PITANJA NA "USMENOME" DIJELU ISPITA IZ KOLEGIJA VJEROJATNOST I STATISTIKA

12. Testiranje hipoteze o nepoznatoj aritmetičkoj sredini. Dvosmjerni i jednosmjerni testovi.

Postavljaju se ove hipoteze:

$H_0 : .....\overline{X} = \overline{X}_0$

$H_1 : .....\overline{X} \neq \overline{X}_0$

Interval prihvaćanja hipoteze nulte hipoteze glasi:

$\overline{X}_0 \pm z \cdot Se(\overline{x})$

Ako je veličina uzorka n ≤ 30, tada se upotrebljava "t" iz Studentove distribucije umjesto koeficijenta "z" iz normalne distribucije.
Ako se aritmetička sredina uzorka nalazi u navedenome intervalu, prihvaćamo hipotezu H0.

Testiranje hipoteze o nepoznatoj aritmetičkoj sredini osnovnoga skupa može se postaviti i **jednosmjerno**.
Pri testiranju na donju granicu postavlja se alternativna hipoteza da je aritmetička sredina osnovnoga skupa manja od neke pretpostavljene vrijednosti. Hipoteza H0 glasi da je aritmetička sredina osnovnoga skupa veća ili jednaka nekoj pretpostavljenoj vrijednosti.

$H_0 : .....\overline{X} \geq \overline{X}_0$

$H_1 : .....\overline{X} \langle \overline{X}_0$

$DG = \overline{X}_0 - z \cdot Se(\overline{x})$

13. Testiranje hipoteze o nepoznatoj proporciji (relativnoj frekvenciji). Dvosmjerni i jednosmjerni testovi.

Testiranje hipoteze o nepoznatoj proporciji osnovnoga skupa vrši se na sličan način kao i testiranje hipoteze o nepoznatoj aritmetičkoj sredini osnovnoga skupa s obzirom na činjenicu da su im sampling distribucije iste.

$H_0 : .....P = P_0$ $\qquad$ $H_0 : .....P \leq P_0$

$H_1 : .....P \neq P_0$ $\qquad$ $H_1 : .....P > P_0$

Interval prihvaćanja nulte hipoteze glasi:

$P_0 \pm z \cdot Se(p)$

DG i GG princio isti kao u i za aritmetičku sredinu.

Ovo ne znam triba li:

Testiranje hipoteze o razlici proporcija dvaju osnovnih skupova vrši se na sličan način kao i testiranje hipoteze o razlici aritmetičkih sredina dvaju osnovnih skupova.

$H_0 : .....P_1 = P_2$

$H_1 : .....P_1 \neq P_2$

Interval prihvaćanja nulte hipoteze glasi:

$$0 \pm z \cdot Se(p_1 - p_2) \qquad Se(p_1 - p_2) = \sqrt{\hat{\bar{p}} \cdot (1 - \hat{\bar{p}}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})} \qquad \hat{\bar{p}} = \frac{m_1 + m_2}{n_1 + n_2}$$

14. Testiranje hipoteze o razlici aritmetičkih sredina dvaju nezavisnih osnovnih skupova.

Postavlja se nulta hipoteza da nema značajne razlike između aritmetičkih sredina dviju nezavisnih populacija.

Nezavisni uzorci potječu iz različitih populacija te među njima ne postoji povezanost (korelacija). Ukoliko se na istome uzorku vrše različita mjerenja pod različitim uvjetima, onda govorimo o **zavisnim uzorcima**. Kod zavisnih uzoraka postoji korelacija između rezultata prije i poslije eksperimenta. Kod nezavisnih uzoraka standardna greška je veća negoli kod zavisnih.

$$H_0 : .....\bar{X}_1 - \bar{X}_2 = 0$$

$$H_1 : .....\bar{X}_1 - \bar{X}_2 \neq 0$$

Test-veličina je Studentov t-omjer. S porastom veličine uzorka sampling distribucija teži ka normalnome obliku pa se može uporabiti i z-omjer.

$$t^* = \frac{\hat{\bar{X}}_1 - \hat{\bar{X}}_2}{Se(\bar{x}_1 - \bar{x}_2)} \qquad Se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Zaključci o statističkoj signifikantnosti donose se na osnovu greške tipa I.
Pretpostavke su da uzorci potječu iz normalnih populacija i da su im varijance jednake, međutim t-test je prilično robustan na neispunjavanje tih pretpostavki, naročito ako su uzorci istih veličina i ako su dovoljno veliki.
Dakle, snaga testa je veća što je greška tipa II manja. Prema tomu, može se zaključiti da **jednosmjerni testovi** imaju veću snagu od dvosmjernih testova, pa ih je pod određenim uvjetima bolje primjenjivati.
Jednosmjerni test izgledao bi ovako:

$$H_0 : .....\bar{X}_1 \geq \bar{X}_2$$

$$H_1 : .....\bar{X}_1 < \bar{X}_2$$

Jednosmjerni test može se postaviti kao test na donju ili kao test na gornju granicu. Npr. ako kod dvosmjernoga testa dobijemo signifikantnost od 8%, to znači da nema signifikantne razlike u aritmetičkim sredinama dviju nezavisnih populacija, a ako test postavimo jednosmjerno signifikantnost iznosi 4% što je statistički signifikantno.

15. Testiranje hipoteze o razlici aritmetičkih sredina dvaju zavisnih osnovnih skupova.

Ovdje je način postavljanja hipoteza i donošenja zaključaka isti kao i kod i kod testiranja razlike sredina dviju nezavisnih populacija, ali je standardna greška drukčija:

$$Se(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2 \cdot r_{1,2} \cdot Se(\bar{x}_1) \cdot Se(\bar{x}_2)}$$

gdje r1,2 predstavlja Pearsonov koeficijent linearne korelacije između dvaju mjerenja iste slučajne varijable na istome uzorku. Što je jača korelacija među mjerenjima standardna greška je manja. Prema tomu, kod testiranja razlike aritmetičkih sredina dviju zavisnih populacija **lakše se odbacuje nulta hipoteza negoli kod nezavisnih uzoraka.**
Za testiranje razlike između dvaju zavisnih uzoraka može se koristiti i **Sign-test** (test predznaka)
Kod ovoga testa također testira se hipoteza da dvije varijable imaju istu distribuciju. Računaju se razlike između dviju varijabli za sve parove te se svrstavaju u pozitivne, negativne i jednake. Ukoliko su dvije varijable jednako distribuirane, onda broj pozitivnih i negativnih razlika ne će biti signifikantno različit.
Test-veličina također je z-omjer koji pripada normalnoj distribuciji.

16. Testiranje hipoteze o nezavisnosti obilježja.

Nulta hipoteza glasi da ne postoji zavisnost dvaju nominalnih odnosno kategorijalnih obilježja. Hipoteze glase ovako:

$$H_0 : .....P_{ij} = P_{i\bullet} \cdot P_{\bullet j} \qquad \forall_i \forall_j$$
$$H_1 : .....\exists P_{ij} \neq P_{i\bullet} \cdot P_{\bullet j}$$

Test-veličina je sljedeća:

$$\chi^{*2} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(m_{ij} - e_{ij})^2}{e_{ij}} \qquad e_{ij} = \frac{m_{i.} \cdot m_{.j}}{n}$$

mij = originalne frekvencije (empirijske),
eij = očekivane (teorijske) frekvencije koje se izračunavaju pod pretpostavkom da ne postoji zavisnost dvaju obilježja osnovnoga skupa.
Navedena veličina uspoređuje se s kritičnom vrijednošću Hi-kvadrata iz tablica. Broj stupnjeva slobode jednak je df = (r-1) · (c-1), gdje oznaka r označava broj redaka, a oznaka c broj stupaca u tabeli kontingence.
Ukoliko empirijska vrijednost Hi-kvadrata ne premaši kritičnu vrijednost Hi-kvadrata iz tablica, prihvaćamo nultu hipotezu kao istinitu, tj. zaključujemo da nema ovisnosti obilježja elemenata osnovnog skupa. U drugom slučaju, ako vrijednost empirijskog Hi-kvadrata premaši kritičnu vrijednost prihvaćamo alternativnu hipotezu, tj. zaključujemo da postoji zavisnost dvaju obilježja elemenata osnovnog skupa. Tabela iz koje se izračunava vrijednost Hi-kvadrata naziva se tabelom kontingence.

17. Testiranje hipoteze da distribucija ima neki određeni oblik.

Može se testirati hipoteza da distribucija slučajne varijable X ima oblik neke teorijske distribucije. Ovdje će se pokazati testiranja da li distribucija ima oblik jednolike, binomne, Poissonove ili normalne distribucije. Kod svih testiranja test-veličina je vrijednost empirijskoga Hi-kvadrata:

$$\chi^{*2} = \sum_{i=1}^{k} \frac{(f_i - f_{ti})^2}{f_{ti}}$$

fi = originalne frekvencije (iz distribucije uzorka),
fti = teorijske frekvencije koje se izračunavaju pod pretpostavkom da distribucija ima oblik neke teorijske distribucije. Teorijske frekvencije za diskontinuirane (diskretne) slučajne varijable dobivaju se ovako:

$$f_{ti} = (\sum_{i=1}^{k} f_i) \cdot p(x_i)$$

Empirijska vrijednost Hi-kvadrata se uspoređuje s kritičnom vrijednošću Hi-kvadrata iz tablica na određenom nivou signifikantnosti testa α i uz određeni broj stupnjeva slobode df=ν.

Broj stupnjeva slobode izračunava se na sljedeći način:
- za jednoliku distribuciju df = k - 1
- za binomnu distribuciju df = k - 2
- za Poissonovu distribuciju df = k - 2
- za normalnu distribuciju df = k - 3

Veličina "k" predstavlja broj frekvencija.

Ukoliko empirijska vrijednost Hi-kvadrata ne premaši kritičnu vrijednost iz tablica, prihvaćamo hipotezu da empirijska distribucija ima pretpostavljeni oblik. Ukoliko je empirijska vrijednost Hi-kvadrata veća od kritične vrijednosti iz tablica, prihvaćamo alternativnu hipotezu da distribucija nema pretpostavljeni oblik.

18. Analiza varijance s jednim i s dva promjenjiva faktora. Analiza varijance na osnovu rangova.

(ovo je isječak iz knjige, maknite što mislite da je nepotrebno :)

U svrhu ispitivanja djelovanja faktora A na vrijednost slučajne varijable X potrebito je uzeti onoliko uzoraka koliko ima varijacija faktora A (A_1, A_2, A_3,....., A_k). Dakle, imamo k uzoraka od kojih svaki ima n_j elemenata. Navedeni uzorci mogu, ali ne moraju imati isti broj elemenata.

Ovo testiranje svodi se na testiranje značajnosti razlike aritmetičkih sredina više osnovnih skupova (populacija), odnosno testiramo hipotezu da svi uzorci potječu iz iste populacije.

U postupku testiranja značajnosti djelovanja promjenjivoga faktora A na slučajnu varijablu X postavljaju se sljedeće hipoteze:

$$H_0 : .....\sigma_A^2 = 0$$
$$H_1 : .....\sigma_A^2 \neq 0$$

Testiranje se vrši F-testom. F-omjer predstavlja omjer varijance koja se pripisuje djelovanju promjenjivoga faktora A i varijance unutar uzoraka, koja nije uzrokovana djelovanjem promjenjivoga faktora A. Naime, unutar svakoga pojedinog uzorka djeluje samo jedna varijacija faktora A, tj. A_i.

Pretpostavke za primjenu analize varijance s jednim promjenjivim faktorom su da uzorci potječu iz normalnih populacija i da imaju jednake varijance. Međutim ti uvjeti ne moraju biti ispunjeni ako su uzorci jednake ili slične veličine i ako su populacije međusobno slične u odstupanju od normalne distribucije.

Ocjene varijance dobivaju se iz odgovarajućih uzoraka na čije vrijednosti varijable X su djelovale varijacije faktora A.

Zbroj kvadrata ukupnih odstupanja vrijednosti varijable X od očekivane vrijednosti (aritmetičke sredine) može se raščlaniti na sljedeći način:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij}-\overline{X}_{\bullet\bullet})^2=\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij}-\overline{X}_{\bullet j})^2+\sum_{j=1}^{k}n_j\cdot(\overline{X}_{\bullet j}-\overline{X}_{\bullet\bullet})^2$$

(ukupno) (unutar uzoraka) (između uzoraka)

Izraz s lijeve strane predstavlja ukupan zbroj kvadrata odstupanja vrijednosti slučajne varijable X od aritmetičke sredine svih uzoraka. Prvi dio izraza s desne strane predstavlja zbroj kvadrata odstupanja vrijednosti varijable X od aritme-tičke sredine unutar uzoraka. Drugi dio izraza s desne strane predstavlja zbroj kvadrata odstupanja aritmetičkih sredina uzoraka od zajedničke aritmetičke sredine, odnosno zbroj kvadrata odstupanja između uzoraka. Zbroj kvadrata odstupanja vrijednosti numeričke varijable između uzoraka predstavlja u stvari efekat djelovanja faktora A.

Testiranje značajnosti djelovanja faktora A na varijablu X može se vršiti F-testom. Kod F-omjera stavlja se u odnos ocjena varijance između uzoraka (dakle varijanca koja se odnosi na djelovanje faktora A) i ocjena varijance unutar uzoraka:

$$F^*=\frac{\sum_{j=1}^{k}n_j\cdot(\overline{X}_{\bullet j}-\overline{X}_{\bullet\bullet})^2/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij}-\overline{X}_{\bullet j})^2/(n-k)}=\frac{s_A^2}{s_u^2}$$

Empirijski F-omjer uspoređuje se s tabličnom vrijednošću F-varijable koja se distribuira prema F distribuciji s (k-1) stupnjeva slobode u brojniku i (n-k) stupnjeva slobode u nazivniku. Ukoliko vrijednost empirijskoga (izračunatoga) F-omjera premaši tabličnu vrijednost za određene stupnjeve slobode i nivo signifikantnosti testa α, prihvaćamo hipotezu da je djelovanje faktora A značajno. U novije vrijeme za analizu varijance koriste se računalni programi (statistički paketi) koji daju precizno izračunatu signifikantnost testa umjesto da se ograničavamo na graničnu vrijednost od 5%.

Ukoliko je $F^*>F^{\alpha}_{(k-1;n-k)}\Rightarrow$ djelovanje faktora A je značajno, odnosno ako je signifikantnost testa manja od 5% prihvaćamo alternativnu hipotezu da je djelovanje faktora A statistički značajno.

Ukoliko se prihvati hipoteza da svi uzorci potječu iz iste populacije nas dalje ne zanimaju međusobne razlike pojedinih uzoraka. Međutim, ukoliko F-testom odbacimo nultu hipotezu može nas zanimati razlika aritmetičkih sredina između pojedinih uzoraka. Ovdje nije dobro primjenjivati standardni t-test za pojedine grupe. Glavni razlog je u tome što se povećanjem broja t-omjera povećava i vjerojatnost da se pojave slučajno značajni t-omjeri. Drugi razlog je što t-test vrijedi za slučajne uzorke. Ako iz niza aritmetičkih sredina izaberemo najveću i najmanju pa njihovu razliku testiramo t-testom onda te dvije aritmetičke sredine nisu izabrane po zakonu slučaja.

Dva faktora:

Na vrijednost neke slučajne varijable X može djelovati više od jednoga promjenjivog faktora. Ovdje će biti prikazano statističko testiranje značajnosti djelovanja dvaju promjenjivih faktora na vrijednost slučajne varijable X. Promjenjive faktore čiji utjecaj testiramo nazovimo A i B.

Analiza varijance s dva promjenjiva faktora također se svodi na hipotezu o razlici aritmetičkih sredina više osnovnih skupova, ali sada po dva različita faktora koji djeluju na istu numeričku

varijablu. Pretpostavke koje bi trebale biti ispunjene prigodom testiranja razlika u aritmetičkim sredinama vrijede kao i kod analize varijance s jednim promjenjivim faktorom.

U postupku testiranja značajnosti djelovanja varijacija u faktorima A i B postavljaju se sljedeće hipoteze:

$$H_0 : .....\sigma_A^2 = 0 \quad \text{odnosno} \quad H_0 : .....\sigma_B^2 = 0$$
$$H_1 : .....\sigma_A^2 \neq 0 \quad \text{odnosno} \quad H_1 : .....\sigma_B^2 \neq 0$$

Testiranje gornjih hipoteza vrši se standardnim F-testom. Naime, omjer dviju varijanci uvijek pripada F-distribuciji s određenim stupnjevima slobode u brojniku i nazivniku i uz određenu signifikantnost α.

Dekompozicija ukupnoga zbroja kvadrata odstupanja varijable X od aritmetičke sredine sada se može prikazati ovako:

$$\begin{pmatrix}\text{ukupan zbroj}\\ \text{kvadrata}\end{pmatrix} = \begin{pmatrix}\text{zbroj kvadrata}\\ \text{izmedu redaka}\end{pmatrix} + \begin{pmatrix}\text{zbroj kvadrata}\\ \text{izmedu stupaca}\end{pmatrix} + \begin{pmatrix}\text{rezidualni}\\ \text{zbroj kvadrata}\end{pmatrix}$$

$$\sum_{j=1}^{k}\sum_{i=1}^{c}(X_{ij} - \overline{X}_{\bullet\bullet})^2 = \sum_{i=1}^{c} n_i \cdot (\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 + \sum_{j=1}^{k} n_j \cdot (\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})^2 +$$
$$+ \sum_{J=1}^{k}\sum_{i=1}^{c}(X_{ij} - \overline{X}_{i\bullet} - \overline{X}_{\bullet j} + \overline{X}_{\bullet\bullet})^2$$

Testiranje značajnosti djelovanja faktora A (testiranje varijabiliteta između redaka):

$$F^* = \frac{s_A^2}{s_R^2}$$

Navedeni empirijski F-omjer uspoređuje se s kritičnom vrijendošću iz tablica F-distribucije s (c-1) stupnjeva slobode u brojniku i (n-k-c+1) stupnjeva slobode u nazivniku uz određeni nivo signifikantnosti testa α. Ukoliko empirijski F- omjer premaši kritičnu vrijednost iz tablica, zaključujemo da je djelovanje faktora A na slučajnu varijablu X značajno.

Testiranje značajnosti djelovanja faktora B:

$$F^* = \frac{s_B^2}{s_R^2}$$

Navedeni F-omjer uspoređuje se s kritičnom vrijednošću iz F-distribucije sa (k-1) stupnjeva slobode u brojniku i (n-k-c+1) stupnjeva slobode u nazivniku uz određeni nivo značajnosti α. Ukoliko empirijska vrijednost F-omjera premaši tabličnu vrijednost, zaključujemo da je djelovanje faktora B na slučajnu varijablu X značajno.

KRUSKAL-WALLIS TEST (ANALIZA VARIJANCE S JEDNIM PROMJENJIVIM FAKTOROM POMOĆU RANGOVA)

U prethodnim testiranjima (1. i 2.) testirali smo jednakost aritmetičkih sredina triju ili više osnovnih skupova. U situaciji kada neki od uvjeta za provedbu tih testova nisu ispunjeni (npr. ako su uzorci uzeti iz populacija koje nisu normalno distribuirane ili nisu s jednakim varijancama ili kada se podatci za statističku analizu sastoje samo od **rangova**) upotrebljavamo neparametrijski Kruskal-Wallis test. Test veličina je ova:

$H = \frac{12}{n(n+1)} \cdot \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3 \cdot (n+1)$, gdje su oznake sljedeće:

k = broj uzoraka,
n_j = veličina j-toga uzorka,
n = veličina svih uzoraka zajedno.
R_j = zbroj rangova u j-tome uzorku.

Postupak testiranja provodi se ovim koracima:

1. Opservacije iz uzoraka n_1, n_2,....., n_k iz k uzoraka kombiniraju se u jedinistveni niz veličine n i poredaju prema veličini od najmanjega prema najvećemu. Zatim se opservacije zamjenjuju njihovim rangovima, počevši od 1, koji se daje najmanjoj opservaciji do n, koji se daje najvećoj opservaciji. Ukoliko dvije opservacije imaju istu vrijednost, onda im se daje aritmetička sredina rangova (isto kao i kod računanja Spearmanovoga koeficijenta korelacije ranga).

2. Rangovi koji su dodijeljeni opservacijama u svakoj od k grupa zbrajaju se posebno te dobivamo k suma rangova.

3. Izračunava se test statistika prema prije navedenoj formuli.

4. Ukoliko se radi o tri uzorka i pet ili manje opservacija, signifikantnost navedene test-veličine određuje se iz posebnih tablica. Ukoliko ima više od pet opservacija i jednome ili više uzoraka, vrijednost H se uspoređuje s tabličnim vrijednostima χ^2 s k-1 stupnjeva slobode.

19. Linearna korelacija. Testiranje značajnosti koeficijenta linearne korelacije.

Pod pojmom korelacije podrazumijevamo međuzavisnost, odnosno povezanost slučajnih varijabli. Mjera stupnja podudarnosti (slaganja) slučajnih varijabli predstavlja mjeru korelacije.

Korelacija može biti pozitivna i negativna po smjeru. Pozitivna korelacija je onda kada rast jedne varijable prati rast druge varijable, odnosno pad jedne varijable prati pad druge. Negativna korelacija znači da rast jedne varijable prati pad druge varijable.

Ukoliko je korelacija potpuna, radi se o funkcionalnoj povezanosti varijabli. Tada se vrijednost jedne slučajne varijable može se s potpunom sigurnošću odrediti pomoću vrijednosti druge varijable. Ukoliko je korelacija djelomična, govorimo o stohastičkoj ili statističkoj vezi.

Najvažnija mjera linearne korelacije među slučajnim varijablama naziva se Pearsonov koeficijent linearne korelacije. Primjenjuje se na omjerne i intervalne skale mjerenja (**numeričke varijable**).

Pretpostavka za njegovu primjenu je da su obje slučajne varijable **normalno distribuirane**. Međutim, ako varijable X i Y nisu normalno distribuirane svejedno se može primijeniti Pearsonov koeficijent linearne korelacije ako je broj parova dovoljno velik.

$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})\cdot(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\cdot\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \qquad r=\frac{\sum_{i=1}^{n}X_i\cdot Y_i-n\cdot\overline{X}\cdot\overline{Y}}{n\cdot\sigma_x\cdot\sigma_y}$$

Izraz u brojniku podijeljen s *n* naziva se kovarijancom, ili prvim mješovitim momentom oko vrijednosti aritmetičkih sredina slučajnih varijabli X i Y.
Testiranje značajnosti koeficijenta linearne korelacije provodi se t-testom:

$$H_0 : .....r = 0$$
$$H_1 : .....r \neq 0$$

Ukoliko je koeficijent linearne korelacije osnovnoga skupa jednak nuli, onda sampling distribucija koeficijenata linearne korelacije teži normalnome obliku ako broj parova vrijednosti slučajnih varijabli X i Y teži k beskonačnosti. Ukoliko je broj parova mali (n < 30) može se koristiti Studentova distribucija.
Test veličina je ova:

$$t^* = \frac{\hat{r}}{Se(\hat{r})}$$

Za velike uzorke standardna greška računa se ovako:

$$Se(\hat{r}) = \sqrt{\frac{1}{n-1}}$$

dok se za male uzorke izračunava ovako:

$$Se(\hat{r}) = \sqrt{\frac{1-\hat{r}^2}{n-2}}$$

Zaključak o statističkoj značajnosti izvodi se kao i kod ostalih testova. Ako je signifikantnost testa manja od 5% onda se prihvaća alternativna hipoteza, odnosno koeficijent linearne korelacije je statistički značajan.
Kao i kod ostalih testova koji koriste t–test, može se također provesti i **jednosmjerno testiranje** ukoliko a priori znamo predznak koeficijenta linearne korelacije a potrebno je samo testirati njegovu statističku značajnost.

Jednosmjerno testiranje ide u korist alternativne hipoteze jer je za njeno prihvaćanje potrebita upola manja signifikantnost negoli kod dvosmjernoga testa. **Jednosmjerni testovi imaju veću snagu**.

VAŽNA NAPOMENA: Kod koeficijenta korelacije treba gledati i njegovu praktičnu značajnost, a ne samo statističku značajnost.

20. Korelacija ranga. Testiranje značajnosti koeficijenta korelacije ranga.

Od svih koeficijenata koji koriste rangove umjesto originalnih vrijednosti najviše u uporabi je **Spearmanov koeficijent korelacije ranga**. Njegova vrijednost kreće se u istome rasponu kao i koeficijent linearne korelacije.

Najčešće se koristi za mjerenje korelacije između redosljednih obilježja ili obilježja ranga, ali se može koristiti i kod numeričkih varijabli ukoliko nisu zadovoljene pretpostavke

za primjenu Pearsonovoga koeficijenta linearne korelacije. U tome slučaju ne koriste se numeričke vrijednosti varijabli X i Y nego njihovi rangovi. Računa se ovako:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Spearmanov koeficijent korelacije ranga spada u neparametrijsku statistiku te samim time ima slabiju snagu negoli Pearsonov koeficijent linearne korelacije.

Značajnost Spearmanovoga koeficijenta korelacije ranga testira se na isti način kao i Pearsonovoga koeficijenta linearne korelacije.

21. Linearna regresija.

(ovde ima viška, ne znam što točno treba)

Za razliku od korelacijske analize, kod regresijske analize treba odrediti koja varijabla će biti regresand (zavisna) varijabla. Nezavisne varijable u modelu obično se nazivaju regresorskim varijablama. U velikom broju primjena veza između varijabli je **linearna**, ali postoje također i slučajevi **nelinearne regresije**.

Dakle, trebamo odabrati matematičku funkciju koja će najbolje aproksimirati vezu među varijablama.

Ako imamo jednu zavisnu i jednu nezavisnu varijablu onda se radi o **jednostrukoj regresiji**, a ako imamo više nezavisnih varijabli onda govorimo o **višestrukoj** (multiploj regresiji).

Regresijski model izgleda ovako:

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

U gornjemu modelu Y je regresand (zavisna) varijabla, X je regresorska (nezavisna) varijabla, *e* je slučajna grješka, dok su *β* populacijski regresijski parametri (koeficijenti). Zadaća regresijske analize je **ocijeniti regresijske parametre na osnovu uzorka.**
Pri tomu slučajna grješka "e" mora udovoljavati određenim uvjetima (tzv. Gauss-Markovljevi uvjeti).
Prvi uvjet:

$$E(e) = 0 \quad \forall \mathrm{i}$$

To znači da je očekivana vrijednost slučajne grješke jednaka nuli. Slučajna grješka je čas pozitivna, čas negativna, ali ne smije imati nikakvo sistematsko kretanje u bilo kojemu smjeru. Ako se radi s konstantnim članom, onda je gornji uvjet ispunjen automatski jer imamo:

$$\sum_{i=1}^{N} (Y_i - \hat{Y}_i) = 0$$

Iz ovoga slijedi da je zbroj originalnih vrijednosti varijable Y jednak zbroju očekivanih (po regresiji) vrijednosti varijable Y.

Drugi uvjet:

$$E(e_i, e_j) = \sigma_e^2 < +\infty \quad \text{za i} = \text{j}$$

Radi se o uvjetu da varijanca reziduala bude konačna i čvrsta, tj. da se ne mijenja od opservacije do opservacije. Taj uvjet naziva se često uvjetom **homoskedastičnosti varijance reziduala.**

Ukoliko taj uvjet nije ispunjen radi se o tzv. **heteroskedastičnosti reziduala**. U tome slučaju varijanca može sustavno kovarirati s regresorskom varijablom.
Ukoliko ovaj uvjet nije ispunjen ocjena parametara standardnom metodom najmanjih kvadrata bit će neefikasna.

Treći uvjet:
$E(e_i, e_j) = 0 \quad \forall \ (i \neq j)$
odnosno $Cov(e_i, e_j) = 0 \quad (i \neq j)$.
Ovaj uvjet se odnosi na nepostojanje sustavnosti u slučajnim grješkama. Naime, ako je grješka doista slučajna ne smije postojati nikakva korelacija između vrijednosti varijable "e" s pomakom.

Ukoliko ovaj uvjet nije ispunjen, standardna metoda najmanjih kvadrata dat će neefikasne ocjene.

Četvrti uvjet:
Slučajna varijabla mora biti distribuirana nezavisno od regresorske (eksplanatorne) varijable:
$E(e_i, X_i) = 0$
Kao dodatak gornjim uvjetima još je važno napomenuti da slučajna varijabla mora biti distribuirana po normalnoj distribuciji,
$N(0; \sigma_e^2)$