Seminarski zadatak iz Statističkog praktikuma 1
Zadatak 5.

Mia Rošić

Siječanj, 2023.

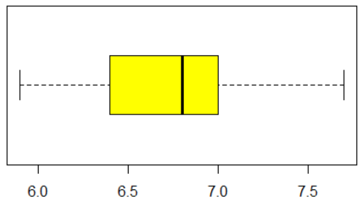
0 Podaci

U ovom seminarskom radu promatrat ćemo količine srebra u Bizantskom kovanom novcu, izražene u postotcima. Podaci su klasificirani u četiri grupe koje označavaju četiri razdoblja.
U svrhu lakšeg praćenja podataka pojedinih grupa, žuta boja označava prvu grupu na grafovima, narančasta drugu, zelena treću i plava četvrtu.

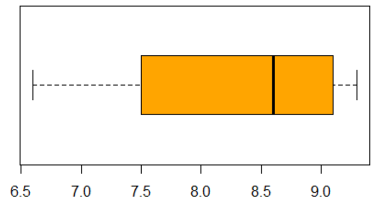
Podaci su navedeni u tablici:

1	2	3	4
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

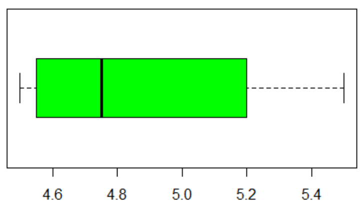
Radi bolje vizualizacije podataka koristimo dijagrame pravokutnika posebno napravljene za svaku pojedinu grupu:



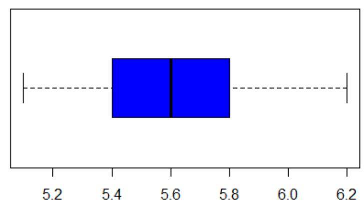
Slika 1: Grupa 1



Slika 2: Grupa 2



Slika 3: Grupa 3



Slika 4: Grupa 4

Na temelju ovih dijagrama možemo zaključiti da se dijagrami međusobno razlikuju. Medijani podataka izračunanih za svaku grupu su 6.8, 8.6, 4.75 i 5.6, redom. Na grafovima su medijani označeni kao zadebljane crte unutar obojanih pravokutnika. Medijan razdvaja podatke tako da je 50% podataka manje od medijana i 50% veće. Grafički to znači da na dijagramu pravokutnika lijevo od medijana i desno od medijana ima jednak broj podataka. Gledajući ovaj primjer, vidimo da se medijani i njihov položaj unutar pravokutnika razlikuju. Razlika se vidi i u duljini obojenih pravokutnika koja ovisi o tome kako su podaci grupirani oko medijana.

1 Distribucije podataka

U ovom odjeljku ispitujemo dolaze li podaci svakog razdoblja iz normalno distriburianih populacija. Ispitvanje provodimo pomoću dva kriterija: grafičkog, koji se sastoji od grafa normalnih vjerojatnosti i Lillieforsove inačice Kolmogorov-Smirnovljevog testa.

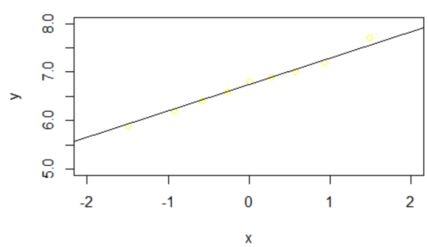
1.1 Graf normalnih vjerojatnosti

Neka je y_1,...,y_n sortirani uzorak. Za i=1,...,n definiramo kvantile jedinične normalne razdiobe formulom:

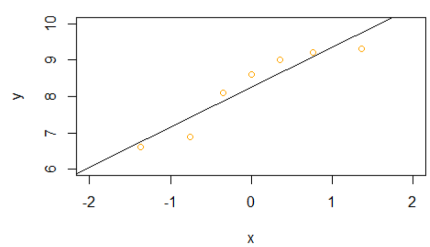
$$q_i = \Phi^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right)$$

Normalni vjerojatnosni graf je prikaz točaka (q_i, y_i). Pomoću njega uspoređujemo kvantile uzorka i teoretske razdiobe. Ako uzorak dolazi iz normalne razdiobe N(μ, σ), tada točke (q_i, y_i) su dobro aproksimirane pravcem y=μ + σq

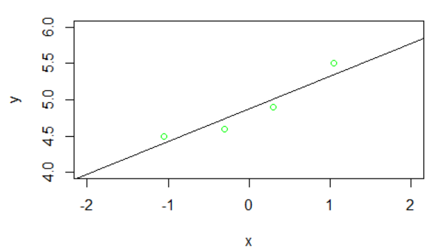
Normalni grafovi za sva četiri razdoblja:



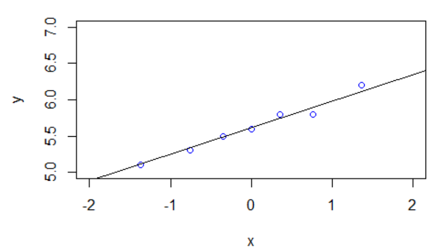
Slika 5: Grupa 1



Slika 6: Grupa 2



Slika 7: Grupa 3



Slika 8: Grupa 4

Gledajući grafove, ne možemo odbaciti pretpostavku da su podaci normalno distribuirani. Podaci grupa 1 i 4 prate pravac koji bi trebao aproksimirati podatke u slučaju da su normalno distribuirani. Kod grupa 2 i 3 su podaci malo udaljeniji od pravca, ali su i dalje poredani oko njega. Da ima malo više podataka za grupe 2 i 3 možda bi se bolje uočile nekakve pravilnosti.

1.2 Lillieforsova inačica Kolmogorov-Smirnovljevog testa

Neka je $Z_1, ..., Z_n$ slučajni uzorak varijable Z s nepoznatom funkcijom distribucije F, i neka je F_0 funkcija distribucije jedinične normalne razdiobe. Pretpostavke ovog testa glase:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

Empirijska funkcija distribucije je

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{Z_i < x\}}$$

Promatramo koliko empirijska distribucija odstupa od distribucije jedinične normalne razdiobe. Testna statistika je dana s:

$$D_n = \sup_{x \in R} \left| \hat{F}_n(x) - F_0(x) \right|$$

Na početku u varijable $z_{(1)}, ..., z_{(n)}$ spremamo sortirane standardizirane podatke određene grupe na način da procjenimo parametre očekivanja i standardne devijacije. Za svaku grupu zasebno radimo ovaj postupak. Računamo testnu statistiku D_n po sljedećoj formuli:

$$D_n = \max_{1 \leq i \leq n} \left\{ \max \left\{ \left| \frac{i-1}{n} - F_0(z_{(i)}) \right|, \left| \frac{i}{n} - F_0(z_{(i)}) \right| \right\} \right\}$$

Realizacija testne statistike D_n u ovom radu je 0.2009168, 0.1987394, 0.4889953 i 0.1613733, za svaku grupu redom.
P-vrijednosti dobivamo pozivom naredbe `ks.test(x = z[[i]], y = "pnorm")$p.value` u R-u te one iznose 0.1201804, 0.1352278, 0.4046568, 0.156016, za svaku grupu, redom. P-vrijednosti za sve grupe su veće od svih standardnih razina značanosti (1%, 5%, 10%) tako da u odnosu na svaku od njih ne možemo odbaciti hipotezu H_0, to jest ne možemo odbaciti pretpostavku da su podaci normalno distribuirani.

Zaključujemo da, na temelju oba kriterija, možemo smatrati da su distribucije svake grupe približno jedinične normalne.

2 Pouzdani intervali populacijske srednje vrijednosti μ

Neka su $\bar{X}_i$, i=1,2,3,4 procjenitelji parametara srednjih vrijednosti μ_i u populaciji i. Neka je $\hat{\sigma^2}$ nepristran procjenitelj zajedničkog parametra varijance na osnovi sva četiri uzorka zajedno.

Pokazala sam da su podaci po grupama iz normalnog modela. Tada vrijedi

$$T = \frac{\bar{X}_n - \mu}{S_n}\sqrt{n} \sim t(n-1)$$

df=$n-1$ označava stupnjeve slobode, odnosno broj nezvisnih opservacija u uzroku ili koliko stvari može varirati. Veličina uzorka je n pa imamo vrijednosti realizacija X_1,...,X_n. Ako pretpostavimo da je očekivanje uzorka $\bar{X}$, zadnju realizaciju X_n možemo dobiti na sljedeći način:

$$X_n = n\bar{X} - \sum_{i=1}^{n-1} X_i$$

Iz tog razloga je broj stupnjeva slobode $n-1$, jer X_n ne može varirati uz pretpostavku na očekivanje uzorka.

Ako sada promatramo $\bar{X}_i$, i=1,2,3,4, aritmetička sredina i-tog uzorka, $\hat{\sigma}$ standardna devijacija sva četiri uzorka zajedno i n=n_1+n_2+n_3+n_4 onda vrijedi:

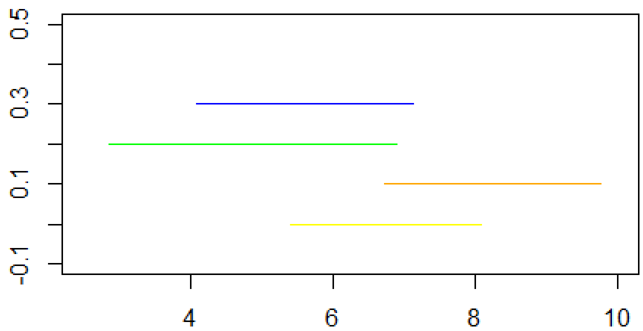
$$T_i = \frac{\bar{X}_i - \mu_i}{\hat{\sigma}}\sqrt{n_i} \sim t(n-4)$$

Ovdje imamo df=$n-4$ stupnja slobode iz razloga što je $\hat{\sigma^2}$ procjenitelj varijance na osnovi sva četiri uzorka skupa, to jest da bi izračunali $\hat{\sigma^2}$ trebaju nam sva četiri uzorka. Broj nepoznanica je n_1+n_2+n_3+n_4=n. Za svaku i-tu grupu smo pretpostavili da je očekivanje $\bar{X}_i$ pa onda za svaku grupu imamo n_i-1 stupnjeva slobode što rezultira:

$$df = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1) = \sum_{i=1}^{4} n_i - 4 = n - 4$$

95% pouzdane intervale za populacijske srednje vrijednosti μ_i određujemo računajući po sljedećoj formuli:

$$\bar{X}_i - t_{\alpha/2}(n-4)\frac{\hat{\sigma}}{\sqrt{n_i}} \leq \mu_i \leq \bar{X}_i + t_{\alpha/2}(n-4)\frac{\hat{\sigma}}{\sqrt{n_i}}$$



Slika 9: Grafički prikaz 95% pouzdanih intervala za μ_i

3 Jednostrana analiza varijanci

Pretpostavimo kako promatramo dva statistička obilježja. Neka je Z diskretna varijabla koja može poprimiti najviše konačno mnogo vrijednosti (razina). U ovom radu su te razine četiri razdoblja u kojima su se mjerile količine srebra. Z je nezavisna varijabla (faktor). Označimo s X skup svih podataka za sva četiri razdoblja. X je neprekidna varijabla koju nazivamo zavisnom varijablom (odazivnom).

Nas zanima postoji li razlika između očekivanja od X za različite razine faktora Z.

Ideja je primjeniti jednofaktorsku ANOVA-u u kojoj promatramo varijaciju između različitih razdoblja te odbacujemo nultu hipotezu ako je ona dovoljno velika. Kako bi taj test imao smisla, potrebno je provjeriti jesu li sve pretpostavke zadovoljene.

Pretpostavke:

1. **Za svaku razinu faktora, odazivna varijabla mora biti normalno distribuirana.** U prvom poglavlju je pokazano da su sve grupe normalno distribuirane.

2. **Uzorci između dvije različite razine faktora moraju biti nezavisni.** Budući da svaki uzorak iz različite razine faktora dolazi iz različitog razdoblja razumno je pretpostaviti da su uzorci nezavisni.

3. **Mora vrijediti homogenost varijance (homoskedastičnost).** Drugim riječima varijanca uzorka mora biti jednaka za svaki uzorak (ne smije ovisiti o razini faktora). Za pokazivanje homogenosti varijance sam provela Leveneov test koji testira hipoteze:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$$
$$H_1 : \neg H_0$$

Za Leveneov test nam je potrebno da su podaci između dvije grupe nezavisni, što imamo iz 2. pretpostavke. P-vrijednost dobivamo pozivom naredbe `leveneTest(uzorak faktor_uzorka)` u R-u te ona iznosi 0.1052. P-vrijednost je veća od svih standardnih razina značajnosti pa ne možemo odbaciti hipotezu H_0.

Sad kad imamo sve pretpostavke zadovoljene, provodimo ANOVA-u. Hipoteze za ANOVA-u su:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1 : \neg H_0$$

Za svako razdoblje računamo aritmetičke sredine $\bar{X}_i$ te uzoračke varijance S_i^2, i=1,...,m gdje m označava broj razina faktora koji je u ovom radu jednak 4. Varijacija između razina faktora se mjeri statistikom

$$SST = \sum_{i=1}^{m} n_i(\bar{X}_i - \bar{X})$$

te ju nazivamo **suma kvadrata zbog tretmana** (suma kvadrata odstupanja između grupa). U ovom radu SST iznosi 37.74753.
Varijacija unutar razine faktora se mjeri statistikom

$$SSE = \sum_{i=1}^{m} (n_i - 1)S_i^2$$

te ju nazivamo **suma kvadrata pogreške** (suma kvadrata odstupanja unutar grupa) te s našim podacima SSE iznosi 11.01544.
Varijanca između grupa se mjeri statistikom:

$$MST = \frac{SST}{m-1}$$

te ju nazivamo **srednjekvadratno odstupanje zbog tretmana**. Ona u ovom slučaju iznosi 12.58251
Varijanca unutar grupe se mjeri statistikom:

$$MSE = \frac{SSE}{n-m}$$

te ju nazivamo **srednjekvadratna pogreška**. Ona u ovom slučaju iznosi 0.478932
Testna statistika koju promatramo je

$$F = \frac{MST}{MSE}$$

koja u uvjetima nulte hipoteze ima F distribuciju s $m-1$ i $n-m$ stupnjeva slobode. Kritično područje je oblika

$$[f_\alpha(m-1, n-m), +\infty\rangle$$

U ovom radu je kritično područje za razinu značajnosti α=5% jednako

$$[9.276628, +\infty\rangle$$

dok testna statistika F iznosi 26.27201 i kao takva upada u kritično područje. Na razini značajnosti od 5% odbacujemo hipotezu H_0, hipotezu o jednakosti srednjih vrijednosti promatranih razdoblja. Do zaključka ovog testa smo mogli doći i računanjem p-vrijednosti koja kod ovog testa iznosi $1.305986 \cdot 10^{07}$ što je manje od svih standardnih razina značajnosti pa odbacujemo hipotezu H_0.