Περιγραφική Στατιστική με R

- Τα δεδομένα, σε μια γλώσσα προγραμματισμού, συνήθως τα αναπαριστούμε χρησιμοποιώντας έναν **n x p** πίνακα, του οποίου οι γραμμές αντιστοιχούν στις μονάδες του πληθυσμού και οι στήλες αντιπροσωπεύουν τις μεταβλητές (χαρακτηριστικά του πληθυσμού) για τις οποίες ενδιαφερόμαστε. Άρα έχουμε πληροφορία (δείγμα) για **p** μεταβλητές, για **n** μονάδες του πληθυσμού.

Κωδικοποίηση

- Πολλές φορές κωδικοποιούμε τις μεταβλητές, ειδικά αν αυτές είναι κατηγορικές. Σε περίπτωη όμως που η μεταβλητή είναι ονομαστική, είναι λάθος να αντικαταστήσουμε τις κατηγορίες με αριθμητικές τιμές, γιατί έτσι οι κατηγορίες αποκτούν προσδιορισμένη σχέση και διάταξη.
- Στις διατάξιμες μεταβλητές αντίθετα, δεν εμφανίζονται τέτοια προβλήματα. Απλά χρειάζεται να υπάρχει συμφωνία μεταξύ των αποστάσεων των κατηγοριών της διατάξιμης μεταβλητής και της διακριτής μεταβλητής που την αντικαθιστά.
- Τέλος, στις δίτιμες κατηγορικές μεταβλητές, χρησιμοποιούμε την κωδικοποίηση “0” και “1”.

Περιγραφική Στατιστική

- Μας δίνει μια συνοπτική παρουσίαση του δείγματος και ελέγχει την ορθότητα των τιμών του.
- Χρησιμοποιεί διάφορες Αριθμητικές και Γραφικές Μεθόδους.
- Η επιλογή των κατάλληλων αριθμητικών και γραφικών μεθόδων γίνεται με βάση τον τύπο της προς παρουσίαση μεταβλητής.

- Σκοπός της Περιγραφικής Στατιστικής είναι να δώσει μια συνοπτική παρουσίαση του δείγματος, καθώς επίσης και να ελέγξει την ορθότητα των τιμών του.

- Αποτελείται από διάφορες Αριθμητικές και Γραφικές Μεθόδους.

- Η επιλογή των κατάλληλων αριθμητικών και γραφικών μεθόδων γίνεται με βάση τον τύπο της μεταβλητής που θέλουμε να παρουσιάσουμε.

Ποσοτικές Μεταβλητές

Αριθμητικές Μέθοδοι

- 1. Μέτρα Θέσης:

1. Δειγματικός Μέσος (Sample Mean).

Ο Δειγματικός μέσος είναι το συνηθέστερο μέτρο θέσης για παρατηρήσεις από μια ποσοτική μεταβλητή. Έχει το μειονέκτημα όμως ότι επηρεάζεται από ακραίες παρατηρήσεις.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 2. Δειγματική Διάμεσος (Sample Median).

  Η μεσαία παρατήρηση από το δείγμα είναι η δειγματική διάμεσος. Αν το μέγεθος του δείγματος είναι n=2m-1 (περιττό) τότε η δειγματική διάμεσος ισούται με ym, όπου y1,…,yn είναι το διατεταγμένο δείγμα. Όταν n=2m (άρτιο) τότε η δειγματική διάμεσος ισούται με (ym+ ym+1)/ 2. Έχει το πλεονέκτημα ότι δεν επηρεάζεται από ακραίες παρατηρήσεις.

- 3. Δειγματική Κορυφή (Sample Mode).

  Η παρατήρηση με την μεγαλύτερη συχνότητα. Ως μέτρο έχει νόημα να υπολογιστεί σε περιπτώσεις όπου έχουμε επαναλήψεις ίδιων τιμών, γεγονός που συνήθως συμβαίνει μόνο για διακριτά δεδομένα.

- 2. Μέτρα Μεταβλητότητας:
- 1. Δειγματική Διασπορά – Τυπική Απόκλιση (Sample Variance – Sample Standard Deviation).

  Για να εκφράσουμε πόσο μακριά είναι οι παρατηρήσεις από τον δειγματικό μέσο συνήθως υπολογίζουμε την δειγματική διασπορά s2 ή την θετική τετραγωνική της ρίζα που καλείται δειγματική τυπική απόκλιση s. Έχει το μειονέκτημα ότι επηρεάζεται από ακραίες παρατηρήσεις.

$$s^2 = (n-1)^{-1} \sum_1^n (x_i - \bar{x})^2$$

Ποσοτικές Μεταβλητές

2. Εύρος Δείγματος (Range).

Η διαφορά μεταξύ της μεγαλύτερης και μικρότερης παρατήρησης. Προφανώς επηρεάζεται από ακραίες παρατηρήσεις.

3. Ενδοτεταρτημοριακό Εύρος (interquartile range – IQR).

Η διαφορά του τρίτου από το πρώτο τεταρτημόριο. Το τρίτο τεταρτημόριο (3rd quartile) είναι η παρατήρηση εκείνη που είναι μεγαλύτερη ή ίση από το 75% ακριβώς των παρατηρήσεων ενώ το πρώτο τεταρτημόριο (1st quartile) είναι η παρατήρηση εκείνη που είναι μεγαλύτερη ή ίση από το 25% ακριβώς των παρατηρήσεων. Το ενδοτεταρτημοριακό εύρος έχει το πλεονέκτημα ότι δεν επηρεάζεται από ακραίες παρατηρήσεις.

Παράδειγμα 1:

Τα δεδομένα στον πίνακα δεξιά εκφράζουν την διάρκεια ζωής (σε ώρες) 20 ηλεκτρονικών εξαρτημάτων του αυτού τύπου.

46	104	94	114	35
70	120	29	19	135
200	222	89	100	55
214	15	81	118	193

- Εισάγουμε τα δεδομένα στην R:

  x<-c(46, 104, 94, 114, 35, 70, 120, 29, 19, 135, 200, 222,89, 100, 55, 214, 15, 81, 118, 193)

- Εναλλακτικά θα μπορούσαμε να τα είχαμε διαβάσει από ένα αρχείο.

- Με την εντολή summary( ) παίρνουμε κάποια από τα αριθμητικά μέτρα που συζητήσαμε πριν.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   52.75   97.00  102.70  123.80  222.00
```

Εντολή	Σημασία
mean(x)	Δειγματικός Μέσος
min(x)	Μικρότερη παρατήρηση
max(x)	Μεγαλύτερη Παρατήρηση
median(x)	Δειγματική Διάμεσος
var(x)	Δειγματική Διασπορά
sd(x)	Δειγματική Τυπική Απόκλιση
quantile(x,p)	Επιστρέφει το p ποσοστημόριο. Για p=0.25 και p=0.75 έχουμε το 1ο και 3ο τεταρτημόριο

Παρατηρήσεις

- Αν τα δεδομένα που διαθέτουμε έχουν ελλιπείς τιμές, τότε για τον υπολογισμό των αριθμητικών μέτρων πρέπει να προσθέσουμε και το όρισμα na.rm=T. Π.χ.

```
> x<-c(1,2,4,5,6,7,10,35,NA,56,NA)
> x
[1] 1 2 4 5 6 7 10 35 NA 56 NA
> mean(x)
[1] NA
> mean(x, na.rm=TRUE)
[1] 14
```

- Υπάρχουν αρκετοί αλγόριθμοι υπολογισμού ποσοστημορίων στην R. Για περισσότερες λεπτομέρειες πληκτρολογήστε help(“quantile”).

Ποσοτικές Μεταβλητές

Γραφικές Μέθοδοι

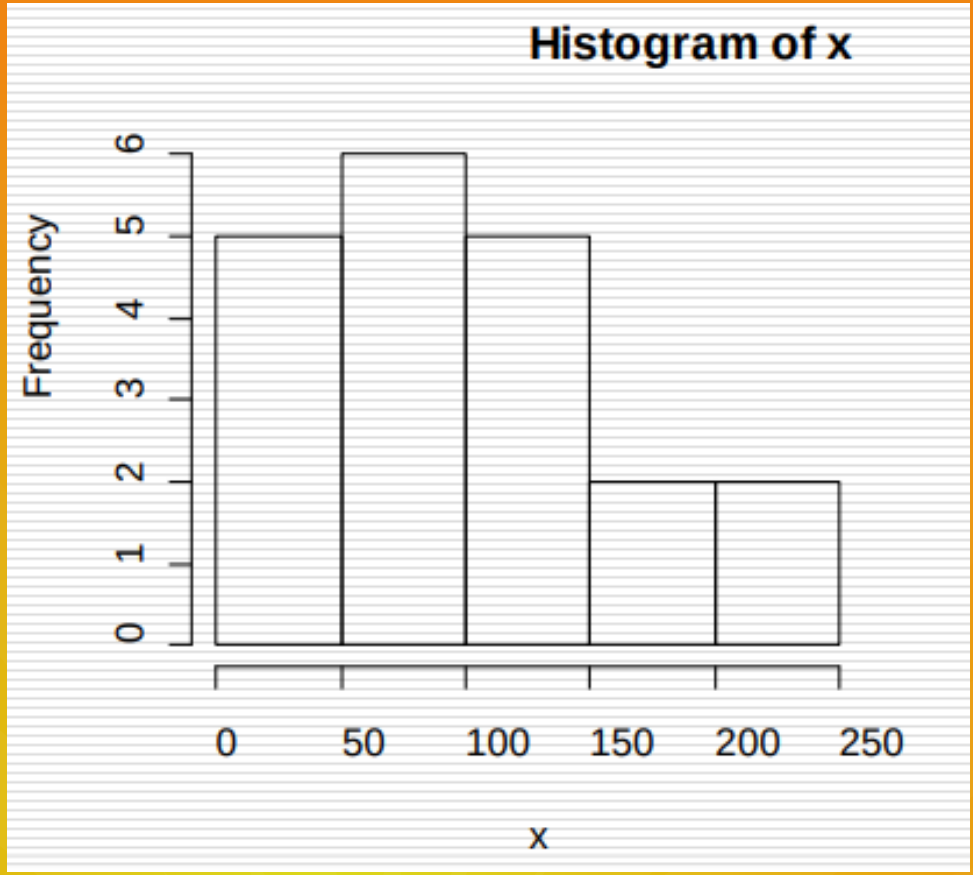
1. Ιστόγραμμα. Για την κατασκευή ενός ιστογράμματος συχνοτήτων

(frequency histogram), χρειάζεται να ομαδοποιήσουμε τα δεδομένα μας, και εν συνεχεία να σχηματίσουμε διαδοχικά ορθογώνια των οποίων οι βάσεις είναι τα διαστήματα των κλάσεων που δημιουργήσαμε και το ύψος τους είναι ίσο με

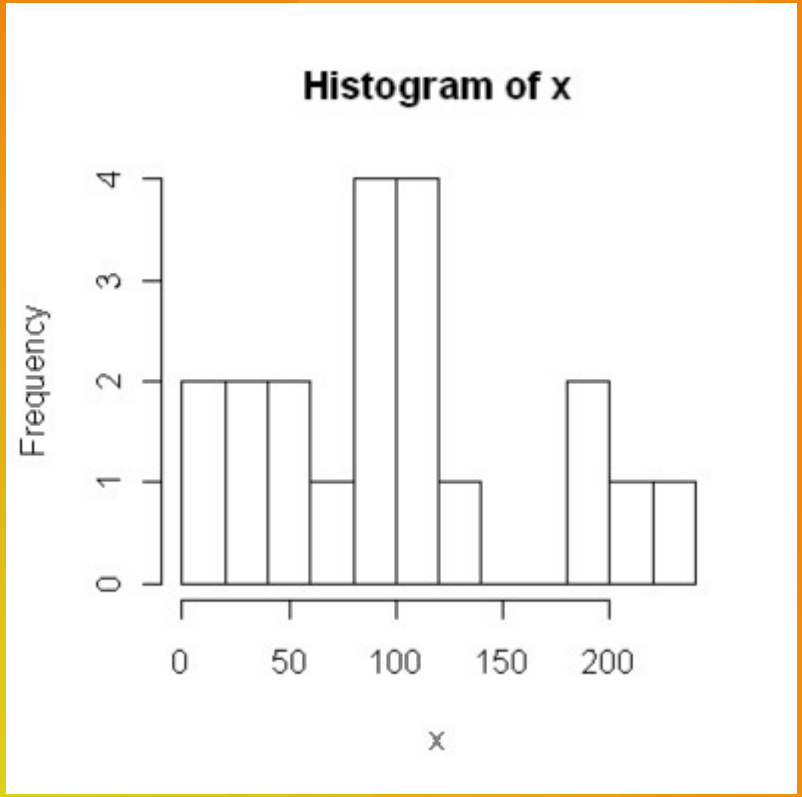
την συχνότητα των παρατηρήσεων στην αντίστοιχη κλάση.

Στις περισσότερες περιπτώσεις, δημιουργούμε κλάσεις ίδιου εύρους οπότε τα ορθογώνια έχουν τότε εμβαδά ανάλογα των αντίστοιχων συχνοτήτων.

> hist(x)

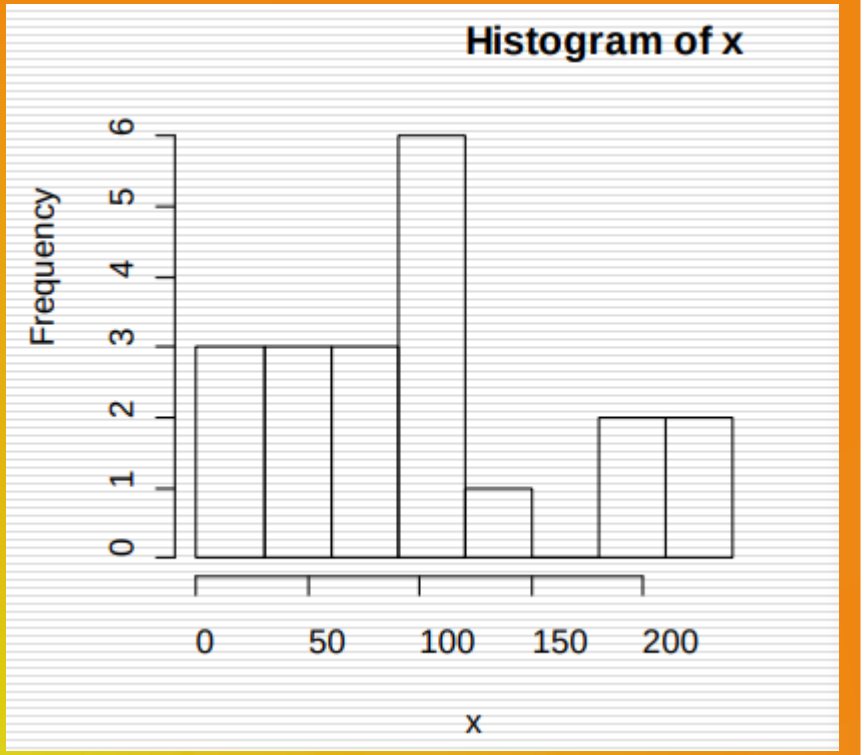


- Αν θέλουμε μπορούμε εμείς να προ-επιλέξουμε τον αριθμό των κλάσεων με τη βοήθεια του ορίσματος nclass. Η R δεν θα τηρήσει πάντα την επιλογή μας, θα κατασκευάσει το ιστόγραμμα με τον κοντινότερο αριθμό κλάσεων με αυτόν που ζητήσαμε, έτσι ώστε να μπορέσει να διατηρήσει το ίδιο πλάτος στις κλάσεις.



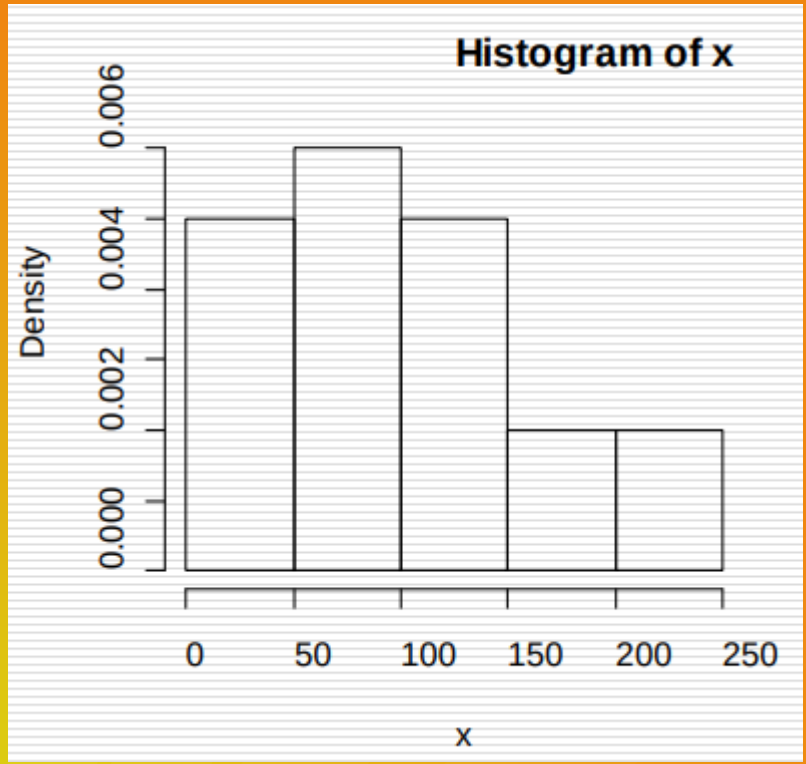

```
> hist(x, nclass=10)
```

> hist(x,breaks=seq(from=0, to=240,by=30))



 Μπορούμε επίσης αν επιθυμούμε να ορίσουμε τα όρια των κλάσεων

Τέλος μπορούμε στον yy’ άξονα αντί για συχνότητες να έχουμε πυκνότητα, και το συνολικό εμβαδόν του ιστογράμματος να ολοκληρώνει στην μονάδα. Έτσι παίρνουμε μια εκτίμηση της κατανομής της μεταβλητής.



> hist(x, probability=T)

2. Θηκοδιαγράμματα (boxplot). Για να παρουσιάσουμε

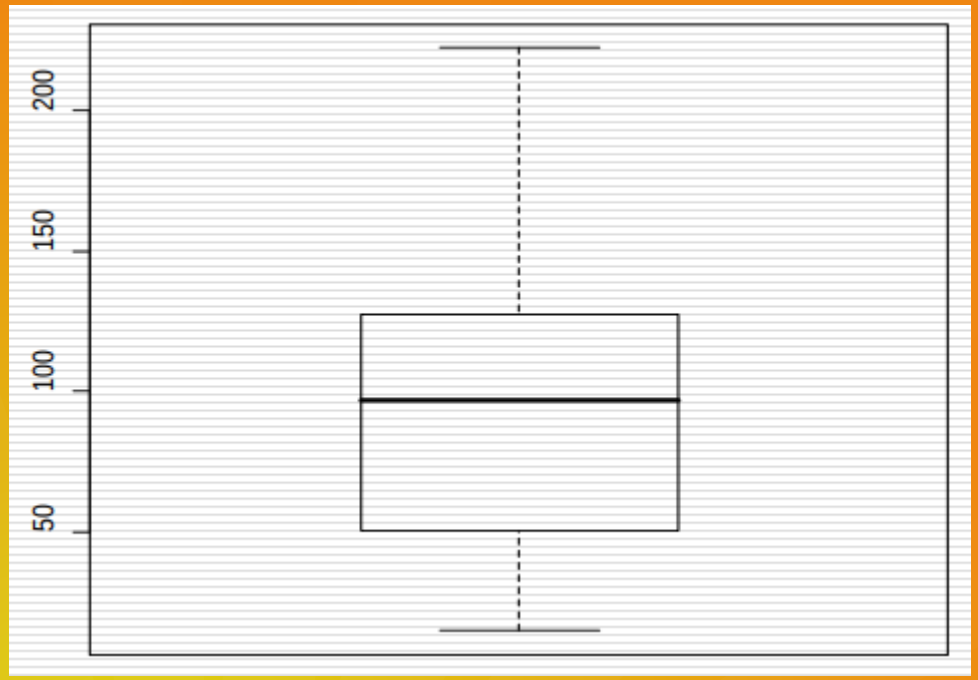
τα κυριότερα χαρακτηριστικά μιας κατανομής συνήθως δημιουργούμε ένα θηκοδιάγραμμα. Για την κατασκευή του δημιουργούμε ένα ορθογώνιο με κάτω βάση στο πρώτο και άνω βάση στο τρίτο τεταρτημόριο. Εν συνεχεία παριστάνουμε την διάμεσο

με ένα ευθύγραμμο τμήμα μέσα στο ορθογώνιο. Έπειτα φέρουμε ευθύγραμμα τμήματα στις 2 οριακές τιμές που ορίζονται ως το 30 (αντίστοιχα 1ο) τεταρτημόριο συν (αντίστοιχα μείον) 1.5 φορές το

ενδοτεταρτημοριακό εύρος. Αν δεν υπάρχουν παρατηρήσεις τόσο απομακρυσμένες, οι γραμμές τοποθετούνται πιο κοντά στο 1ο και 3ο τεταρτημόριο. Τέλος πιο ακραίες τιμές (αν υπάρχουν) παριστάνονται

με μια κουκκίδα, ενώ υπερβολικά έκτροπες τιμές παριστάνονται με αστερίσκο.

> boxplot(x)

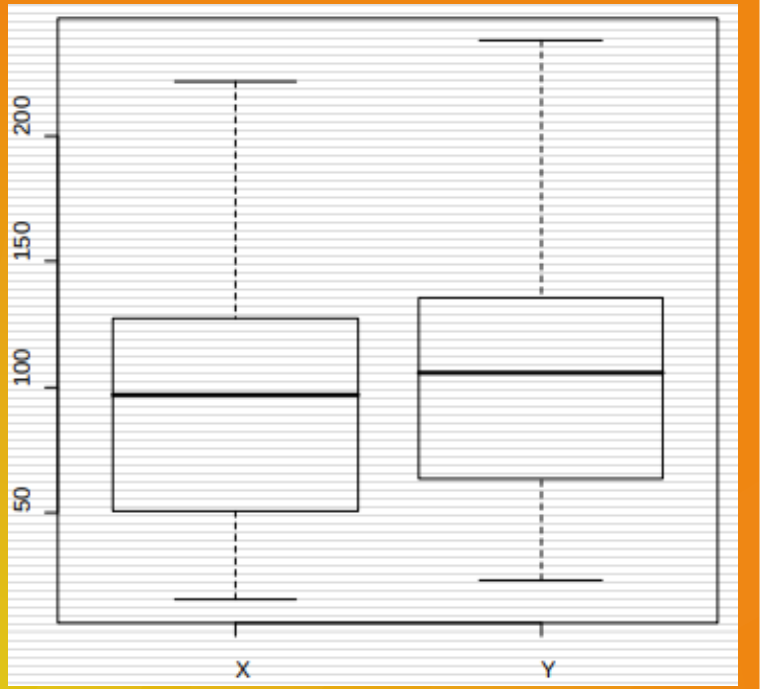


Τα θηκοδιαγράμματα είναι χρήσιμα για να συγκρίνουμε δύο δείγματα. Έστω ότι επιπλέον
με τα δεδομένα του 1ου Παραδείγματος έχουμε και τις διάρκειες ζωής (σε ώρες) 20 Ηλεκτρονικών εξαρτημάτων Κάποιου άλλου τύπου.

60	119	100	130	43
227	23	91	128	199
85	125	40	26	141
212	238	94	111	67

```
> y<-
c(60,119,100,130,43,227,23
,91,128,199,85,125,40,26,1
41,212,238,94,111,67)

> boxplot(x,y, names=c("X", "Y"))
```



 Τις τιμές των πέντε στατιστικών που χρησιμοποιούμε για την κατασκευή ενός θηκοδιαγράμματος μπορούμε να τις πάρουμε στην R με χρήση της εντολής fivenum().

```
> fivenum(y)
[1] 23.0 63.5 105.5 135.5 238.0
```

Κατηγορικές Μεταβλητές

Αριθμητικές Μέθοδοι.

Πίνακες Συχνοτήτων.

Παράδειγμα 2.

Τα δεδομένα δεξιά, αφορούν
τον τρόπο (αυτοκίνητο=C, μετρό=M,
λεωφορείο=B και πόδια=F)
που επιλέγουν 20 Αθηναίοι
για να πάνε κάθε πρωί στην
δουλειά τους.

C	C	B	M	M
C	M	M	F	C
F	B	B	M	M
C	C	C	M	C

 Περνάμε τα δεδομένα στην R

```
> A<-c("C", "C", "B", "M", "M", "C", "M", "M", "F", "C",
"F", "B", "B", "M", "M", "C", "C", "C", "M", "C")
```

 Με την εντολή table βλέπουμε τις συχνότητες σε κάθε κατηγορία.

```
> table(A)
A
B C F M
3 8 2 7
```

 Μπορούμε να δούμε και τις σχετικές συχνότητες

```
> prop.table(table(A))
A
B C F M
0.15 0.40 0.10 0.35
```

 Έστω ότι στο προηγούμενο παράδειγμα οι 10 πρώτοι ήταν άντρες και οι υπόλοιποι 10 γυναίκες. Έτσι έχουμε και μια άλλη κατηγορική μεταβλητή το φύλο.

 Μπορούμε τότε να κατασκευάσουμε τον πίνακα συνάφειας (contingency table), όπου απεικονίζει τη διμεταβλητή κατανομή συχνοτήτων για τις δύο κατηγορικές μεταβλητές.

```
> mytable<-table(A,Gender)
> mytable
   Gender
A   F M
  B 2 1
  C 4 4
  F 1 1
  M 3 4
> margin.table(mytable, 1)
A
B C F M
3 8 2 7
> margin.table(mytable, 2)
Gender
 F  M
10 10
```

συχνότητες για το μεταφ. μέσο

συχνότητες για το φύλο

```
> prop.table(mytable)
     Gender
A         F    M
  B 0.10 0.05
  C 0.20 0.20
  F 0.05 0.05
  M 0.15 0.20
> prop.table(mytable, 1)
            Gender
A           F         M
  B 0.6666667 0.3333333
  C 0.5000000 0.5000000
  F 0.5000000 0.5000000
  M 0.4285714 0.5714286
> prop.table(mytable, 2)
     Gender
A     F   M
  B 0.2 0.1
  C 0.4 0.4
  F 0.1 0.1
  M 0.3 0.4
```

Σχετικές συχνότητες κελιών

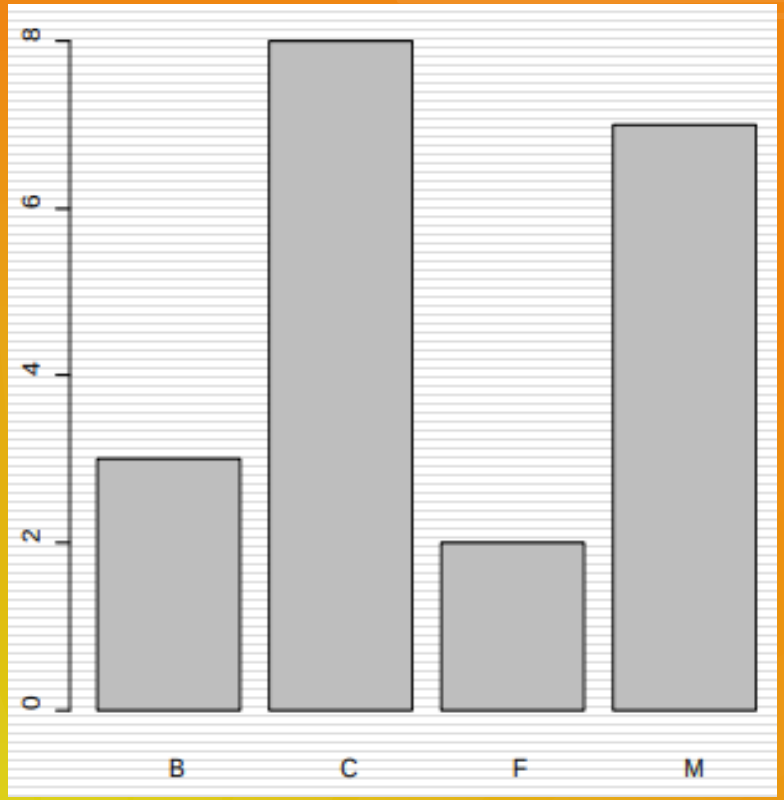
Σχετικές συχνότητες γραμμών

Σχετικές συχνότητες στηλών

- **Γραφικές Μέθοδοι**

  1. Ραβδόγραμμα. Στο ραβδόγραμμα οι κατηγορίες της μεταβλητής παρουσιάζονται στον ένα άξονα και οι αντίστοιχες συχνότητές τους στον άλλο άξονα, και εν συνεχεία κατασκευάζονται ορθογώνια πάνω από κάθε κατηγορία με ύψος ίσο με την αντίστοιχη συχνότητα της.
  2. Τομεόγραμμα. Στο τομεόγραμμα διαιρούμε ένα κύκλο σε κυκλικούς τομείς με εμβαδά ανάλογα προς τις σχετικές συχνότητες των κατηγοριών.

```
> AA<-table(A)
> AA
A
B C F M
3 8 2 7
> barplot(AA)
```



> pie(AA)