ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Θέμα Εργασίας: Εισαγωγή στην R και Περιγραφική Στατιστική

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.

i) Για την ποσοτική μεταβλητή age δίνοντας την εντολή

> summary(data$age)

παίρνουμε το αποτέλεσμα

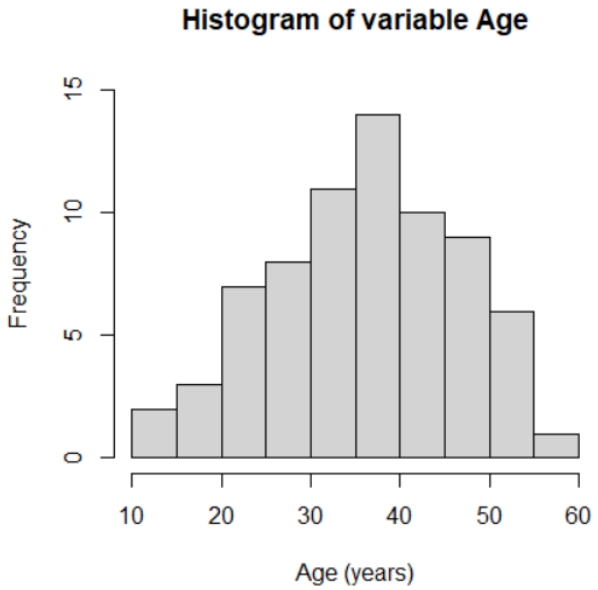
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
13.00	28.00	36.00	36.41	44.00	59.00

Το νεαρότερο άτομο του δείγματος μας είναι 13 χρονών. Το πρώτο τεταρτημόριο (1st Qu.) δείχνει την τιμή για την οποία το 25% των παρατηρήσεων είναι μικρότερες από αυτή. Το τρίτο τεταρτημόριο (3rd Qu.) δείχνει την τιμή για την οποία το 75% των παρατηρήσεων μας είναι μικρότερες από αυτή. Το δεύτερο τεταρτημόριο που ταυτίζεται και με τη διάμεσο του δείγματος (Median), δείχνει την τιμή για την οποία το 50% των παρατηρήσεων είναι μικρότερες ή το 50% των παρατηρήσεων είναι μεγαλύτερες από αυτή. Η μέγιστη τιμή (Max.) της μεταβλητής age στο δείγμα είναι 59 χρονών. Τέλος, ο δειγματικός μέσος (Median) μας δείχνει την τιμή για τον μέσο όρο των παρατηρήσεων.

Δίνοντας στη συνέχεια την εντολή

> hist(data$age, main='Histogram of variable Age',xlab='Age (years)', ylim=c(0,15))

παίρνουμε το ιστόγραμμα:



Ο οριζόντιος άξονας εκφράζει την ηλικία και ο κατακόρυφος τις συχνότητες των παρατηρούμενων τιμών για τις κλάσεις που δημιουργήθηκαν. Τα περισσότερα άτομα του δείγματος έχουν ηλικίες από 35 μέχρι 40 χρονών (κορυφή ιστογράμματος). Η κλάση των ηλικιών από 55 μέχρι 60 χρονών περιέχει τα λιγότερα άτομα.

Δίνοντας την εντολή

> boxplot(data$age,main='Boxplot of variable Age')

παίρνουμε το θηκοδιάγραμμα

Τις τιμές των πέντε στατιστικών μεγεθών που χρησιμοποιούμε για την κατασκευή ενός θηκοδιαγράμματος μπορούμε να τις πάρουμε μέσω της εντολής

> fivenum(data$age)

[1] 13 28 36 44 59

Η πρώτη τιμή (13) αντιστοιχεί στη μικρότερη παρατηρούμενη τιμή του δείγματος για τη μεταβλητή age. Η δεύτερη τιμή (28) αντιστοιχεί στο πρώτο τεταρτημόριο των τιμών της μεταβλητής, ενώ η τέταρτη (44) στο τρίτο τεταρτημόριο. Η τρίτη τιμή (36) αντιστοιχεί στη διάμεσο των παρατηρήσεων. Η τελευταία τιμή (59) αντιστοιχεί στη μέγιστη τιμή των παρατηρήσεων.

Όπως είδαμε και προηγουμένως, το νεαρότερο άτομο του δείγματος μας είναι 13 χρονών και το μεγαλύτερο 59 χρονών. Τα μισά άτομα του δείγματος έχουν ηλικία μικρότερη ή ίση με 36 χρονών. Το 25% του πληθυσμού του δείγματος έχει ηλικία το πολύ 28 χρονών, ενώ το 75% έχει ηλικία το πολύ 44 χρονών.

Στο θηκοδιάγραμμα αυτό (αλλά και στα επόμενα) η μικρότερη παρατηρούμενη τιμή της ποσοτικής μεταβλητής που εξετάζουμε παρουσιάζεται ως η χαμηλότερη οριζόντια γραμμή του σχήματος. Οι τιμές που αντιστοιχούν στο πρώτο και τρίτο τεταρτημόριο παρουσιάζονται στο θηκοδιάγραμμα ως η δεύτερη και τέταρτη οριζόντια γραμμή (αντίστοιχα) ξεκινώντας από κάτω. Η διάμεσος αναπαρίσταται ως η έντονη οριζόντια γραμμη του διαγράμματος. Η μέγιστη τιμή παρουσιάζεται ως η υψηλότερη οριζόντια γραμμή του διαγράμματος.

Για την κατηγορική μεταβλητή **category**, δίνοντας την εντολή

```
> table(data$category)

  cosmetics  healthcare     other
         19          21        31
```

Βλέπουμε τις συχνότητες των κατηγοριών, ενώ με την εντολή

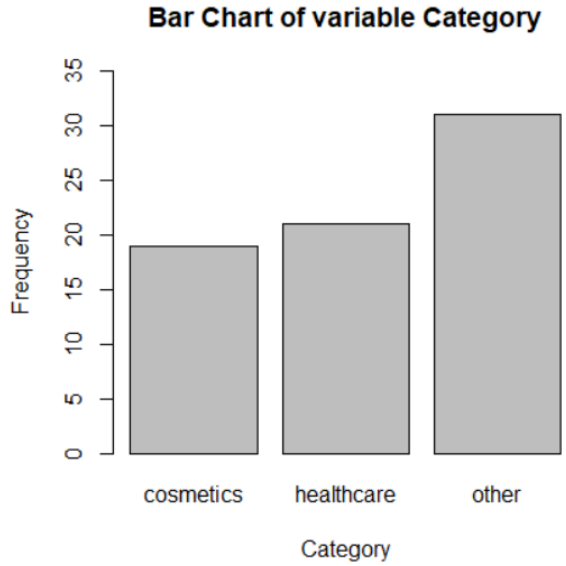
```
> prop.table(table(data$category))

  cosmetics   healthcare          other
0.2676056    0.2957746    0.4366197
```

Βλέπουμε τις αντίστοιχες σχετικές συχνότητες.

Με την εντολή

```
> barplot(table(data$category),xlab='Category',ylab='Frequency', ylim=c(0,35))
```

Παίρνουμε το ραβδοδιάγραμμα:

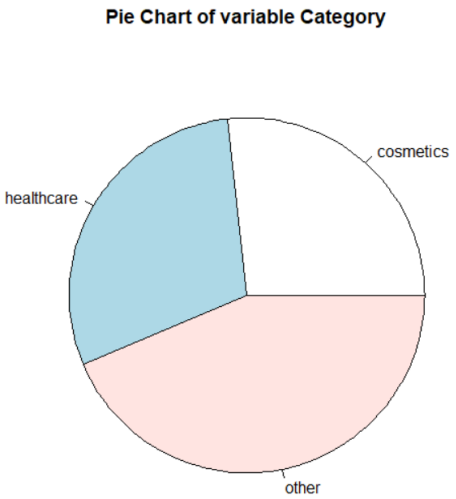


Όπου στον οριζόντιο άξονα βλέπουμε τις κατηγορίες του μη φαρμακευτικού προϊόντος και στον κατακόρυφο τις αντίστοιχες συχνότητες. Παρατηρούμε ότι μεγαλύτερη συχνότητα εμφανίζει η κατηγορία “other” ενώ μικρότερη συχνότητα η κατηγορία “cosmetics”.

Με την εντολή

```
> pie(table(data$category),main='Pie Chart of variable Category')
```

Παίρνουμε το τομεόγραμμα:



Όπου και πάλι το εμβαδόν κάθε τομέα είναι ανάλογο της σχετικής συχνότητας της αντίστοιχης κατηγορίας. Από τις γραφικές μεθόδους επιβεβαιώνεται αυτό που παρατηρούμε και από τις αριθμητικές μεθόδους, δηλαδή ότι μεγαλύτερη συχνότητα εμφανίζει η κατηγορία “other” ενώ μικρότερη συχνότητα η κατηγορία “cosmetics”.

Αντίστοιχα, για την ποιοτική μεταβλητή **sex** με την εντολή

> table(data$sex)

Man Woman

37 34

Βλέπουμε τις συχνότητες των κατηγοριών της μεταβλητής, ενώ με την εντολή

> prop.table(table(data$sex))

Man Woman

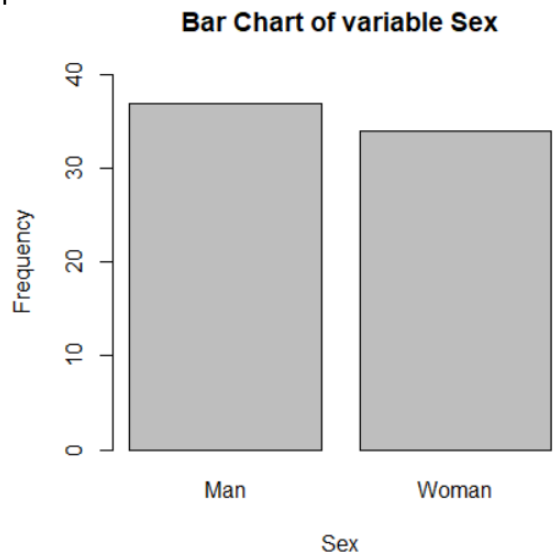
0.5211268 0.4788732

Βλέπουμε τις αντίστοιχες σχετικές συχνότητες.

Με την εντολή

> barplot(table(data$sex),main='Bar Chart of variable Sex', xlab='Seasons', ylab='Frequency', ylim=c(0,40))

παίρνουμε το ραβδοδιάγραμμα:

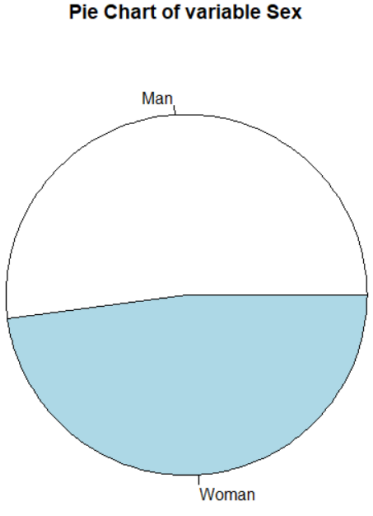


Όπου στον οριζόντιο άξονα βλέπουμε το φύλο και στον κατακόρυφο τις αντίστοιχες συχνότητες. Παρατηρούμε ότι οι άντρες είναι περισσότεροι από τις γυναίκες.

Με την εντολή

> pie(table(data$sex),main='Pie Chart of variable Sex')

παίρνουμε το τομεόγραμμα:



Όπου το εμβαδόν κάθε κυκλικού τομέα είναι ανάλογο της σχετικής συχνότητας της αντίστοιχης κατηγορίας της μεταβλητής sex.

Τόσο από τις αριθμητικές όσο και από τις γραφικές μεθόδους προκύπτει ότι οι περισσότεροι από τους πελάτες που έχουν αγοράσει μόνο ένα μη φαρμακευτικό προϊόν τον προηγούμενο μήνα είναι άντρες παρά γυναίκες.

Για την ποσοτική μεταβλητή **med** δίνοντας την εντολή

```
> summary(data$med)
```

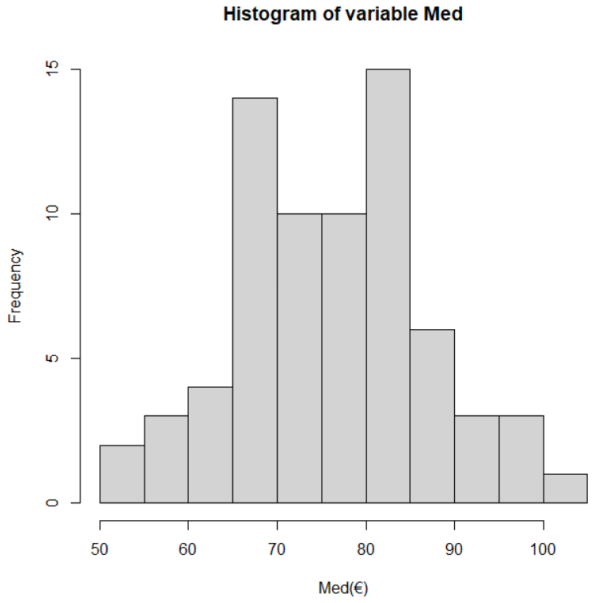
```
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
  50.10   68.90    76.40   76.19    82.10  103.10
```

Το μικρότερο ποσό χρημάτων που ξόδεψε το φαρμακείο σε φαρμακευτικά προϊόντα των προηγούμενο μήνα που παρατηρήθηκε στο δείγμα μας είναι €50,10 ενώ το μεγαλύτερο €103,10. Το 25% των χρημάτων που ξοδεύτηκαν είναι μικρότερο από €68,90, το 50% μικρότερο από €76,40, ενώ το 75% μικρότερο από €82,10. Η μέση τιμή των χρημάτων είναι €76,19.

Δίνοντας την εντολή

```
> hist(data$med, main='Histogram of variable Med', xlab='Med(€)')
```

παίρνουμε το ιστόγραμμα:

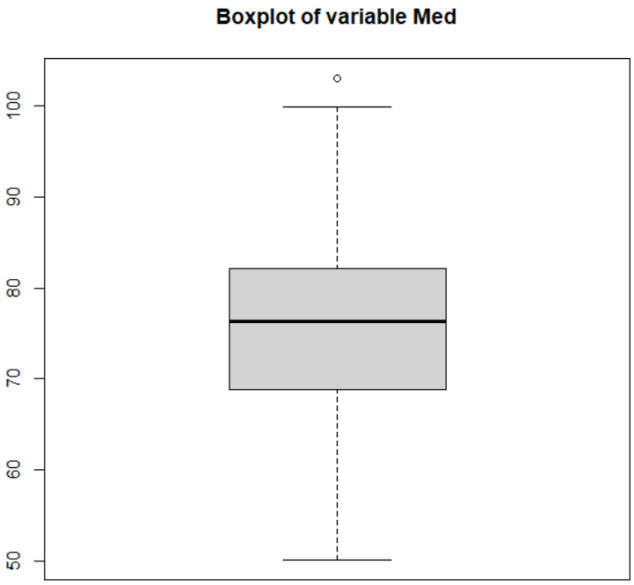


Ο οριζόντιος άξονας εκφράζει τα χρήματα που ξόδεψε το φαρμακείο σε φαρμακευτικά προϊόντα των προηγούμενο μήνα ενώ ο κατακόρυφος τις συχνότητες των αντίστοιχων κλάσεων. Στο δείγμα το περισσότερο ποσό χρημάτων που ξοδεύτηκε για φαρμακευτικά προϊόντα ήταν €80-€85. Το λιγότερο ποσό χρημάτων που ξοδεύτηκε για φαρμακευτικά προϊόντα ήταν πάνω από €100.

Δίνοντας την εντολή

```
> boxplot(data$med, main='Boxplot of variable Med')
```

παίρνουμε το θηκοδιάγραμμα:



Τις τιμές των πέντε στατιστικών μεγεθών που χρησιμοποιούμε για την κατασκευή ενός θηκοδιαγράμματος μπορούμε να τις πάρουμε μέσω της εντολής

> fivenum(data$med)

[1] 50.1 68.9 76.4 82.1 103.1

Η ερμηνεία τους έχει δοθεί προηγουμένως (εντολή summary).

Επίσης, για την ποσοτική μεταβλητή **population**

Δίνοντας την ετνολή

> summary(data$population)

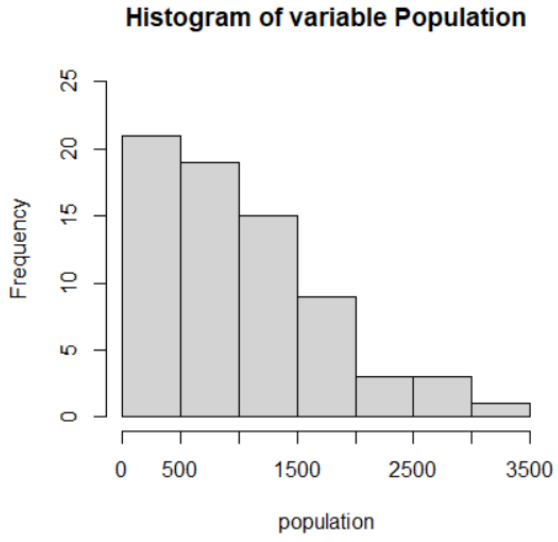
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17.0	420.0	885.0	994.8	1388.5	3344.0

Ο μικρότερος αριθμός κατοίκων που έχει η περιοχή που βρίσκεται το κατάστημα όπου έγινε η αγορά είναι 17.0 ενώ ο μεγαλύτερος 3344.0. Το 25% των αριθμών κατοίκων που έχει η περιοχή που βρίσκεται το κατάστημα όπου έγινε η αγορά είναι λιγότεροι από 420.0, το 50% λιγότεροι από 885.0 και το 75% λιγότεροι από 1388.5. Ο μέσος όρος των παρατηρούμενων πληθυσμών είναι 994.8.

Δίνοντας την εντολή

> hist(data$population, main='Histogram of variable Population',xlab='Population', ylim=c(0,25))

παίρνουμε το ιστόγραμμα:

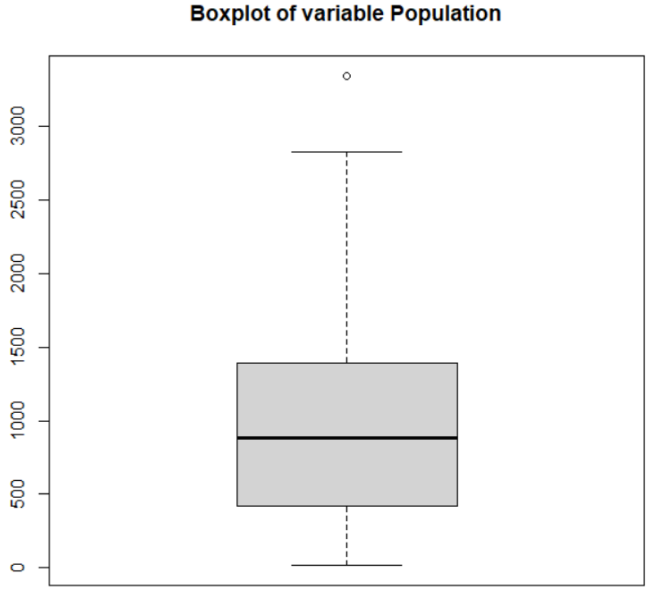


Ο οριζόντιος άξονας εκφράζει τον αριθμό κατοίκων που έχει η περιοχή που βρίσκεται το κατάστημα όπου έγινε η αγορά, ενώ ο κατακόρυφος τις συχνότητες των αντίστοιχων κλάσεων. Ο συχνότερος πληθυσμός που παρατηρήθηκε ήταν από 0 μέχρι 500 και ο σπανιότερος μεταξύ 3000 και 35000.

Δίνοντας την εντολή

> boxplot(data$population, main='Boxplot of variable Population')

παίρνουμε το θηκοδιάγραμμα:



Τις τιμές των πέντε στατιστικών μεγεθών που χρησιμοποιούμε για την κατασκευή ενός θηκοδιαγράμματος μπορούμε να τις πάρουμε μέσω της εντολής

> fivenum(data$population)

[1] 17.0 420.0 885.0 1388.5 3344.0

Η τελευταία ποσοτική μεταβλητή είναι η **money**

Δίνοντας την εντολή

> summary(data$money)

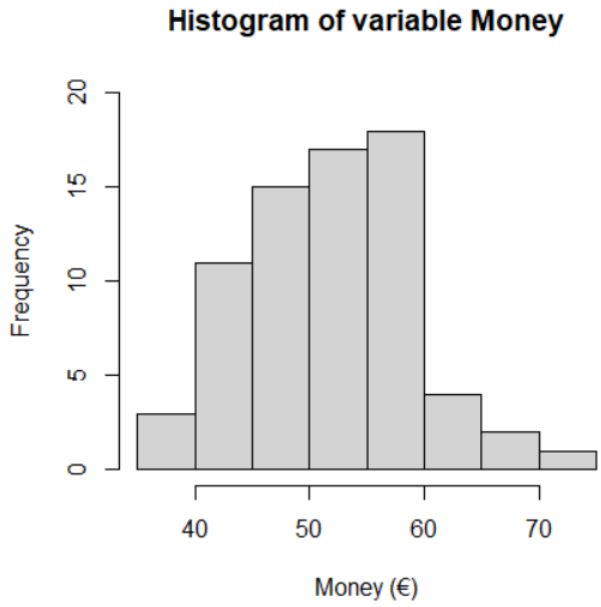
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
35.32	46.16	51.56	51.75	56.98	71.60

Το μικρότερο ποσό χρημάτων που ξόδεψε το φαρμακείο στην αγορά μη φαρμακευτικού προϊόντος που παρατηρήθηκε στο δείγμα μας είναι €35,12 ενώ το μεγαλύτερο €71,60. Το 25% των χρημάτων που ξοδεύτηκαν είναι μικρότερο από €46,16, το 50% μικρότερο από €51,56, ενώ το 75% μικρότερο από €56,98. Η μέση τιμή των χρημάτων είναι €51,75.

Δίνοντας την εντολή

> hist(data$money, main='Histogram of variable Money', xlab='Money(€)', ylim=c(0,20))

παίρνουμε το ιστόγραμμα:

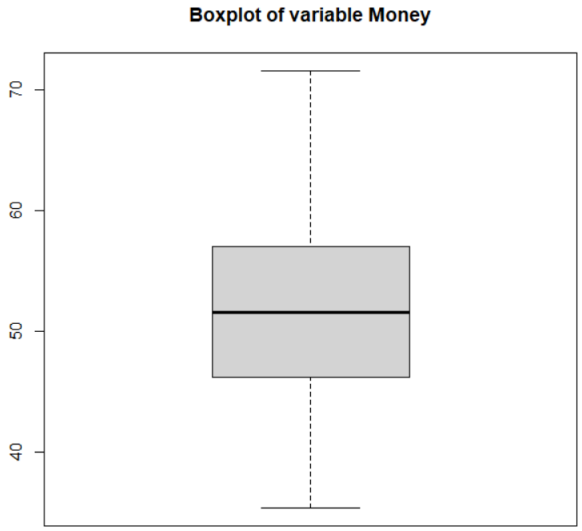


Ο οριζόντιος άξονας εκφράζει τα χρήματα που ξόδεψε το φαρμακείο σε μη φαρμακευτικά προϊόντα ενώ ο κατακόρυφος τις συχνότητες των αντίστοιχων κλάσεων. Στο δείγμα το περισσότερο ποσό χρημάτων που ξοδεύτηκε για μη φαρμακευτικά προϊόντα ήταν €55-€60. Το λιγότερο ποσό χρημάτων που ξοδεύτηκε για φαρμακευτικά προϊόντα ήταν πάνω από €70.

Δίνοντας την εντολή

> boxplot(data$money, main='Boxplot of variable Money')

παίρνουμε το θηκοδιάγραμμα



Τις τιμές των πέντε στατιστικών μεγεθών που χρησιμοποιούμε για την κατασκευή ενός θηκοδιαγράμματος μπορούμε να τις πάρουμε μέσω της εντολής

> fivenum(data$money)

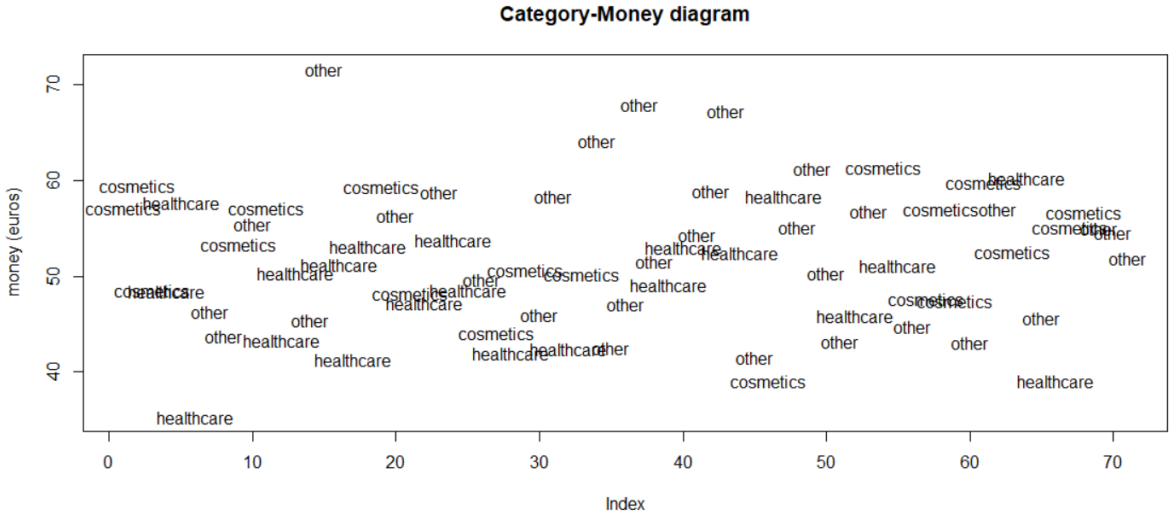
[1] 35.320 46.165 51.560 56.975 71.600

Η ερμηνεία τους έχει δοθεί προηγουμένως (εντολή summary).

ii) Για να εξετάσουμε ταυτόχρονα το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα και την κατηγορία αγορών σε μη φαρμακευτικά προϊόντα γράφουμε τον παρακάτω κώδικα:

```
>plot(data$money,type='n',main='Category-Money diagram',ylab='money (euros)')
>text(data$money, label=data$category)
```

και λαμβάνουμε το πιο κάτω διάγραμμα:



Όπου στον κατακόρυφο άξονα παρουσιάζεται το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα. Στο διάγραμμα τα σημεία στα οποία αναγράφεται το cosmetics αντιστοιχούν στην cosmetics που είναι τα καλλυντικά, τα σημεία που αναγράφεται healthcare αντιστοιχούν στην κατηγορία healthcare που είναι τα υγειονομικά προϊόντα ενώ στα σημεία που αναγράφεται other αντιστοιχούν στην κατηγορία other που είναι άλλου είδους προϊόντα.

Παρατηρούμε, σύμφωνα με το παραπάνω διάγραμμα, ότι το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα δεν εξαρτάται από την κατηγορία, αφού παρατηρούμε ότι σε κάθε ύψος του διαγράμματος έχουμε περίπου ίδιο πλήθος σημείων από κάθε κατηγορία.

Για να εξετάσουμε ταυτόχρονα το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα και την ηλικία γράφουμε τον παρακάτω κώδικα

```
> plot(data$age,data$money,xlab='Age (years)',ylab='money (euros)',main='Age-Money diagram')
```

και παίρνουμε το παρακάτω διάγραμμα:

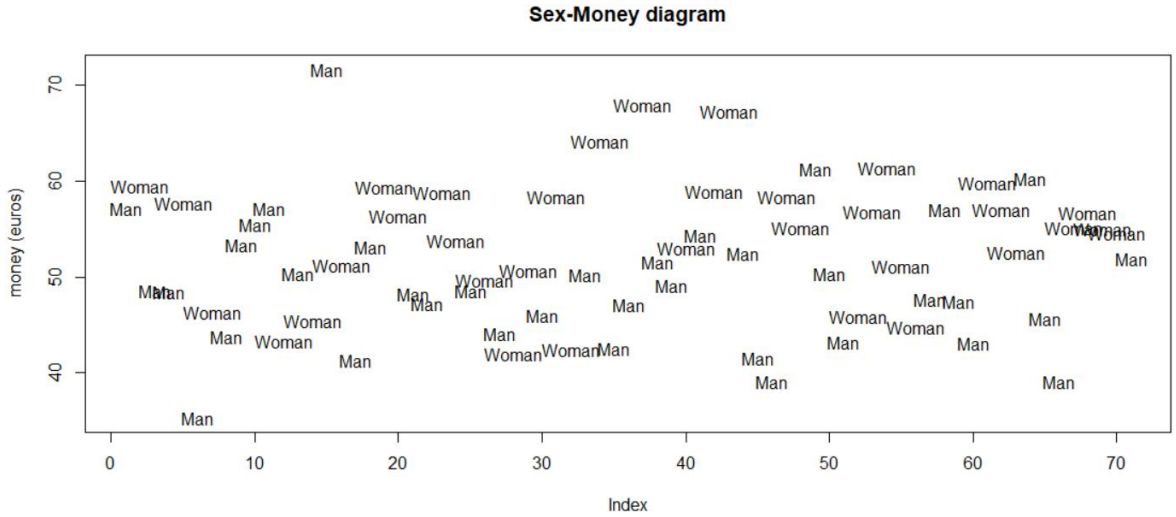
από το οποίο παρατηρούμε ότι για όλες τις ηλικίες το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα κυμαίνεται στα ίδια επίπεδα (€40-€60). Συνεπώς το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα δεν διαφοροποιείται πολύ ανάλογα με την ηλικία.

Για να εξετάσουμε ταυτόχρονα το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα και το φύλο γράφουμε τον παρακάτω κώδικα

```
>plot(data$money,type='n',main='Sex-Money diagram',ylab='money (euros)')
```

```
>text(data$money,label=data$sex)
```

και λαμβάνουμε το πιο κάτω διάγραμμα:

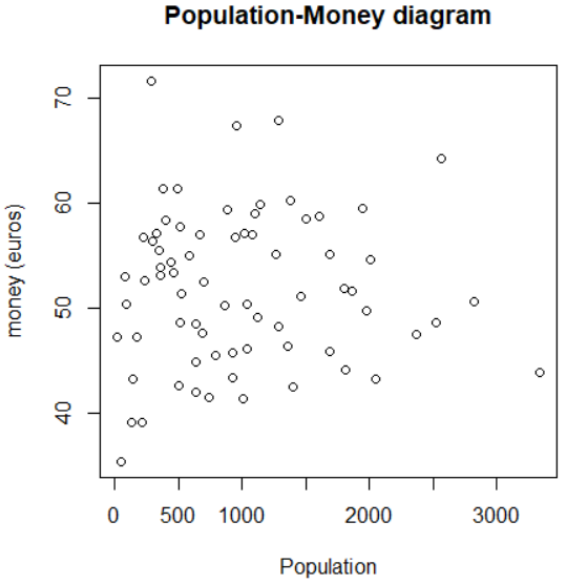


Σύμφωνα με το παραπάνω διάγραμμα, παρατηρούμε ότι το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα δεν εξαρτάται ούτε από το φύλο, αφού παρατηρούμε ότι σε κάθε ύψος του διαγράμματος έχουμε περίπου ίδιο πλήθος σημείων από κάθε φύλο.

Για να εξετάσουμε ταυτόχρονα το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα και τον πληθυσμό της περιοχής του καταστήματος όπου έγινε η αγορά γράφουμε τον παρακάτω κώδικα

```
>plot(data$population,data$money,xlab='Population',ylab='money(euros)',main='Population-Money diagram')
```

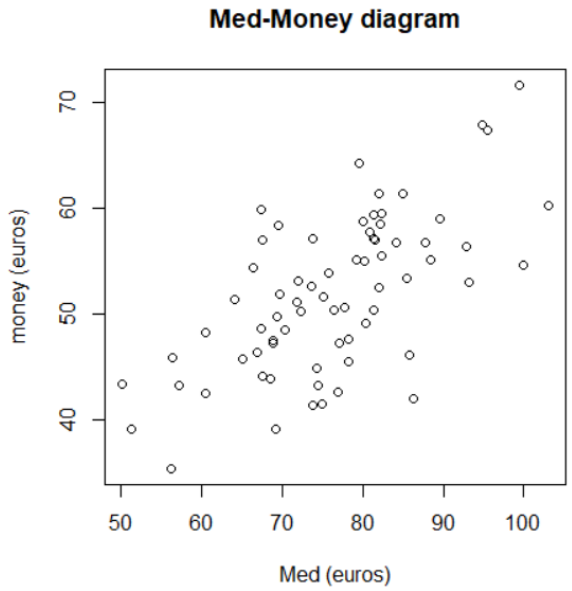
και λαμβάνουμε το πιο κάτω διάγραμμα:



Σε αυτό το διάγραμμα παρατηρούμε ότι το ποσό που ξοδεύτηκε σε μη φαρμακευτικά προϊόντα δεν εξαρτάται ούτε από τον αριθμό των κατοίκων τις περιοχής, αφού πάλι το ποσό κυμαίνεται στα ίδια επίπεδα (€40-€60)

Για να εξετάσουμε ταυτόχρονα το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα και το ποσό που ξοδεύτηκε στο φαρμακείο για φαρμακευτικά προϊόντα γράφουμε τον παρακάτω κώδικα

>plot(data$med,data$money,xlab='Med (euros)',ylab='money (euros)',main='Med-Money diagram')
και λαμβάνουμε το παρακάτω διάγραμμα



Σε αυτό το διάγραμμα παρατηρούμε πως όσο αυξάνεται το ποσό που ξοδεύει κάποιος πελάτης σε φαρμακευτικά προϊόντα αυξάνεται και το ποσό που ξοδεύει σε μη φαρμακευτικά προϊόντα. Επομένως, το ποσό που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα διαφοροποιείται ανάλογα με το ποσό που ξοδεύτηκε σε μη φαρμακευτικά προϊόντα.

iii) Για να δημιουργήσουμε τον πίνακα συχνοτήτων και σχετικών συχνοτήτων για τις κλάσεις [18-30), [30-50), [50 και άνω) θα πρέπει να δημιουργήσουμε μια κατηγορική μεταβλητή (f_age) με τις τρεις αυτές κατηγορίες. Χρησιμοποιούμε τον εξής κώδικα

> f_age<-rep('0',44)

>f_age[data$age>=18&data$age<30]<- '[18-30)'

>f_age[data$age>=30&data$age<50]<- '[30-50)'

>f_age[data$age>=50]<- '>=50'

Ο πίνακας συχνοτήτων λαμβάνεται από την παρακάτω εντολή

> table(f_age)

f_age

[18-30) [30-50) >=50   0

17   43   9   1

Ενώ ο πίνακας σχετικών συχνοτήτων

> prop.table(table(f_age))

f_age

[18-30) [30-50) >=50 0

0.24285714 0.61428571 0.12857143 0.01428571

Για να δημιουργήσουμε τον πίνακα συχνοτήτων και σχετικών συχνοτήτων για τις κλάσεις [0,q1), [q1,q2), [q2,q3), [q3 και άνω) θα πρέπει να δημιουργήσουμε μια κατηγορική μεταβλητή (f_pop) με τις τρεις αυτές κατηγορίες. Για να το κάνουμε αυτό χρησιμοποιούμε τον εξής κώδικα

```
> f_pop<-rep('0',44)
>q1<-quantile(data$population,0.25,na.rm=T)
>q2<-quantile(data$population,0.5,na.rm=T)
>q3<-quantile(data$population,0.75,na.rm=T)
>f_pop[data$population>=0&data$population<q1]<-'[0,q1) '
>f_pop[data$population>=q1&data$population<q2]<-'[q1,q2)'
>f_pop[data$population>=q2&data$population<q3]<-'[q2,q3)'
>f_pop[data$population>q3]<- '>=q3 '
>f_pop[f_pop=='0']<-NA
```

Ο πίνακας συχνοτήτων λαμβάνεται από την παρακάτω εντολή

```
> table(f_pop)
f_pop
[0,q1)  [q1,q2) [q2,q3)   >=q3
   18         17         18      18
```

Ενώ ο πίνακας σχετικών συχνοτήτων

```
> prop.table(table(f_pop))
f_pop
 [0,q1)          [q1,q2)       [q2,q3)          >=q3
0.2535211   0.2394366   0.2535211   0.2535211
```

Για να δημιουργήσουμε πίνακα συνάφειας συχνοτήτων των μεταβλητών f_pop και f_age γράφουμε την παρακάτω εντολή

```
> table(f_pop,f_age)
      f_age
f_pop     [18-30)  [30-50)  >=50
 [0,q1)          3         12       3
 [q1,q2)         5         10       1
 [q2,q3)         5         10       2
 >=q3            4         11       3
```

Ενώ για πίνακα συνάφειας σχετικών συχνοτήτων

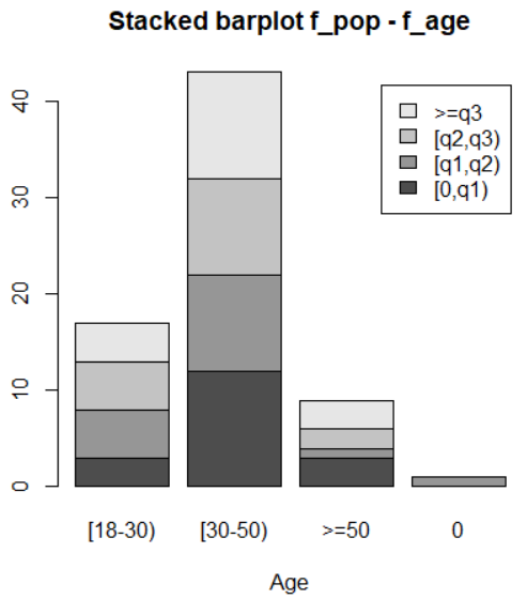
```
> prop.table(table(f_pop,f_age))
          f_age
f_pop          [18-30)     [30-50)       >=50
 [0,q1)    0.04285714  0.17142857  0.04285714
 [q1,q2)   0.07142857  0.14285714  0.01428571
 [q2,q3)   0.07142857  0.14285714  0.02857143
 >=q3      0.05714286  0.15714286  0.04285714
```

Συνεπώς ο πληθυσμός εξαρτάται από την ηλικία των ατόμων, κάτι το οποίο παρατηρήσαμε και στο αντίστοιχο διάγραμμα (Pop-Age diagram).

Για να δημιουργήσουμε στοιβαγμένο ραβδοδιάγραμμα εισάγουμε τον κώδικα

```
>barplot(table(f_pop,f_age),xlab='Age',legend=levels(factor(f_pop)),main='Stacked barplot f_pop - f_age')
```

Και λαμβάνουμε το εξής διάγραμμα:



Τα ύψη των ορθογωνίων μας δείχνουν τις συχνότητες για τους συνδυασμούς των κατηγοριών των δύο μεταβλητών.

Άσκηση 2

(α) Τέλειος αριθμός ονομάζεται ένας φυσικός αριθμός του οποίου το άθροισμα των διαιρετών του εκτός του εαυτού του είναι ίσο με τον αριθμό αυτό (ο αριθμός 1 δεν είναι τέλειος). Πιο κάτω παρατίθεται η συνάρτηση find_perfect_numbers η οποία δέχεται ένα διάνυσμα από θετικούς ακέραιους αριθμούς και εξάγει ένα διάνυσμα μόνο με αυτούς που είναι τέλειοι.

```
 find_perfect_numbers <- function(numbers) {
  # Έλεγχος αν το διάνυσμα περιέχει μόνο θετικούς ακέραιους αριθμούς
if (!all(numbers > 0) || !all(numbers==floor(numbers))) {
   stop("Το διάνυσμα πρέπει να περιέχει μόνο θετικούς ακέραιους αριθμούς.")
  }

  perfect_numbers <- c()

  # Έλεγχος για κάθε αριθμό στο διάνυσμα
  for (num in numbers) {
   divisors <- 1:(num-1)  # Διαιρέτες του αριθμού

   # Υπολογισμός του άθροισματος των διαιρετών
   sum_divisors <- sum(divisors[which(num %% divisors == 0)])

   # Έλεγχος αν ο αριθμός είναι τέλειος
   if (sum_divisors == num && num!=1) {
    perfect_numbers <- c(perfect_numbers, num)
   }
  }

  # Έλεγχος αν βρέθηκαν τέλειοι αριθμοί
  if (length(perfect_numbers) == 0) {
   print("Δεν βρέθηκαν τέλειοι αριθμοί στο διάνυσμα.")
  } else {
   return(perfect_numbers)
  }
}
```

Παράδειγμα χρήσης

```
>numbers <- c(6, 12, 28, 496, 1)
>find_perfect_numbers(numbers)
[1]   6  28 496

>numbers <- c(6, 12, -28, 496, 1)
>find_perfect_numbers(numbers)
Error in find_perfect_numbers(numbers) :
  Το διάνυσμα πρέπει να περιέχει μόνο θετικούς ακέραιους αριθμούς.

>numbers <- c(6, 12.5, 28, 496, 1)
>Find_perfect_numbers(numbers)
Error in find_perfect_numbers(numbers) :
  Το διάνυσμα πρέπει να περιέχει μόνο θετικούς ακέραιους αριθμούς
```

(β) Πιο κάτω παρατίθεται η συνάρτηση simulate_exponential η οποία δέχεται ως όρισμα θετικό ακέραιο αριθμό n και προσομοιώνει n τυχαίες τιμές από την εκθετική κατανομή με μέση τιμή ½.

```
simulate_exponential <- function(n) {

 # Έλεγχος αν ο αριθμός n είναι θετικός ακέραιος
 if (n!=floor(n) || n <= 0) {
  numb stop("Ο αριθμός n πρέπει να είναι θετικός ακέραιος.")
 }

 # Προσομοίωση τυχαίων τιμών
 u <- runif(n) # Τυχαίες τιμές u ~ U(0,1)
 simulated_values <- -log(1 - u) * 2 # Υπολογισμός των προσομοιωμένων τιμών

 return(simulated_values)
}
```

(γ) Για την προσομοίωση τεσσάρων διαφορετικών δειγμάτων μεγέθους n_1=10, n_2=100, n_3=1000, n_4=10000 από την εκθετική κατανομή με μέση τιμή ½, χρησιμοποιούμε τις εντολές του ερωτήματος (β) και στη συνέχεια τις εξής εντολές:

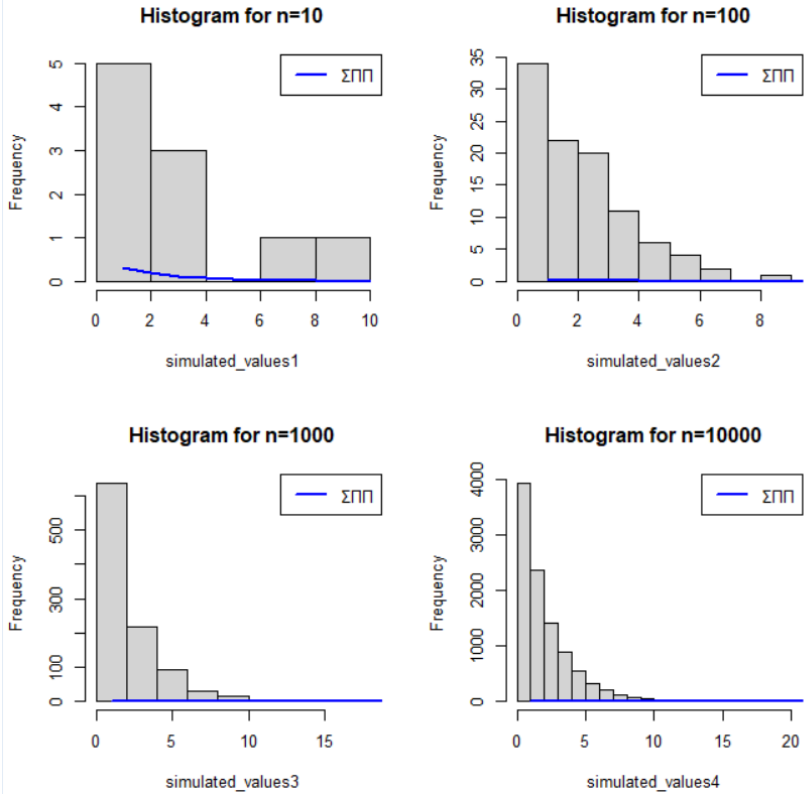
```
>n <- 10
>simulated_values1 <- simulate_exponential(n)
>print(simulated_values1)
>n <- 100
>simulated_values2 <- simulate_exponential(n)
>print(simulated_values2)
>n <- 1000
>simulated_values3 <- simulate_exponential(n)
>print(simulated_values3)
>n <- 10000
>simulated_values4 <- simulate_exponential(n)
>print(simulated_values4)
```

Για την παρουσίαση των αντίστοιχων ιστογραμμάτων και των συναρτήσεων πυκνότητας πιθανότητας της αντίστοιχης εκθετικής κατανομής για τις πιο πάνω τιμές χρησιμοποιούμε τις εξής εντολές:

```
> par(mfrow=c(2,2))
>hist(simulated_values1, main="Histogram for n=10")
>lines(dexp(1:10,1/2),col="blue",lwd=2)
>legend("topright",legend="ΣΠΠ", col="blue", lwd=2)
>hist(simulated_values2, main="Histogram for n=100")
>lines(dexp(1:100,1/2),col="blue",lwd=2)
>legend("topright",legend="ΣΠΠ", col="blue", lwd=2)
>hist(simulated_values3, main="Histogram for n=1000")
>lines(dexp(1:1000,1/2),col="blue",lwd=2)
>legend("topright",legend="ΣΠΠ", col="blue", lwd=2)
>hist(simulated_values4, main="Histogram for n=10000")
>lines(dexp(1:10000,1/2),col="blue",lwd=2)
>legend("topright",legend="ΣΠΠ", col="blue", lwd=2)
```

Η εντολή par(mfrow=c(2,2)) χρησιμοποιείται για την εμφάνιση των τεσσάρων ιστογραμμάτων σε ένα γραφικό παράθυρο χωρισμένο σε 4 τμήματα. Με την εντολή dexp( ) υπολογίζεται η συνάρτηση πυκνότητας πιθανότητας της αντίστοιχης εκθετικής κατανομής για δείγματα μεγέθους 10, 100, 1000,

10000 αντίστοιχα. Με την εντολή lines προστίθεται σε κάθε γράφημα και η συνάρτηση πυκνότητας πιθανότητας. Επίσης, με την εντολή col= “blue” χρωματίζεται το γράφημα της συνάρτησης πυκνότητας πιθανότητας μπλε και με την εντολή lwd=2 διπλασιάζουμε το πάχος των γραμμών της συνάρτησης σε όλες τις περιπτώσεις. Με την εντολή legend προσθέσαμε τη λεζάντα για την ΣΠΠ στο πάνω δεξιό μέρος του κάθε γραφήματος. Πιο κάτω παρατίθεται το γραφικό πλαίσιο:



Παρατηρούμε πως όσο αυξάνονται τα δείγματα που χρησιμοποιούμε οι κορυφές των στηλών στα ιστογράμματα σχηματίζουν καμπύλη όπως και η καμπύλη της εκθετικής κατανομής, δηλαδή μειώνονται εκθετικά.

Πιο κάτω παρατίθενται και ξεχωριστά τα γραφήματα για τις συναρτήσεις πυκνότητας πιθανότητας που έγιναν με χρήση των εντολών:

```
> plot(dexp(1:10,1/2),type='l',col="blue",lwd=2, main="ΣΠΠ για n=10",ylab="dexp")
> plot(dexp(1:100,1/2),type='l',col="blue",lwd=2, main="ΣΠΠ για n=100",ylab="dexp")
> plot(dexp(1:1000,1/2),type='l',col="blue",lwd=2, main="ΣΠΠ για n=1000",ylab="dexp")
> plot(dexp(1:10000,1/2),type='l',col="blue",lwd=2, main="ΣΠΠ για n=10000",ylab="dexp")
```