UNIVERZITET U BEOGRADU

MATEMATIČKI FAKULTET

SEMINARSKI RAD IZ STATISTIČKOG SOFTVERA 4 NA TEMU:

ISHRANA MLADIH

Asistent:

Bojana Todić

Studenti:

Jelena Cicvarić 158/2012

Tijana Molerović 217/2012

Jovana Protić 162/2012

Beograd, maj 2016. godine

Sadržaj:

1. UVOD

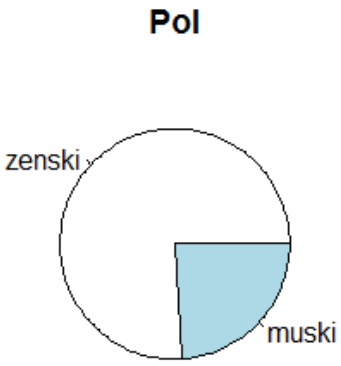
Ispitujemo bazu *Domaci* koja predstavlja istraživanje kako fizička aktivnost utiče na način ishrane, gojaznost, kao i količinu vode koju unosimo u organizam. Sproveli smo anketu nad 168 ljudi, uzrasta između 17 i 30 godina, i te rezultate smo dalje ispitivali.

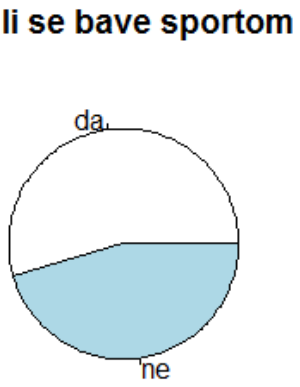
Faktori baze su sledeći:

- **Pol**
- **Godine**
- **Visina**
- **Težina**
- **Fizička aktivnost-** da li se bave fizičkom aktivnosti ili ne
- **Sport-** tip fizičke aktivnost ukoliko se bave njom
- **Minuti sporta-** koliko dnevno vremena provode vežbajući
- **Ishrana-**da li se zdravo hrane
- **Kalorije-** koliki je dnevni unos kalorija
- **Voda-**koliki je dnevni unos vode

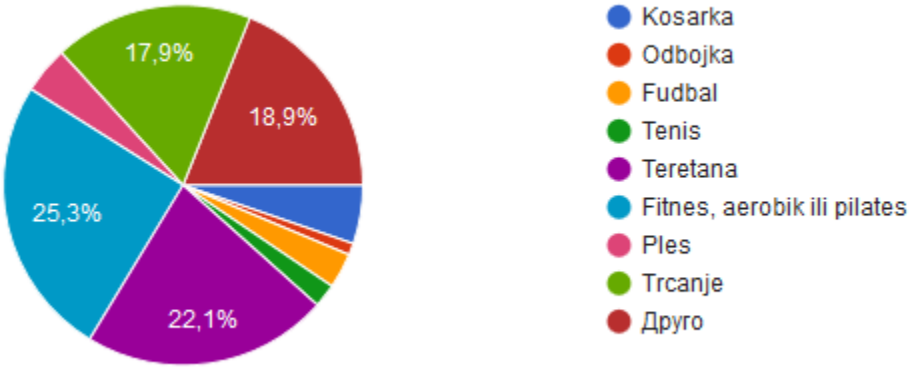
Prvo ćemo se malo bolje upoznati sa našom bazom. Možemo grafički prikazati naše promenljive.

Na primer, prikazaćemo grafički odnos muškaraca i žena koji su učestvovali u anketi, kao i da li se bave sportom ili ne. U daljem radu sa bazom, označavaćamo sa 0 muškarce, sa 1 žene. Takođe, 0 ukoliko se ne bave sportom, a 1 ukoliko se bave sportom.

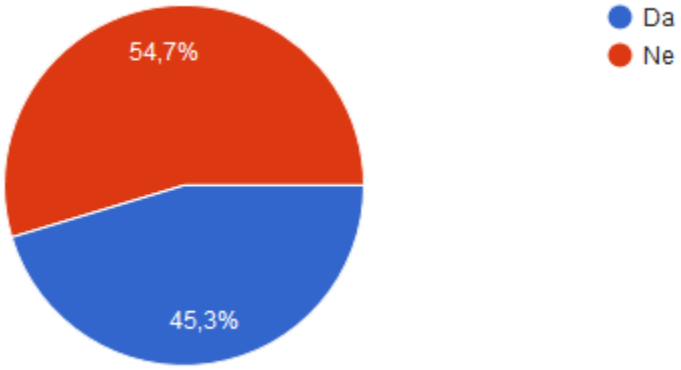




Grafički prikaz vrsta sporta kojim se bave:



Da li se zdravo hrane ili ne:



2. CILJ ISTRAŽIVANJA

U našem istraživanju odgovorićemo na nekoliko pitanja:

1. Koliki je BMI svih učesnika ankete
2. Uticaj bavljenja sporta na težinu ispitanika
3. Zavisnost težine od visine, starosti i bavljenja sportom
4. Da li se ljudi zdravo hrane u odnosu na unos vode i hrane

2.1. BMI UČESNIKA ANKETE

Indeks telesne mase računa se po sledećoj formuli:

Težina(kg):Visina(m)^2=BMI

Računanjem BMI-a možemo odrediti stepen neuhranjenosti naših ispitanika. Odnos je dat sledećom tablicom:

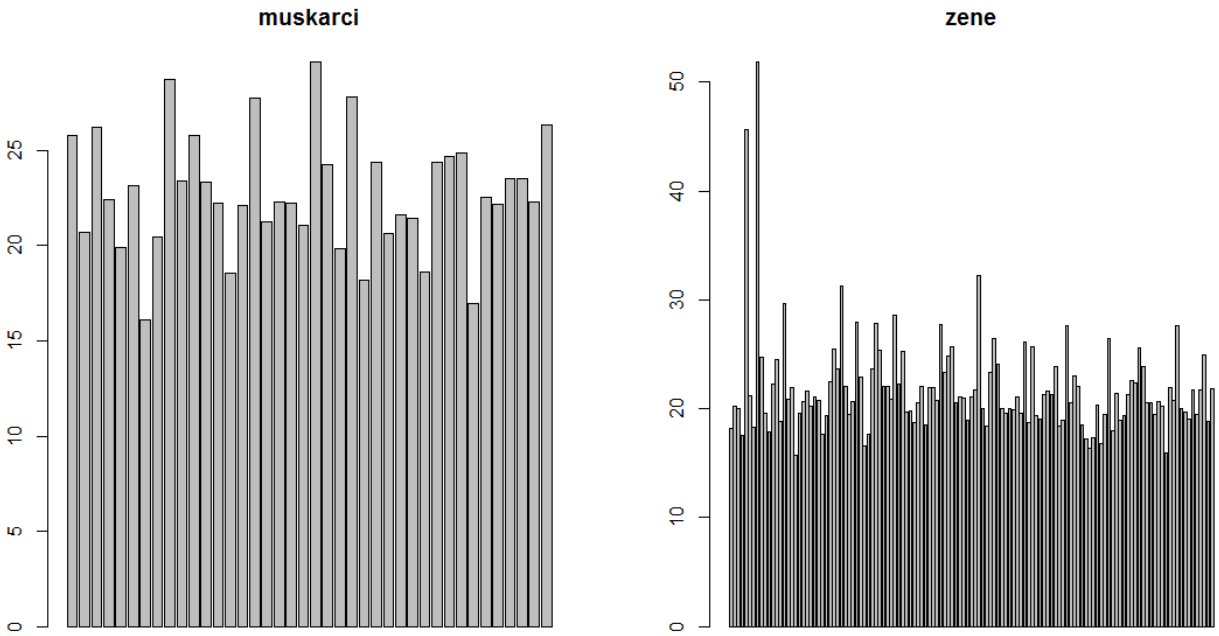
Muškarci	Žene	
< 20.7	19.1	BMI prenizak
20.7 - 26.4	19.1 - 25.8	BMI idealan
26.5 - 27.8	25.9 - 27.3	BMI malo iznad normale
27.9 - 31.1	27.4 - 32.2	BMI visok
31.2 - 45.4	32.3 - 44.8	BMI previsok
> 45.4	> 44.8	BMI izrazito visok

Sada računamo indeks telesne mase naših ispitanika:

Kod u R-u:

```
domaci.pol<-split(domaci,domaci$Pol,drop=FALSE)
domaci.Fizicka.aktivnost<-split(domaci,
domaci$Fizicka.aktivnost,drop=FALSE)
domaci.voda<-split(domaci,domaci$Voda,drop=FALSE)
par(mfrow=c(1,2))
pie(c(nrow(domaci.pol$`0`),nrow(domaci.pol$`1`)),
labels=c("muski","zenski"), main="Pol")
pie(c(nrow(domaci.Fizicka.aktivnost$`0`),nrow(domaci.Fizicka.aktivnost
$`1`)),labels = c("ne","da"),main="Da li se bave sportom")
bmi_muskarci<-domaci.pol$`0`$Tezina/(domaci.pol$`0`$Visina/100)^2
bmi_muskarci
bmi_zene<-domaci.pol$`1`$Tezina/(domaci.pol$`1`$Visina/100)^2
bmi_zene
par(mfrow=c(1,2))
barplot(bmi_muskarci)
barplot(bmi_zene)
```

Pri čemu dobijamo sledeće barplotove:



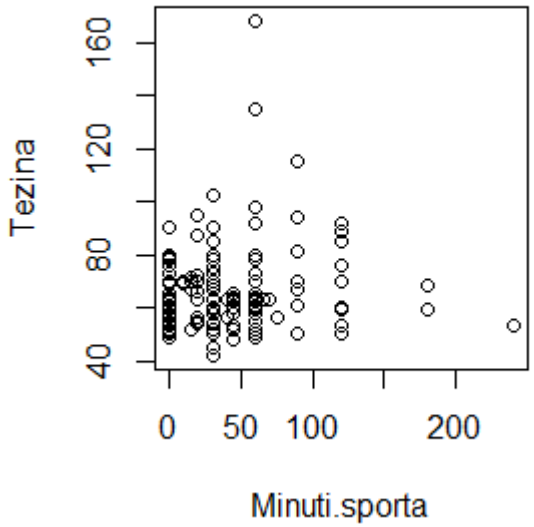
Kod muškaraca možemo zaključiti da većina ima idealan BMI, dok mali broj ispitanika ima prenizak BMI, kao i da u anketi nije učestvovao niko sa izrazito visokim BMI-om.

Kod žena vidimo da postoje dve osobe sa izrazito visokim BMI-om među ispitanicima, kao i da ima veći broj žena sa preniskim BMI-om.

2.2. UTICAJ BAVLJENJA SPORTA NA TEŽINU ISPITANIKA

Da bismo ispitali kako bavljenje sportom utiče na težinu ispitanika koristićemo prostu linearnu regresiju. Dakle, imamo jednu zavisnu promenljivu što je Težina, i jednu nezavisnu promenljivu što je dužina bavljenja sportom.

Prvo ćemo nacrtati dijagram raspršivanja:



Kako program R ima ugrađene funkcije koje se koriste za regresioni linearni model, na sledeći način možemo izračunati parametre:

Kod u R-u:

x<-lm(Tezina~Minuti.sporta)

summary(x)

Dobijamo:

```
Call:
lm(formula = Tezina ~ Minuti.sporta)

Residuals:
    Min      1Q  Median      3Q     Max
-24.131  -8.884  -4.378   5.557 100.622

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.88390    1.63658  39.646   <2e-16 ***
Minuti.sporta  0.04158    0.02980   1.395    0.165
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.55 on 165 degrees of freedom
Multiple R-squared:  0.01166,   Adjusted R-squared:  0.005672
F-statistic: 1.947 on 1 and 165 DF,  p-value: 0.1648
```

Odavde vidimo da je: a=64.88390, b=0.04158.

Kako je formula za prost linearni model: $Y=a+bX+\varepsilon$, jednačina regresije je oblika:

$$Y=64.88390+0.04158*x.$$

Takođe, parametre a i b možemo i ručno izračunati pomoću formula:

$$\widehat{a} = \overline{\mathbf{Y}} - \widehat{b\overline{\mathbf{X}}}$$

$$\widehat{b} = \frac{\operatorname{cov}(\mathbf{X}, \mathbf{Y})}{S^2_{n\mathbf{X}}}$$

.

Pomoću funkcije *cor.test()* dobijamo vrednost test statistike i *p* vrednost za testiranje nulte hipoteze da ne postoji linearna korelacija, interval poverenja i Pirsonov koeficijent:

```
> cor.test(Minuti.sporta,Tezina)

        Pearson's product-moment correlation

data:  Minuti.sporta and Tezina
t = 1.3953, df = 165, p-value = 0.1648
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04460396  0.25566189
sample estimates:
      cor
0.1079911
```

Kako je vrednost *p* statistike 0.1648, prihvatamo nultu hipotezu, tj da ne postoji linearna korelacija između težine i vremena bavljenja sportom.

2.3. ZAVISNOST TEŽINE OD VISINE, STAROSTI I BAVLJENJA SPORTOM

Kako ovde imamo zavisnost jedne pojave, tj. težine od više faktora, tj. od visine, starosti i bavljenja sportom koristićemo višestruku linearnu regresiju.

Računamo paramatre višestruke linearne regresije:

```
> y<-lm(Tezina~Minuti.sporta+Visina+Godine)
> summary(y)

Call:
lm(formula = Tezina ~ Minuti.sporta + Visina + Godine)

Residuals:
    Min      1Q  Median      3Q     Max
-21.912  -7.163  -1.865   4.746  92.902

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -117.98676   21.67058  -5.445 1.88e-07 ***
Minuti.sporta    0.01442    0.02522   0.572  0.56831
Visina           0.89902    0.11721   7.670 1.48e-12 ***
Godine           1.32160    0.46815   2.823  0.00535 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.04 on 163 degrees of freedom
Multiple R-squared:  0.3134,    Adjusted R-squared:  0.3008
F-statistic:  24.8 on 3 and 163 DF,  p-value: 2.846e-13
```

Dobijamo koeficijente: a=-117.98676, $b_1 = 0.01442, b_2 = 0.89902, b_3 = 1.32160.$

Regresiona jednačina je:

$$Y=-117.98676+0.01442X_1+0.89902X_2+1.32160X_3$$

Vidimo i da je vrednost testa manja od 0.05 što znači da odbacujemo nultu hipotezu, odnosno tezina zavisi od visine,godina i minuta bavljenja sportom . Koeficijent determinacije definisanog modela je

0.3134 sto znaci da 31.34% promenljiva Tezina objasnjena promenljivama Visina,Godine i Minuti.sporta.

Takođe možemo vršiti predviđanje na osnovu višestruke linearne regresije:

```
> newdata <- data.frame(Minuti.sporta=60, Visina=180, Godine=18, Pol=1)
> predict(lm(Tezina ~ Minuti.sporta + Visina + Godine), newdata)
       1
68.49047
```

Predvideli smo da žena stara 18 godina, visine 180cm i koja se bavi 60min sportom treba da ima oko 68kg.

2.4. DA LI SE LJUDI ZDRAVO HRANE U ODNOSU NA UNOS VODE I HRANE

Na osnovu rezultata naše ankete napravićemo model linearne regresije koji može da predvidi da li se ispitanici zdravo hrane.

Baza sadrži 168 opservacija i razmatramo 3 promenljive: ishranu, kalorije i vodu.

```
> model<-glm(Ishrana~Kalorije + Voda,family = binomial)
> summary(model)

Call:
glm(formula = Ishrana ~ Kalorije + Voda, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8378  -1.0631  -0.8502   1.1911   1.6673

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.6355592  0.5832503  -1.090  0.27585
Kalorije    -0.0002040  0.0002372  -0.860  0.38974
Voda         0.0005056  0.0001751   2.887  0.00389 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 230.50  on 166  degrees of freedom
Residual deviance: 220.59  on 164  degrees of freedom
AIC: 226.59

Number of Fisher Scoring iterations: 4
```

Koeficijenti modela:

```
> coef(model)
  (Intercept)      Kalorije          Voda
-0.6355591584 -0.0002039845  0.0005056195
```

Intervali poverenja za parametre:

```
> confint(model)
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept) -1.7989210541 0.5030156107
Kalorije    -0.0006784403 0.0002590112
Voda         0.0001795120 0.0008686442
```

Kako su p vrednosti Valodovih testova velike zakljucujemo da se hipoteza da je neki od od parametra jednak nuli prihvata, tj. nezdravo se hrane.

Sada ćemo izvršiti predviđanje na osnovu dobijenog modela. Proveravamo da li se ispitanik zdravo hrani ako unosi 2000 kalorija i 3000ml vode dnevno.

Kod u R-u:

p.X <- predict(model)

y <- rep(0, length(Ishrana))

y[p.X>0.5] <- 1

newdata <- data.frame(Kalorije=2000, Voda=3000)

predict(model,newdata)

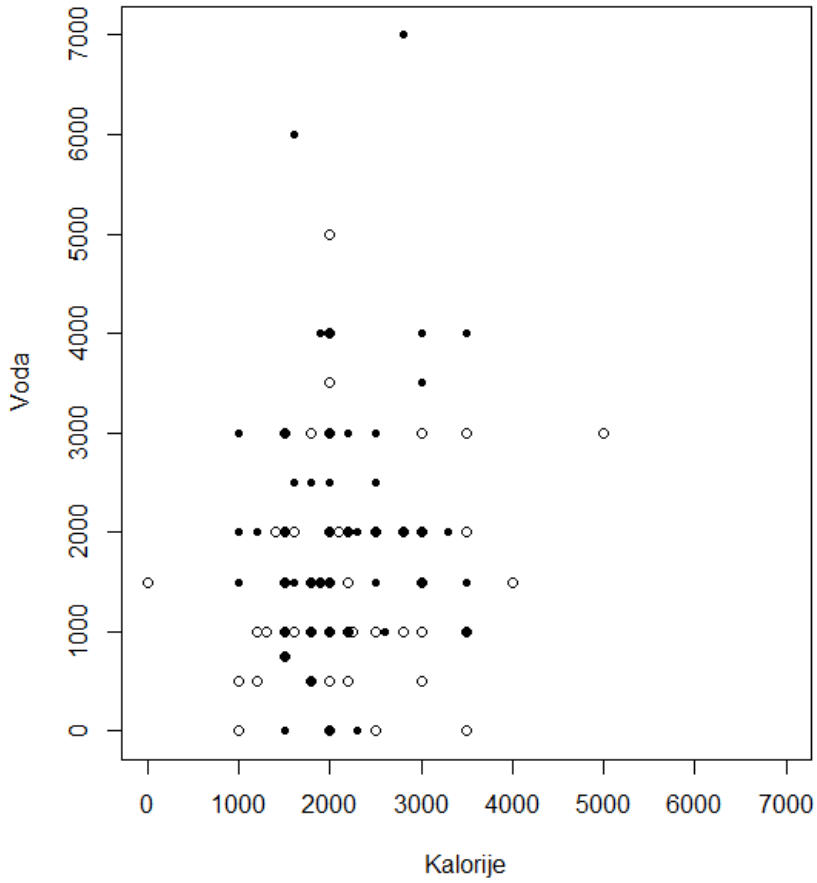
plot(Kalorije[Ishrana==0], Voda[Ishrana==0], xlab = „Kalorije“, ylab = „Voda“, xlim= c(0,7000), ylim=c(0,7000))

points(Kalorije[Ishrana==1], Voda[Ishrana==1], pch = 20)

Dobijamo:

```
> predict(model,newdata)
        1
0.4733303
```

Ovde mozemo videti da se konkretno ovaj ispitanik zdravo hrani.



3.1. TEST NORMALNOSTI(ŠAPIRO-VILK)

Ispitaćemo da li promenljive imaju normalnu raspodelu. Normalnost možemo videti odmah sa grafika, ali da bismo bili sigurniji pozvaćemo Šapiro-Vilkov test:

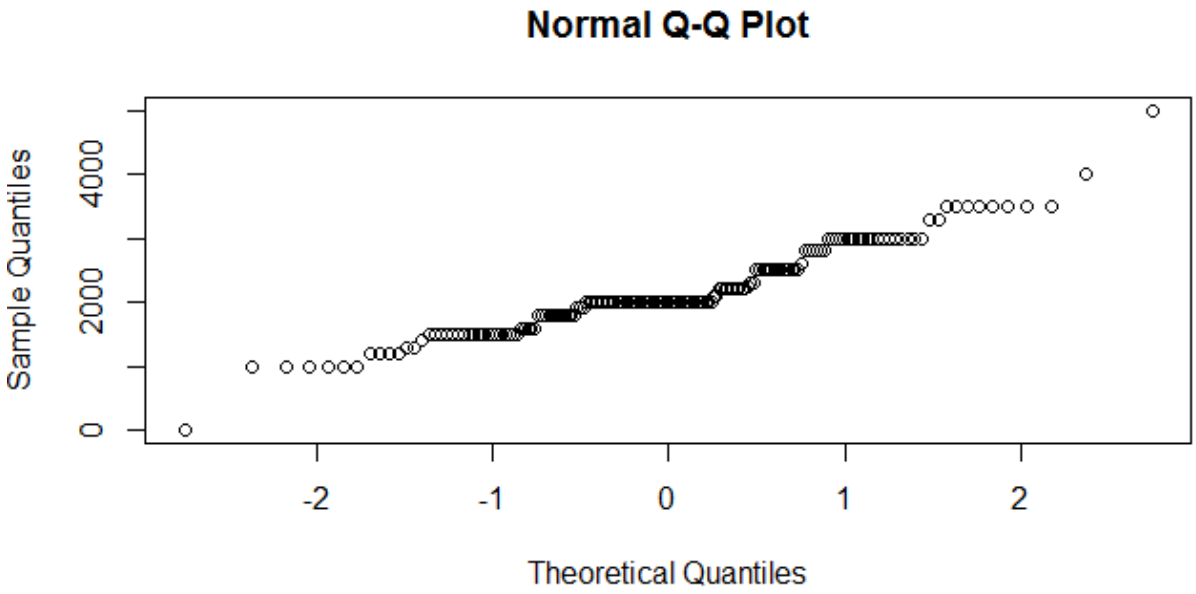
```
> shapiro.test(Kalorije)

        Shapiro-Wilk normality test

data:  Kalorije
W = 0.9434, p-value = 3.289e-06
```

Kako je vrednost p statistike mala, odbacujemo nultu hipotezu.

Kada pozovemo qqnorm() dobijamo:



I sa grafika vidimo da nije normalno raspodeljena.

Proverićemo i za Težinu:

```
> shapiro.test(Tezina)

        Shapiro-Wilk normality test

data:  Tezina
W = 0.8039, p-value = 1.064e-13
```

Takođe vidimo da je p-vrednost mala što znači da odbacujemo nultu hipotezu, tj. nije normalno raspodeljena.

3.2. TEST NEZAVISNOSTI(MANN-WHITNEY-WILCOXON)

Kako smo videli da promenljive nisu normalno raspodeljene, sada ćemo ispitati nezavisnost naših promenljivih. Nulta hipoteza će nam biti da jesu nezavisne, dok alternativna da nisu.

```
> wilcox.test(Kalorije~Pol, data=domaci)

        Wilcoxon rank sum test with continuity correction

data:  Kalorije by Pol
W = 3470, p-value = 0.0004112
alternative hypothesis: true location shift is not equal to 0
```

Kako je p vrednost mala, odbacujemo nultu hipotezu, što znači da su nam kalorije i pol zavisne promenljive.

```
> wilcox.test(Visina~Ishrana, data=domaci)

        Wilcoxon rank sum test with continuity correction

data:  Visina by Ishrana
W = 3570.5, p-value = 0.7357
alternative hypothesis: true location shift is not equal to 0
```

Odavde vidimo da je p-vrednost velika, što znači da prihvatamo nultu hipotezu, tj. Visina i Ishrana su nezavisne promenljive.

Možemo test primenjivati i na ostale promenljive, s tim što je bitno da bude jedna numerička i jedna kategorijska promenljiva.

3.3. TEST SLUČAJNOSTI(TEST TAČAKA ZAOKRETA)

Pomoću testa tačaka zaokreta ispitaćemo da li su naše numeričke promenljive slučajne. Nulta hipoteza će nam biti da je niz brojeva slučajan, dok je alternativna hipoteza da nije slučajan.

Test tačaka zaokreta je već definisan u R-u, što znači da možemo samo pozvati funkciju turning.point.test().

Kod u R-u:

```
> library(randtests)
> turning.point.test(Kalorije)

        Turning Point Test

data:  Kalorije
statistic = 0.71925, n = 148, p-value = 0.472
alternative hypothesis: non randomness
```

Odakle vidimo da je p-vrednost testa velika, što znači da prihvatamo nultu hipotezu.

```
> turning.point.test(Voda)

        Turning Point Test

data:  Voda
statistic = 1.9181, n = 135, p-value = 0.0551
alternative hypothesis: non randomness
```

I u ovom slučaju prihvatamo nultu hipotezu.

Test tačaka zaokreta može se i ručno pozvati na sledeći način:

Kod u R-u:

```
testzaokreta<-function(x,alfa)
{
n=length(x)
s=0
for(i in 1:(n-2))
{
a=(x[i+1]-x[i])*(x[i+2]-x[i+1])
if(a<0) s=s+1
}
```

m=(2(n-2))/3*
*dz=(16*n-29)/90*
cat("srednja vrednost :",m," uzoracka disperzija: ", dz,"\n")

print("vrednost test statistike je:")
print(s)
print(" standardizovana vrednost test statistike je:")
print((s-m)/sqrt(dz))
c=qnorm(p=(1-(alfa/2)))
*cat((1-alfa)*100,"%-tni interval poverenja stand stat je (",-c," ,",c,")")*
}

Na primer, računaćemo za 95% interval poverenja pri čemu dobijamo:

```
> testzaokreta(domaci$Kalorije, 0.05)
srednja vrednost : 110  uzoracka disperzija:  29.36667
[1] "vrednost test statistike je:"
[1] 89
[1] " standardizovana vrednost test statistike je:"
[1] -3.875181
95 %-tni interval poverenja stand stat je ( -1.959964  , 1.959964 )
```

4. ZAKLJUČAK

Na početku, videli smo da muškarci u većini slučajeva imaju idealan BMI, dok žene uglavnom imaju prenizak BMI, a u nekim slučajevima I previsok.

Primetili smo i da, bavljenje sportom ne utiče na težinu naših ispitanika. Međutim, kada smo ispitivali zavisnost težine u odnosu na starost, visinu i bavljenje sportom, primetili smo da težina zavisi od ovih promenljivih kod naših ispitanika.

Predviđanjem za osobu koja unosi 2000 kalorija i 3000ml vode dnevno, dobili smo da se ta osoba zdravo hrani.

Testiranjem smo dobili da promenljive nisu iz normalne raspodele. Pomoću testa Mann-Whitney-Wilcoxon videli smo da su kalorije i pol zavisne promenljive, dok su visina i ishrana nezavisne. Međutim, pozivanjem ovog testa na naše ostale promenljive, dobili smo da, na primer, težina i ishrana su nezavisne promenljive, što nam može doneti zaključak da naši ispitanici i nisu bili najiskreniji prilikom popunjavanja ankete.

5. LITERATURA

- http://www.matf.bg.ac.rs/p/bojana-todic/kurs/361/%D0%A1%D0%A14/
- http://www.matf.bg.ac.rs/p/files/42-VS2cas.pdf