Življenjska doba avtomobilskih gum

1. Opis podatkov in raziskovalne domneve

Želimo preveriti trditev, da nova metoda izdelave avtomobilskih gum vpliva na njihovo življenjsko dobo.

Zamislili smo si, da bomo izmerili življenjsko dobo avtomobilskih gum, narejenih po novi metodi, in življenjsko dobo avtomobilskih gum, narejenih po stari metodi (kontrolna skupina).

Zbrali smo vzorec 25 avtomobilskih gum, narejenih po novi metodi in 25 avtomobilskih gum, narejenih po stari metodi. Podatke smo zapisali v dokument, ki ima dva stolpca.

1. *metoda* je nominalna spremenljivka (faktor), ki ima vrednosti "1" in "2", ki opisuje pripadnost skupini.
2. *zdoba* je numerična zvezna spremenljivka, ki predstavlja življenjsko dobo avtomobilskih gum, merjeno v 1000 kilometrih.

Baza podatkov se imenuje *avgume.csv*. Najprej bomo prebrali podatke v R s pomočjo funkcije *read.csv*, in zatem pogledali strukturo podatkov.

```
zivdoba<-read.csv("G:\\Documents\\Temp\\avgume.csv", header=TRUE)
str(zivdoba)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ metoda: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ zdoba : num  61.3 64 61.3 62.4 60.2 ...
```

Obstaja več načinov za razdelitev podatkov na dva numerična vektorja, ki vsebujeta vrednosti življenjskih dob vsake skupine. Lahko uporabimo funkcijo *split*, s katero bomo dobili listo teh vektorjev.

```
zivdoba1<-split(zivdoba$zdoba, zivdoba$metoda)
zivdoba1
```

```
## $`1`
##  [1] 61.27 63.96 61.26 62.36 60.16 60.85 63.95 67.20 60.23 59.01 64.60 64.56
## [13] 64.53 63.80 61.51 58.45 62.91 63.25 62.48 57.63 58.63 59.11 62.35 60.56
## [25] 59.89
##
## $`2`
##  [1] 54.82 54.55 65.42 54.71 61.97 67.03 57.40 47.62 58.69 47.85 61.48 59.30
## [13] 57.06 54.70 58.85 60.62 58.53 63.52 64.76 58.09 55.66 57.51 55.34 56.53
## [25] 51.07
```

Vektor *zivdoba1$1* predstavlja življenjsko dobo avtomobilsih gum, narejenih po novi metodi in vektor *zivdoba1$2* življenjsko dobo gum, narejenih po stari metodi.

2. Grafični prikaz podatkov – histogram in škatla z brki

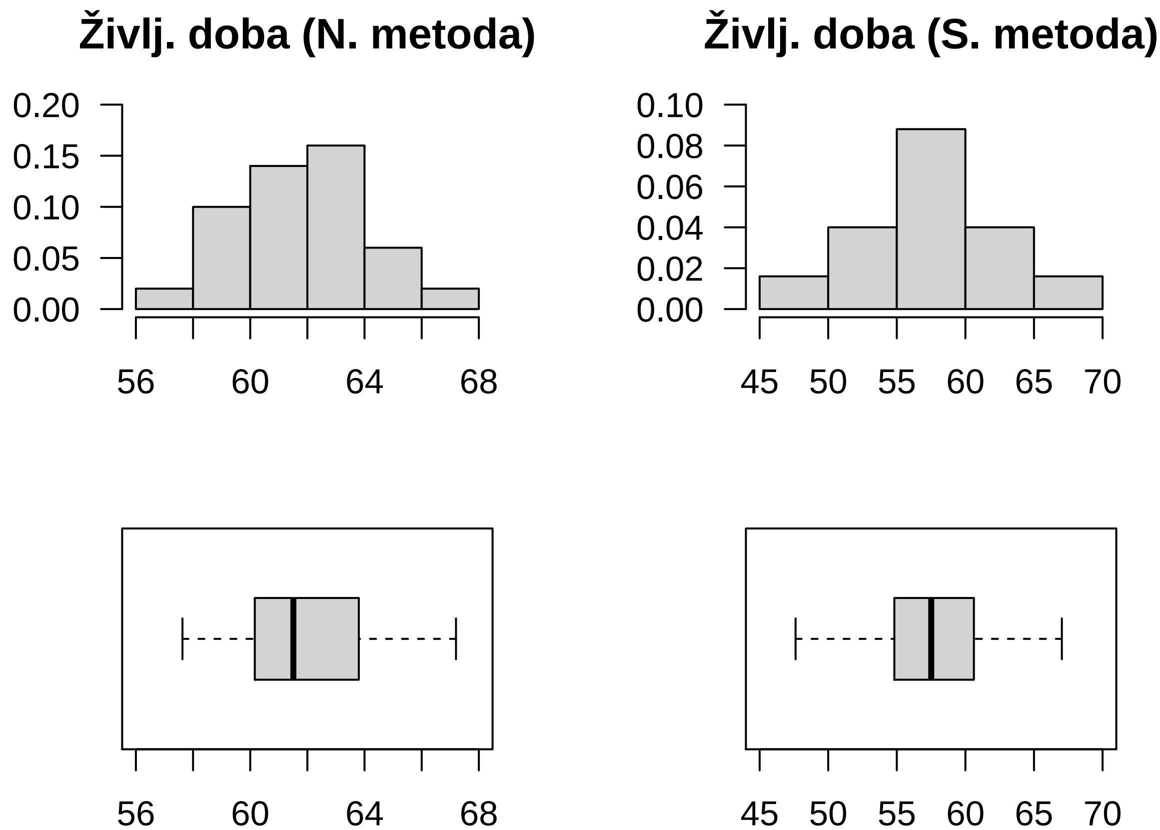
Podatke lahko grafično predstavimo s pomočjo histograma gostote (funkcija *hist*) in škatle z brki (funkcija *boxplot*).

```
par(mfrow=c(2,2),las=1,cex=1.1,mar=c(2,3,3,3))
hist(zivdoba1$`1`,freq=FALSE,ylim=c(0,0.2), main="Življ. doba (N. metoda)",xlab="",ylab="")

hist(zivdoba1$`2`,freq=FALSE,ylim=c(0,0.1),main="Življ. doba (S. metoda)",xlab="",ylab="")

boxplot(zivdoba1$`1`,horizontal=TRUE,ylim=c(56,68))

boxplot(zivdoba1$`2`,horizontal=TRUE,ylim=c(45,70))
```



Funkcija *par* omogoča postavljanje različnih nastavitev na grafih, preko velikega števila parametrov: *mfrow=c(2,2)* nastavlja risanje 4 grafov na eni sliki (grafi v dveh vrsticah in dveh stolpcih), različice parametra *cex* povečanje simbolov, teksta, osi, itn., *mar* pa prostor na robu, izražen s številom vrstic (privzeta vrednost je *mar=c(5.1,4.1,4.1,2.1)* za spodnji, levi, zgornji in desni rob grafa).

Parameter *main* služi za definiranje naziva grafa, *xlab* in *ylab* pa za nazive na x in y osi, ter *xlim* in *ylim* za postavljanje mej na oseh.

Histogram gostote smo predstavili s pomočjo parametra *freq=FALSE*. Na ta način ploščina vsakega pravokotnika histograma predstavlja relativno frekvenco razreda. Skupna ploščina histograma gostote je 1 in je primeren za primerjanje skupin in risanje pričakovane gostote porazdelitve.

Izbrali smo horizontalno škatlo z brki (privzeta vrednost je vertikalna), da bi lahko primerjali obliko porazdelitve s histogramom. Zaradi rotacije škatle z brki se *ylim* nanaša na meje na x-osi. Meje na x-oseh histograma in škatle z brki so za iste podatke enake.

3. Opisna statistika

Zdaj bomo izračunali opisno statistiko za dobljene podatke – povzetek s petimi števili (minimum, maksimum, prvi in tretji kvartil, mediano), vzorčno povprečje in popravljeni vzorčni standardni odklon. Potem pa bomo interpretirali grafe skupaj z opisno statistiko. Za prikaz opisne statistike lahko uporabimo funkciji *summary* in *sd*.

```
summary(zivdoba1$`1`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.63   60.16   61.51   61.78   63.80   67.20
```

```
sd(zivdoba1$`1`)
```

```
## [1] 2.37471
```

```
summary(zivdoba1$`2`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   47.62   54.82   57.51   57.72   60.62   67.03
```

```
sd(zivdoba1$`2`)
```

```
## [1] 4.862611
```

4. Interpretacija grafov in opisne statistike

Poglejmo opisno statistiko, histogram in škatlo z brki za življenjsko dobo avtomobilskih gum. Za oba vzorca sta vrednosti vzorčne mediane in vzorčnega povprečja blizu. Porazdelitev življenjske dobe gum, narejenih po novi metodi zgleda asimetrična, zato moramo preveriti, ali je asimetrija zanemarljiva.

```
library(e1071)
skewness(zivdoba1$`1`)
```

```
## [1] 0.1802499
```

Kot vidimo, je asimetrija porazdelitve življenjske dobe gum, narejenih po novi metodi zanemarljiva, za porazdelitev življenjske dobe gum, narejenih po stari metodi pa lahko iz histograma in škatle z brki razberemo, da je približno simetrična.

Ker sta obe porazdelitvi življenjske dobe gum simetrični ali zelo blago asimetrični, lahko izberemo povprečno vrednost kot dobrega predstavnika centralne vrednosti porazdelitve, kar pomeni, da bomo primerjali življenjsko dobo dveh skupin s pomočjo njihovih vzorčnih povprečij.

5. Definiranje ničelne in alternativne domneve

Zdaj lahko definiramo ničelno domnevo kot $H_0 : \mu_X = \mu_Y$ (povprečna vrednost življenjske dobe gum, narejenih po novi metodi, je enaka povprečni vrednosti življenjske dobe gum, narejenih po stari metodi) proti alternativi domnevi $H_1 : \mu_X > \mu_Y$ (povprečna vrednost življenjske dobe gum, narejenih po novi metodi, je večja kot povprečna vrednost življenjske dobe gum, narejenih po stari metodi).

6. Preverjanje predpostavke t-testa za neodvisna vzorca o normalnosti porazdelitev

Preden uporabimo t-test za testiranje ničelne domneve, moramo preveriti, ali je predpostavka njegove uporabe izpolnjena: normalnost porazdelitve življenjskih dob avtomobilskih gum obeh populacij.

Na osnovi oblike histograma (kot zvon) lahko zaključimo, da sta obe porazdelitvi življenjske dobe gum normalni, kar bomo še preverili s Shapiro-Wilkovim testom normalnosti (funkcija *shapiro.test*). Ničelna domneva tega testa je, da je porazdelitev slučajne spremenljivke normalna, proti alternativni domnevi, da ni normalna.

```
shapiro.test(zivdoba1$`1`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  zivdoba1$‘1‘
## W = 0.97567, p-value = 0.7882
```

```
shapiro.test(zivdoba1$`2`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  zivdoba1$‘2‘
## W = 0.969, p-value = 0.6197
```

Na osnovi rezultatov Shapiro-Wilkovega testa (nova metoda: testna statistika $W = 0.976$, p-vrednost $p = 0.7882 > 0.05$; stara metoda: testna statistika $W = 0.969$, p-vrednost $p = 0.6197 > 0.05$), lahko zaključimo, da ima življenjska doba obeh skupin gum (vsaj) približno normalno porazdelitev. Odstopanja vzorčnih porazdelitev od normalne porazdelitve niso statistično pomembna.

7. Testiranje enakosti varianc dveh populacij

Na osnovi normalnosti porazdelitev lahko uporabimo t-test za neodvisna vzorca za testiranje ničelne domneve. Vprašanje je edino, kateri t-test bomo uporabili: standardni t-test za neodvisna vzorca, ki predpostavlja enakost varianc dveh populacij, ali Welchov t-test, ki se lahko uporabi pri neenakih variancah.

Za testiranje enakosti varianc dveh populacij bomo uporabili Levenov test (funkcija *levene.test*) v paketu *lawstat*). Ničelna domneva tega testa je, da sta varianci dveh populacij enaki, proti alternativni domnevi, da nista enaki.

```
library(lawstat)
levene.test(zivdoba$zdoba,zivdoba$metoda,location="mean")
```

```
##
##  Classical Levene’s test based on the absolute deviations from the mean
##  ( none not applied because the location is not set to median )
##
## data:  zivdoba$zdoba
## Test Statistic = 6.1855, p-value = 0.01641
```

Na osnovi rezultata Levenovega testa (testna statistika $W = 6.1855$, p-vrednost $p = 0.01641 < 0.05$), lahko zaključimo, da varianci življenjskih dob avtomobilskih gum obeh populacij nista približno enaki.

Zaradi neenakosti varianc moramo uporabiti Welchov t-test za neodvisna vzorca za testiranje ničelne domneve o enakosti povprečnih vrednosti življenjskih dob gum.

8. Rezultati t-testa za neodvisna vzorca in njihova interpretacija

Za Welchov t-test lahko uporabimo funkcijo *t.test*.

```
t.test(zivdoba1$`1`,zivdoba1$`2`,var.equal=FALSE,alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  zivdoba1$‘1‘ and zivdoba1$‘2‘
## t = 3.7487, df = 34.832, p-value = 0.0003224
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.228339      Inf
## sample estimates:
## mean of x mean of y
##   61.7804   57.7232
```

V t-testu smo označili, da varianci življenjskih dob gum obeh populacij nista enaki (*var.equal=FALSE*) in smer razlike v alternativni domnevi *alternative="greater"*).

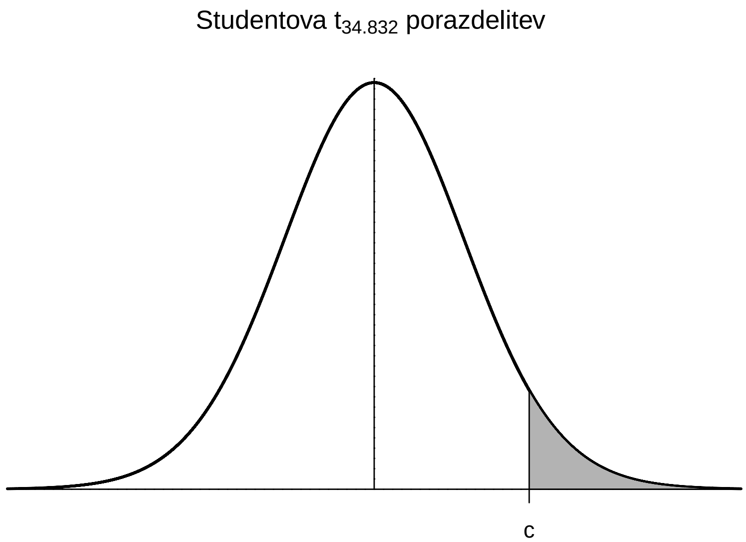
Zaključek lahko napišemo na naslednji način. Na osnovi dobljenih rezultatov Welchovega t-testa ($t = 3.7487$, $df = 34.832$, $p = 0.0003224 < 0.05$), zavrnemo ničelno domnevo, oziroma potrdimo alternativno domnevo, da je življenjska doba avtomobilskih gum, narejenih po novi metodi večja od življenjske dobe gum, narejenih po stari metodi, za dano stopnjo značilnosti $\alpha = 0.05$ in dobljena vzorca. Potrdili smo našo raziskovalno domnevo, da nova metoda proizvodnje avtomobilskih gum vpliva na povečanje njihove življenjske dobe.

9. Dodatek: računanje kritične vrednosti in p-vrednosti testa

Računanje kritične vrednosti

Za alternativno domnevo $H_1 : \mu_X > \mu_Y$ je kritično območje K oblike $K : T \geq c$, oziroma $K = [c, \infty)$. Ploščina pod krivuljo gostote od $-\infty$ do c je enaka 0.95. Kritično območje je pobarvano sivo na naslednji sliki.

```
source("G:\\Documents\\Temp\\tdftails.R")
plottdf.rtail(df=34.832,alpha=0.05)
```



Kritično vrednost c izračunamo v R-ju kot 0.95-kvantil Studentove porazdelitve s 34.832 prostostnimi stopnjami

```
(c<-qt(0.95, 34.832))
```

```
## [1] 1.689794
```

Potem je kritično območje $K = [1.689, \infty)$.

Vrednost testne statistike $t = 3.7487 \in K = [1.689, \infty)$, oziroma vrednost testne statistike pade v kritično območje, zato zavrnemo ničelno domnevo.

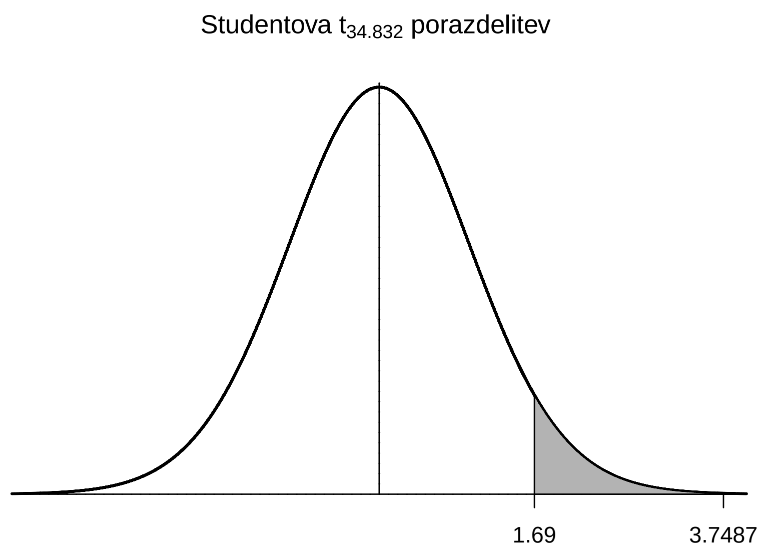
Računanje p-vrednosti

V R-ju se odločitev o nulti domnevi naredi na osnovi p-vrednosti testa, ne na osnovi kritične vrednosti. Za izbrano alternativno domnevo je p-vrednost enaka

$$p - \text{vrednost} = P(T \geq 3.7487)$$

in predstavlja ploščino pod krivuljo gostote od ∞ do vrednosti testne statistike 3.7487. Na naslednji sliki je prikazana Studentova porazdelitev s 34.832 prostostnimi stopnjami, vrednost testne statistike in kritično območje.

```
plottdf.rtail(df=34.832,alpha=0.05,ind=TRUE,t=3.7487)
```



Izračunajmo p-vrednost v R-ju

```
1-pt(3.7487,34.832)
```

```
## [1] 0.0003223751
```

P-vrednost je manjša od $\alpha = 0.05$, kar pomeni, da vrednost testne statistike pade v kritično območje, zato zavrnemo ničelno domnevo. Pridemo do enakih zaključkov, kot v primeru računanja kritične vrednosti.