Seminar iz Primjenjene statistike

Zadatak 16

Iva Tutiš

1.Zadatak

Hotellingov T^2 -test za dva uzorka

Podaci: ALM, Exercise 1.8.3, str. 68. i 70.

(a) Opišite test omjera vjerodostojnosti za hipotezu o jednakosti oˇcekivanja (*TMS 8.9, Problem 1, str. 137 - 138.*) dva normalna uzorka.

(b) Ispitajte normalnost podataka.

(c) Sprovedite test iz (a) na usporedbu skupina ”Thyroxine” i ”Control” iz zadatka.

2. Opis testa omjera vjerodostojnosti za hipotezu o jednakosti očekivanja dva normalna uzorka

Neka su $x_1, \dots, x_n$ i $y_1, \dots, y_m$ dva slučajna uzorka za međusobno nezavisne slučajne varijable $X \sim N(\mu, \sigma^2)$ i $Y \sim N(\tau, \sigma^2)$. Želimo testirati hipoteze:

1. $\mathrm{H}_0: \mu = \tau$
2. $\mathrm{H}_1: \mu \neq \tau$

Definiramo varijance uzoraka sa:

$S_x = [\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{y})^T]/(n-1)$

$S_y = [\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T]/(m-1)$

Te definiramo:

$$S = \left[(n-1)S_x + (m-1)S_y\right]/(n+m-2)$$

Statistika koju koristimo za testiranje gornjih hipoteza je dana sa:

$$T^2 = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1} (\bar{x} - \bar{y})^T S^{-1} (\bar{x} - \bar{y})$$

Sada, želimo odrediti

1. Distribuciju od S
2. Distribuciju od T^2
3. Sada je omjer vjerodostojnosti za testiranje gornje hipoteze

3.1 Distribucija od S

Sa predavanja znamo da vrijedi $(n-1)S_x \sim W_p(n-1, \sigma^2)$ i $(m-1)S_y \sim W_p(m-1, \sigma^2)$,
Po definiciji Wishartove distribucije postoje nezavisni
$X_i \sim N_p(0, \sigma^2)$, gdje je $i = 1, \dots n-1$
$Y_i \sim N_p(0, \sigma^2)$, gdje je $i = 1, \dots m-1$
za koje vrijedi

$$(n-1)S_x = \sum_{i}^{n-1} X_i {X_i}^T$$

$$(m-1)S_y = \sum_{i}^{m-1} Y_i {Y_i}^T$$

Stavimo da je $X_{i+n-1} = Y_i$ za sve $i = 1, \dots m-1$ dobivamo da je
$(n-1)S_x + (m-1)S_y = \sum_i^{n+m-2} X_i {X_i}^T$
Pa iz definicije Wishartove distribucije je:

$$(n-1)S_x + (m-1)S_y \sim W_p(n+m-2, \sigma^2)$$

To jest distribucija od S je

$$(n+m-2)S \sim W_p(n+m-2, \sigma^2)$$

2.2 Distribucija od T^2

Statistika od T^2 je dana sa

$$T^2 = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1} (\bar{x} - \bar{y})^T S^{-1} (\bar{x} - \bar{y})$$

Gdje su x i y zapisi uzoraka (u matričnom obliku).

Podijelimo li izraz sa $(n+m-2)$ te izrazimo prvo zagradu na desnoj strani, dobivamo:

$$\frac{1}{n+m-2} T^2 = \frac{mn}{n+m} (\bar{x} - \bar{y})^T (\,(n+m-2)S)^{-1} (\bar{x} - \bar{y})$$

Izrazimo li to preko modela, dobivamo $x = 1_n \mu + E_x$ i $y = 1_m \mu + E_y$, gdje je

$$vec(E_x) \sim N_{np}(0, \sigma^2 \otimes \mathrm{I}_n)$$

$$vec(E_y) \sim N_{mp}(0, \sigma^2 \otimes \mathrm{I}_m)$$

Sada, koristeći metodu najmanjih kvadrata, možemo procijeniti očekivanja iz hipoteze.

Pa je:

$\hat{\mu}^T = (1^T 1)^{-1} 1^T x = \bar{x}^T$

$\hat{\tau}^T = (1^T 1)^{-1} 1^T y = \bar{y}^T$

E_x i E_y imaju normalnu razdiobu (po pretpostavci iz zadatka), pa slijedi da je

$\epsilon_x = \frac{1}{n} 1_n E_x \sim N_n(0, \frac{1}{n}\sigma^2)$

$\epsilon_y = \frac{1}{m} 1_m E_y \sim N_m(0, \frac{1}{m}\sigma^2)$

Te iz toga i činjenica da je

$$\bar{x} - \mu = \epsilon_x$$

$$\bar{y} - \tau = \epsilon_y$$

Slijedi da je

$$\frac{1}{n+m-2}T^2 = \frac{mn}{n+m}\left(\epsilon_x - \epsilon_y + \mu - \tau\right)^T\left((n+m-2)S\right)^{-1}(\epsilon_x - \epsilon_y + \mu - \tau)$$

Nadalje, iz nezavisnost ϵ_x i ϵ_y i regularnosti matrice $\sigma^2 > 0$:

$$\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\left(\epsilon_x - \epsilon_y + \mu - \tau\right) \sim N_p\left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}(\mu - \tau), \sigma^2\right)$$

Definiramo, uz oznaku $\delta = \sigma^{-1}\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}(\mu - \tau)$,

- $H := \sigma^{-1}\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\left(\epsilon_x - \epsilon_y + \mu - \tau\right) \sim N_p(\delta, I_p)$
- $K := (m+n-2)\sigma^{-1}S\sigma^{-1}$

Gdje se distribucija od K može opisati sa:

$$K = (m+n-2)\sigma^{-1}S\sigma^{-1} = \sigma^{-1}\left(\sum_i^{n+m-2} X_i {X_i}^T\right)\sigma^{-1} = \sum_i^{n+m-2}(\sigma^{-1}X_i)(\sigma^{-1}X_i)^T$$

Sada očito vrijedi $\sigma^{-1}X_i \sim N_p(0, I_p)$, pa iz definicije Wishartove distribucije slijedi

$$K \sim W_p(n+m-2), \text{ gdje je } n+m-2 > p$$

Pa slijedi

$$\begin{aligned}\frac{1}{n+m-2}T^2 &= \frac{mn}{n+m}\left(\epsilon_x - \epsilon_y + \mu - \tau\right)^T\sigma^{-1}\sigma\left((n+m-2)S\right)^{-1}\sigma\sigma^{-1}\left(\epsilon_x - \epsilon_y + \mu - \tau\right)\\ &= H^T K^{-1} H\end{aligned}$$

Koristeći propoziciju 1.7 sa predavanja, sada možemo zaključiti da je

$$\frac{m+n-p-1}{p}\frac{1}{n+m-2}T^2 = \frac{m+n-p-1}{p}H^T K^{-1} H \sim \mathrm{F(p, m+n-p-1, \delta^T\delta)}$$

$$(gdje\ je\ m+n-p-1 = m+n-2-p+1)$$

2.3 Omjer vjerodostojnosti

Neka su $X_1, \ldots, X_n$ slučajni vektori distribucije $N(\mu, \sigma^2)$.

Neka su dani vektori nezavisni.

Logaritamska vjerodostojnost je zadana sa:

$$l(\hat{\mu}, \hat{\sigma})) = \left(-\frac{np}{2}\right)log(2\pi) + \left(\frac{-n}{2}\right)log(det\hat{\sigma}) + (-\frac{np}{2})$$

Vjerodostojnost je sada zadana sa:

$$L\left(\hat{\mu}, \widehat{\sigma^2}\right) = \exp\left(l(\hat{\mu}, \hat{\sigma})\right) = (2\pi)^{-\frac{np}{2}} det\hat{\sigma}^{\frac{-n}{2}} e^{-\frac{np}{2}}$$

Podsjetimo se, tražimo omjer vjerodostojnosti za testiranje hipoteze

$$H_0: \theta \in \Theta_0$$

$$H_1: \theta \in \Theta_1$$

Gdje je

$$\Theta = \Theta_0 \cup \Theta_1 = \{\mu_0, \mu_1 \in \mathrm{M}_{p,1}, \sigma^2 \in \mathrm{M}_{q,q}, \sigma^2 > 0\}$$

$$\Theta_0 = \{\mu_0 \in \mathrm{M}_{p,1}, \sigma^2 \in \mathrm{M}_{q,q}, {\sigma_0}^2 > 0\}$$

Omjer vjerodostojnosti je sada zadan sa statistikom Λ:

$$\Lambda = \frac{\max\{L(\Theta) : \Theta \in \Theta_0\}}{\max\{L(\Theta) : \Theta \in \Theta\}}$$

Označimo nezavisne varijable $X \sim N(\mu, \sigma^2)$ i $Y \sim N(\tau, \sigma^2)$.

Uz notaciju kao u dijelu 2.1, imamo da za $\Theta \in \Theta$ vrijedi

$$L(\mu_0, \mu_1, \sigma^2)$$
$$= (2\pi)^{-\frac{(m+n)p}{2}} det\hat{\sigma}^{\frac{-(m+n)}{2}} e^{-0.5\mathrm{tr}((n-1)S_x\sigma^2-1)-0.5\,\mathrm{tr}((m-1)S_y\sigma^2-1)(-0.5\mathrm{n})(\mathrm{x}-\mu_0)^T\sigma^{-2}\sigma^2(-0.5\mathrm{m})(\mathrm{y}-\mu_1)^T\sigma^{-2}(\mathrm{y}-\mu_1)}$$

Logaritmiranjem izraza i traženjem njegovog maksimuma dobivamo da se maksimum postiže za:

1. $\widehat{\mu_0} = \bar{x}$
2. $\widehat{\mu_1} = \bar{y}$
3. $\widehat{\sigma^2} = \frac{1}{m+n}((n-1)S_x + (m-1)S_y)$

Uvrštavanjem tih rezultata za procjenitelje maksimalne vjerodostojnosti, dobivamo:

$$L(\widehat{\mu_0}, \widehat{\mu_1}, \widehat{\sigma^2}) = (2\pi)^{-\frac{(m+n)p}{2}} det\hat{\sigma}^{\frac{-(m+n)}{2}} e^{-\frac{(m+n)p}{2}}$$

Provedbom sasvim analognog računa, ali uz pretpostavku $\Theta \in \Theta_0$ umjesto $\Theta \in \Theta$, dobivamo:

- $\widehat{{\sigma_o}^2} = \frac{1}{m+n}\left((n-1)S_x + (m-1)S_y + \frac{mn}{m+n}(\bar{x}-\bar{y})(\bar{x}-\bar{y})^T\right)$
- $L(\widehat{\mu_0}, \widehat{\mu_1}, \widehat{{\sigma_o}^2}) = (2\pi)^{-\frac{(m+n)p}{2}} det\widehat{\sigma_0}^{\frac{-(m+n)}{2}} e^{-\frac{(m+n)p}{2}}$

Pa to napokon uvrštavamo:

$$\Lambda = \frac{\max\{L(\Theta) : \Theta \in \Theta_0\}}{\max\{L(\Theta) : \Theta \in \Theta\}} = \frac{L(\widehat{\mu_0}, \widehat{\mu_1}, \widehat{{\sigma_o}^2})}{L(\widehat{\mu_0}, \widehat{\mu_1}, \widehat{\sigma^2})} = \frac{det\widehat{\sigma_0}^{\frac{-(m+n)}{2}}}{det\hat{\sigma}^{\frac{-(m+n)}{2}}} = \left(\frac{det\widehat{\sigma_0}}{det\hat{\sigma}}\right)^{\frac{-(m+n)}{2}}$$

Pa se sada možemo prisjetiti Leme s predavanja koja tvrdi:

Za $A \in M_{p,q}$ *i* $B \in M_{q,p}$ slijedi $det(I_p + AB) = det\,(I_q + BA)$

Označimo li:

$$A = ((n-1)S_x + (m-1)S_y)^{-1}(\bar{x}-\bar{y})$$

$$B = \frac{mn}{m+n}(\bar{x} - \bar{y})^T$$

Iz leme direktno slijedi (budući $I - AB = \Lambda$)da

$$\Lambda = \det(I - BA) = \det\left(I - \frac{mn}{m+n}(\bar{x} - \bar{y})^T \left((n-1)S_x + (m-1)S_y\right)^{-1} (\bar{x} - \bar{y})\right)$$

Sada je omjer za testiranje H_0 u odnosu na H_1 jednak

$$\Lambda = \left(1 + \frac{1}{m+n-2}T^2\right)^{\frac{-(m+n)}{2}}$$

Još nam je prostalo izračunati p-vrijednost testa omjera vjerodostojnosti. Definirajmo pomoćnu funkciju:

$$h(t) = \left(1 + \frac{1}{m+n-2}t\right)^{\frac{-(m+n)}{2}}$$

Takva funkcija je očito padajuća i postiže vrijednost Λ za $t = T^2$. Koristeći te činjenice, dobivamo:

$$p = \mathrm{P}(\Lambda \leq c) = P(h(T^2) \leq c) = \mathrm{P}\left(T^2 \geq h^{-1}(\mathrm{c})\right)$$
$$= 1 - \mathrm{P}\left(\frac{m+n-p-1}{p(n+m-2)}T^2 \leq \frac{m+n-p-1}{p(n+m-2)} \leq h^{-1}(\mathrm{c})\right)$$

$$(gdje\ je\ m+n-p-1 = m+n-2-p+1)$$

Budući da iz dijela 2.2 znamo odgovarajuću distribuciju varijable $\frac{m+n-p-1}{p(n+m-2)}T^2$, uz pretpostavku da je početna hipoteza istinita se može izračunati željena p-vrijednost.

3.Podaci

U donjim tablicama su dani podaci o težinama tri grupe štakora, kojima je dana voda infuzirana kemikalijom – u prvoj grupi je to bio hormon štitnjače (Thyroxine), u drugoj grupi kemikalija koja demotivira proizvodnju hormona štitnjače (Thiouracil), dok je treće grupa štakora bila kontrolna.

Mi ćemo ispitati normalnost podataka grupa tretiranih Thyroxinom i kontrolne (Control) grupe Lillieforsovim testom (što je test normalnosti baziran na Kolmogorov-Smirnoffljevom testu), te ćemo ispitati jednakost očekivanja testom sa statistikom T^2 kako je opisan u **(2.)** dijelu zadatka.

Rezultate i grafove temeljene na podacima ćemo generirati uz pomoć programskog jezika R.

Tablica 2.1 Grupa Thyroxine

Thyroxine				
Tjedan 0	Tjedan 1	Tjedan 2	Tjedan 3	Tjedan 4
59	85	121	156	191
54	71	90	110	138
56	75	108	151	189
59	85	116	148	177
57	72	97	120	144
52	73	97	116	140

Tablica 2.2 Kontrolna grupa

Control				
Tjedan 0	Tjedan 1	Tjedan 2	Tjedan 3	Tjedan 4
57	86	114	139	172
60	93	123	146	177
52	77	111	144	185
49	67	100	129	164
56	81	104	121	151
46	70	102	131	153
51	71	94	110	141
63	91	112	130	154
49	67	90	112	140

Naredbe kojima vršimo unos koda u R su:

```
thyroxine <- matrix(c(59,85,121,156,191, 54,71,90,110,138,56,75,108,151,189,
59,85,116,148,177, 57,72,97,120,144,52,73,97,116,140,52,70,105,138,171),
nrow=7,ncol=5,byrow=TRUE)

control <- matrix(c(57,86,114,139,172,60,93,123,146,177,52,77,111,144,185,
49,67,100,129,164,56,81,104,121,151,46,70,102,131,153,51,71,94,110,141,63,
91,112,130,154,49,67,90,112,140,57,82,110,139,169),nrow=10,ncol=5,byrow=TRUE)
i in 1:5
```

3.1 Ispitivanje normalnosti za grupe Thyroxin i Control

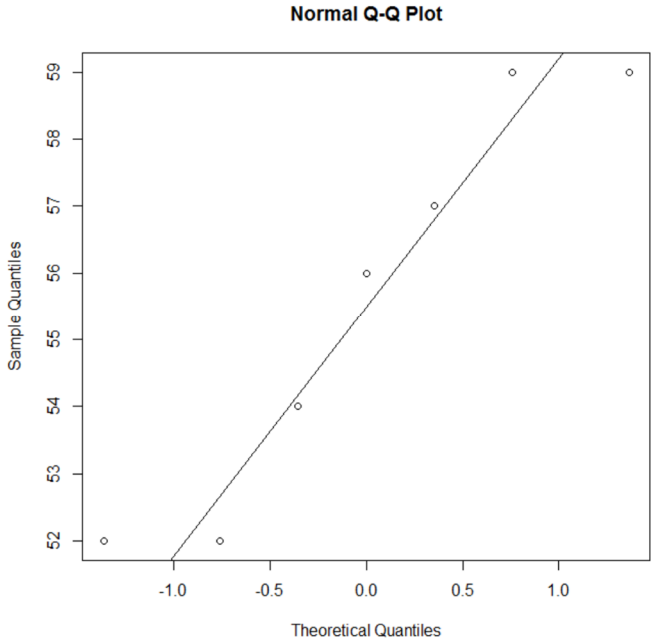
Kod u R-u kojim nad unesenim podacima izvršavamo Lillieforsov test te crtamo 5 grafova, od kojih svaki ispituje normalnost podataka u tjednu i (gdje $i = 0, \dots 4$) :

```
qqnorm(thyroxine[,i])
qqline(thyroxine[,i])
lillie.test(thyroxine[,i])
qqnorm(control[,i])
qqline(control[,i], col = 2)
lillie.test(control[,i])
```

Time dobivamo rezultate:

Tjedan 0

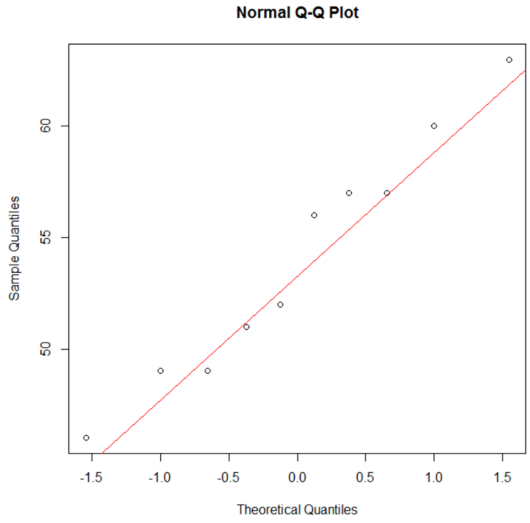
Grupa Thyroxine



Sa pripadnim rezultatima

- D = 0.1694
- p-vrijednost = 0.7868

Grupa Control

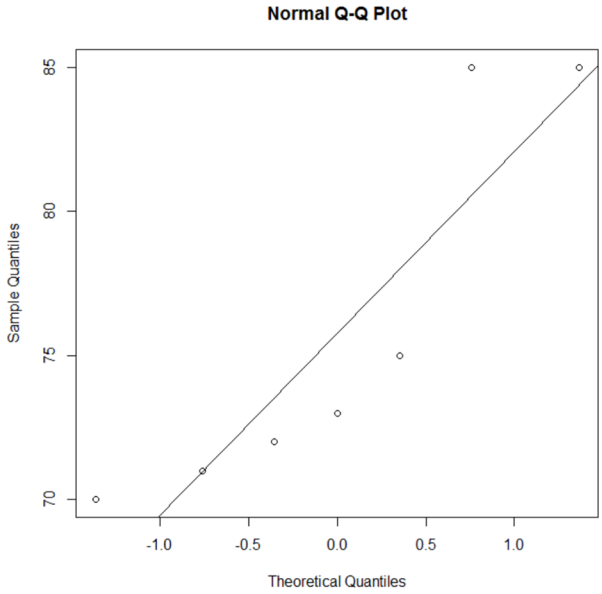


Sa pripadnim rezultatima

- D = 0.1435
- p-vrijednost = 0.8096

Tjedan 1

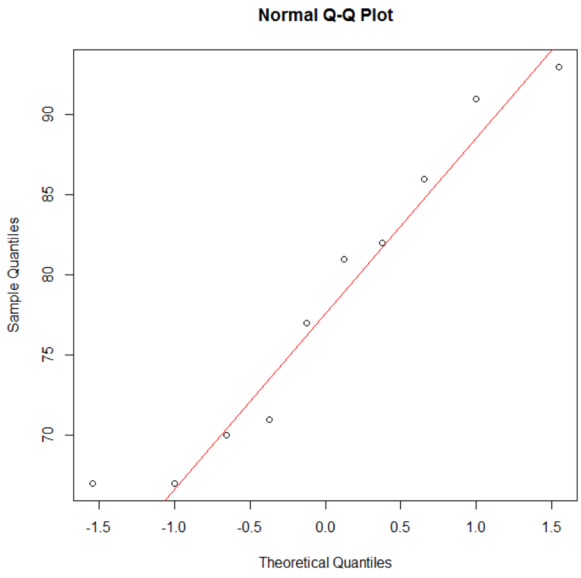
Grupa Thyroxine



Sa pripadnim rezultatima

- D = 0.2672
- p-vrijednost = 0.1363

Grupa Control

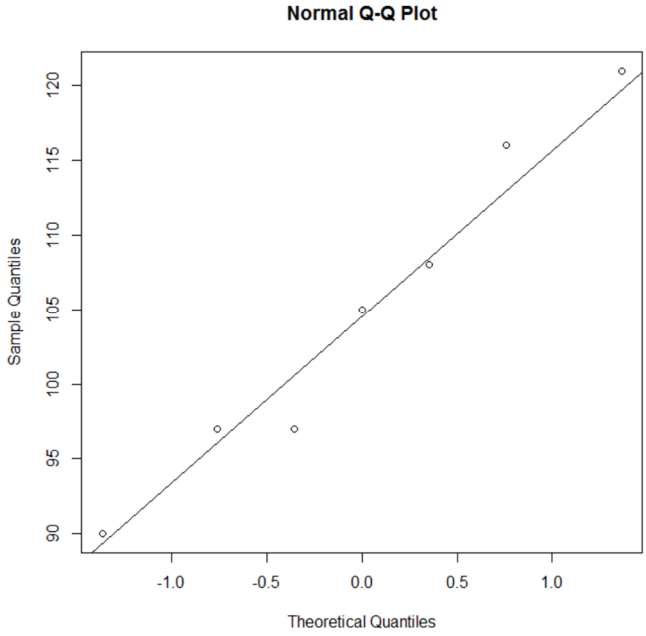


Sa pripadnim rezultatima

- D = 0.1817
- p-vrijednost = 0.4592

Tjedan 2

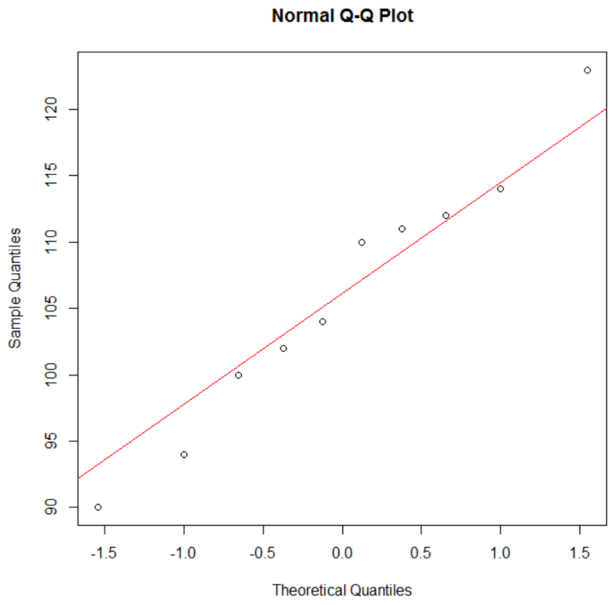
Grupa Thyroxine



Sa pripadnim rezultatima

- D = 0.1891
- p-vrijednost = 0.6303

Grupa Control

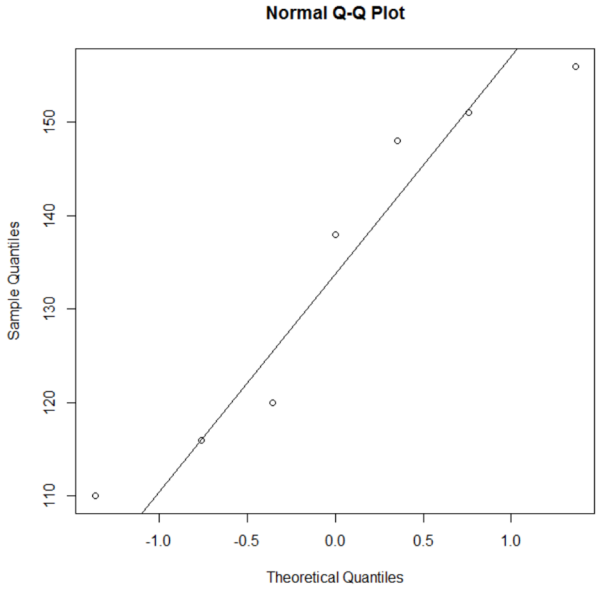


Sa pripadnim rezultatima

- D = 0.1566
- p-vrijednost = 0.6955

Tjedan 3

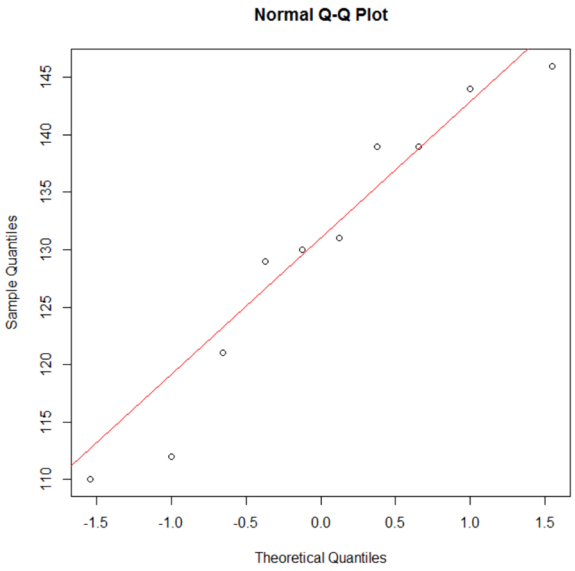
Grupa Thyroxine



Sa pripadnim rezultatima

- D = 0.2048
- p-vrijednost = 0.5021

Grupa Control

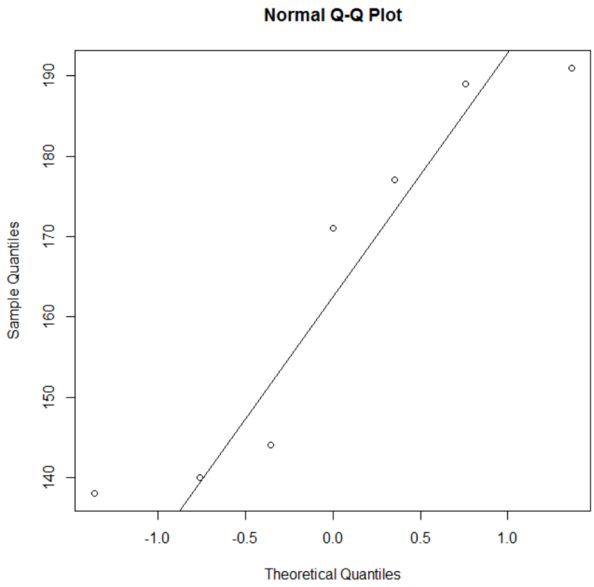


Sa pripadnim rezultatima

- D = 0.2048
- p-vrijednost = 0.5021

Tjedan 4

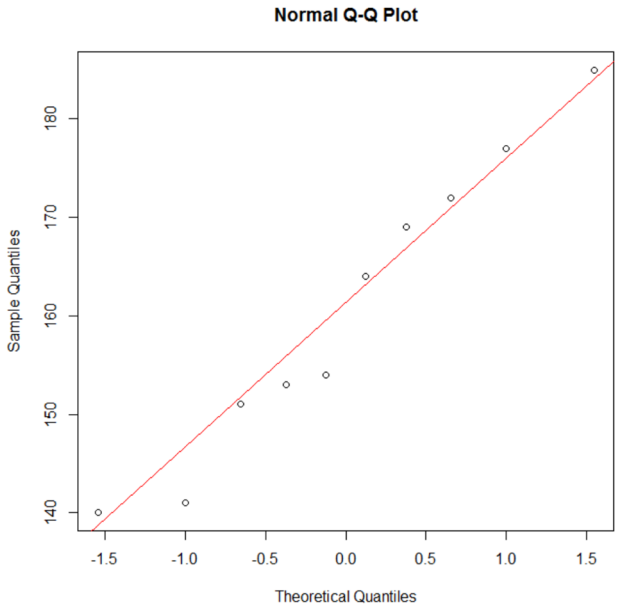
Grupa Thyroxine



Sa pripadnim rezultatima

- D = 0.2378
- p-vrijednost = 0.2696

Grupa Control



Sa pripadnim rezultatima

- D = 0.168
- p-vrijednost = 0.5872

Komentar na rezultate:

Sve p-vrijednosti su velike (> 0.05, ovdje i >0.1), što nam ukazuje da ne da ne možemo odbaciti hipotezu H_o Lillieforsovog testa normalnosti da su podaci normalno distribuirani.

Dakle, podaci su normalno distribuirani, što bi iz ovakvih grafova lako interpretirali po činjenici da točke (podaci) na grafu imaju vrlo mala odstupanja od crnih (crvenih) linija koje opisuju normalnu razdiobu.

3.2 Provedba testa opisanog u (a) dijelu zadatka

Računamo očekivane tjedne vrijednosti težina štakora ovisno o pripadnosti grupama Thyroxine i Control, te spremamo podatke u x i y varijable u R-u.

```
#x = thyroxine
#y = control
x_p<-c(mean(thyroxine[1,]),mean(thyroxine[,]),mean(thyroxine[3,]),
mean(thyroxine[4,]),mean(thyroxine[5,]))
x_p
y_p<-c(mean(control[1,]),mean(control[,]),mean(control[3,]),
mean(control[4,]),mean(control[5,]))
y_p
```

Računamo vrijednosti statistika S_X , S_Y i S

```
Sx <- matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
nrow=5,ncol=5,byrow=TRUE)
Sy <- matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
nrow=5,ncol=5,byrow=TRUE)
n <- 7
m <- 10
p <- 5
for(i in 1:7)
Sx <- Sx+(t(t(thyroxine[i,]-x_p)))%*%(t(t(t(thyroxine[i,]-x_p))))
Sx <- Sx/(n-1)
for(i in 1:m)
Sy <- Sy+(t(t(control[i,]-y_p)))%*%(t(t(t(control[i,]-y_p))))
Sy <- Sy/(m-1)
S <- 1/(m+n-2)*((n-1)*Sx+(m-1)*Sy)
```

Raučunamo i ispisujemo vrijednosti T^2 , Λ i p-vrijednost.

```
Tkvadrat <- (1/m+1/n)^(-1)*t(t(t((x_p-y_p))))%*%solve(S)%*%(t(t(x_p-y_p)))
Tkvadrat = 22.46916
lambda<-( 1+( 1/(m+n-2)*Tkvadrat ) )^( -(m+n)/2 )
lambda = 0.0004173945
fil=(lambda^(-2/(m+n))-1)*(m+n-2)
ts=(11/(15*5))*(fil)
pvrijednost <- 1- pf(ts,5,11)
pvrijednost = 0.04617933
```

Prema dobivenoj p – vrijednosti (to jest činjenici da p<0.05) na nivou značajnosti od 5% mogu odbaciti početnu hipotezu H_0 u korist alternativne hipoteze H_1.

Dakle, slijedi da su očekivane vrijednosti težina štakora za grupu Thyroxine i grupu Control jednake.

4. Izvori

1. *R. Christensen, Advanced Linear Modeling, 2nd edition, Springer-Verlag, 2001.*
2. *M. Bilodeau, D. Brenner, Theory of Multivariate Statistics, Springer-Verlag, 1999.*
3. *https://web.math.pmf.unizg.hr/nastava/ps/*
4. *https://web.math.pmf.unizg.hr/nastava/stat/*