**Zadatak:**

U bazi podataka **BMI.csv** dani su podaci o dobi, spolu, tjelesnoj visini, masi te BMI uzorka ljudi iz jedne populacije. U varijabli DOB 1, dob osoba je kategorizirana po desetljećima.

1. Analizirajte tjelesnu masu, visinu i BMI po spolu za sve dobne skupine zajedno. Da li su razlike po spolu statistički značajne?
2. Analizirajte razlike varijabli TM i BMI po spolu samo za osobe u dvadesetim godinama i provjerite da li su statistički značajne.
3. Provjerite da li distribucija spola po godištima odgovara zahtjevu da su muškarci i žene ravnomjerno zastupljeni u uzorku iz svakog desetljeća.

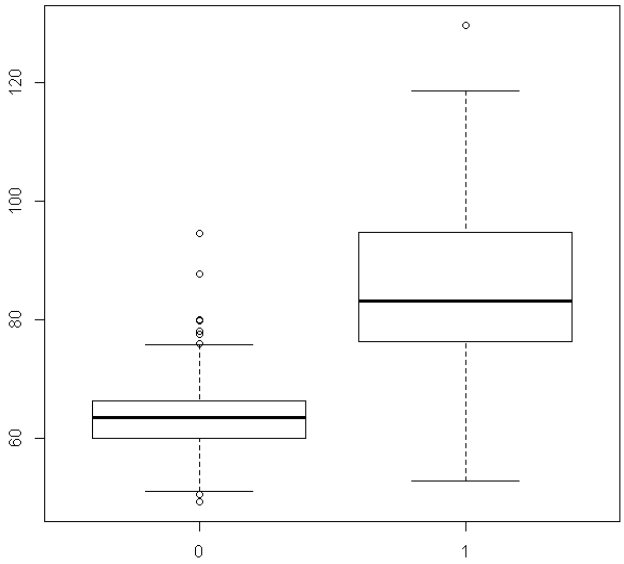
Rjesenja

**1. dio**

```
> table(baza1$SPOL)

 0  1
57 43
> prop.table( table(baza1$SPOL))

   0    1
0.57 0.43

> summary(baza1[baza1$SPOL==0,])
     SPOL     DOB.1        DOB.2            TM              TV            BMI
 Min.   :0   20-29:11   Min.   :20.00   Min.   :49.30   Min.   :142.0   Min.   :17.47
 1st Qu.:0   30-39:10   1st Qu.:31.00   1st Qu.:60.10   1st Qu.:160.0   1st Qu.:21.49
 Median :0   40-49:13   Median :45.00   Median :63.60   Median :164.0   Median :23.41
 Mean   :0   50-59: 8   Mean   :45.35   Mean   :64.38   Mean   :164.5   Mean   :23.84
 3rd Qu.:0   60-69:10   3rd Qu.:60.00   3rd Qu.:66.40   3rd Qu.:169.0   3rd Qu.:25.18
 Max.   :0   70-79: 5   Max.   :79.00   Max.   :94.50   Max.   :177.0   Max.   :38.83
> summary(baza1[baza1$SPOL==1,])
     SPOL     DOB.1        DOB.2            TM              TV            BMI
 Min.   :1   20-29: 9   Min.   :20.0   Min.   : 52.90   Min.   :160.0   Min.   :17.88
 1st Qu.:1   30-39:10   1st Qu.:30.0   1st Qu.: 76.40   1st Qu.:172.5   1st Qu.:23.95
 Median :1   40-49: 9   Median :42.0   Median : 83.10   Median :180.0   Median :26.08
 Mean   :1   50-59: 5   Mean   :45.4   Mean   : 85.61   Mean   :179.9   Mean   :26.35
 3rd Qu.:1   60-69: 2   3rd Qu.:58.5   3rd Qu.: 94.80   3rd Qu.:184.0   3rd Qu.:27.77
 Max.   :1   70-79: 8   Max.   :79.0   Max.   :129.70   Max.   :198.0   Max.   :34.82
>
```

Minimalan dio koji mora ući u izvještaj iz deskriptivne statistike:

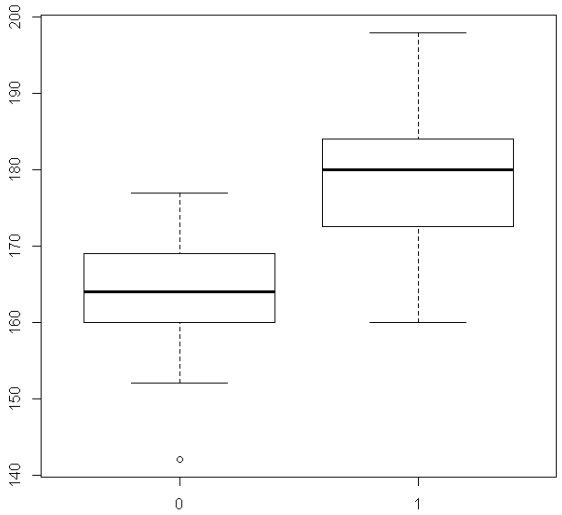
U uzorku je bilo 57 (57%) muškaraca i 43 (43%) žene. Među ženama prosječna tjelesna masa je 64.38, a među muškarcima 85.61. Kutijasti dijagrami nalaze se u nastavku.
Među ženama prosječna tjelesna visina je164.5, a među muškarcima 179.9. Kutijasti dijagrami nalaze se u nastavku.
Među ženama prosječni BMI je 23.84, a među muškarcima 26.35. Kutijasti dijagrami nalaze se u nastavku.




Dimenzija uzorka je dovoljna za provođenje neke od varijanti t-testa o jednakosti očekivanja. Da bi testirali jednakost očekivanja testiramo prvo jednakost varijanci.

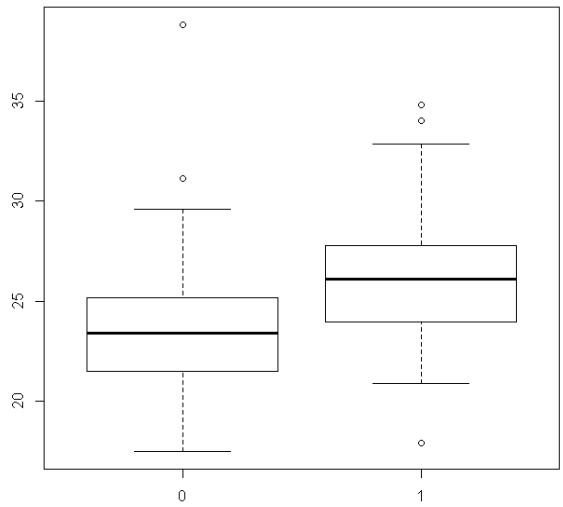
```
> var.test(baza1$TM~baza1$SPOL)
        F test to compare two variances

data:  baza1$TM by baza1$SPOL
F = 0.3383, num df = 56, denom df = 42, p-value = 0.0001728
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1885573 0.5927303
sample estimates:
ratio of variances
         0.3382883
> t.test(baza1$TM~baza1$SPOL,var.equal = FALSE)
        Welch Two Sample t-test

data:  baza1$TM by baza1$SPOL
t = -8.081, df = 63.09, p-value = 2.55e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.48531 -15.98356
sample estimates:
mean in group 0 mean in group 1
       64.37719        85.61163
```

Zaključak: Welch t-test potvrđuje postojanje statistički značajne razlike u očekivanim vrijednostima mase žena i muškaraca (p= 2.55e-11). Očekivanje mase muškaraca je veće.




```
> var.test(baza1$TV~baza1$SPOL)

        F test to compare two variances

data:  baza1$TV by baza1$SPOL
F = 0.5548, num df = 56, denom df = 42, p-value = 0.03958
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3092486 0.9721235
sample estimates:
ratio of variances
         0.5548189

> t.test(baza1$TV~baza1$SPOL,var.equal = FALSE)

        Welch Two Sample t-test

data:  baza1$TV by baza1$SPOL
t = -9.7976, df = 74.701, p-value = 4.783e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.53662 -12.27203
sample estimates:
mean in group 0 mean in group 1
       164.4561        179.8605
```

Zaključak: Welch t-test potvrđuje postojanje statistički značajne razlike u očekivanim vrijednostima visine žena i muškaraca (p= 4.783e-15). Očekivanje visine muškaraca je veće.




```
> var.test(baza1$BMI~baza1$SPOL)

        F test to compare two variances

data:  baza1$BMI by baza1$SPOL
F = 0.9709, num df = 56, denom df = 42, p-value = 0.9079
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5411588 1.7011337
sample estimates:
ratio of variances
          0.970886

> t.test(baza1$BMI~baza1$SPOL,var.equal = T)

        Two Sample t-test

data:  baza1$BMI by baza1$SPOL
t = -3.5125, df = 98, p-value = 0.0006732
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.930031 -1.092441
sample estimates:
mean in group 0 mean in group 1
       23.83737        26.34860
```

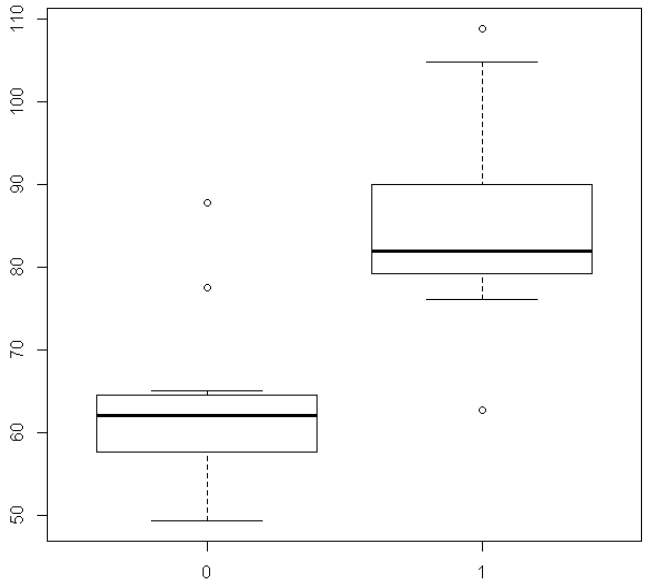
Zaključak: Standardni t-test potvrđuje postojanje statistički značajne razlike u očekivanim vrijednostima BMI žena i muškaraca (p=0.0006732). Očekivanje BMI muškaraca je veće.

**2. dio**

```
> table(dvadesete$SPOL)

 0  1
11  9
> prop.table(table(dvadesete$SPOL))

   0    1
0.55 0.45

> summary(dvadesete[dvadesete$SPOL==0,])
     SPOL     DOB.1       DOB.2            TM              TV            BMI
 Min.   :0   20-29:11   Min.   :20.00   Min.   :49.30   Min.   :160.0   Min.   :19.26
 1st Qu.:0   30-39: 0   1st Qu.:22.50   1st Qu.:57.60   1st Qu.:162.0   1st Qu.:20.71
 Median :0   40-49: 0   Median :25.00   Median :62.00   Median :167.0   Median :22.22
 Mean   :0   50-59: 0   Mean   :24.45   Mean   :63.55   Mean   :166.4   Mean   :22.78
 3rd Qu.:0   60-69: 0   3rd Qu.:26.00   3rd Qu.:64.55   3rd Qu.:169.0   3rd Qu.:22.71
 Max.   :0   70-79: 0   Max.   :29.00   Max.   :87.80   Max.   :177.0   Max.   :31.11
> summary(dvadesete[dvadesete$SPOL==1,])
     SPOL     DOB.1       DOB.2            TM              TV            BMI
 Min.   :1   20-29:9   Min.   :20.00   Min.   : 62.70   Min.   :169.0   Min.   :21.95
 1st Qu.:1   30-39:0   1st Qu.:22.00   1st Qu.: 79.20   1st Qu.:178.0   1st Qu.:23.88
 Median :1   40-49:0   Median :27.00   Median : 82.00   Median :183.0   Median :24.44
 Mean   :1   50-59:0   Mean   :25.22   Mean   : 85.24   Mean   :184.1   Mean   :24.98
 3rd Qu.:1   60-69:0   3rd Qu.:28.00   3rd Qu.: 90.00   3rd Qu.:189.0   3rd Qu.:26.73
 Max.   :1   70-79:0   Max.   :29.00   Max.   :108.90   Max.   :198.0   Max.   :28.64
```

Minimalan dio koji mora ući u izvještaj iz deskriptivne statistike:

U uzorku je bilo 11 (55%) muškaraca i 9 (45%) žene. Među ženama prosječna tjelesna masa je 63.55 , a među muškarcima 85.24. Kutijasti dijagrami nalaze se u nastavku.
Među ženama prosječni BMI je 22.78, a među muškarcima 24.98. Kutijasti dijagrami nalaze se u nastavku.



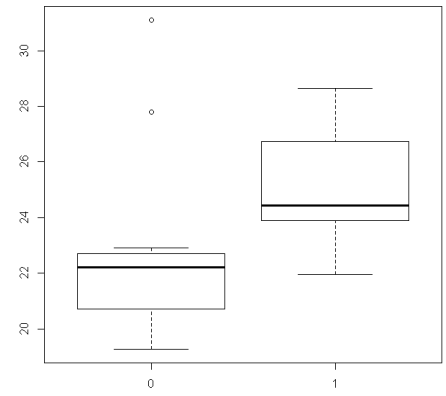

Dimenzija uzorka nije dovoljna za provođenje neke od varijanti t-testa o jednakosti očekivanja. Provodimo egzaktni Wicoxonov test.

> wilcox.test(dvadesete$TM~dvadesete$SPOL)

Wilcoxon rank sum test

data: dvadesete$TM by dvadesete$SPOL
W = 11, p-value = 0.002286
alternative hypothesis: true location shift is not equal to 0

Zaključak:
Egzaktni Wicoxonov test sume rangova potvrđuje postojanje razlike u masi među mučkarcima i ženama u dvadesetim godinama života (p=0.002286). Muškarci u dvadesetim godinama života u pravilu imaju veću masu od žena.




Wilcoxon rank sum test

data: dvadesete$BMI by dvadesete$SPOL
W = 24, p-value = 0.0281
alternative hypothesis: true location shift is less than 0
Zaključak:
Egzaktni Wicoxonov test potvrđuje postojanje razlike u BMI među mučkarcima i ženama u dvadesetim godinama života (p=0.0281). Muškarci u dvadesetim godinama života u pravilu imaju veći BMI nego žene.

**3. dio**

```
> (t1<-table(baza1$SPOL,baza1$DOB.1)  )

   20-29 30-39 40-49 50-59 60-69 70-79
 0    11    10    13     8    10     5
 1     9    10     9     5     2     8

> prop.table(t1,2)

       20-29     30-39     40-49     50-59     60-69     70-79
 0 0.5500000 0.5000000 0.5909091 0.6153846 0.8333333 0.3846154
 1 0.4500000 0.5000000 0.4090909 0.3846154 0.1666667 0.6153846
```

Zaključak: Ako analziramo zastupljenost spola po godištima vidimo da su razlike najizraženije u skupini 60-69 godina. Treba testirati da li su te razlike statistički značajne.

To ste mogli najednostavnije tako da za svaku dobnu skupinu napravite binomni test posebno:

```
> pvrBin<-rep(0,6)

> for (i in 1:6) {
+ pvrBin[i]<-binom.test(t1[1,i],sum(t1[,i]),p=0.5)$p.value
+ }

> pvrBin
[1] 0.82380295 1.00000000 0.52346706 0.58105469 0.03857422 0.58105469
```

Odavde se vidi da su razlike značanje u dobnoj skupini 60-69 godina (p=0.038 57422).

Alternativa je hi-kvadrat test ali ne za svaku skupinu posebno jer onda nema dovoljno podataka nego zajednički. To nije baš to što treba ali bi isto priznala:

```
> ocekivane<-matrix(0,2,6)
> for (i in 1:2) {
+ for (j in 1:6) ocekivane[i,j]<-sp[i]*dob[j]*100}
> ocekivane
     [,1] [,2]  [,3] [,4] [,5] [,6]
[1,] 11.4 11.4 12.54 7.41 6.84 7.41
[2,]  8.6  8.6  9.46 5.59 5.16 5.59

> chisq.test(t1)

        Pearson's Chi-squared test

data:  t1
X-squared = 5.7989, df = 5, p-value = 0.3263
```

Međutim, ovaj test ne potvrđuje postojanje razlika. U pravoj analizi bilo bi bolje koristiti rezultat binomnog testa.