ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF APPLIED MATHEMATICAL AND PHYSICAL SCIENCES

Στατιστική Συμπερασματολογία

Ονοματεπώνυμο φοιτητή: Γεώργιος Λεβής
Αριθμός μητρώου: ge19120
Έτος: 8ο
Εξάμηνο: 4ο
Email: ge19120@mail.ntua.gr / levgiorg@gmail.com

Άσκηση 1

Εισάγουμε τα δοθέντα στοιχεία στην R με την παρακάτω εντολή :

```
gifts <-
read.table("http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/gifts.txt", header = TRUE,na.strings="*")
```

Έτσι δηλώνουμε το path του αρχείου .txt, με το argument (header = TRUE) ενημερώνουμε ότι στην πρώτη γραμμή δηλώνονται ονόματα και με το argument (na.strings="*") δηλώνουμε ότι ο χαρακτήρας "*" είναι οι αγνοούμενες τιμές. Η R μας δίνει το πλαίσιο δεδομένων gifts το οποίο είναι ένα data frame. Συνεχίζουμε με τον υπόλοιπο κώδικα.

```
str(gifts)
gifts <- na.omit(gifts)
gifts <- subset(gifts, age >= 18)
```

Χρησιμοποιώ την εντολή (str(gift)) για να δω την δομή του 'gifts'. Έπειτα χρησιμοποιώ την εντολή (gifts <- na.omit(gifts)) για να αφαιρέσω οποιαδήποτε γραμμή περιέχει αγνοουμένη τιμή NA. Η εντολή (gifts <- subset(gifts, age >= 18)) με βοηθάει να αφαιρέσω οποιαδήποτε γραμμή που περιέχει τιμές από μη ενήλικα άτομα. Χρησιμοποιούμε την εντολή subset() γιατί μας επιτρέπει να πάρουμε ένα συγκεκριμένο subset από ένα data frame. Έτσι, στο πρώτο argument επιλέγουμε το data frame που θέλουμε να πάρουμε ('gifts') και στο δεύτερο argument γράφουμε τα conditions, στην περίπτωση μας είναι age >= 18 με το οποίο κρατάμε όλες τις σειρές που έχουν τιμή ίση ή πάνω από 18.

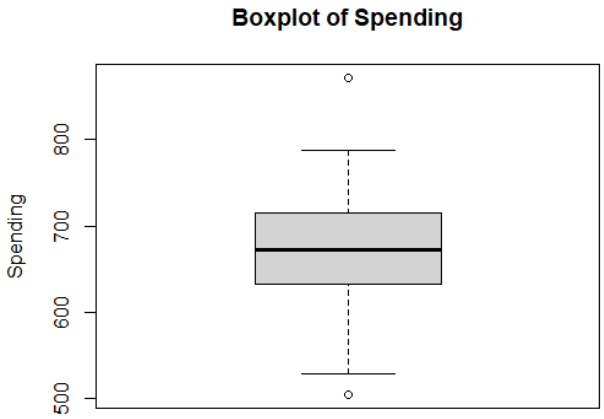
i)

Αρχικά θα ορίσουμε τις 6 μεταβλητές από τις στήλες εξαρχής για δική μας ευκολία.

```
spend <- gifts$spend
age <- gifts$age
holiday <-gifts$holiday
sex <- gifts$sex
time <- gifts$time
salary <- gifts$salary
```

Θα κάνουμε περιγραφική στατιστική για να πάρουμε μια ιδέα για το αν οι άνθρωποι ξοδεύουν, κατά μέσο όρο, περισσότερα από 500 ευρώ κατά τη διάρκεια των εορταστικών περιόδων. Θα δημιουργήσουμε ένα θηκόγραμμα και θα υπολογίσουμε και τις πέντε σημαντικές τιμές του.

```
boxplot(gifts$spend, ylab = "Spending", main = "Boxplot of Spending")
fivenum(gifts$spend)
```



>fivenum(gifts$spend)

[1] 504.250 632.305 672.410 715.480 872.420

Με την πρώτη ματιά, μπορούμε να δούμε ότι η μέση δαπάνη μπορεί να είναι μεγαλύτερη από 500 ευρώ. Τώρα, μπορούμε να προχωρήσουμε σε περεταίρω αναλυση:

Γράφουμε:

```
hist(gifts$spend, xlab = "Spending", ylab = "Frequency", main =
"Histogram of Spending")
qqnorm(gifts$spend, main = "Normal Q-Q Plot of Spending", ylab = "Sample
Quantiles", xlab = "Theoretical Quantiles")
qqline(gifts$spend)
```

Από τα παραπάνω διαγράμματα βλέπουμε ότι η υπόθεση της κανονικότητας δεν είναι παράλογη. Σε κάθε περίπτωση, κάνουμε και ένα Shapiro και Wilk normality test για να είμαστε περισσότερο σίγουροι. Γράφουμε:

> shapiro.test(gifts$spend)

Shapiro-Wilk normality test

data: gifts$spend

W = 0.98978, p-value = 0.937

Η τιμή p για το τέστ Shapiro-Wilk είναι μεγαλύτερη από 0,01. Επομένως, δεν έχουμε ισχυρά στοιχεία για να απορρίψουμε τη μηδενική υπόθεση της κανονικότητας. Στη συνέχεια, μπορούμε να one-sample t-test με μια μονόπλευρη εναλλακτική υπόθεση για να ελέγξουμε εάν το μέσο ποσό που δαπανήθηκε κατά τη διάρκεια της εορταστικής περιόδου είναι μεγαλύτερο από 500 ευρώ. Η μηδενική υπόθεση είναι ότι το μέσο ποσό που δαπανήθηκε είναι ίσο με 500 ευρώ και η εναλλακτική υπόθεση είναι ότι το μέσο ποσό που δαπανήθηκε είναι μεγαλύτερο από 500 ευρώ. Κατασκευάζουμε ένα (1-α)% = 99% διάστημα εμπιστοσύνης για τη μέση τιμή του

damage και ελέγχουμε αν η υποτιθέμενη τιμή μ_0 ανήκει $(\bar{x} + \frac{t_{n-1,a}S}{\sqrt{n}}, +\infty)$. Δίνουμε, λοιπόν, στην R την εντολή :

```
t.test(gifts$spend, mu = 500, alternative = "greater", conf.level =
0.99)
```

```
> t.test(gifts$spend, mu = 500, alternative = "greater", conf.level = 0.99)

        One Sample t-test

data:  gifts$spend
t = 16.706, df = 50, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 500
99 percent confidence interval:
 646.5072      Inf
sample estimates:
mean of x
 671.1253
```

Η τιμή t είναι θετική και η τιμή p είναι μικρότερη από 0,01, γεγονός που παρέχει ισχυρές ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Εν κατακλείδι, με βάση τα δεδομένα που έχουν δοθεί, έχουμε ισχυρά στοιχεία για την απόρριψη της μηδενικής υπόθεσης, ότι τα άτομα ξοδεύουν, κατά μέσο όρο, 500 ευρώ κατά τη διάρκεια των εορταστικών περιόδων. Τα αποτελέσματα της ανάλυσής μας υποδηλώνουν ότι το μέσο ποσό που δαπανήθηκε είναι μεγαλύτερο από 500 ευρώ.

ii)

Για να διερευνήσουμε εάν οι καταναλωτές των Χριστουγέννων ξοδεύουν περισσότερα σε σύγκριση με άλλες εορταστικές περιόδους, θα δημιουργήσουμε μια νέα κατηγορική μεταβλητή που ονομάζεται "Χριστούγεννα" με δύο κατηγορίες: 0 για τις μη χριστουγεννιάτικες εορταστικές περιόδους και 1 για τις εορταστικές περιόδους των Χριστουγέννων. Στη συνέχεια θα περιγράψουμε τις τιμές του ποσού που δαπανήθηκε ξεχωριστά για κάθε κατηγορία. Γράφουμε :

```
gifts$Christmas <- ifelse(gifts$holiday == "Christmas", 1, 0)
non_christmas_spend <- gifts[gifts$Christmas == 0, ]$spend
summary(non_christmas_spend)
```

```
fivenum(non_christmas_spend)
christmas_spend <- gifts[gifts$Christmas == 1, ]$spend
summary(christmas_spend)
fivenum(christmas_spend)
boxplot(non_christmas_spend, ylab = "Spending", main = "Boxplot of
Spending (Non-Christmas)")
hist(non_christmas_spend, breaks = 10, xlab = "Spending", ylab =
"Frequency", main = "Histogram of Spending (Non-Christmas)")
boxplot(christmas_spend, ylab = "Spending", main = "Boxplot of Spending
(Christmas)")
hist(christmas_spend, breaks = 10, xlab = "Spending", ylab =
"Frequency", main = "Histogram of Spending (Christmas)")
boxplot(spend ~ Christmas, data = gifts, main = "Expenditure by festive
period", xlab = "Festive period", ylab = "Expenditure", names = c("Non-
Christmas", "Christmas"))
```

Η πρώτη γραμμή του παραπάνω κώδικα μας δημιουργεί μια νέα μεταβλητή στο dataframe gifts που ονομάζεται Christmas. Η συνάρτηση ifelse() ελέγχει εάν το holiday για κάθε σειρά είναι ίση με "Christmas". Εάν είναι, αντιστοιχίζει ένα 1 στην αντίστοιχη σειρά στη στήλη του Christmas. Εάν δεν είναι, εκχωρεί ένα 0.

Η δεύτερη γραμμή δημιουργεί μια νέα μεταβλητή non_christmas_spend που εξάγει όλες τις τιμές από το spend όπου τα Χριστούγεννα είναι 0 (περίοδοι εκτός Χριστουγέννων). Το gifts$Christmas == 0 φιλτράρει τις σειρές όπου τα Χριστούγεννα είναι 0.

Η τρίτη και τέταρτη γραμμή υπολογίζουν το summary και το fivenum της non_christmas_spend μεταβλητής, η οποία περιλαμβάνει το ελάχιστο, 1ο, διάμεσο, μέσο όρο, 3ο και μέγιστο και αλλα.

```
summary(non_christmas_spend)
   Min. 1st Qu.  Median    Mean 3rd Qu.
Max.
  504.2   593.3   654.7   653.6   700.5
788.4
```

```
> fivenum(non_christmas_spend)
[1] 504.25 593.29 654.71 700.47
788.44
```

Αντίστοιχα στην πέμπτη γραμμή κάνουμε ότι κάναμε και στην δεύτερη, δηλαδή δημιουργεί μια νέα μεταβλητή christmas_spend που εξάγει όλες τις τιμές από το spend όπου το Christmas είναι 1 (περίοδοι Χριστουγέννων).

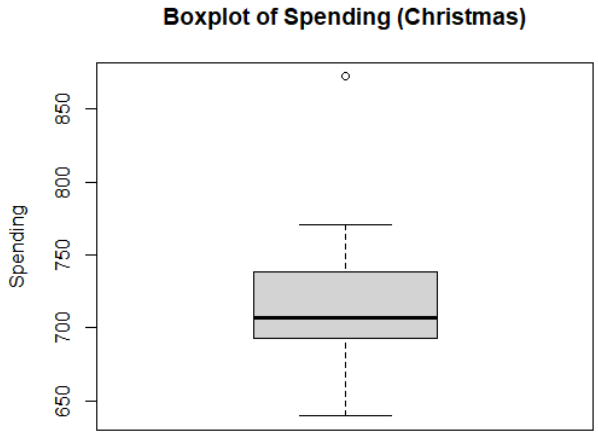
Όμοια στην έκτη και την εβδόμη γραμμή υπολογίζουμε το summary και το fivenum:

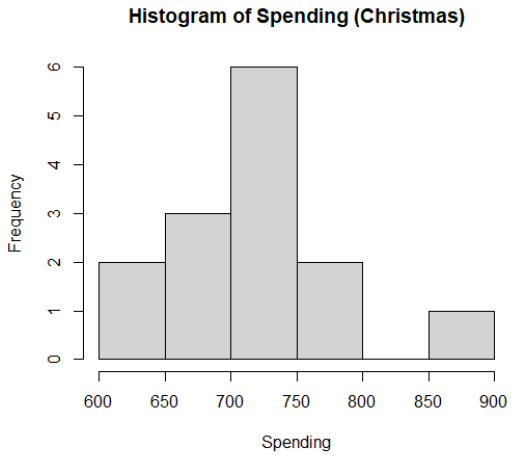
```
> summary(christmas_spend)
   Min. 1st Qu.  Median    Mean 3rd Qu.
Max.
  639.5   694.0   707.2   717.5   734.4
872.4
```

```
> fivenum(christmas_spend)
[1] 639.48 692.57 707.25 738.66
872.42
```

Τέλος στο τελευταίο κομμάτι του κώδικα δημιουργούμε διάφορες γραφικές παραστάσεις συγκρίνει τις δαπάνες σε περιόδους εκτός Χριστουγέννων, Χριστουγέννων και τις 2 μαζί.

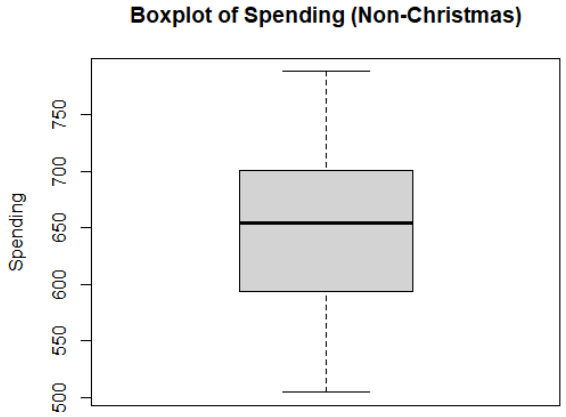
Christmas

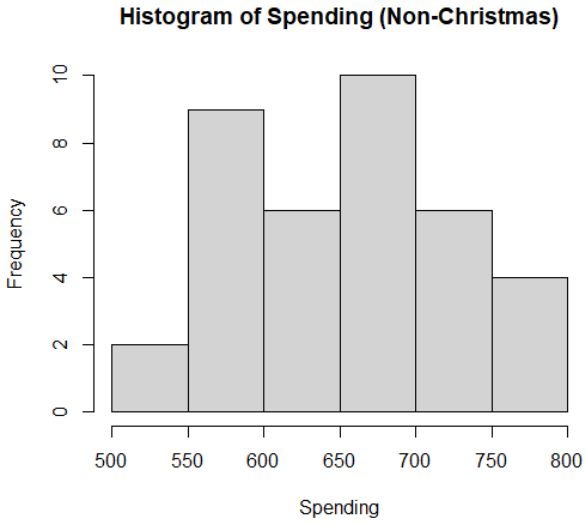




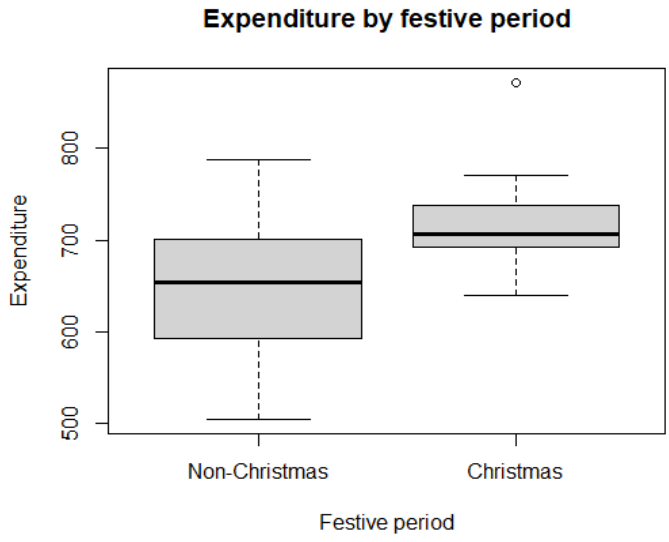
Από το ιστόγραμμα βλέπουμε ότι η πλειοψηφία των ανθρώπων τείνουν να ξοδεύουν μεταξύ 700 και 750 ευρώ κατά την εορταστική περίοδο των Χριστουγέννων. Αυτό το εύρος δαπανών είναι το πιο κοινό, καθώς έχει την υψηλότερη συχνότητα εμφανίσεων στο ιστόγραμμα. Επίσης βλέπουμε ότι η ελάχιστη δαπάνη κατά τη διάρκεια των Χριστουγέννων αναφέρεται ότι είναι κάτω από 650 ευρώ, γεγονός που δείχνει ότι υπάρχουν ορισμένα άτομα που ξοδεύουν σχετικά λιγότερα. Το πρώτο τεταρτημόριο είναι περίπου 700 ευρώ, υποδεικνύοντας ότι το 25% των αξιών δαπανών συγκεντρώνεται γύρω από αυτό το εύρος. Ο διάμεσος είναι λίγο μεγαλύτερος από 700 ευρώ, υποδηλώνοντας ότι η κατανομή των δαπανών τείνει προς το χαμηλότερο εύρος. Το τρίτο τεταρτημόριο είναι περίπου 750 ευρώ, υποδηλώνοντας ότι το 75% των αξιών δαπανών πέφτει κάτω από αυτήν την τιμή. Έχουν χρησιμοποιηθεί αριθμητικοί μέθοδοι (summary & fivenum) για επιβεβαιωθούν τα λεγομενά μου

Non-Christmas





Από το ιστόγραμμα, μπορούμε να παρατηρήσουμε ότι η πλειοψηφία των ανθρώπων τείνουν να ξοδεύουν μεταξύ 600 και 700 ευρώ σε περιόδους εκτός Χριστουγγένων. Η συχνότητα των δαπανών μειώνεται καθώς το ποσό αποκλίνει από αυτό το εύρος. Το σχήμα του ιστογράμματος υποδηλώνει μια σχετικά συμμετρική κατανομή, χωρίς εμφανείς ακραίες τιμές. Συνοπτικά, τα στοιχεία δείχνουν ότι σε περιόδους εκτός των Χριστουγέννων, τα περισσότερα άτομα τείνουν να ξοδεύουν μεταξύ 600 και 700 ευρώ, με μέση δαπάνη 654,7 ευρώ. Η κατανομή των δαπανών κατά τη διάρκεια αυτών των περιόδων φαίνεται να είναι σχετικά συνεπής, χωρίς σημαντικές αποκλίσεις ή ακραίες τιμές.



Φαίνεται ότι η περίοδος των Χριστουγέννων περιλαμβάνει γενικά υψηλότερα επίπεδα δαπανών, όπως υποδεικνύεται από το υψηλότερο διάμεσο, μέσο όρο και μέγιστη τιμή δαπανών σε σύγκριση με την περίοδο εκτός των Χριστουγέννων.

iii)

Θα κάνουμε ένα two-sample t-test για να ελέγξουμε τον ισχυρισμό της ομάδας.

```
t.test(spend ~ Christmas, data = gifts, alternative = "two.sided")
```

```
> t.test(spend ~ Christmas, data = gifts, alternative = "two.sided")

        Welch Two Sample t-test

data:  spend by Christmas
t = -3.3221, df = 29.32, p-value = 0.002403
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -103.23521  -24.58298
sample estimates:
mean in group 0 mean in group 1
       653.5816        717.4907
```

Από το output παρατηρούμε ότι:

t = -3,3221: Αυτή είναι η t-value. Η αρνητική τιμή υποδεικνύει ότι ο μέσος όρος της ομάδας 0 είναι μικρότερος από τον μέσο όρο της ομάδας 1.

df = 29,32: Πρόκειται για τους βαθμούς ελευθερίας για το t-test, που σχετίζεται με τον αριθμό των παρατηρήσεων στα δεδομένα μας. Οι βαθμοί ελευθερίας χρησιμοποιούνται για τον υπολογισμό της τιμής p.

p-value = 0,002403: Η p-value είναι ένα μέτρο της ισχύος των αποδεικτικών στοιχείων που υποστηρίζουν μια μηδενική υπόθεση. Η τιμή του p εδώ είναι μικρότερη από 0,05, γεγονός που υποδεικνύει ισχυρές ενδείξεις έναντι της μηδενικής υπόθεσης στο επίπεδο σημαντικότητας 5%. Ως εκ τούτου, θα απορρίψουμε τη μηδενική υπόθεση και θα συμπεράνουμε ότι υπάρχει στατιστικά σημαντική διαφορά στη μέση δαπάνη μεταξύ των περιόδων Χριστουγέννων και μη Χριστουγέννων.

Διάστημα εμπιστοσύνης 95 %: [-103,23521 -24,58298]. Δεδομένου ότι το διάστημα δεν περιέχει 0, ενισχύει την περίπτωση κατά της μηδενικής υπόθεσης. mean in group 0 mean in group 1 (Non-Christmas) = 653,5816, mean in group 1 (Christmas) = 717,4907: Αυτοί είναι οι μέσοι όροι του δείγματος για τις δύο ομάδες. Η μέση δαπάνη κατά τη διάρκεια των Χριστουγέννων είναι υψηλότερη από ό,τι σε περιόδους εκτός των Χριστουγέννων, γεγονός που ευθυγραμμίζεται με την πεποίθηση της ερευνητικής ομάδας. Συνοπτικά, τα αποτελέσματα των δοκιμών υποστηρίζουν τον ισχυρισμό της ερευνητικής ομάδας: οι άνθρωποι φαίνεται να ξοδεύουν πολύ περισσότερα κατά την περίοδο των Χριστουγέννων σε σύγκριση με τις περιόδους εκτός των Χριστουγέννων.

iv)

Εάν ο προϊστάμενος του τμήματος πιστεύει ότι δεν υπάρχει διαφορά στο μέσο χρηματικό ποσό που δαπανάται κατά τη διάρκεια της περιόδου των διακοπών σε σχέση με το φύλο ενός ατόμου, μπορούμε να ελέγξουμε αυτόν τον ισχυρισμό χρησιμοποιώντας ένα two sample t-test. Η μηδενική υπόθεση (H_0) είναι ότι δεν υπάρχει διαφορά στη μέση δαπάνη μεταξύ των φύλων, ενώ η εναλλακτική υπόθεση (H_1) είναι ότι υπάρχει διαφορά. Τρέχουμε τον παρακάτω κώδικα

```
t.test(spend ~ sex, data = gifts, alternative = "two.sided", conf.level
= 0.90)
```

και παίρνουμε

```
> t.test(spend ~ sex, data = gifts, alternative = "two.sided" , conf.level = 0.90)

        Welch Two Sample t-test

data:  spend by sex
t = 0.99837, df = 48.724, p-value = 0.323
alternative hypothesis: true difference in means between group Man and group
Woman is not equal to 0
90 percent confidence interval:
 -13.88547  54.75677
sample estimates:
  mean in group Man mean in group Woman
          681.9442          661.5085
```

Τα αποτελέσματα της ανάλυσης έδειξαν ότι η t-statistic είναι 0,99837 με βαθμούς ελευθερίας 48,724. Με βάση το επιλεγμένο επίπεδο σημαντικότητας 10%, και με το p-value να είναι 0,323 βλέπουε ότι το p-value είναι μεγαλύτερο από το επίπεδο . Επομένως, δεν έχουμε επαρκή στοιχεία για να απορρίψουμε τη μηδενική υπόθεση.

Αυτό δείχνει ότι δεν υπάρχει στατιστικά σημαντική διαφορά στο μέσο χρηματικό ποσό που δαπανάται κατά την περίοδο των διακοπών μεταξύ ανδρών και γυναικών.

v)

Για να διερευνήσουμε εάν η πιθανότητα να ξοδέψουν ένα μεγάλο ποσό κατά την περίοδο των εορτών είναι μεγαλύτερη για τις γυναίκες σε σχέση με τους άνδρες, μπορούμε να δημιουργήσουμε μια νέα κατηγορική μεταβλητή που ονομάζεται "spend_bin" με δύο κατηγορίες: 0 για ένα μικρό ποσό και 1 για ένα μεγάλο ποσό. Θα ορίσουμε το threshold στα 673 ευρώ, όπου ποσά μικρότερα από 673 ευρώ θα κατηγοριοποιούνται ως 0 (μικρό ποσό) και ποσά μεγαλύτερα ή ίσα με 673 ευρώ θα ταξινομούνται ως 1 (μεγάλο ποσό). Για να πραγματοποιήσουμε την ανάλυση, θα χρησιμοποιήσουμε ένα τεστ X^2 ανεξαρτησίας για να εξετάσουμε εάν υπάρχει συσχέτιση μεταξύ του φύλου και της πιθανότητας να ξοδευτεί ένα μεγάλο ποσό κατά την περίοδο των γιορτών. Η μηδενική υπόθεση (H_0) υποθέτει ότι δεν υπάρχει συσχέτιση του φύλου και της πιθανότητας να ξοδευτεί ένα μεγάλο ποσό κατά την περίοδο των διακοπών. Η εναλλακτική υπόθεση (H_1), από την άλλη πλευρά, υποδηλώνει ότι υπάρχει συσχέτιση ή σχέση μεταξύ των μεταβλητών. Χρησιμοποιούμε τον ακόλουθο κώδικα:

```
gifts$spend_bin <- ifelse(gifts$spend < 673, 0, 1)
cont_table <- table(gifts$sex, gifts$spend_bin)
chi_result <- chisq.test(cont_table)
chi_result
```

και παίρνουμε :

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 0.17027, df = 1, p-value = 0.6799
```

Τα αποτελέσματα του τεστ ανεξαρτησίας X^2 δείχνουν ότι το X^2 είναι 0,17027 με 1 βαθμό ελευθερίας και το p-value είναι 0,6799. Με βάση το επιλεγμένο επίπεδο σημαντικότητας 2% (α = 0,02), η τιμή του p 0,6799 είναι πολύ μεγαλύτερη από το επίπεδο σημαντικότητας. Επομένως, δεν έχουμε επαρκή στοιχεία για να απορρίψουμε τη μηδενική υπόθεση. Αυτό υποδηλώνει ότι δεν υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ του φύλου και της πιθανότητας να ξοδευτεί ένα μεγάλο ποσό κατά την περίοδο των διακοπών.

vi)

Για να ελέγξουμε εάν το φύλο του ατόμου είναι ανεξάρτητο από την περίοδο των εορτών κατά την οποία έγιναν οι αγορές, μπορούμε να χρησιμοποιήσουμε το τεστ X^2 ανεξαρτησίας, παρόμοιο με την προηγούμενη ερώτηση. Με μηδενική υπόθεση (H_0) να είναι οι μεταβλητές ανεξάρτητες η μία από την άλλη. Δηλαδή ότι το φύλο ενός ατόμου είναι ανεξάρτητο από την περίοδο των εορτών κατά την οποία έγιναν οι αγορές. Η εναλλακτική υπόθεση (H_1) είναι η αντίθετη από τη μηδενική υπόθεση, θα είναι ότι το φύλο ενός ατόμου δεν είναι ανεξάρτητο από την περίοδο των εορτών κατά την οποία έγιναν οι αγορές. Παρακάτω δίνεται ο κώδικας:

```
cont_table <- table(gifts$sex, gifts$holiday)
chi_result1 <- chisq.test(cont_table)
chi_result1
```

Συγκρίνουμε το p-value με το επίπεδο σημαντικότητας. Δεδομένου ότι η τιμή p (0,1026) είναι μεγαλύτερη από το επίπεδο σημαντικότητας (0.05), δεν έχουμε επαρκή στοιχεία για να απορρίψουμε τη μηδενική υπόθεση. Αυτό υποδηλώνει ότι δεν υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ του φύλου ενός ατόμου και της περιόδου διακοπών κατά την οποία έγιναν οι αγορές. Με βάση την ανάλυση σε επίπεδο σημαντικότητας 5%, δεν βρίσκουμε σημαντικά στοιχεία που να υποδηλώνουν ότι το φύλο ενός ατόμου εξαρτάται από την περίοδο των εορτών κατά την οποία έγιναν οι αγορές. Ωστόσο, είναι σημαντικό να σημειωθεί ότι η τιμή p είναι σχετικά κοντά στο επίπεδο σημαντικότητας, υποδεικνύοντας μια οριακή συσχέτιση.

Άσκηση 2

1)

Πριν ξεκινήσουμε, χρειάζεται να ξεκαθαρίσουμε το πλαίσιο. Εξετάζουμε τον αριθμό των προσπαθειών από 10 αθλητές στο τμήμα μπάσκετ ενός δημοτικού σχολείου μέχρι να πετύχουν ένα τρίποντο από μια συγκεκριμένη θέση. Υποθέτουμε ότι οι αριθμοί που δίνονται αντιπροσωπεύουν τον αριθμό των προσπαθειών που χρειάστηκε κάθε αθλητής για να κάνει το πρώτο επιτυχημένο τρίποντο. Σε αυτό το σενάριο, μπορούμε να μοντελοποιήσουμε τα δεδομένα ως μια γεωμετρική κατανομή με μια άγνωστη παράμετρο p, όπου το p αντιπροσωπεύει την πιθανότητα να γίνει ένα επιτυχημένο τρίποντο σε μία μόνο προσπάθεια. Για να βρούμε την Ε.Μ.Π. θα δουλέψουμε την λογαριθμική πιθανοφάνεια. Εκφράζουμε με την τυχαία μεταβλητή x_i τις προσπάθειες που χρειάστηκε ο δοκιμαστής i μέχρι να νικήσει. Έχουμε:

$$l(p) = \log\big(L(p)\big) = \sum_{i=1}^{n} \log((1-p)^{x_i-1}p) =$$

$$= \sum_{i=1}^{n} (\log((1-p)^{x_i-1}) + \log(p)\,) = \log(1-p)\sum_{i=1}^{n}(x_i - 1) + n\log(p)$$

$$= n\log(p) - \log(1-p)\left(n - \sum_{i=1}^{n} x_i\right)$$

Παραγωγίζουμε και εξισώνουμε με το 0. Έχουμε:

$$l'(p) = 0 \Leftrightarrow \frac{n}{p} + \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \Rightarrow n(1-p) + np - p\sum_{i=1}^{n} x_i = 0$$

$$p\sum_{i=1}^{n} x_i = n \Rightarrow p = \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^{-1} \Leftrightarrow p = \frac{1}{\bar{x}}$$

Ελέγχουμε τη δεύτερη παράγωγο για να επιβεβαιώσουμε πως πράγματι είναι μέγιστο. Έχουμε:

$$l''(p) = -\frac{n}{p^2} + \frac{n - \sum_{i=1}^{n} x_i}{(1-p)^2} < 0$$

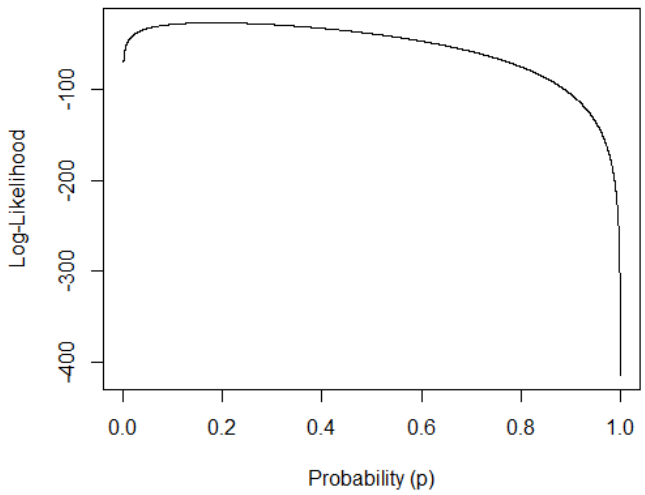
διότι κάθε x_i θα είναι ≥ του 1, άρα $n - \sum_{i=1}^{n} x_i \leq 0$ και n, p^2 >0, άρα $\frac{n}{p^2} < 0$. Άρα, το $\hat{p} = \frac{1}{\bar{x}}$ είναι πράγματι μέγιστο και άρα είναι η Ε.Μ.Π. Για τις συγκεκριμένες παρατηρήσεις, η τιμή της εκτιμήτριας είναι:

$$\hat{p} = \frac{10}{3+2+1+5+2+4+5+2+3+10} \Leftrightarrow \hat{p} = 0.27$$

Ο παρακάτω κώδικας στην R χρησιμοποιείται για τον υπολογισμό της λογαριθμικής γεωμετρικής πιθανότητας για το δεδομένο δείγμα και 10.000 διαφορετικές τιμές του p. Η συνάρτηση geom_loglikelihood παίρνει δύο ορίσματα: "data" (το δείγμα των παρατηρούμενων τιμών) και p (το διάνυσμα των πιθανοτήτων προς αξιολόγηση). Υπολογίζει τη log-γεωμετρική πιθανότητα για κάθε τιμή του p και επιστρέφει ένα διάνυσμα αποτελεσμάτων. Για να απεικονίσουμε τα αποτελέσματα, ο κώδικας χρησιμοποιεί τη συνάρτηση "plot" για να δημιουργήσει ένα γράφημα. Ο άξονας x αντιπροσωπεύει τις τιμές του p, που κυμαίνονται από 0,001 έως 1 με 10.000 σημεία. Ο άξονας y αντιπροσωπεύει τις αντίστοιχες τιμές λογαριθμικής πιθανότητας που υπολογίζονται χρησιμοποιώντας το δεδομένο δείγμα. Το γράφημα σχεδιάζεται ως γραμμικό γράφημα (τύπος = "l"):

```r
p <- seq(0.001, 1, length = 10000)
geom_loglikelihood <- function(data, p) {
  results <- rep(NA, 10000)
  for (i in 1:10000) {
    results[i] <- sum(dgeom(data - 1, p[i], log = TRUE))
  }
  return(results)
}

x <- c(3, 2, 1, 5, 2, 4, 5, 2, 3, 10)
results <- geom_loglikelihood(x, p)
plot(p, results, xlab = "Probability (p)", ylab = "Log-Likelihood", type
= "l")
```



Για να βρούμε εμπειρικά το μέγιστο (με ακρίβεια 4 δεκαδικών) και το πραγματικό μέγιστο, γράφουμε τις εντολές:

```r
p_results <- round(p[order(results)[10000]], 4)
theoretical_p= round(1/mean(x),4)
```

και παίρνουμε :

```
> P_result
[1] 0.2703
```

```
> theoretical_p
[1] 0.2703
```

Ταυτίζονται!! Κάτι το οποίο επιβεβαιώνει τους υπολογισμούς μας.

2)

Δίνουμε τον κώδικα και μετά θα τον περιγράψουμε:

```
three_dice_rolls <- function(N) {
  if (!is.numeric(N) || N <= 0 || round(N) != N) {
    stop("N must be a positive integer.")
  }

  success_count <- 0

  for (i in 1:N) {
    dice_rolls <- sample(1:6, size = 3, replace = TRUE)

    if (dice_rolls[1] == dice_rolls[3] && dice_rolls[1] !=
dice_rolls[2]) {
      success_count <- success_count + 1
    }
  }

  relative_frequency <- success_count / N
  return(relative_frequency)
}
```

Ορίζουμε το function three_dice_rolls παίρνει ένα όρισμα, το N, το οποίο αντιπροσωπεύει τον αριθμό των δοκιμών που θα εκτελεστούν. Ελέγχουμε πρώτα αν το N ειναι θετικός ακέραιος. Εάν δεν είναι θετικός ακέραιος, η συνάρτηση σταματά και επιστρέφει ένα μήνυμα σφάλματος: "N must be a positive integer." Στη συνέχεια αρχικοποιούμε έναν μετρητή, success_count, για να παρακολουθούμε τον αριθμό των επιτυχημένων δοκιμών. Μια επιτυχημένη δοκιμή είναι αυτή όπου το πρώτο και το τρίτο ζάρι είναι το ίδιο και το δεύτερο παιχνίδι είναι διαφορετικό από το πρώτο. Για κάθε δοκιμή (από 1 έως N), προσομοιώνουμε τρία ζάρια χρησιμοποιώντας τη συνάρτηση sample. Αυτή η συνάρτηση επιλέγει τυχαία 3 αριθμούς από το 1 έως το 6 (που αντιπροσωπεύουν τις όψεις των ζαριών). Μετά από κάθε δοκιμή, ελέγχουμε εάν οι πρώτες και οι τρίτες ζαριές είναι ίδιες και διαφορετικές από την δεύτερη ζαριά. Εάν συμβαίνει αυτό, αυξάνετε κατά ένα το success_count. Αφού ολοκληρωθούν όλες οι δοκιμές, υπολογίζουμε τη σχετική συχνότητα των επιτυχημένων δοκιμών διαιρώντας το success_count με το N. Αυτή η τιμή στη συνέχεια επιστρέφεται ως αποτέλεσμα της συνάρτησης.