množenju cene svakog od n proizvoda u tekućem i u baznom periodu količinom istog proizvoda iz tekućeg perioda. Dakle količine se uzimaju nepromenjene, tako da će dobijeni rezultat pokazati samo relativnu promenu cena.

Slično, pri računanju agregatnog indeksa količine, ponderisanje sprovodimo tako što količinu svakog od n proizvoda u tekućem i baznom periodu množimo odgovarajućom cenom iz tekućeg perioda. Dakle cene se uzimaju nepromenjene, tako da će dobijeni rezultat pokazati samo relativnu promenu količina.

$$I_p = \frac{\sum p_t q_t}{\sum p_0 q_t} \cdot 100 \qquad I_q = \frac{\sum q_t p_t}{\sum q_0 p_t} \cdot 100$$

Kod indeksa vrednosti nema potrebe uzimati bilo kakav ponder:

$$I_{pq} = \frac{\sum p_t q_t}{\sum p_0 q_0} \cdot 100$$

76.Konstrukcija grupnih indeksa metodom prosečnih odnosa

Po metodu prosečnih odnosa grupni indeks se računa kao prosečna vrednost individualnih indeksa n sastavnih serija. Zato se ovi grupni indeksi zovu još i srednji grupni indeksi, a mogu biti aritmetički, geometrijski, harmonijski i dr., u zavisnosti od toga koju smo srednju vrednost individualnih indeksa koristili. Na primer, aritmetički grupni indeksi se računaju na sledeći način:

$$\bar{I}_p = \frac{\sum I_p}{n} \qquad \bar{I}_q = \frac{\sum I_q}{n} \qquad \bar{I}_{pq} = \frac{\sum I_{pq}}{n}$$

gde su $I_p, I_q \; i \; I_{pq}$ individualni indeksi cena, količina i vrednosti.

Ako prilikom konstrukcije grupnih indeksa koristimo metod prosečnih odnosa, onda se postupak ponderisanja sprovodi po pravilima izračunavanja ponderisanih srednjih vrednosti. Svaki individualni indeks se množi svojim ponderom, a suma ovih proizvoda se deli sumom pondera:

$$\bar{I}_{(t)} = \frac{\sum I_{i(t)} \alpha_i}{\sum \alpha_i}$$

gde su $I_{i(t)}$ individualni indeksi (količine ili cena). Kao ponderacioni faktor, α_i, najčešće se koristi vrednost iz baznog perioda $(\alpha = p_0 q_0)$, pa se dobija:

$$\bar{I}_{(t)} = \frac{\sum I_{i(t)} p_{i0} q_{i0}}{\sum p_{i0} q_{i0}}$$

77. Konstrukcija grupnih indeksa metodom agregata

Ideja metoda agregata je da se grupni indeks konstruiše kao odnos zbira podataka svih n sastavnih serija u tekućem periodu t i zbira podataka istih n serija u izabranom baznom periodu:

$$I_p = \frac{\sum p_t}{\sum p_0} \cdot 100 \qquad I_q = \frac{\sum q_t}{\sum q_0} \cdot 100 \qquad I_{pq} = \frac{\sum p_t q_t}{\sum p_0 q_0} \cdot 100$$

gde su p_0, q_0 i $p_0 q_0$ cena količina i vrednost u baznom periodu, a p_t, q_t i $p_t q_t$ cena količina i vrednost u tekućem periodu.

U postupku ponderisanja grupnih indeksa cena dobijenih po metodu agregata kao pondere koristimo količine proizvoda, a u slučaju grupnih indeksa količina ponderi su cene. U zavisnosti od toga da li se ponderi određuju prema strukturi sastavnih serija u baznom ili tekućem periodu razlikujemo Laspejresov metod i Pašeov metod ponderisanja.

Laspejresov metod ponderisanja kao pondere koristi cene ili količine iz baznog perioda. To znači da u konstrukciji agregatnog indeksa cena, postupak ponderisanja po Laspejresovom metodu se sastoji u množenju cene svakog od n proizvoda u tekućem i u baznom periodu količinom istog proizvoda iz baznog perioda. Dakle količine se uzimaju nepromenjene, tako da će dobijeni rezultat pokazati samo relativnu promenu cena.

Slično, pri računanju agregatnog indeksa količine, ponderisanje sprovodimo tako što količinu svakog od n proizvoda u tekućem i baznom periodu množimo odgovarajućom cenom iz baznog perioda. Dakle cene se uzimaju nepromenjene, tako da će dobijeni rezultat pokazati samo relativnu promenu količina.

$$I_p = \frac{\sum p_t q_0}{\sum p_0 q_0} \cdot 100 \qquad I_q = \frac{\sum q_t p_0}{\sum q_0 p_0} \cdot 100$$

Po Pašeovom metodu ponderisanja, ponderi se određuju prema strukturi sastavnih serija u tekućem periodu. Kao ponderi se koriste cene ili količine iz tekućeg perioda. To znači da u konstrukciji agregatnog indeksa cena, postupak ponderisanja po Pašeovom metodu se sastoji u množenju cene svakog od n proizvoda u tekućem i u baznom periodu količinom istog proizvoda iz tekućeg perioda. Dakle količine se uzimaju nepromenjene, tako da će dobijeni rezultat pokazati samo relativnu promenu cena.

Slično, pri računanju agregatnog indeksa količine, ponderisanje sprovodimo tako što količinu svakog od n proizvoda u tekućem i baznom periodu množimo odgovarajućom cenom iz tekućeg perioda. Dakle cene se uzimaju nepromenjene, tako da će dobijeni rezultat pokazati samo relativnu promenu količina.

$$I_p = \frac{\sum p_t q_t}{\sum p_0 q_t} \cdot 100 \qquad I_q = \frac{\sum q_t p_t}{\sum q_0 p_t} \cdot 100$$

Kod indeksa vrednosti nema potrebe uzimati bilo kakav ponder:

$$I_{pq} = \frac{\sum p_t q_t}{\sum p_0 q_0} \cdot 100$$

78. Prosečna stopa rasta i bazni indeksi – interpretacija

Individualni indeksi pokazuju relativne promene samo jedne pojave u posmatranom (tekućem) u odnosu na izabrani bazni period. Individualni indeksi se dele na bazne i lančane.

Bazni indeksi (indeksi sa stalnom bazom) dobijaju se tako što podatke vremenske serije u svakom posmatranom period stavimo u odnos sa podatkom u periodu koji smo izabrali kao bazu koju tokom vremena ne menjamo.

$$I_t = \frac{Y_t}{Y_0} \cdot 100$$

Bazni indeks se kao i svi indeksi tumači u odnosu na 100. Ako je bazni indeks veći od 100 to pokazuje da je pojava u tekućem periodu toliko procenata koliko je indeks veći od 100, veća nego što je u baznom periodu. Ako je bazni indeks manji od 100 to pokazuje da je pojava u tekućem periodu toliko procenata koliko je indeks manji od 100, manja nego što je u baznom periodu. Ako je bazni indeks 100 to znači da je nivo pojave isti u tekućem i baznom periodu.

Individualne indekse možemo izračunati i kao odnos nivoa pojave u tekućem i prethodnom periodu. Takvi indeksi se zovu lančanim indeksima (indeksima sa promenljivom bazom):

$$L_t = \frac{Y_t}{Y_{t-1}} \cdot 100$$

Lančani indeksi izražavaju tempo promene nivoa pojave u sukcesivnim vremenski periodima: indeks iznad 100 ukazuje na rast, a ispod 100 na pad nivoa pojave u tekućem u odnosu na prethodni period.
Geometrijska sredina lančanih indeksa predstavlja njihov prosek i pokazuje srednji tempo rasta:

$$G = \sqrt[T-1]{L_2 \cdot L_3 \cdot \cdots \cdot L_T}$$

Da bismo utvrdili za koliko u procentima pojava u proseku raste ili opada, na osnovu geometrijske sredine G izračunaćemo prosečnu (geometrijsku) stopu rasta, r_g:

$$r_g = G - 100$$

Prosečna stopa rasta je stopa po kojoj se nivo pojave prosečno (godišnje, kvartalno, mesečno,…) u relativnom iznosu povećava ili smanjuje u periodu obuhvaćenom vremenskom serijom.
Na primer, ako je prosečna stopa rasta 23%, to znači da je posmatrana pojava u posmatranom periodu u proseku godišnje(mesečno, kvartalno,…) rasla po stopi od 23%, a ako je prosečna stopa rasta -17,2%, to znači da je posmatrana pojava u posmatranom periodu u proseku godišnje(mesečno, kvartalno,…) opadala po stopi od 17,2%.
Ukoliko raspolažemo originalnim podacima vremenske serije, prosečnu stopu rasta možemo jednostavnije izračunati direktno iz tih podataka koristeći sledeći izraz:

$$r_g = \left(\sqrt[T-1]{\frac{Y_T}{Y_1}} - 1 \right) \cdot 100$$

79. Pristupi u analizi vremenskih serija

Metode analize vremenskih serija možemo podeliti na kvalitativne i kvantitativne.
Kvalitativni metodi se koriste kada podaci o nekoj pojavi nisu dostupni ili se ne mogu kvantifikovati. Zasnivaju se na procesu usklađivanja mišljenja eksperata. Cilj ovih metoda nije da se dođe do jedinstvenog mišljenja, već da se nađe što uži interval u kome će se naći mišljenje većine članova ekspertskog tima u vezi sa kretanjem posmatrane pojave.
Nasuprot kvalitativnim metodima, preduslovi primene kvantitativnih metoda su da se informacije o pojavi koju analiziramo mogu kvantifikovati, da su podaci o prošlom i sadašnjem periodu dostupni i da odslikavaju pravu prirodu posmatrane pojave. Promina metoda se zasniva na generalnoj pretpostavci da će se pojava u budućnosti ponašati na približno isti način kao i u prošlom periodu.
Sve kvantitativne metode možemo svrstati u dve osnovne grupe:
- metode statističke analize vremenskih serija
- kauzalne (uzročne) metode.

Metodi statističke analize vremenskih serija (statistički pristup) orijentisani su na analizu osnovnih karakteristika pojedinačne vremenske serije i na prognoziranje njenih budućih vrednosti isključivo na osnovu sopstvenih vrednosti iz prošlog i sadašnjeg perioda. U ovu grupu metoda spadaju metod dekompozicije, različiti metodi izravnanja i Boks-Dženkinsonova metodologija.
Kauzalni metodi spadaju u domen regresione analize vremenskih serija. Koristimo ih ako varijacije posmatrane serije objašnjavamo varijacijama u drugim vremenskim serijama. Primenom ovih metoda ocenjujemo regresioni model jedne vremenske serije (zavisne promenljive) u funkciji drugih vremenskih serija, koje predstavljaju objašnjavajuće promenljive.

80. Vremenske serije – pojam i podela

Vremenska serija je hronološki uređeni niz podataka koji prikazuje varijacije pojave tokom sukcesivnih, jednakih vremenskih intervala, odnosno predstavlja jednu realizaciju uređenog niza slučajnih promenljivih $Y_1, Y_2, Y_3, \dots$ s obzirom na vreme.

Podaci vremenske serije se mogu posmatrati u vremenskim momentima ili u vremenskim intervalima. U zavisnosti od toga, vremenske serije možemo klasifikovati na momentne i intervalne serije. Ova podela vremenskih serija analogna je podeli ekonomskih veličina na promenljive stanja i promenljive toka.

Momentna serija sadrži podatke o nivou pojave koja je posmatrana u određenim sukcesivnim vremenskim trenucima. Na primer, broj zaposlenih se menja tokom meseca, ali se stanje zaposlenosti registruje samo na kraju meseca. Primeri momentnih serija su: zalihe, štednja, stanje na računu, broj lica koje traže posao (stanje krajem perioda),…

Podaci intervalne serije odnose se na određeni vremenski interval (dan, nedelju, mesec, kvartal, godinu). Primeri intervalnih vremenskih serija su vrednost prometa u trgovini na malo, lična potrošnja, vrednost izvoza i uvoza, fizički obim industrijske proizvodnje i slično.

Podatke momentne serije nema smisla sabirati, zato šro prikazuju stanje pojave na kraju vremenskog perioda. U slučaju intervalne vremenske serije ima smisla da kumuliramo podatke. Dakle, na primer, godišnji izvoz možemo dobiti sabiranjem mesečnih izvoza.

S obzirom na domen analize, vremenske serije možemo analizirati u vremenskom domenu i u frekventnom domenu. Analiza u vremenskom domenu podrazumeva analizu vremenskih serija u funkciji vremena. U domenu frekvencija, vremenska serija se posmatra kao kompozicija sinusoida frekvencija, od kojih svaka nosi određene informacije. Cilj ove analize jeste da sumira te informacije u funkciji od frekvencija, odnosno da ustanovi koliki je pojedinačni doprinos komponenata na različitim frekvencijama ukupnom varijabilitetu vremenske serije. U tu svrhu se koristi spektralna funkcija gustine, te se otuda analiza u domenu frekvencija naziva još i spektralnom analizom. Krajnji cilj analize jeste prognoza buduće vrednosti posmatrane vremenske serije.

81. Dekompozicija vremenskih serija

Klasičan metod dekompozicije polazi od pretpostavke da na razvojnu tendenciju vremenske serije izvesni faktori utiču postojano u određenom pravcu, dok ostali (nepostojani) faktori uzrokuju odstupanja od te osnovne putanje serije. Svi faktori koji utiču na varijacije posmatrane serije tokom vremena se mogu svrstati u četiri grupe, odnosno varijacije u vremenskoj seriji mogu se dekomponovati (razložiti) na četiri sastavne komponente: trend, ciklične varijacije, sezonske varijacije i rezidualnu komponentu.

Trend (T) predstavlja razvojnu rastuću ili opadajuću tendenciju neke pojave u određenom vremenskom periodu. Veliki broj ekonomskih vremenskih serija pokazuje rastuću tendenciju tokom vremena, uzrokovanu tehničkim progresom i posledično ekonomskim rastom.

Ciklične varijacije (C) predstavljaju naizmenično smenjivanje višegodišnjih odstupanja posmatrane pojave iznad i višegodišnjih odstupanja ispod njenog prosečnog kretanja. Svaki ciklus ima svoj prosperitet (vrh), opadanje (recesiju), depresiju (najnižu tačku) i oporavak (ekspanziju). Ciklusi se razlikuju među sobom kako po dužini trajanja, tako i po intenzitetu oscilacija oko prosečnog nivoa pojave. Ove oscilacije nastaju usled delovanja više kombinacija faktora. Primer privrednih ciklusa su fluktuacije društvenog proizvoda kao pokazatelja privredne aktivnosti zemlje.

Sezonske varijacije (S) su pravilnosti u kretanju serije koje se ponavljaju u vremenskim razmacima kraćim od godinu dana. Sezonske varijacije se javljaju pod uticajem vremenskih uslova, kalendarskih faktora (praznici, raspust) i slično.
Rezidualna komponenta (R) sadrži sve ostale varijacije koje postoje u vremenskoj seriji. Ovom komponentom obuhvaćena su slučajna kolebanja, kao i varijacije koje nastaju zbog nepredvidivih događaja (vremenske nepogode, ratovi, štrajkovi).
Svaka vremenska serija ne mora imati sve četiri komponente, što zavisi od prirode analizirane pojave i od učestalosti intervala u kojima se ona posmatra. Na primer, godišnje vremenske serije nemaju sezonsku komponentu, jer uticaj sezonskih faktora možemo identifikovati samo unutar godine. Takođe, ciklične varijacije možemo identifikovati samo na podacima u dužem vremenskom periodu (na dužem nizu godina).
U pogledu načina delovanja faktora na kretanje vremenske serije, razlikujemo tri osnovne grupe modela: aditivni, multiplikativni i kombinovani modeli.
Kod aditivnog modela svaka opservacija vremenske serije jednaka je zbiru komponenti:

$$Y = T + C + S + R$$

Kod multiplikativnog modela opservacija je jednaka proizvodu komponenti:

$$Y = T \cdot C \cdot S \cdot R$$

Kod kombinovanog modela, varijacije vremenske serije su rezultat različitih kombinacija delovanja komponenata:

$$Y = T \cdot C + S \cdot R \qquad ili \qquad Y = T \cdot C \cdot S + R \;\; \ldots$$

Kod aditivnog modela vremenske serije sve komponente su izražene u istim jedinicama mere kao i originalna serija.
Kod multiplikativnog modela samo je trend izražen u istim jedinicama mere kao i originalna serija, dok je uticaj ostalih komponenti dat relativno, u procentima od trenda ili neke druge komponente.
Cilj metode dekompozicije vremenske serije je da se identifikuje uticaj svake komponente zasebno. Na osnovu originalne vremenske serije i poznatih vrednosti nekih od komponenata utvrđujemo uticaj ostalih faktora sadržanih u nepoznatim komponentama. Ako, na primer, želimo da identifikujemo sezonska i slučajna kolebanja, to postižemo na sledeći način:

$$S \cdot R = \frac{Y}{T \cdot C} = \frac{T \cdot C \cdot S \cdot R}{T \cdot C}$$

Dakle, isključenje jedne grupe komponenata u isto vreme znači izdvajanje preostalih komponenata.
Identifikacija komponenata je značajna naročito kod vremenskih serija sa izraženim periodičnim varijacijama koje mogu da zamagle pravu sliku o osnovnom toku posmatrane pojave.

82. Pokretni proseci – značenje i upotreba

Pokretni proseci su transformacija vremenske serije u kojoj se svaki originalni podatak zamenjuje aritmetičkom sredinom tog podatka, nekoliko prethodnih i istog broja sledećih podataka.
U slučaju neparnog broja podataka u grupi, na primer tri, prvi pokretni prosek će biti:

$$\bar{y}_2 = \frac{y_1 + y_2 + y_3}{3}$$

Ovaj prosek zamenjuje drugi podatak y_2 u originalnoj vremenskoj seriji. Drugi pokretni prosek dobijamo na sledeći način:

$$\bar{y}_3 = \frac{y_2 + y_3 + y_4}{3}$$

i on zamenjuje treći originalni podatak y_3. Sukcesivnim pomeranjem ka kraju serije, dolazimo do T-1 pokretnog proseka:

$$\bar{y}_{T-1} = \frac{y_{T-2} + y_{T-1} + y_T}{3}$$

koji zamenjuje pretposlednji (T-1) originalni podatak u seriji. Na ovaj način, polazeći od originalne vremenske serije $y_1, y_2, \dots, y_T$ dobijamo transformisanu seriju $\bar{y}_2, \bar{y}_3, \dots, \bar{y}_{T-1}$.
Vidimo da u ovom slučaju ukupno dva podatka (prvi i poslednji) u originalnoj vremenskoj seriji ostaje bez svog pokretnog proseka. Slično, kod petočlanih pokretnih proseka ukupno četiri podatka (po dva na krajevima) ostaje bez pokretnih proseka, ili u opštem slučaju, ako su grupe sa neparnim brojem podataka veličine n, ukupno n-1 podataka ostaće bez svog pokretnog proseka.
Ako su grupe sa parnim brojem podataka postupak formiranja pokretnih proseka je sledeći (primer sa 4 člana):

$$y_{2\_3} = \frac{y_1 + y_2 + y_3 + y_4}{4}$$

i nalazi se između drugog i trećeg člana grupe. Drugi pokretni prosek je:

$$y_{3\_4} = \frac{y_2 + y_3 + y_4 + y_5}{4}$$

i tako redom sve do poslednjeg četvoročlanog proseka:

$$y_{T-2\_T-1} = \frac{y_{T-3} + y_{T-2} + y_{T-1} + y_T}{4}$$

Za razliku od pokretnih proseka sa neparnim brojem članova, gde je jasno koji originalni podatak zamenjujemo pokretnim prosekom, ovde se javlja problem da se izračunati proseci nalaze između dva srednja podatka u grupi. Zbog toga se radi dodatni korak, centriranje, računanjem aritmetičke sredine svaka dva uzastopna proseka. Na primer, aritmetička sredina proseka y_{2_3} **i** y_{3_4} zamenjuje treći originalni podatak y_3, zatim sredina proseka y_{3_4} **i** y_{4_5} menja podatak y_4 i tako dalje. Na kraju dobijamo seriju sastavljenu od četvoročlanih centriranih pokretnih proseka:

$$\bar{y}_3, \bar{y}_4, \dots, \bar{y}_{T-2}$$

Dakle, gube se ukupno 4 podatka, dva na svakom kraju, ili u opštem slučaju n podataka ako su pokretni proseci sa parnim brojem podataka veličine n.
Pokretni proseci se koriste kod izbora funkcije trenda. Naime, često originalna serija zbog izraženih sezonskih uticaja, ne ukazuje na grafičkom prikazu na oblik funkcije trenda koji joj odgovara. Međutim, ako se grafički prikažu centrirani pokretni proseci, eliminisaće se uticaj sezonskih faktora i lakše će moći da se uoči koja funkcija trenda je najbolja za konkretnu vremensku seriju.
Druga primena pokretnih proseka je za prognozu pojave u budućnosti. Ukoliko, na primer, imamo seriju od 1990. do 2006. godine i želimo prognozu pojave za 2007. godinu, postupak je sledeći:

$$\hat{y}_{2007} = \frac{y_{2004} + y_{2005} + y_{2006}}{3}$$

ako prognozu vršimo pomoću trogodišnjih pokretnih proseka.

83. Eksponencijalna i geometrijska stopa rasta

1) Geometrijska stopa rasta

Geometrijska sredina lančanih indeksa predstavlja njihov prosek i pokazuje srednji tempo rasta:

$$G = \sqrt[T-1]{L_2 \cdot L_3 \cdot \dots \cdot L_T}$$

Da bismo utvrdili za koliko u procentima pojava u proseku raste ili opada, na osnovu geometrijske sredine G izračunaćemo prosečnu (geometrijsku) stopu rasta, r_g:

$$r_g = G - 100$$

Prosečna stopa rasta je stopa po kojoj se nivo pojave prosečno (godišnje, kvartalno, mesečno,…) u relativnom iznosu povećava ili smanjuje u periodu obuhvaćenom vremenskom serijom.
Na primer, ako je prosečna stopa rasta 23%, to znači da je posmatrana pojava u posmatranom periodu u proseku godišnje(mesečno, kvartalno,…) rasla po stopi od 23%, a ako je prosečna stopa rasta -17,2%, to znači da je posmatrana pojava u posmatranom periodu u proseku godišnje(mesečno, kvartalno,…) opadala po stopi od 17,2%.
Ukoliko raspolažemo originalnim podacima vremenske serije, prosečnu stopu rasta možemo jednostavnije izračunati direktno iz tih podataka koristeći sledeći izraz:

$$r_g = \left(\sqrt[T-1]{\frac{Y_T}{Y_1}} - 1\right) \cdot 100$$

2) Eksponencijalna stopa rasta
Na osnovu ocene b_1 eksponencijalnog trenda možemo izračunati prosečnu stopu rasta, r_e.
Eksponencijalna prosečna stopa rasta pokazuje za koliko procenata pojava u proseku raste (opada) u posmatranom periodu (npr. mesečno, kvartalno, godišnje).

$$r_e = (b_1 - 1) \cdot 100$$

Na primer, ako smo ocenili eksponencijalni trend godišnje vremenske serije za period 1996-2005 i dobili ocenjenu vrednost $b_1 = 1{,}05$, to bi značilo da eksponencijalna stopa rasta iznosi 5%. Dakle, pojava je u posmatranom periodu od 1996. do 2005. u proseku godišnje rasla po stopi od 5%. Da smo dobili ocenjenu vrednost $b_1 = 0{,}90$ to bi značilo da je eksponencijalna stopa rasta -10%, što znači da je pojava u posmatranom periodu od 1996. do 2005. u proseku godišnje opadala po stopi od 10%.

84. Izbor funkcije trenda

Postoji više načina za izbor funkcije trenda.
1) Trend se može uočiti pomoću grafičkog prikazivanja vremenske serije. Serija prikazana na aritmetičkom dijagramu svojim izgledom ukazuje na korišćenje linearnog, eksponencijalnog ili paraboličnog trenda. Ukoliko serija prikazana na polulogaritamskom dijagramu ima približno oblik prave linije, treba koristiti eksponencijalni trend.
2) Metod diferencija (razlika) - Najpre se formira nova kolona u tabeli koja se označava sa 1 i u koju se upisuju apsolutne razlike susednih podataka iz kolone y. Ukoliko su te razlike približno jednake, onda je najbolji linearni trend, a ako nisu jednake formira se nova kolona u tabeli 2 u koju se upisuju apsolutne razlike susednih podataka iz kolone 1. Ako su te razlike približno jednake onda je u pitanju parabolični trend, a ako nisu formira se nova kolona logy u koju se upisuju apsolutne razlike svih susednih podataka it kolone logy, pa ako su te razlike približno jednake onda je to eksponencijalni trend. Obratiti pažnju da su razlike uvek pozitivne. Znači nije bitno koji je podatak veći
3) Metod pokretnih sredina (proseka) - Najpre se formira kolona u koju se upisuju pokretni proseci. Pokretni proseci mogu da budu sa dva podatka, tri četiri , itd. Ukoliko se pokretni proseci prave sa neparnim brojem podataka, npr. trogodišnji pokretni proseci, vrši se sabiranje svake tri susedne godine i deljenje sa tri. Rezultat koji se dobije se piše naspram srednje od te tri godine, znači naspram druge. Ukoliko se pokretni proseci prave sa parnim brojem podataka, npr. četvorogodišnji pokretni proseci,

vrši se sabiranje svake četiri susedne godine i deljenje sa 4. Rezultat koji se dobije se piše na mestu između drugog i trećeg podatka. Međutim, kako sad pokretni proseci nisu centrirani, vrši se njihovo centriranje, tako što se svaka dva susedna podatka saberu, rezultat podeli sa dva i napiše između ta dva podatka. Kada su formirani pokretni proseci, na te pokretne proseke se primenjuje grafički metod.

4) Srednja kvadratna greška - Najpre se pronađu sve tri funkcije trenda, pa se za svaku od tih funkcija izračuna srednja kvadratna greška po sledećoj formuli:

$$SKG = \frac{\sum(y - \hat{y}_t)^2}{T}$$

Bira se ona funkcija trenda koja ima najmanju srednju kvadratnu grešku.

85. Linearni i eksponencijalni trend

1) Linearni trend – Ako se vremenska serija iz godine u godinu povećava (ili smanjuje) u približno istom iznosu, odnosno ako pokazuje pravolinijsku tendenciju, njeno kretanje možemo opisati pomoću linearnog trenda. Model linearnog trenda ima sledeći oblik:

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

Modelom se opisuje linearno kretanje serije Y_t u funkciji vremena t. Parametri β_0 i β_1 su nepoznati i ocenjuju se na osnovu uzorka. U uzorku ocenjena linija trenda glasi:

$$\hat{Y}_t = b_0 + b_1 t$$

gde je $\hat{Y}_t$ ocena prosečne vrednosti vremenske serije, a b_0 i b_1 ocene parametara β_0 i β_1.

Ocene b_0 i b_1 dobijamo po metodu najmanjih kvadrata, koji podrazumeva minimiziranje kvadrata vertikalnih odstupanja podataka od funkcije trenda. Ocene b_0 i b_1 se dobijaju pomoću sledećih formula:

$$b_1 = \frac{\sum ty - \frac{\sum t \sum y}{T}}{\sum t^2 - \frac{(\sum t)^2}{T}}, \qquad b_0 = \bar{y} - b_1 \bar{t}$$

Vreme t se izražava tako što se prvom podatku dodeli 1, drugom 2, itd.

Ocena nagiba b_1 u modelu linearnog trenda pokazuje srednji apsolutni porast, odnosno prosečnu promenu vremenske serije u sukcesivnim vremenski intervalima posmatranog perioda.

2) Eksponencijalni trend – Kad serija tokom vremena pokazuje približno isti relativni rast ili pad, tada razvojni tok serije možemo aproksimirati eksponencijalnim trendom. Teorijski model eksponencijalnog trenda glasi:

$$Y_t = \beta_0 \cdot {\beta_1}^t \cdot \varepsilon_t$$

Parametri β_0 i β_1 su nepoznati i ocenjuju se na osnovu uzorka. U uzorku ocenjena funkcija trenda glasi:

$$\hat{Y}_t = b_0 \cdot {b_1}^t$$

gde je $\hat{Y}_t$ ocena prosečne vrednosti vremenske serije, a b_0 i b_1 ocene parametara β_0 i β_1.

Ocene b_0 i b_1 dobijamo po metodu najmanjih kvadrata, koji podrazumeva minimiziranje kvadrata vertikalnih odstupanja podataka od funkcije trenda. Ocene b_0 i b_1 se dobijaju pomoću sledećih formula:

$${b_1}^* = \frac{\sum ty^* - \frac{\sum t \sum y^*}{T}}{\sum t^2 - \frac{(\sum t)^2}{T}}, \qquad {b_0}^* = \overline{y^*} - {b_1}^* \bar{t}$$

gde je ${b_1}^* = log b_1,\ y^* = log y$ itd.

Ocena nagiba b_1 u modelu eksponencijalnog trenda pomnožena sa 100 se zove srednji relativni porast

ili srednji tempo rasta.
Na osnovu ocene b_1 eksponencijalnog trenda možemo izračunati prosečnu stopu rasta, r_e.
Eksponencijalna prosečna stopa rasta pokazuje za koliko procenata pojava u proseku raste (opada) u posmatranom periodu (npr. mesečno, kvartalno, godišnje).

$$r_e = (b_1 - 1) \cdot 100$$

Na primer, ako smo ocenili eksponencijalni trend godišnje vremenske serije za period 1996-2005 i dobili ocenjenu vrednost $b_1 = 1{,}05$, to bi značilo da eksponencijalna stopa rasta iznosi 5%. Dakle, pojava je u posmatranom periodu od 1996. do 2005. u proseku godišnje rasla po stopi od 5%. Da smo dobili ocenjenu vrednost $b_1 = 0{,}90$ to bi značilo da je eksponencijalna stopa rasta -10%, što znači da je pojava u posmatranom periodu od 1996. do 2005. u proseku godišnje opadala po stopi od 10%.

86. Sezonske varijacije – merenje i interpretacija

Sezonske varijacije su periodične fluktuacije vremenske serije u vremenskim intervalima unutar godine, a ponavljaju se tokom više godina u isto doba i u približno istom intenzitetu i smeru. Uticaj sezonskih faktora ne možemo identifikovati kod godišnjih vremenskih serija, jer se sezonske varijacije ispoljavaju samo unutar godine. Poznavanje prirode i intenziteta sezonskih varijacija je veoma važno, jer od toga zavisi i kvalitet prognoze budućih vrednosti posmatrane pojave. Kod nekih vremenskih serija sezonske fluktuacije se izražavaju jednoobrazno tokom vremena. Tada je dejstvo sezonskih faktora stabilno i ispoljava se na približno isti način iz godine u godinu. Međutim, ako sezonska varijacija menja intenzitet i periodičnost ponavljanja u svakom narednom periodu, onda govorimo o nestabilnom sezonskom ritmu.
Postoje dva pristupa merenju sezonskih varijacija: modelski i empirijski pristup. Kod modelskog pristupa sezonska priroda se meri uključivanjem sezonskih objašnjavajućih promenljivih u regresioni model vremenske serije (ako je stabilan sezonski ritam) , ili posebnom vrstom modela, tzv. Sezonskim ARIMA modelima (ako je sezonski ritam nestabilan).
Modelski pristup je dosta složen, tako da pri analizi koristimo isključivo empirijski pristup. Postoji više metoda u okviru empirijskog pristupa, a jedan od njih je metod odnosa prema pokretnim prosecima. Budući da polazimo od multiplikativnog modela, uticaj sezonskih varijacija ćemo pratiti na osnovu relativnih odstupanja od prosečnog nivoa, tj. Pomoću sezonskih indeksa. Sezonski indeksi su relativni brojevi, koji mere jačinu uticaja sezone u određenom kvartalu (mesecu) tokom više godina.
Kod sezonskih indeksa, prosečan kvartalni (mesečni) nivo pojave izražava se indeksom 100. Sezonski indeks iznad 100 pokazuje viši nivo pojave od kvartalnog (mesečnog) proseka, zbog uticaja sezone. Ako sezonski faktori ne bi značajno uticali na pojavu, tada bi svi sezonski indeksi iznosili približno 100, odnosno pojava ne bi značajnije odstupala od opšteg kvartalnog (mesečnog) proseka. Zbir sezonskih indeksa kod kvartalne serije je jednak 400, a kod mesečne 1200.
Merenje sezonskih varijacija po metodu odnosa prema pokretnim prosecima zasniva se na njuihovoj izolaciji (izdvajanju) od ostalih nesezonskih komponenata (T,C i R). Osnovna ideja ovog metoda je da se:
1) pokretnim prosecima združeno obuhvate trend i ciklična komponenta
2) deljenjem originalne serije pokretnim prosecima eliminiše uticaj trend i ciklične komponente i u seriji ostane sezonska i rezidualna komponenta

$$S \cdot R = \frac{Y}{T \cdot C}$$

Da bi se eliminisao uticaj i rezidualne komponente računa se kvartralni (mesečni) prosek komponenata

S i R, odnosno:

$$Kvartalni\ prosek = \frac{\sum S \cdot R}{n}$$

gde je n broj istih kvartala tokom posmatranih godina.

U poslednjem koraku, dobijeni kvartalni proseci se pomnože sa 100 i dobijaju se sezonski indeksi.

Najveći uticaj sezone je kvartal čiji sezonski indeks najviše odstupa od 100. Najveći pozitivan uticaj sezone je kvartal sa najvećim sezonskim indeksom, a najveći negativanj uticaj sezone je kvartal sa najmanjim sezonskim indeksom.

87. Desezoniranje – postupak i značaj

Analiza sezonske komponente u vremenskoj seriji je veoma značajna sa stanovišta kvaliteta prognoze. Kada bi se zanemarila sezonska komponenta u analizi vremenske serije, prognoze bi česti bile nerealne. Osim toga, prisustvo sezonske komponente zamagljuje osnovnu razvojnu tendenciju pojave, pa samo njenim eliminisanjem možemo sagledati osnovni tok, odnosno nesezonske karakteristike serije. Zato se u praksi najčešće pored originalne serije prikazuje i serija iz koje je eliminisana sezonska komponenta. Ovaj postupak eliminisanja sezonske komponente iz vremenske serije zove se sezonsko izravnanje ili desezoniranje.

Kod sezonskih vremenskih serija ima smisla porediti nivo serije u jednom kvartalu (mesecu) u odnosu na prethodni kvartal (mesec), samo ako je serija desezonirana, odnosno ako je iz nje prethodno isključena sezonska komponenta.

Kada koristimo originalnu, a ne desezoniranu vremensku seriju, tada možemo jedino porediti njen nivo u određenom kvartalu (mesecu) tekuće godine u odnosu na isti kvartal (mesec) prethodnih godina.

Desezoniranje se sprovodi na sledeći način:

$$Y_{St} = \frac{Y_t}{I_s} \cdot 100$$

88. Metodi prognoziranja vremenskih serija

Jedan od najvažnijih ciljeva analize vremenskih serija jeste prognoziranje budućeg toka vremenske serije sa što manjom greškom prognoze. Postoji više pristupa u prognoziranju vremenskih serija. Neki od njih su:

1) Prognoziranje vremenske serije po metodu dekompozicije – ekstrapolacija trenda

Ekstrapolacija trenda podrazumeva produžavanje ocenjene funkcije trenda izvan uzoračkog perioda. Osnovna pretpostavka prilikom prognoziranja vremenske serije korišćenjem metoda dekompozicije jeste da će faktori koji su delovali na nivo serije u prošlosti i sadašnjosti delovati i u budućem periodu na isti način, približno istim intenzitetom, u istom smeru i bez značajnijeg uticaja novih faktora. Radi se, dakle, o mehaničkoj projekciji ponašanja pojave iz prošlog i sadašnjeg perioda u budućnost. Navedena pretpostavka ekstrapolacije trenda ujedno predstavlja i osnovno ograničenje, zbog čega prognozu po ovom metodu možemo sprovoditi samo u neposrednoj budućnosti i samo kod pojava koje pokazuju relativno stabilnu razvojnu tendenciju u dužem vremenskom periodu.

2) Metodi izravnanja

a) Metod pokretnih proseka – Kod ovog metoda se pokretni prosek ne pridružuje središnjem članu kao kod određivanja tipa funkcije trenda, već predstavlja prognozu za jedan period unapred. Izbor broja članova u grupi je subjektivan (187ročlani, petočlani proseci itd.)

b) Eksponencijalno izravnanje – metod se koristi za vremenske serije koje nemaju izražen trend. Takve serije karakterišu slučajna kolebanja oko svog prosečnog nivoa, koji je konstantan ili se tokom vremena sporo menja. Prognoza po metodu jednostavnog eksponencijalnog izravnanja dobija se polazeći od sledećeg opšteg izraza:

$$F_{t+1} = \alpha \cdot Y_t + (1 - \alpha) \cdot F_t$$

Jednostavno eksponencijalno izravnanje podrazumeva takvu transformaciju vremenske serije u kojoj su podaci zamenjeni ponderisanim prosekom svih prethodnih podataka, sa ponderima koji se eksponencijalno smanjuju sa starošću podataka.

89. Ekstrapolacija trenda (postupak, problemi i ograničenja)

Ekstrapolacija trenda podrazumeva produžavanje ocenjene funkcije trenda izvan uzoračkog perioda. Osnovna pretpostavka prilikom prognoziranja vremenske serije korišćenjem metoda dekompozicije jeste da će faktori koji su delovali na nivo serije u prošlosti i sadašnjosti delovati i u budućem periodu na isti način, približno istim intenzitetom, u istom smeru i bez značajnijeg uticaja novih faktora. Radi se, dakle, o mehaničkoj projekciji ponašanja pojave iz prošlog i sadašnjeg perioda u budućnost. Navedena pretpostavka ekstrapolacije trenda ujedno predstavlja i osnovno ograničenje, zbog čega prognozu po ovom metodu možemo sprovoditi samo u neposrednoj budućnosti i samo kod pojava koje pokazuju relativno stabilnu razvojnu tendenciju u dužem vremenskom periodu.

Ukoliko se vrši ekstrapolacija trenda godišnjih vremenskih serija, princip je da se u formuli ocenjenog trenda umesto t zameni vrednost za t koja bi odgovarala godini koju prognoziramo , da seriji pripada ta godina za koju se vrši prognoza.

Ukoliko se vrši ekstrapolacija trenda kvartalnih vremenskih serija, mora se uzeti u obzir i uticaj sezone. Postupak je sledeći:

1) Sprovodi se desezoniranje serije, odnosno, eliminiše se uticaj sezonskih faktora

2) Ocenjuje se funkcija trenda na dobijenim desezoniranim podacima, a ne na originalnim podacima vremenske serije, jer prisustvo sezonske komponente značajno zamagljuje osnovni tok serije

3) Ekstrapolira se ocenjeni trend za i-ti kvartal u nekoj budućoj godini

4) Dobijena ekstrapolisana vrednost se koriguje uticajem sezone pomoću sledeće formule:

$$\hat{y}_t^* = \frac{\hat{y}_t I_S}{100}$$

Ekstrapolacija trenda podrazumeva produžavanje ocenjene funkcije trenda izvan uzoračkog perioda. Osnovna pretpostavka prilikom prognoziranja vremenske serije korišćenjem metoda dekompozicije jeste da će faktori koji su delovali na nivo serije u prošlosti i sadašnjosti delovati i u budućem periodu na isti način, približno istim intenzitetom, u istom smeru i bez značajnijeg uticaja novih faktora. Radi se, dakle, o mehaničkoj projekciji ponašanja pojave iz prošlog i sadašnjeg perioda u budućnost. Pretpostavlja se da je reč o globalnom trendu, koji u celom periodu ostaje nepromenljiv, što često nije slučaj, pa ovaj vid prognoze može da bude dosta neprecizan. U praksi se primenjuje uglavnom bliska ekstrapolacija, tj. prognoza za nekoliko perioda unapred.

90. Predviđanje nivoa pojave pomoću trenda i sezonskih indeksa

Isto kao pitanje 89