**SEMINARSKI ZADATAK**

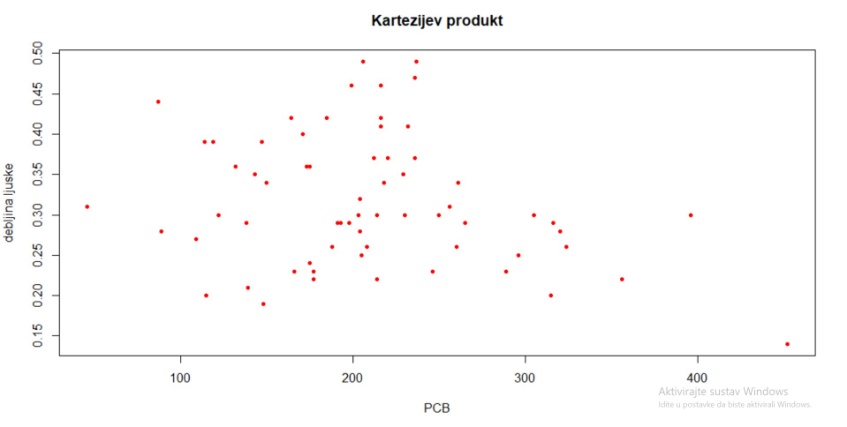
**ZADATAK 26 – IVAN ČULIN**

**OPIS ZADATKA:**

-U svrhu istraživanja utjecaja zagađenja okoliša na životinje, sprovedeno je sljedeće istraživanje. Na slučajan način prikupljeno je 65 pelikanskih jaja. Na svakom jajetu izmjerena je koncentracija PCB-a (vrste industrijskog zagađivača, u ppm) i debljina ljuske jajeta (u mm). Cilj našeg zadatka je provjeriti normalnost podataka, te uz pretpostavku da podaci dolaze iz bivarijantne normalne razdiobe, procijeniti parametre modela, te sprovesti različite statističke testove.

**a)Prikaz podataka u Kartezijevom koordinatnom sustavu**

Pogledajmo kako debljina ljuske jajeta ovisi o koncentraciji PCB-a.



**b)Provjera normalnosti podataka**

Neka X predstavlja koncentraciju PCB-a, a Y debljinu ljuske

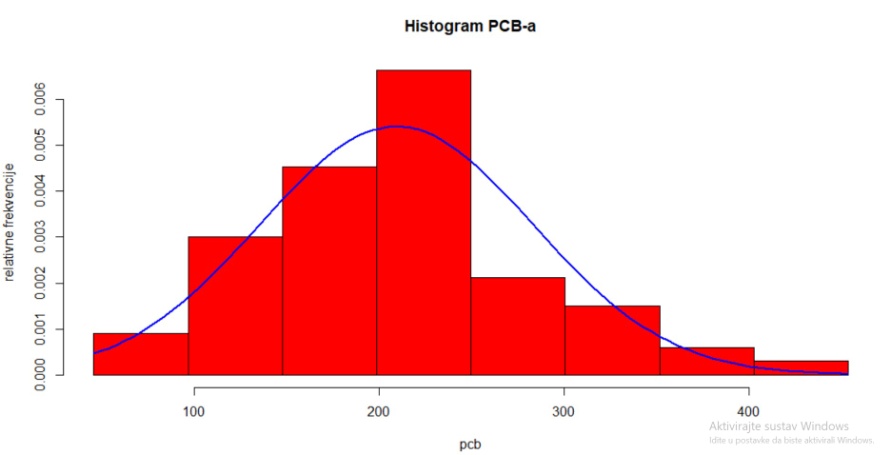
Za svaku od varijabli treba ispitati je li normalno distribuirana, i to upotrebom tri kriterija: grafičkog, koji se sastoji od grafa normalnih vjerojatnosti i histograma, Lillieforsove inačice Kolmogorov-Smirnovljevog testa i Pearsonovog χ2 -testa.

Prvo ćemo napraviti sve testove za varijablu X, a potom za varijablu Y, pri čemu ćemo postupak objašnjavati samo prilikom testiranja varijable X s obzirom da je za varijablu Y postupak identičan

**Testovi normalnosti za varijablu X**

Histogram

Prvo ćemo nacrtati histogram podataka, i to tako da su na y-osi relativne frekvencije kako bi taj histogram mogli usporediti sa procijenjenom funkcijom gustoće(funkcijom gustoće normalne razdiobe s očekivanjem koje je jednako aritmetičkoj sredini podataka, te standardnom devijacijom, koja je jednaka uzoračkoj standardnoj devijaciji). S obzirom da imamo ukupno 65 podataka, a običaj je da je broj razreda jednak sqrt(n) ili log(n), imat ćemo ukupno 8 razreda.



Vidimo da procijenjena funkcija gustoće poprilično dobro leži uz histogram tako da ne možemo odbaciti pretpostavku o normalnosti podataka

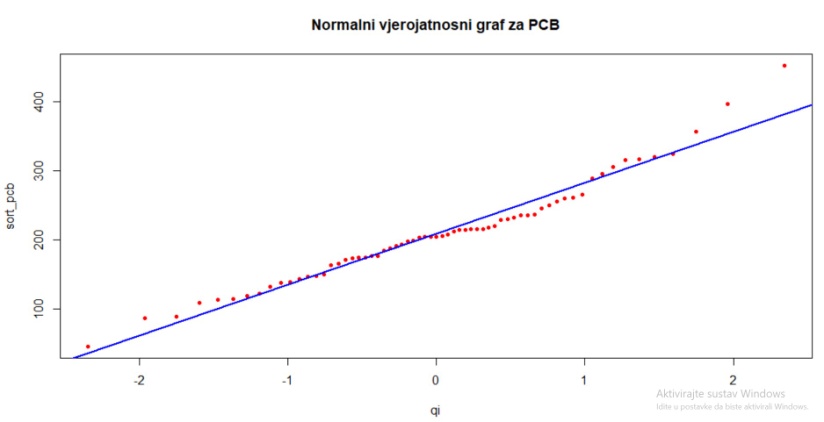
Normalni vjerojatnosni graf

Ideja vjerojatnosnih grafova jest uspoređivati kvantile uzorka i teoretske razdiobe. Neka je y(1), . . . , y(n) sortirani uzorak. Za i = 1, . . . , n definiramo kvantile jedinične normalne razdiobe na načinSnimka zaslona 2021-12-21 122807.jpg.

Graf dobiven prikazom točaka (qi , y(i)) nazivamo normalni vjerojatnosni graf. Ako uzorak dolazi iz Y ∼ N(µ, σ2 ), tada točke (qi , y(i)) aproksimativno moraju biti na pravcu

y = µ + σq.

Nakon napravljenog postupka usporedili smo normalni vjerojatnosti graf s pravcem y = µ + σq, gdje smo µ procijenili sa mean(X), a σ sa sd(X). Dobili smo sljedeći rezultat.



Vidimo da se ni na temelju normalnog vjerojatnosnog grafa ne može odbaciti pretpostavku da podaci o PCB-u dolaze iz normalne razdiobe.

Lillieforsov test

Ukoliko želimo testirati složenu hipotezu o pripadnosti uzorka nekoj normalnoj distribuciji, postupak je sljedeći:

-na osnovi uzorka procijenimo parametre očekivanja µ = mean(x) i standardne devijacije s = sd(x)

-početne podatke standardiziramo na način yi = (xi − µ) /s te za takve podatke računamo testnu statistiku Dn kao i u KS testu gdje za F0 uzimamo funkciju distribucije jedinične normalne razdiobe.

Nakon pokretanja koda u R-u dobili smo vrijednost testne statistike D=0.107, te p-vrijednost 0.064, pa ćemo na razini značajnosti od 0.1 odbaciti pretpostavku o pripadnosti podataka normalnoj distribuciji, a na razinama značajnosti od 0.05 i 0.01 nećemo odbaciti pretpostavku o pripadnosti podataka normalnoj distribuciji.

Pearsonov χ2 -test

S obzirom da Pearsonov χ2 –test promatra diskretna obilježja, naše podatke ćemo podijeliti u razrede kao što smo ih podijelili prilikom formiranja histograma. Nakon analize podataka vidimo da su frekvencije podataka unutar razreda sljedeće:[ 3 10 15 22 7 5 2 1]. Sada računamo teorijske frekvencije(razdiobe N(mean(X), sd2(X)) i dobijemo sljedeće rezultate:[ 4.1229017 8.9923980 15.6635678 17.2357444 11.9820363 5.2608154 1.4577921 0.2847444].S obzirom da sve teorijske frekvencije nisu ≥5 spojit ćemo prva 2 i zadnja 3 razreda. Sada dobijemo sljedeće rezultate:

Frekvencije: [13 15 22 7 8]

Teorijske frekvencije : [13.115300 15.663568 17.235744 11.982036 7.003352]

S obzirom da imamo 2 procijenjena parametra, broj stupnjeva slobode df=5-1-2=2

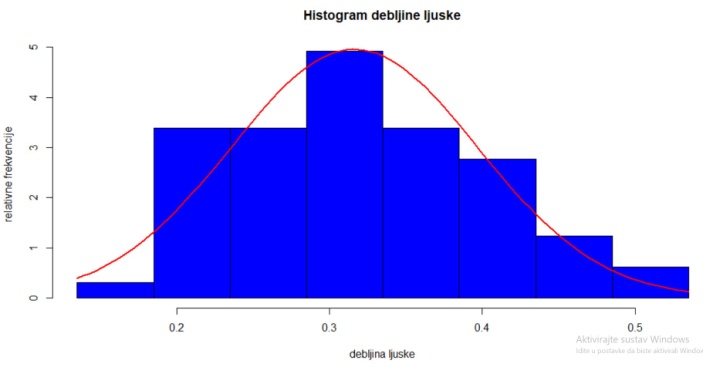
Vrijednost testne statistike H=3.56

p- vrijednost:0.168

Nakon analize dobivenih rezultata zaključujemo da ne možemo odbaciti pretpostavku da podaci dolaze iz N(mean(X), sd2(X)) razdiobe na razinama značajnosti 0.1, 0.05 i 0.01.

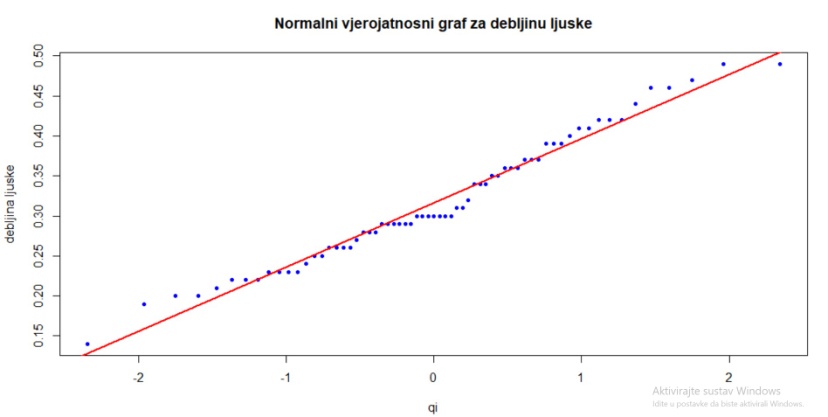
**Testovi normalnosti za varijablu Y**

Histogram

****

Vidimo da procijenjena funkcija gustoće poprilično dobro leži uz histogram tako da ne možemo odbaciti pretpostavku o normalnosti podataka

Normalni vjerojatnosni graf



Vidimo da ni na temelju normalnog vjerojatnosnog grafa ne može odbaciti pretpostavku da podaci o debljini ljuske dolaze iz normalne razdiobe.

Lillieforsov test

Nakon pokretanja koda u R-u dobili smo vrijednost testne statistike D=0.13414, te p-vrijednost 0.054, pa ćemo na razini značajnosti od 0.1 odbaciti pretpostavku o pripadnosti podataka normalnoj distribuciji, a na razinama značajnosti od 0.05 i 0.01 nećemo odbaciti pretpostavku o pripadnosti podataka normalnoj distribuciji.

Pearsonov χ2 -test

Radimo potpuno istu stvar kao sa varijablom X, te dobivamo sljedeće rezultate:

Frekvencije:[ 1 11 11 16 11 9 4 2]

Teorijske frekvencije:[ 3.339902 6.811461 12.509640 15.808196 13.746770 8.225670 3.386030 1.172331]

Ponovno moramo spojiti prva 2 i zadnja 3 razreda, te dobijemo sljedeće rezultate:

Frekvencije:[ 12 11 16 11 15]

Teorijske frekvencije:[ 10.15136 12.50964 15.80820 13.74677 12.78403]

Broj stupnjeva slobode df=2

Realizacija testne statistike H=1.454

p- vrijednost=0.4833

Iz dobivenih rezultata zaključujemo da ne možemo odbaciti pretpostavku da podaci dolaze iz N(mean(Y), sd2(Y)) razdiobe

**c)Procjena parametara i pouzdani intervali**

Ako pretpostavimo da vektor (X,Y) ima bivarijantnu normalnu razdiobu, onda vrijedi da je X~N(µx, σx2) i Y~N(µY ,σY2) .

Nadalje, s preddiplomske statistike znamo da je aritmetička sredina nepristrani procjenitelj za µ, a uzoračka standardna devijacija nepristrani procjenitelj za σ. Zbog toga vrijedi:

* Procjena za µx : mean(X)=209.123
* Procjena za σx : sd(X)=73.773
* Procjena za µY : mean(Y)=0.316
* Procjena za σY : sd(Y)=0.08

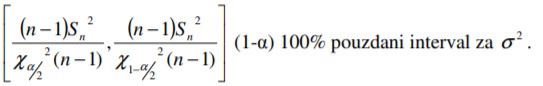
Nadalje, R je nepristrani procjenitelj za ρ, gdje je R Pearsonov koeficijent korelacije

* Procjena za ρ: rXY=-0.244

Konstruirajmo sada 95% pouzdane intervale za µx, σx,µY, σY, i ρ.

Neka je X1,…,Xn~N(µ, σ). Tada vrijedi:

t.jpg, Snimka zaslona 2021-12-21 145013.jpg Snimka zaslona 2021-12-21 145128.jpg

V.jpg, 

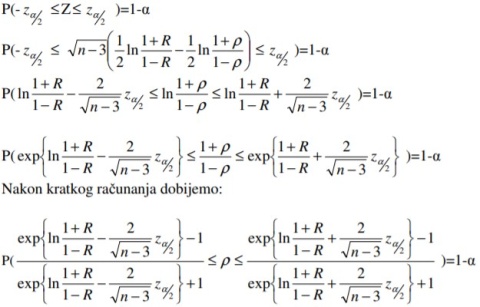
Koristeći napisane rezultate, te naše podatke dobiju se sljedeći rezultati:

* 95% pouzdani interval za µx : [190.842,227.403 ]
* 95% pouzdani interval za σx : [62.913,89.2]
* 95% pouzdani interval za µY : [0.2963,0.3362]
* 95% pouzdani interval za σY : [0.069,0.097]

Pri gornjem rješenju smo pouzdane intervale za σ dobili korjenovanjem pouzdanih intervala za σ2.

Sada još treba odrediti 95% pouzdani interval za ρ.

Snimka zaslona 2021-12-21 150936.jpg



Kada se sve izračuna dobije se :

* 95% pouzdani interval za ρ:[-0.4603,0.00015]

**d)Graf funkcije gustoće s prilagođenim vrijednostima pomoću izohipsi**

Ovaj zadatak ćemo riješiti koristeći se sljedećim teoremom:

**Neka je Z=(X,Y)~N( μ , C ). Tada postoji standardni**

**normalni vektor W t.d. vrijedi Z=AW+μ, gdje je A kvadratni**

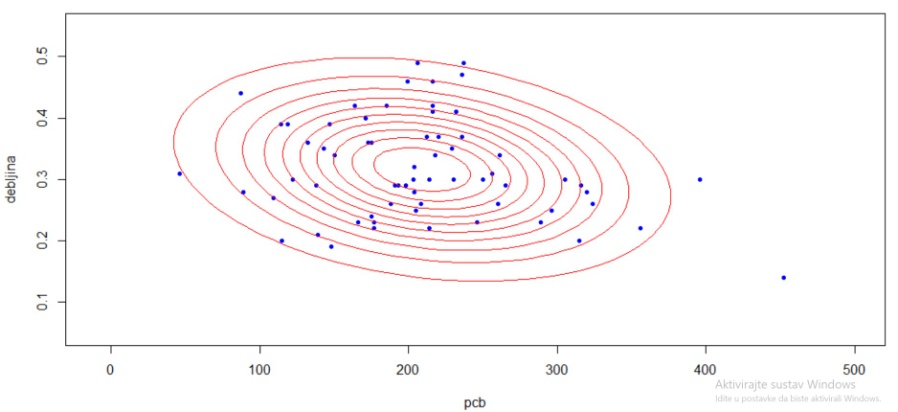
**korijen matrice C tj., C=AA**

Kako ćemo sada nacrtati te izohipse? Promotrimo prvo naš novodobiveni standardni normalni vektor W=(X1,X2)

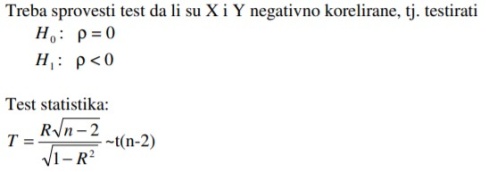
Tada vrijedi da je V= X12 + X22 ~ χ2(2). Sada zapravo želimo naći radijuse kružnica tako da vrijedi da je vjerojatnost da W upadne unutar tog kruga jednaka baš 6i/65. Međutim s obzirom da je X12 + X22 upravo jednadžba kružnice oko ishodišta jasno je da za radijuse upravo trebamo uzeti kvantile χ2(2) razdiobe zadane u zadatku. Naše izohipse ćemo dobiti onda translacijom zadanih kružnica pomoću Aw + μ. U nastavku slijedi prikaz R- koda ovog dijela zadatka s obzirom da je to bio dio zadatka koji nije dosad spominjan na predavanjima i vježbama.



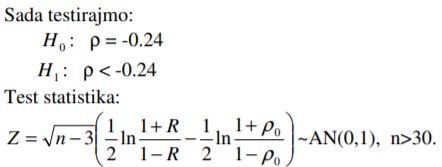
Pokretanjem gore napisanog r-koda dobije se sljedeće rješenje:



**e)Pearsonov koeficijent korelacije**



Za vrijednost testne statistike dobili smo sljedeći rezultat: T=-1.995. Za p-vrijednost dobijemo sljedeći rezultat:p vrijednost=0.0252. To znači da na razini značajnosti od 0.1 i 0.05 odbacujemo hipotezu da je Pearsonov koeficijent korelacije jednak 0, dok na razini značajnosti 0.01 ne odbacujemo spomenutu hipotezu.



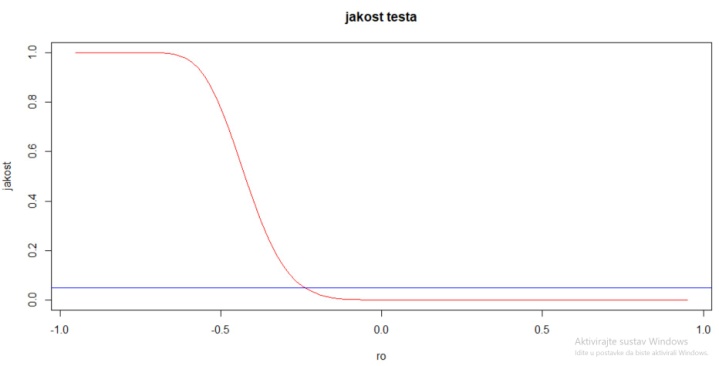
Realizacija testne statistike Z=-0.314.

Za jakost testa će nam trebati i kritično područje C=[-inf,-1.644]

S obzirom da realizacija testne statistike ne upada u kritično područje ne odbacujemo nultu hipoteze u korist alternativne. Zaključak smo također mogli donijeti na temelju p-vrijednosti koja je jednaka 0.487.

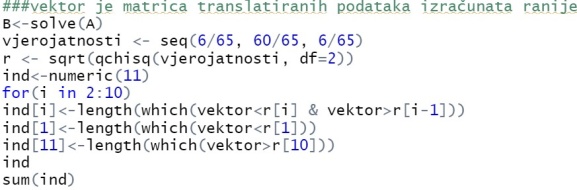
Još je potrebno nacrtati graf jakosti testa. S obzirom da je alternativna hipoteza jednostrana, točka ρ0=-0.24 treba biti točka infleksije, te u njoj vrijednost funkcije jakosti mora biti 0.05, kada ρ teži u -1 vrijednost mora biti 1, a kada ρ teži u 1 vrijednost mora biti 0. U nastavku prilažem r-kod u kojem je rađena jakost testa, te graf funkcije jakosti testa.





**f)** **Pearsonov χ2 –test o pripadnosti bivarijantnoj normalnoj razdiobi**

U d) dijelu zadatka smo nacrtali zadane izohipse sada još moramo izračunati koliko je točaka upalo unutar kojeg područja. Kako bi to napravili opet ćemo se koristiti spomenutim teoremom. S obzirom da je Z=Aw + μ, te je A regularna, vrijedi da je w=A-1(Z- μ). Sada kada smo tako transformirali naše podatke Z=(X,Y), vrlo lako možemo vidjeti unutar kojeg kružnog vijenca se oni nalaze. U nastavku prikazujemo R- kod u kojem smo računali frekvencije od svake od 11 područja.



Frekvencije=[ 6 5 7 7 3 5 13 3 6 6 4]

Teorijske frekvencije=[6 6 6 6 6 6 6 6 6 6 5]

S obzirom da su sve teorijske frekvencije ≥5 ne moramo spajati razrede. Također procijenili smo ukupno 5 parametara pa je broj stupnjeva slobode df=11-1-5=5

Još preostaje izračunati realizaciju testne statistike H i p-vrijednost. Nakon računa u R-u dobijemo sljedeće rezultate:

H=12.03, p-vrijednost=0.034, pa na razini značajnosti, od 0.1 i 0.05 odbacujemo pretpostavku o pripadnosti podataka bivarijantnoj normalnoj razdiobi, a na razini značajnosti od 0.01 ne odbacujemo.