ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Θέμα Εργασίας: Στατιστική Συμπερασματολογία

Μάθημα: Ανάλυση Δεδομένων με Η/Υ

Διδάσκων: Δημήτρης Φουσκάκης

Φοιτήτρια: Ελένη Στυλιανού, ge21708

Email: elenistylianou03@live.com

Άσκηση 1:

Αρχικά, φορτώνω τα δεδομένα μου στην R,δημιουργώντας ένα dataframe που το ονομάζω data, χρησιμοποιώντας την εντολή
>data<na.omit(read.table("http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/pharmacy.txt", header=T, na.strings="$"))

Το τελευταίο όρισμα δηλώνει στην R ότι οι αγνοούμενες τιμές του dataframe είχαν συμβολιστεί με $ στο αρχείο των δεδομένων. Η R θα μετατρέψει το $ σε NA που είναι το δικό της σύμβολο για τις αγνοούμενες τιμές.

Με την εντολή:
>nrow(data)
Επιστράφηκε ο αριθμός των γραμμών του νέου πλαισίου δεδομένων data, δηλαδή ο αριθμός των πελατών που έκαναν μία μόνο αγορά των προηγούμενο μήνα για κάποιο μη φαρμακευτικό προϊόν. Παρατηρούμε ότι αφαιρέθηκε 1 πελάτης από τους 72 του αρχικού δείγματος.

Με την εντολή:
>data<-subset(data,age>=18)
Αφαιρούμε, αν υπάρχει, οποιαδήποτε γραμμή περιέχει τιμές από μη ενήλικα άτομα και με την εντολή:
>nrow(data)
Επιστρέφεται πάλι ο αριθμός των γραμμών του νέου πλαισίου δεδομένων ο οποίο είναι 69.

Προκειμένου να κάνουμε ελέγχους υποθέσεων που αφορούν τις μεταβλητές του data, δημιουργούμε στην R τα διανύσματα που αφορούν τις μεταβλητές αυτές

> attach(data)

(i) Σε ε.σ. 5% θα κάνουμε τον έλεγχο με την υπόθεση ότι η μέση τιμή της τ.μ. money είναι ίση ή μεγαλύτερη από 45 ευρώ με εναλλακτική ότι είναι χαμηλότερη από 45 ευρώ.

Πρώτα όμως θα πρέπει να ελέγξουμε κατά πόσο τα δεδομένα που συλλέξαμε μπορούν να αφορούν δεδομένα που προέρχονται από κανονική κατανομή. Θα κάνουμε, λοιπόν σε ε.σ. 5% τον έλεγχο

H_0: τα δεδομένα της μεταβλητής money προέρχονται από κανονική κατανομή, με εναλλακτική H_1: τα δεδομένα της money δεν προέρχονται από κανονική κατανομή

> shapiro.test(money)

Shapiro-Wilk normality test

data: money

W = 0.98954, p-value = 0.8372

Παρατηρούμε ότι η p-value >5% και συμπεραίνουμε ότι, σε ε.σ. 5%, δεν υπάρχουν σημαντικές ενδείξεις να απορρίψουμε την υπόθεση ότι τα δεδομένα αυτά προέρχονται από κανονική κατανομή, ως προς την εναλλακτική υπόθεση ότι δεν προέρχονται από κανονική κατανομή.

Επίσης, με τις εντολές:

>x<-data$money

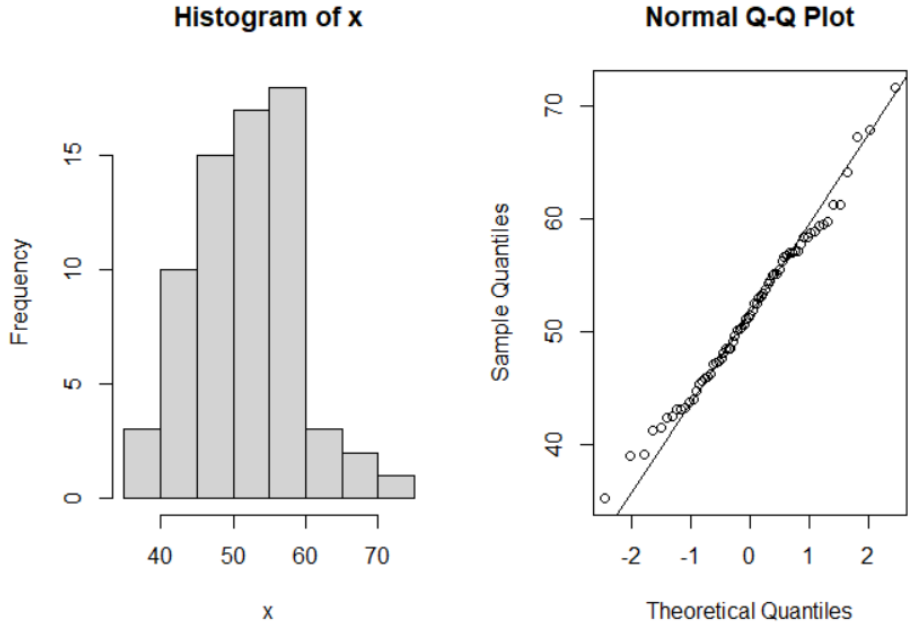
>par(mfrow=c(1,2))

> hist(x)

> qqnorm(x)

> qqline(x)

Ορίζουμε την τυχαία μεταβλητή x που αποτελείτε από τις τιμές της μεταβλητής money και δημιουργούμε ένα ιστόγραμμα και ένα qqplot ώστε να ελέγξουμε την κανονικότητα και γραφικά. Από αυτές τις εντολές παίρνουμε τα εξής αποτελέσματα:



Από αυτά τα γραφήματα παρατηρούμε πως η υπόθεση της κανονικότητας δεν είναι παράλογη αφού δεν μας δίνουν αρκετές ενδείξεις για να την απορρίψουμε.

Συνεπώς για τον έλεγχο

H_0: η μέση τιμή της τ.μ. money είναι μεγαλύτερη ή ίση με 45 ευρώ, με εναλλακτική H_1: η μέση τιμή της τ.μ. money είναι χαμηλότερη από 45 ευρώ, σε ε.σ. 5%, θα προβούμε στον έλεγχο one sample t-test

```
> t.test(x,alternative="less",mu=45,conf.level=0.95)

        One Sample t-test

data:  x
t = 7.7659, df = 68, p-value = 1
alternative hypothesis: true mean is less than 45
95 percent confidence interval:
    -Inf 53.22443
sample estimates:
mean of x
 51.77058
```

Η p-value > 5% και συνεπώς σε ε.σ. 5% δεν έχουμε αρκετές ενδείξεις να απορρίψουμε την υπόθεση H_0, δηλαδή την υπόθεση ότι η μέση τιμή της μεταβλητής money είναι ίση ή μεγαλύτερη από 45 ευρώ ως προς την εναλλακτική ότι είναι χαμηλότερη από 45 ευρώ.

(ii) Προκειμένου να μελετήσουμε αν ξοδεύονται κατά μέσο όρο περισσότερα χρήματα στα καλλυντικά σε σύγκριση με τις άλλες δύο κατηγορίες δημιουργούμε την νέα κατηγορική μεταβλητή

```
> Cosmetics_or_not<-rep(0,69)
> Cosmetics_or_not[category=="cosmetics"]<-1
> Cosmetics_or_not<-factor(Cosmetics_or_not)
```

και τις βοηθητικές μεταβλητές

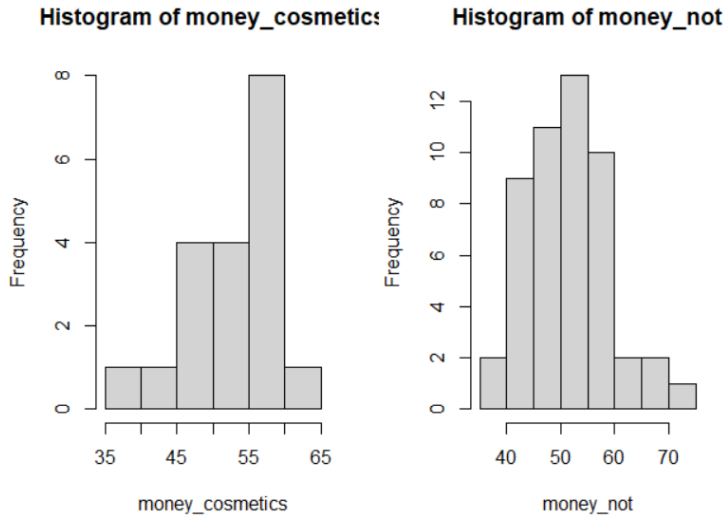
```
> money_cosmetics<-money[Cosmetics_or_not==1]
> money_not<-money[Cosmetics_or_not==0]
```

που αφορούν τη μεταβλητή money σε κάθε μια από τις κατηγορίες της Cosmetics_or_not.

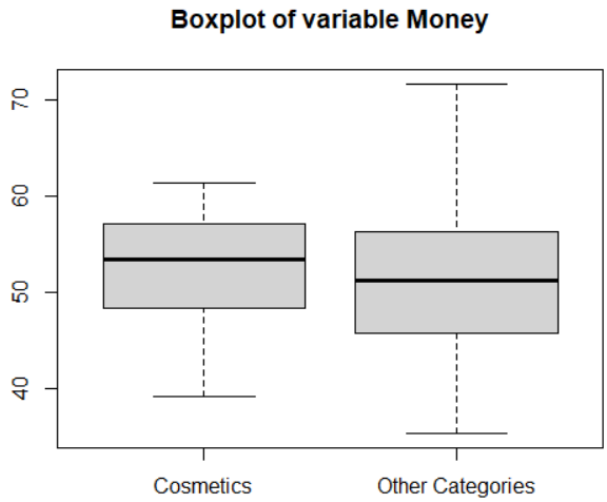
Ο επόμενος κώδικας αφορά στην περιγραφή της μεταβλητής money σε σχέση με τις κατηγορίες της Cosmetics_or_not

```
> summary(money_cosmetics)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 39.11   48.38   53.38   52.87   57.11   61.29
> summary(money_not)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 35.32   45.74   51.20   51.35   56.11   71.60
> var(money_cosmetics)
[1] 36.31467
> var(money_not)
[1] 58.79563
> par(mfrow=c(1,2))
> hist(money_cosmetics)
> hist(money_not)
```




```
>boxplot(money_cosmetics,money_not,names=c('Cosmetics','Other
Categories'),main='Boxplot of variable Money')
```



> fivenum(money_cosmetics)

[1] 39.110 48.375 53.380 57.110 61.290

> fivenum(money_not)

[1] 35.32 45.71 51.20 56.31 71.60

Από τα παραπάνω έχουμε ότι το ελάχιστο ποσό που ξοδεύτηκε στα καλλυντικά είναι 39.11 ευρώ και το μέγιστο 61.29 ευρώ, ενώ στα υπόλοιπα είδη το ελάχιστο ποσό είναι 35.32 ευρώ και ο μέγιστος 71.60 ευρώ. Το 25% το ποσό που ξοδεύτηκε στα καλλυντικά είναι μικρότερο από 48.38 ευρώ, το 50% μικρότερο από 53.38 ευρώ και το 75% μικρότερο από 57.11 δευτερόλεπτα. Για τα υπόλοιπα είδη τα αντίστοιχα ποσοστημόρια είναι 45.74 ευρώ, 51.20 ευρώ και 56.11 ευρώ. Η δειγματική μέση τιμή του ποσού που ξοδεύτηκε στα καλλυντικά είναι 52.87 ευρώ και η δειγματική διασπορά 36.31 ευρώ. Για τα υπόλοιπα είδη τα αντίστοιχα μεγέθη είναι 51.35 ευρώ και 58.79 ευρώ. Το εύρος των ποσών που ξοδεύτηκαν (μέγιστη-ελάχιστη τιμή) στα καλλυντικά είναι είναι 22.18 ευρώ, ενώ για τα υπόλοιπα είδη είναι 36.28 ευρώ.

Από τα ιστογράμματα παρατηρώ ότι για τα καλλυντικά η κλάση με τη μεγαλύτερη συχνότητα παρατηρήσεων της μεταβλητής money είναι αυτή μεταξύ των 55 και 60 ευρώ ενώ για τα υπόλοιπα είδη είναι αυτή μεταξύ των 50 και 55 ευρώ. Επίσης, παρατηρώ πως το ιστόγραμμα που αφορά τα καλλυντικά δεν είναι συμμετρικό ενώ για τα άλλα είδη υπάρχει ένα είδος συμμετρίας. Για τα καλλυντικά παρατηρούμε μια ουρά με μεγάλη (σχετικά) συχνότητα προς τα δεξιά.

Από το θηκοδιάγραμμα και την εντολή fivenum() επιβεβαιώνουμε αυτά που παρατηρήσαμε και προηγουμένως όσο αφορά τις ελάχιστες/μέγιστες παρατηρήσεις, τα ποσοστημόρια και το εύρος των παρατηρήσεων και τη μη ύπαρξη συμμετρίας. Παρατηρούμε επίσης ότι σε καμία από τις δύο περιπτώσεις δεν υπάρχουν ακραίες τιμές.

(iii) Θέλουμε να ελέγξουμε σε ε.σ. 10% τον ισχυρισμό ότι κατά μέσο όρο οι καταναλωτές ξοδεύουν περισσότερα χρήματα στα καλλυντικά απ' ότι ξοδεύουν στα άλλα είδη. Για το σκοπό αυτό αρχικά θα κάνουμε, σε ε.σ. 10% τον έλεγχο

H_0: τα δεδομένα της μεταβλητής money_cosmetics προέρχονται από κανονική κατανομή, με εναλλακτική H_1: τα δεδομένα της money_cosmetics δεν προέρχονται από κανονική κατανομή

>shapiro.test(money_cosmetics)

Shapiro-Wilk normality test

data: money_cosmetics

W = 0.94712, p-value = 0.3527

Συνεπώς σε ε.σ. 10% δεν έχουμε αρκετές ενδείξεις για να απορρίψουμε την υπόθεση ότι τα δεδομένα αυτά προέρχονται από κανονική κατανομή, ως προς την εναλλακτική ότι δεν προέρχονται από κανονική κατανομή, αφού p-value>10%.

Κάνουμε τον αντίστοιχο έλεγχο για την μεταβλητή money_not σε ε.σ. 10%

H_0: τα δεδομένα της μεταβλητής money_not προέρχονται από κανονική κατανομή, με εναλλακτική H_1: τα δεδομένα της money_not δεν προέρχονται από κανονική κατανομή

```
> shapiro.test(money_not)

    Shapiro-Wilk normality test

data:  money_not

W = 0.97961, p-value = 0.5354
```

Συνεπώς σε ε.σ. 10% δεν έχουμε αρκετές ενδείξεις για να απορρίψουμε την υπόθεση ότι τα δεδομένα αυτά προέρχονται από κανονική κατανομή, ως προς την εναλλακτική ότι δεν προέρχονται από κανονική κατανομή, αφού p-value>10%.

Και για τις δύο μεταβλητές ,λοιπόν, θεωρούμε σε ε.σ. 10% ότι προέρχονται από κανονική κατανομή. Επίσης οι υποπληθυσμοί που εξετάζουμε είναι ανεξάρτητοι. Πριν προχωρήσουμε σε έλεγχο για τις μέσες τιμές, θα κάνουμε σε ε.σ. 10% τον έλεγχο

H_0: οι διασπορές των υποπληθυσμών είναι ίσες, με εναλλακτική H_1: οι διασπορές των υποπληθυσμών δεν είναι ίσες

```
> var.test(money_cosmetics,money_not,conf.level=0.9)

    F test to compare two variances

data:  money_cosmetics and money_not

F = 0.61764, num df = 18, denom df = 49, p-value = 0.2637

alternative hypothesis: true ratio of variances is not equal to 1

90 percent confidence interval:

 0.3396395 1.2585506

sample estimates:

ratio of variances

     0.6176424
```

Από τον παραπάνω έλεγχο συμπεραίνουμε ότι σε ε.σ. 10% δεν υπάρχουν αρκετές ενδείξεις για να απορρίψουμε την υπόθεση ισότητας των διασπορών, μιας και p-value > 10% και το 90% Δ.Ε. του λόγου των διασπορών περίεχει το 1. Έτσι, σε ε.σ. 10% θεωρούμε ότι ισχύει η υπόθεση ισότητας των διασπορών.

Συνεπώς, για τον έλεγχο

H_0: η μέση τιμή των ποσών που ξοδεύτηκαν στα καλλυντικά είναι μικρότερη ή ίση από τη μέση τιμή των ποσών που ξοδεύτηκαν στα άλλα είδη, με εναλλακτική H_1: η μέση τιμή των

ποσών που ξοδεύτηκαν στα καλλυντικά είναι μεγαλύτερη από τη μέση τιμή των ποσών που ξοδεύτηκαν στα άλλα είδη

σε ε.σ. 10%, γράφουμε τον παρακάτω κώδικα

```
>t.test(money_cosmetics,money_not,alternative="greater",conf.level=0.9,var.equal=T)

      Two Sample t-test

data:  money_cosmetics and money_not

t = 0.77545, df = 67, p-value = 0.2204

alternative hypothesis: true difference in means is greater than 0

90 percent confidence interval:

 -1.015677      Inf

sample estimates:

mean of x mean of y

 52.87053  51.35260
```

Συμπερασματικά, σε ε.σ. 10% δεν έχουμε ισχυρές ενδείξεις να απορρίψουμε την υπόθεση H_0, την υπόθεση δηλαδή ότι η μέση τιμή των ποσών που ξοδεύτηκαν στα καλλυντικά είναι μικρότερη ή ίση από τη μέση τιμή των ποσών που ξοδεύτηκαν στα άλλα είδη, με εναλλακτική ως προς την εναλλακτική ότι η μέση τιμή των ποσών που ξοδεύτηκαν στα καλλυντικά είναι μεγαλύτερη από τη μέση τιμή των ποσών που ξοδεύτηκαν στα άλλα είδη. Αυτό προκύπτει από το γεγονός ότι το p-value > 10% όπως επίσης από το ότι το 90% Δ.Ε. που προκύπτει περιέχει το 0.0

(iv) Σε ε.σ. 1% θέλουμε να ελέγξουμε κατά πόσο διαφέρει κατά μέσο όρο το ποσό που ξοδεύεται σε μη φαρμακευτικά προϊόντα μεταξύ ανδρών και γυναικών. Για το σκοπό αυτό ορίζουμε τις βοηθητικές μεταβλητές

```
> money_Man<-money[sex=='Man']

> money_Woman<-money[sex=='Woman']
```

Ο επόμενος κώδικας αφορά στην περιγραφή της μεταβλητής money σε σχέση με το φύλο των πελατών (Man, Woman)

```
> summary(money_Man)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  35.32   44.00   48.52   49.01   52.62   71.60

> summary(money_Woman)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  42.44   51.12   55.08   54.78   58.74   67.85
```

> var(money_Man)

[1] 48.12196

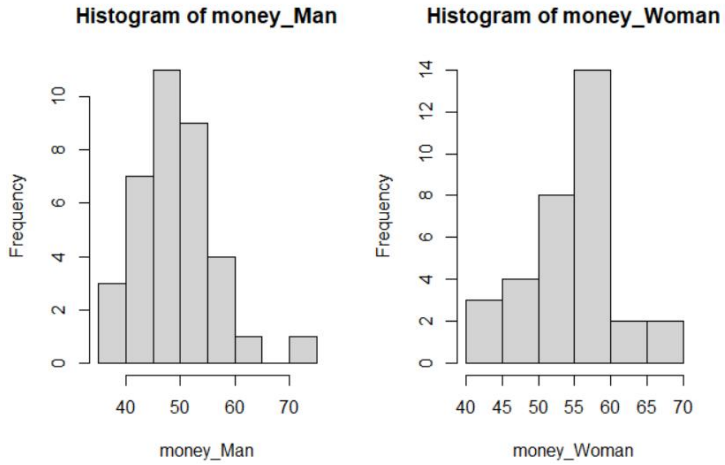
> var(money_Woman)

[1] 40.88595

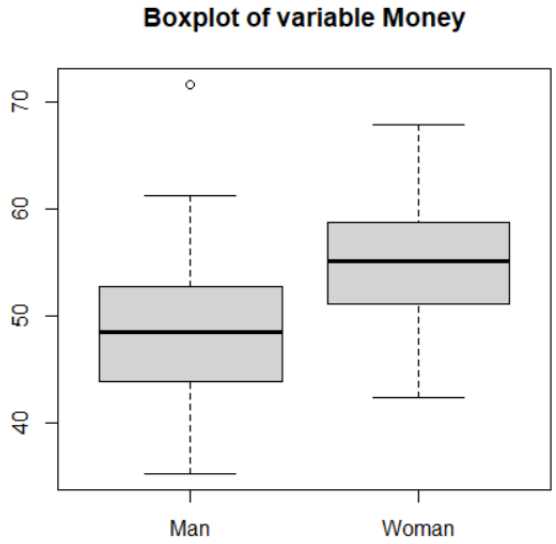
> par(mfrow=c(1,2))

> hist(money_Man)

> hist(money_Woman)



> boxplot(money_Man,money_Woman,names=c('Man','Woman'),main='Boxplot of variable Money')



>fivenum(money_Man)

[1] 35.320 43.930 48.515 52.785 71.600

> fivenum(money_Woman)

[1] 42.44 51.12 55.08 58.74 67.85

Οι μεταβλητές αυτές προέρχονται από ανεξάρτητους πληθυσμούς. Σε ε.σ. 1% θα κάνουμε τους ελέγχους για το κατά πόσο τα δεδομένα τους μπορούν να προέρχονται από κανονική κατανομή. Έχουμε λοιπόν

H_0: τα δεδομένα της money_Man προέρχονται από κανονική κατανομή, με εναλλακτική υπόθεση H_1: τα δεδομένα της money_Man δεν προέρχονται από κανονική κατανομή

```
> shapiro.test(money_Man)

    Shapiro-Wilk normality test

data:  money_Man

W = 0.95832, p-value = 0.1907
```

Αφού p-value > 1% , δεν έχουμε ικανοποιητικά στοιχεία να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από κανονική κατανομή, ως προς την εναλλακτική υπόθεση ότι δεν προέρχονται από κανονική κατανομή.

H_0: τα δεδομένα της money_Woman προέρχονται από κανονική κατανομή, με εναλλακτική υπόθεση H_1: τα δεδομένα της money_Woman δεν προέρχονται από κανονική κατανομή

```
> shapiro.test(money_Woman)

    Shapiro-Wilk normality test

data:  money_Woman

W = 0.96819, p-value = 0.4319
```

Αφού p-value > 1% , δεν έχουμε ικανοποιητικά στοιχεία να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από κανονική κατανομή, ως προς την εναλλακτική υπόθεση ότι δεν προέρχονται από κανονική κατανομή.

Θεωρούμε, λοιπόν, σε ε.σ. 1% ότι τα δεδομένα από τους δύο αυτούς ανεξάρτητους πληθυσμούς προέρχονται από κανονική κατανομή. Κάνουμε τώρα, σε ε.σ. 1%, τον έλεγχο

H_0: οι διασπορές των υποπληθυσμών είναι ίσες, με εναλλακτική H_1: οι διασπορές των υποπληθυσμών δεν είναι ίσες

```
> var.test(money_Man,money_Woman,conf.level=0.99)

    F test to compare two variances

data:  money_Man and money_Woman

F = 1.177, num df = 35, denom df = 32, p-value = 0.6444

alternative hypothesis: true ratio of variances is not equal to 1

99 percent confidence interval:

 0.4687168 2.9085186
```

sample estimates:

ratio of variances

1.17698

Αφού p-value > 1% και το 99% Δ.Ε. περιέχει το 1, δεν έχουμε αρκετές ενδείξεις να απορρίψουμε την υπόθεση ότι οι διασπορές είναι ίσες, ως προς την εναλλακτική ότι δεν είναι ίσες.

Για να κάνουμε λοιπόν τον έλεγχο

H_0: οι μέσες τιμές των υποπληθυσμών είναι ίσες, με εναλλακτική H_1: οι μέσες τιμές των υποπληθυσμών διαφοροποιούντε σε ε.σ. 1%, γράφουμε τον εξής κώδικα

> t.test(money_Man,money_Woman,conf.level=0.99,var.equal=T)

Two Sample t-test

data: money_Man and money_Woman

t = -3.5841, df = 67, p-value = 0.0006366

alternative hypothesis: true difference in means is not equal to 0

99 percent confidence interval:

-10.042923 -1.502481

sample estimates:

mean of x mean of y

49.00972 54.78242

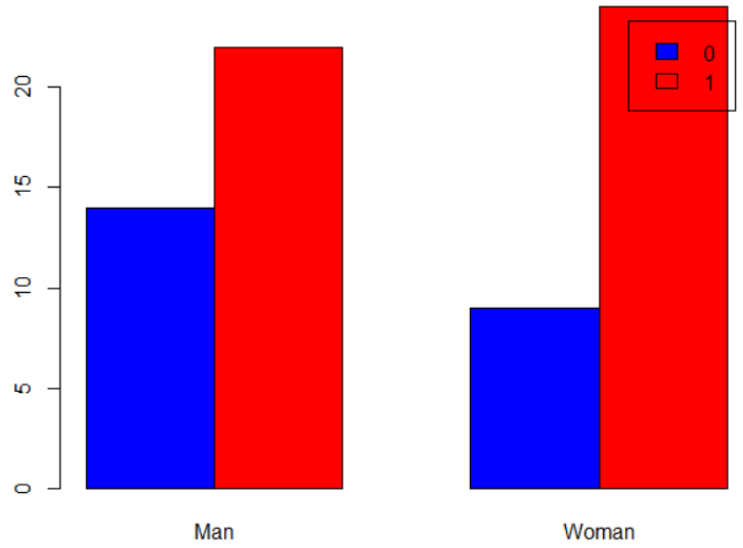
Μιας και p-value < 1% και το 99% Δ.Ε. δεν περιέχει το 0, έχουμε ισχυρές ενδείξεις να απορρίψουμε την υπόθεση ότι οι μέσες τιμές των πληθυσμών είναι ίσες, ως προς την εναλλακτική ότι διαφέρουν τα ποσά που ξοδεύονται από άντρες και γυναίκες σε μη φαρμακευτικά προϊόντα.

v) Για να κατασκευάσουμε τη νέα κατηγορική μεταβλητή med_f που θα έχει τιμές 0 όταν med<70 και ένα διαφορετικά, δίνουμε την εντολή

> data$med_f<-ifelse(data$med<70,0,1)

Ακολούθως, για να ελέγξουμε αν ο ισχυρισμός του ιδιοκτήτη ότι η πιθανότητα να ξοδευτεί μεγάλο ποσό σε φαρμακευτικά προϊόντα είναι μεγαλύτερη στις γυναίκες σε σχέση με τους άντρες χρησιμοποιούμε ένα ποσοτικό ραβδόγραμμα για την κατηγορική μεταβλητή sex και την κατηγορική μεταβλητή med_f με σκοπό να εξετάσουμε τη διαφορά στο ποσοστό των ανδρών και των γυναικών που ξοδεύουν μεγάλα ποσά.

> barplot(table(data$med_f,data$sex),beside=TRUE,col=c("blue","red"),legend=TRUE)



Το ποσοτικό ραβδόγραμμα δείχνει μια σημαντική διαφορά στο ποσοστό των ανδρών και των γυναικών που ξοδεύουν μεγάλα ποσά και συγκεκριμένα των γυναικών είναι μεγαλύτερο από αυτό των ανδρών, σε αντίθεση με το ποσοστό των ανδρών που ξόδεψαν μικρά ποσά που είναι μικρότερο από αυτό των γυναικών.

Προχωράμε στον έλεγχο υπόθεσης με χρήση ενός κατάλληλου στατιστικού τεστ.

Για τον έλεγχο υπόθεσης σε ε.σ. 5% αρχικά ορίζουμε τις βοηθητικές μεταβλητές med_f_man και med_f_woman ως εξής:

```
> med_f_man<- data$med _f[sex=='Man']
> med_f_woman<- data$med_f[sex=='Woman']
```

Έπειτα, κάνουμε τον εξής έλεγχο υπόθεσης:

Η0: η πιθανότητα να ξοδευτεί μεγάλο ποσό σε φαρμακευτικά προϊόντα είναι μικρότερη ή ίση στις γυναίκες σε σχέση με τους άνδρες, με εναλλακτική Η1: η πιθανότητα μα ξοδευτεί μεγάλο ποσό σε φαρμακευτικά προϊόντα είναι μεγαλύτερη στις γυναίκες σε σχέση με τους άνδρες

```
>x<-c(23,24)
>n<-c(36,33)
>prop.test(x,n,alternative='greater')
      2-sample test for equality of proportions with continuity correction
data:  x out of n
X-squared = 0.27918, df = 1, p-value = 0.7014
alternative hypothesis: greater
95 percent confidence interval:
 -0.3007286  1.0000000
sample estimates:
  prop 1    prop 2
0.6388889 0.7272727
```

Το p-value>5% και το 95% ΔΕΝ περιέχει το 0, επομένως δεν έχουμε επαρκή στοιχεία για να απορρίψουμε τη μηδενική υπόθεση, δηλαδή το ότι η πιθανότητα να ξοδευτεί μεγάλο ποσό σε φαρμακευτικά προϊόντα είναι μικρότερη ή ίση στις γυναίκες σε σχέση με τους άνδρες. Επομένως, ο ισχυρισμός του ιδιοκτήτη είναι λανθασμένος.

vi) Για να ελέγξουμε εάν η κατηγορία του μη φαρμακευτικού προϊόντος (category) είναι ανεξάρτητη του φύλου (sex) σε ε.σ. 5% χρησιμοποιούμε και πάλι x^2 independence test

```
> chisq.test(data$category,data$sex)
    Pearson's Chi-squared test
data:  data$category and data$sex
X-squared = 0.42895, df = 2, p-value = 0.807
```

Το p-value>5% επομένως δεν έχουμε επαρκή αποδεικτικά στοιχεία για να απορρίψουμε τη μηδενική υπόθεση, δηλαδή το ότι οι δύο μεταβλητές είναι ανεξάρτητες.

Άσκηση 2:

1) Θεωρούμε ένα τυχαίο δείγμα 11 παιχτών μιας ομάδας ποδοσφαίρου. Τα δεδομένα που συλλέξαμε τα περνάμε στην R

```
> data<-c(3,2,1,1,1,2,5,1,3,2,4)
```

Το δείγμα μας εξ ορισμού ακολουθεί γεωμετρική κατανομή (αφού περιγράφει τη θέση της πρώτης «επιτυχίας») με άγνωστη παράμετρο p (πιθανότητα «επιτυχίας»), εφόσον μας δίνει τον αριθμό των προσπαθειών των παιχτών μέχρι να ευστοχήσουν σε χτύπημα πέναλτι στις προπονήσεις πρώτη φορά.

(i) Ψάχνουμε να βρούμε την ΕΜΠ του p αναλυτικά

$$L(p) = \prod_{i=1}^{n}(1-p)^{xi-1} * p$$

$$= (1-p)^{\sum_{i=1}^{n} xi - n} * p^n$$

$$l(p) = \ln(L(p)) = \left(\sum_{i=1}^{n} xi - n\right) * \ln(1-p) + n * \ln(p)$$

$$\frac{dl}{dp} = 0 \Rightarrow \frac{\sum_{i=1}^{n} xi - n}{1-p} = \frac{n}{p}$$

Λύνοντας ως προς p:

$$\hat{p} = \frac{1}{\bar{x}}$$

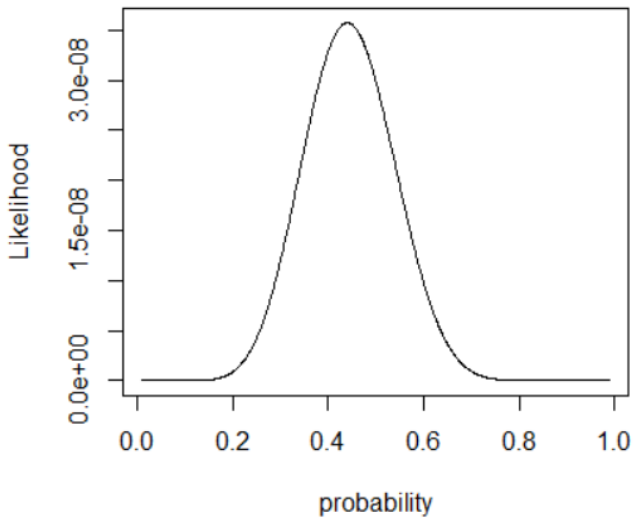
Η ΕΜΠ λοιπόν του p είναι η παραπάνω. Στο συγκεκριμένο δείγμα που συλλέξαμε $\hat{p} = 0.44$

```
> 1/mean(data)
[1] 0.44
```

(ii) Την ΕΜΠ της ίδιας παραμέτρου θα εκτιμήσουμε τώρα με τη βοήθεια της R

```
> data<-c(3,2,1,1,1,2,5,1,3,2,4)
> p<-seq(0.01,0.99,0.001)
> L<-function(data,p){
```

```
+ m=length(p)
+ f<-rep(NA,m)
+ for(i in 1:m){
+ f[i]=prod(dgeom(data,p[i],log=F))}
+ return(f)}
> f<-L(data-1,p)
> plot(p,f,type='l',xlab='probability',ylab='Likelihood')
> p[order(f)[length(f)]]
[1] 0.44
```



Παρατηρώ ότι και με τους δύο τρόπους (αναλυτικά και με τη βοήθεια της R) παίρνουμε το ίδιο αποτέλεσμα.

2) Αρχικά δημιουργούμε μια συνάρτηση που προσομοιώνει τη ρίψη τεσσάρων ζαριών N φορές και στη συνέχεια υπολογίζει τη σχετική συχνότητα με την οποία τα αποτελέσματα όλων των ρίψεων είναι διαφορετικά. Η συνάρτηση στην R χρησιμοποιεί την `sample` για την προσομοίωση των ρίψεων και την `unique` για να ελέγχει αν όλα τα αποτελέσματα είναι διαφορετικά.

```
> simulate_dice_rolls <- function(N) {
+   if (N <= 0 || !is.numeric(N) || N != as.integer(N)) {
+     stop("Το N πρέπει να είναι θετικός ακέραιος")
+   }
```

```
+   count_successes <- 0
+   for (i in 1:N) {
+     rolls <- sample(1:6, 4, replace = TRUE)
+     if (length(unique(rolls)) == 4) {
+       count_successes <- count_successes + 1
+     }
+   }
+   return(count_successes / N)
+ }
```

Στη συνέχεια υπολογίζουμε την θεωρητική πιθανότητα με την εντολή:

```
> theoretical_probability <- (6/6) * (5/6) * (4/6) * (3/6)
```

Με την εντολή:

```
> simulation_result <- simulate_dice_rolls(10000)
```

εκτελούμε τη συνάρτηση για N=10000

Τέλος, συγκρίνουμε την προσομοιωμένη πιθανότητα με τη θεωρητική:

```
> print(paste("Προσομοιωμένη πιθανότητα:", simulation_result))
[1] "Προσομοιωμένη πιθανότητα: 0.2742"
> print(paste("Θεωρητική πιθανότητα:", theoretical_probability))
[1] "Θεωρητική πιθανότητα: 0.277777777777778"
```

Παρατηρούμε ότι είναι ίδιες σε επίπεδο σημαντικότητας δύο δεκαδικών ψηφίων

Η παραπάνω συνάρτηση ελέγχει αν ο αριθμός N είναι θετικός ακέραιος, επαναλαμβάνει τη διαδικασία ρίψης ζαριού N φορές, καταμετρά τις φορές που τα αποτελέσματα είναι όλα διαφορετικά, υπολογίζει την προσομοιωμένη πιθανότητα ως τη σχετική συχνότητα επιτυχιών και υπολογίζει τη θεωρητική πιθανότητα βάσει συνδυαστικής ανάλυσης και συγκρίνει τις δύο τιμές.