Analiza studije razumijevanja riječi

Grupa ***sudo***

3 May 2017

Opis eksperimenta

Nad velikim brojem ispitanika proveden je eksperiment razumijevanja engleskog jezika. Ispitanicima su dana dva zadatka te su oba ponavljana više puta. Prvi zadatak (dalje: **A**) bavi se prepoznavanjem ispravne riječi, prilikom čega je ispitanik za zadani niz znakova morao odrediti radi li se o ispravnoj riječi engleskog jezika, a drugi zadatak (dalje: **B**) se bavi pravilnim izgovaranjem zadane riječi. Za svaku riječ i svakog ispitanika mjereno je vrijeme rješavanja svakog zadatka, te niz podataka o ispitaniku.

Ishodi eksperimenta

Cilj eksperimenta je naučiti kako mjerene veličine ispitanika utječu na vrijeme potrebno za rješavanje pojedinih zadataka. Na temelju tih podataka može se odgovoriti na neka zanimljiva pitanja poput: utječe li dob na brzinu rješavanja zadataka, kako na brzinu rješavanja utječe duljina zadane riječi, je li riječ kraća ukoliko se češće pojavljuje, itd.

Skup podataka

Za određivanje ishoda eksperimenta potreban nam je skup podataka eksperimenta. Programski jezik R sadrži skup podataka već provedenog eksperimenta te nam dopušta uključivanje tog skupa te analizu podataka. Podaci se nalaze u paketu `languageR`. Nakon instaliranja paketa, podaci se mogu učitati naredbom `require(languageR)` te dohvatiti s naredbom `data(english)`. Kompletno dohvaćanje i uključivanje podataka prikazano je kodom ispod.

```
require(languageR, quietly = TRUE)
data(english)
```

Podaci se sada mogu koristiti naredbom `english`, npr. deskriptivna statistika može se dobiti naredbom `summary(english)`, a pregled prvih par redova podataka može se pregledati naredbom `head(english)`.

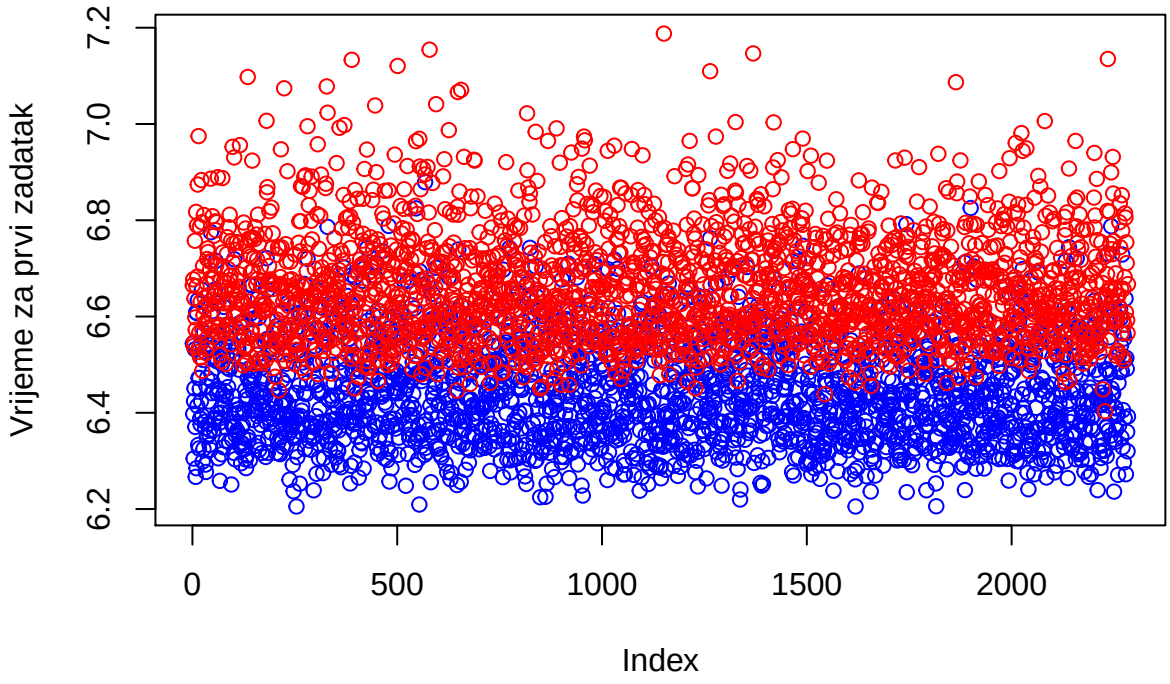
Ishodi eksperimenta

Utjecaj dobi na brzinu rješavanja

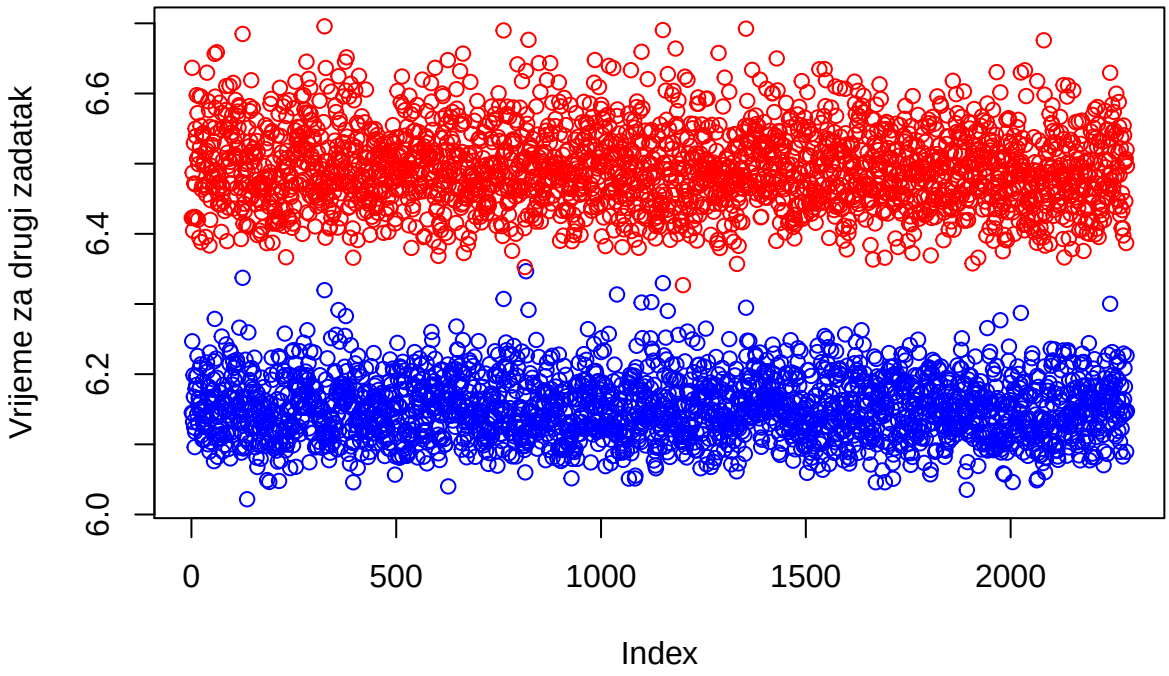
Pitamo se utječe li dobna razlika između starijih i mlađih ispitanika na brzinu rješavanja zadataka? Uspoređujući srednje vrijednosti logaritama vremena za rješavanje A i B zadataka mlađih i starijih ispitanika te gledajući dijagrame, možemo zaključiti da su mlađi u prosjeku brže rješavali oba zadatka. *t-testom* potvrđujemo naš zaključak.

```
young = english[english$AgeSubject == "young", ] # mladi
old = english[english$AgeSubject == "old", ] # stari

 # vrijeme potrebno mlađima za rješavanje prvog zadatka
RTlexdec_young = young[, "RTlexdec"]
```

```r
 # vrijeme potrebno starijima za rješavanje prvog zadatka
RTlexdec_old = old[, "RTlexdec"]

 # vrijeme potrebno mlađima za rješavanje drugog zadatka
RTnaming_young = young[, "RTnaming"]

 # vrijeme potrebno starijima za rješavanje drugog zadatka
RTnaming_old = old[, "RTnaming"]

plot(RTlexdec_young, col = 'blue',
     ylim = c(min(english$RTlexdec), max(english$RTlexdec)),
     ylab = "Vrijeme za prvi zadatak")

points(RTlexdec_old, col='red')
```




```r
plot(RTnaming_young, col = 'blue',
     ylim = c(min(english$RTnaming), max(english$RTnaming)),
     ylab = "Vrijeme za drugi zadatak")

points(RTnaming_old, col='red')
```




```
# testiranje jednakosti varijance prije t-testa
var.test(RTlexdec_young, RTlexdec_old)
```

```
##
##  F test to compare two variances
##
## data:  RTlexdec_young and RTlexdec_old
## F = 0.84625, num df = 2283, denom df = 2283, p-value = 6.737e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7795831 0.9186270
## sample estimates:
## ratio of variances
##          0.8462542
```

```
t.test(RTlexdec_young, RTlexdec_old, alt = "less", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  RTlexdec_young and RTlexdec_old
## t = -67.468, df = 4566, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2163149
## sample estimates:
## mean of x mean of y
```

```
##  6.439237  6.660958
```

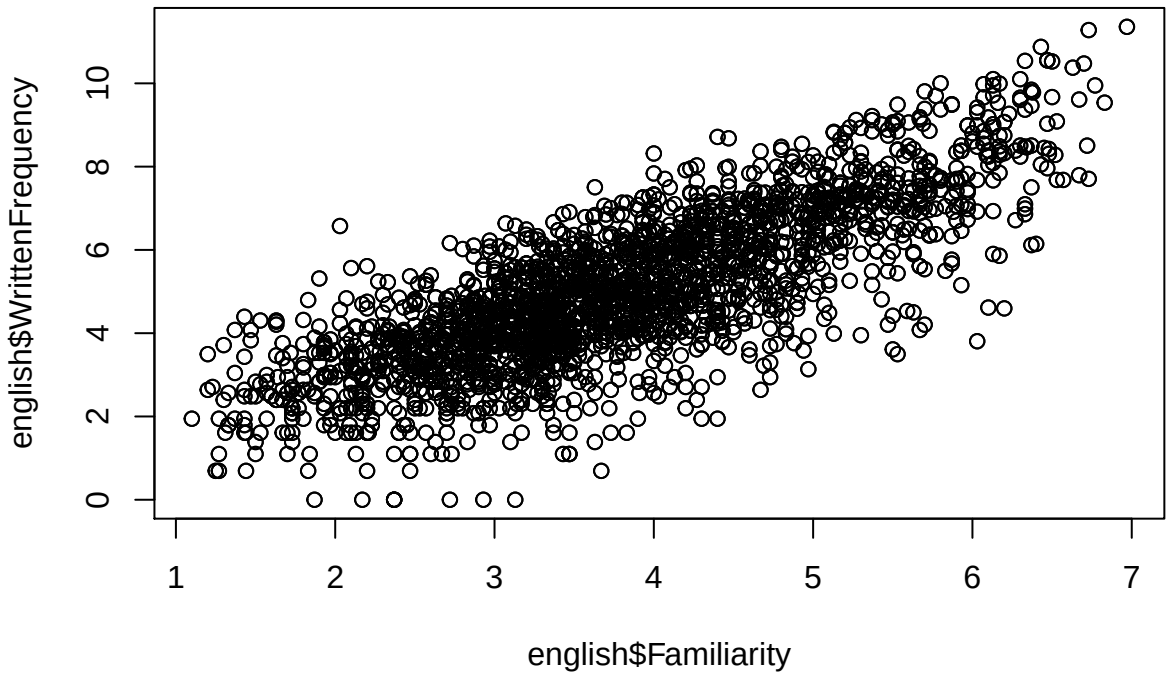
Prepoznatljivost riječi s obzirom na frekvenciju pojavljivanja

Zanima nas jesu li riječi koje se više pojavljuje prepoznatljivije? Računamo korelaciju između prepoznatljivosti riječi i njenog pojavljivanja u tekstovima. Dobivamo korelaciju `~0.8`, što nam potvrđuje da su te dvije stavke povezane, tj. riječi koje se više pojavljuju su prepoznatljivije. To također vidimo i iz dijagrama rasipanja.

```
cor(english$Familiarity, english$WrittenFrequency)
```

```
## [1] 0.7912556
```

```
plot(english$Familiarity, english$WrittenFrequency)
```



Utjecaj glasa prvog slova na prepoznatljivost riječi

Je li riječ koja počinje na samoglasnik u odnosu na suglasnik ljudima prepoznatljivija? Uzimamo skup riječi koje počinju sa samoglasnikom, te skup riječi koje počinju sa suglasnikom te računamo srednju vrijednost. Kod samoglasnika dobivamo srednju vrijednost `4.0`, a kod suglasnika `3.79`, što bi nas moglo dovesti do zaključka da riječi koje počinju sa samoglasnikom su prepoznatljivije. No testiranjem putem *t-testa* zaključujemo da ne postoji razlika između prepoznatljivosti riječi koje počinju samoglasnikom u odnosu na one koje počinju suglasnikom.

```
firstVowel = english[english$CV == "V",] #riječi koje počinju sa samoglasnikom
firstConsonant = english[english$CV == "C",] # riječi koje počinju sa suglasnikom
```

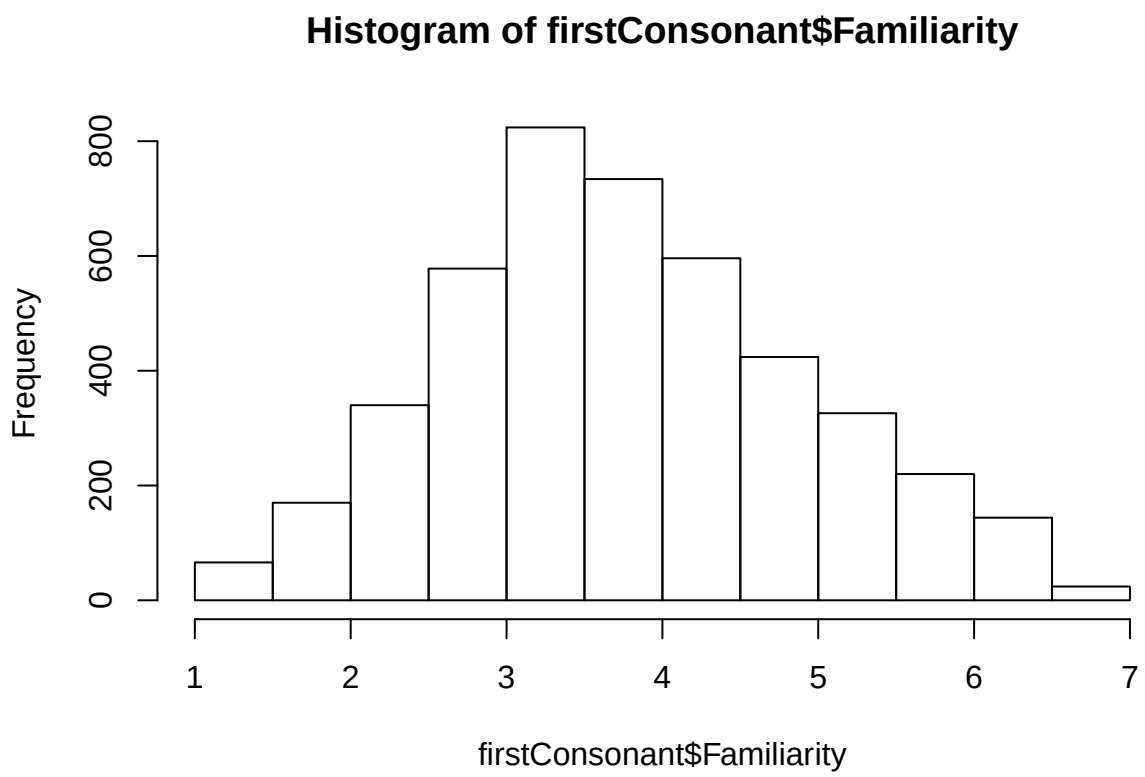
```r
mean(firstConsonant$Familiarity)
```

```
## [1] 3.789892
```

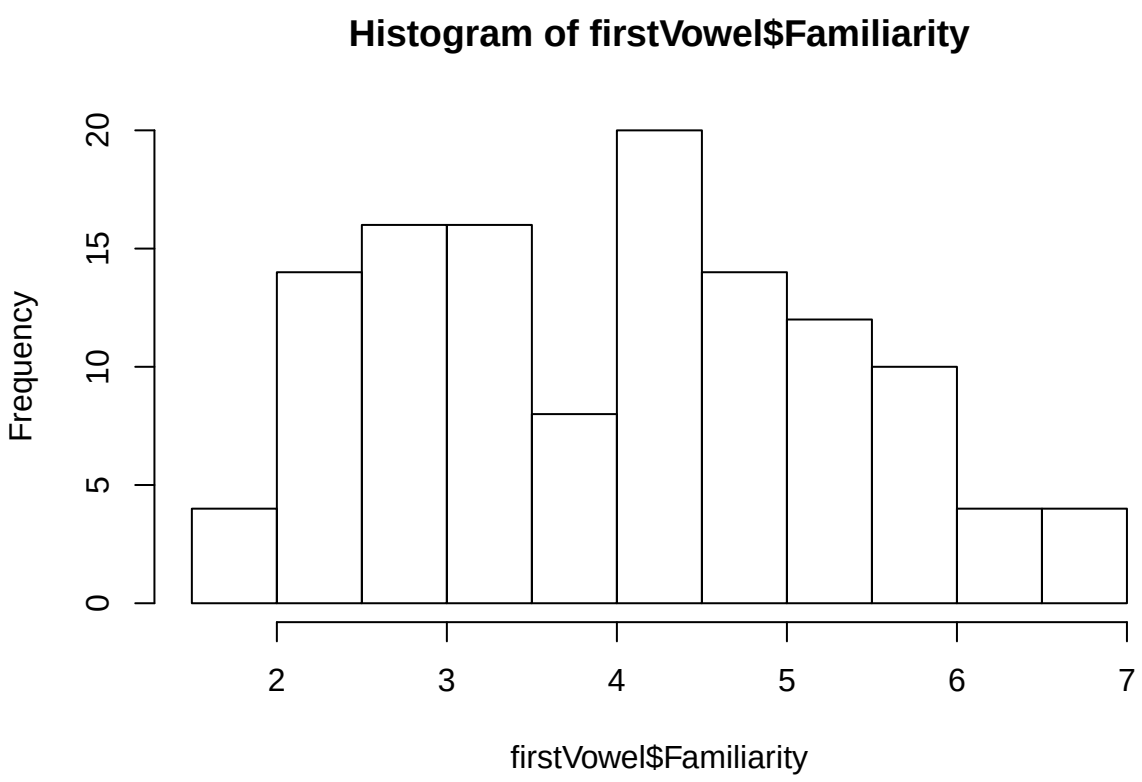
```r
mean(firstVowel$Familiarity)
```

```
## [1] 4.002951
```

```r
hist(firstConsonant$Familiarity)
```




```r
hist(firstVowel$Familiarity)
```




```
# prije testiranja t-testom trebamo zaključiti jesu li varijance jednake u oba slučaja
var.test(firstConsonant$Familiarity, firstVowel$Familiarity)
```

```
##
##  F test to compare two variances
##
## data:  firstConsonant$Familiarity and firstVowel$Familiarity
## F = 0.78386, num df = 4445, denom df = 121, p-value = 0.04732
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5971133 0.9971773
## sample estimates:
## ratio of variances
##          0.7838635
```

```
t.test(firstVowel$Familiarity, firstConsonant$Familiarity, alt = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  firstVowel$Familiarity and firstConsonant$Familiarity
## t = 2.0207, df = 4566, p-value = 0.02168
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.03959434        Inf
## sample estimates:
## mean of x mean of y
```

```
##  4.002951  3.789892
```

Utjecaj broja samoglasnika na brzinu rješavanja zadataka

Pitamo se utječe li broj samoglasnika u riječi na brzinu rješavanja zadataka?

```
#cor(english$WrittenSpokenFrequencyRatio, english$RTlexdec)

moreSpoken = english[english$WrittenSpokenFrequencyRatio > 0, ]
#je li riječi koje se više govore, dakle writespokefrequncyratio < 0 imaju i veći rezultat na RTnamingu
```

```
#numericYoung = young[sapply(english, is.numeric)]
#numericOld = old[sapply(english, is.numeric)]

#cor(numericOld)
#cor(numericYoung)

#diff = cor(numericOld) - cor(numericYoung)
#diff
```