Ανάλυση Δεδομένων με χρήση του Στατιστικού Πακέτου R

Δημήτρης Φουσκάκης,
Καθηγητής,
Τομέας Μαθηματικών,
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών,
Εθνικό Μετσόβιο Πολυτεχνείο.

Περιεχόμενα

Εισαγωγή

- Ας υποθέσουμε ότι έχουμε ένα ερευνητικό ερώτημα που αφορά το αποτέλεσμα ενός τυχαίου πειράματος, και έχουμε συλλέξει με τυχαίο τρόπο δεδομένα, τα οποία θα μας βοηθήσουν να ποσοτικοποιήσουμε την αβεβαιότητά μας. Μεταφέρουμε τα δεδομένα στον Η/Υ και με τρόπους που αναφέραμε πριν διαβάζουμε αυτά τα δεδομένα στην R.

Εισαγωγή

- Τα δεδομένα τις περισσότερες φορές τα αναπαριστούμε με την βοήθεια ενός n×p πίνακα, του οποίου οι γραμμές αποτελούν τα αποτελέσματα που προέκυψαν από κάθε μονάδα του δείγματος 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

- Αρκετές φορές κωδικοποιούμε τις μεταβλητές ειδικά αν αυτές είναι κατηγορικές. Πρέπει να είμαστε όμως προσεκτικοί. Ειδικά στην περίπτωση που η μεταβλητή είναι ονομαστική, είναι λάθος να αντικαταστήσουμε τις κατηγορίες με αριθμητικές τιμές διότι έτσι οι κατηγορίες αποκτούν προσδιορισμένη σχέση και διάταξη.
- Αντίθετα δεν υπάρχει τόσο μεγάλο πρόβλημα αν η μεταβλητή είναι διατάξιμη. Το μόνο ερωτηματικό σε τέτοιου είδους κωδικοποιήσεις είναι αν υπάρχει συμφωνία μεταξύ των αποστάσεων των κατηγοριών της διατάξιμης μεταβλητής και της διακριτής μεταβλητής που την αντικαθιστά.
- Τέλος, αν η κατηγορική μεταβλητή είναι δίτιμη, χρησιμοποιούμε την κωδικοποίηση “0” και “1”.

Ακραίες, Αγνοούμενες και Εσφαλμένες Τιμές

- Είναι αρκετά σημαντικό προτού ξεκινήσουμε οποιαδήποτε Στατιστική Ανάλυση να ελέγξουμε τα δεδομένα μας για τυχόν λάθη ή παραλήψεις, να κάνουμε δηλαδή *διερευνητική ανάλυση δεδομένων* (Exploratory Data Analysis). Με την βοήθεια απλών περιγραφικών πινάκων ή γραφημάτων (όπως θα τα δούμε παρακάτω) μπορούμε να εντοπίσουμε "προβληματικές τιμές ή και μονάδες του δείγματος".

Ακραίες, Αγνοούμενες και Εσφαλμένες Τιμές

- Αρκετά συχνά παρατηρούμε ότι κάποια ή κάποιες τιμές μιας συγκεκριμένης μεταβλητής είναι *ακραίες ή έκτροπες* (outliers), απομακρυσμένες δηλαδή από τις υπόλοιπες τιμές της εν λόγω μεταβλητής. Τέτοιες τιμές δεν πρέπει να τις αντιμετωπίζουμε ως λανθασμένες, παρά μόνο αν είμαστε σίγουροι ότι πράγματι είναι. Ένας τρόπος να μειώσουμε την επιρροή αυτών των τιμών στα τελικά μας αποτελέσματα είναι με την χρήση κατάλληλων στατιστικών τεχνικών ή με κάποιον μετασχηματισμό των δεδομένων.

Ακραίες, Αγνοούμενες και Εσφαλμένες Τιμές

- Αρκετά συχνά επίσης ερχόμαστε αντιμέτωποι με *αγνοούμενες ή ελλειπείς τιμές* (missing values), δηλαδή με κάποιες μονάδες του δείγματος που έχουνε ελλιπή πληροφορία μιας και απουσιάζουν οι τιμές κάποιων μεταβλητών. Συχνά προσπαθούμε να εκτιμήσουμε την αγνοούμενη τιμή με την βοήθεια των υπόλοιπων τιμών (imputation). Σε μία τέτοια λύση θα πρέπει να καταλήγουμε μόνο αν το δείγμα μας είναι πολύ μικρό και δεν έχουμε την πολυτέλεια να χάσουμε επιπλέον πληροφορία λόγω των αγνοούμενων τιμών.
- Είναι σημαντικό να χρησιμοποιούμε το ίδιο σύμβολο για όλες τις αγνοούμενες τιμές. Το σύμβολο αυτό πρέπει να συμφωνεί με τον κωδικό που χρησιμοποιεί το πακέτο στο οποίο θα γίνει η ανάλυση. Στην R, π.χ. το σύμβολο αυτό είναι το NA.

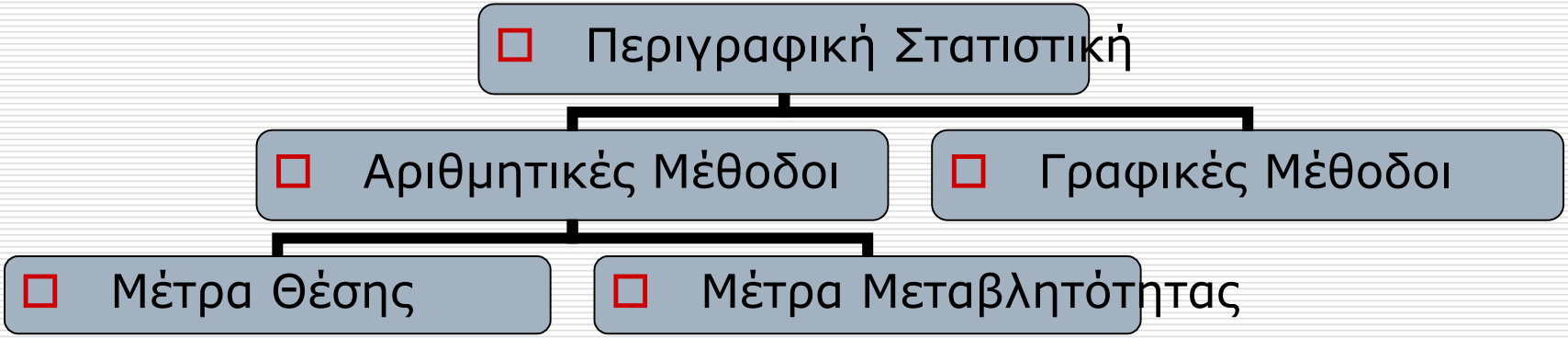
Ακραίες, Αγνοούμενες και Εσφαλμένες Τιμές

- Τέλος υπάρχουν περιπτώσεις που με βεβαιότητα αντιλαμβανόμαστε ότι μια τιμή είναι *εσφαλμένη*. Σε αυτές τις περιπτώσεις πρέπει να ελέγξουμε αν το λάθος προήλθε από την μεταφορά των δεδομένων στον Η/Υ και ρωτάμε αυτόν που σύλλεξε το δείγμα αν γνωρίζει την σωστή τιμή. Αν δεν μάθουμε την σωστή τιμή αντικαθιστούμε την εσφαλμένη τιμή με μια αγνοούμενη.
- Συνηθισμένα λάθη που γίνονται κατά την μεταφορά των δεδομένων στον Η/Υ είναι η αντιστροφή ψηφίων, και οι διπλοεγγραφές.

Περιγραφική Στατιστική

- 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.

Περιγραφική Στατιστική



Ποσοτικές Μεταβλητές

A. 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.

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

Ποσοτικές Μεταβλητές

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 y_m, όπου $y_1,\ldots,y_n$ είναι το διατεταγμένο δείγμα. Όταν n=2m (άρτιο) τότε η δειγματική διάμεσος ισούται με $(y_m + y_{m+1})/2$. 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 s^2 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.

Ποσοτικές Μεταβλητές

- Παράδειγμα 1:

  Τα παρακάτω δεδομένα 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

Ποσοτικές Μεταβλητές

- 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
```

Ποσοτικές Μεταβλητές

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

Ποσοτικές Μεταβλητές

- 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").

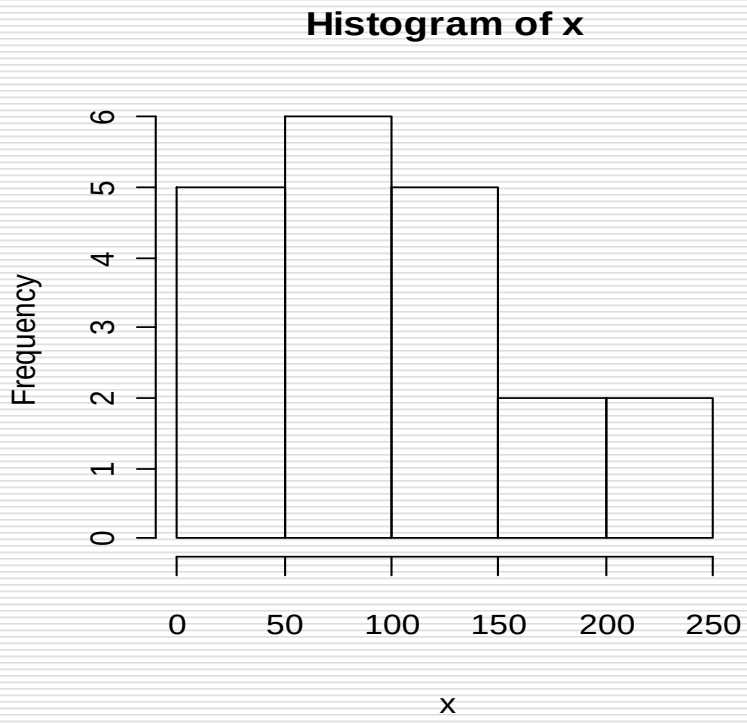
Ποσοτικές Μεταβλητές

Β. 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.

Ποσοτικές Μεταβλητές

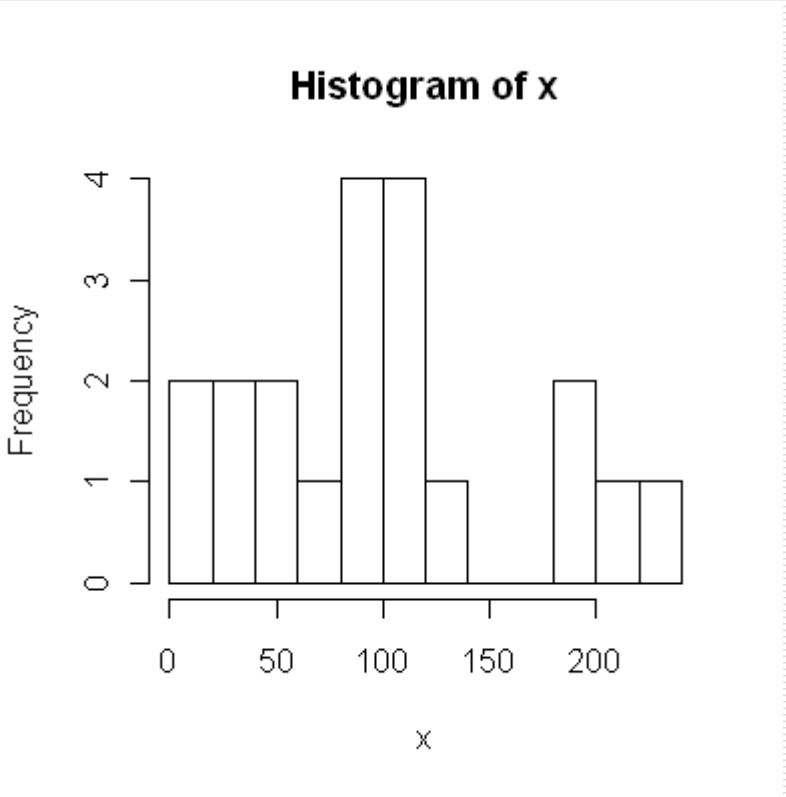
```
> hist(x)
```



Ποσοτικές Μεταβλητές

- 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.

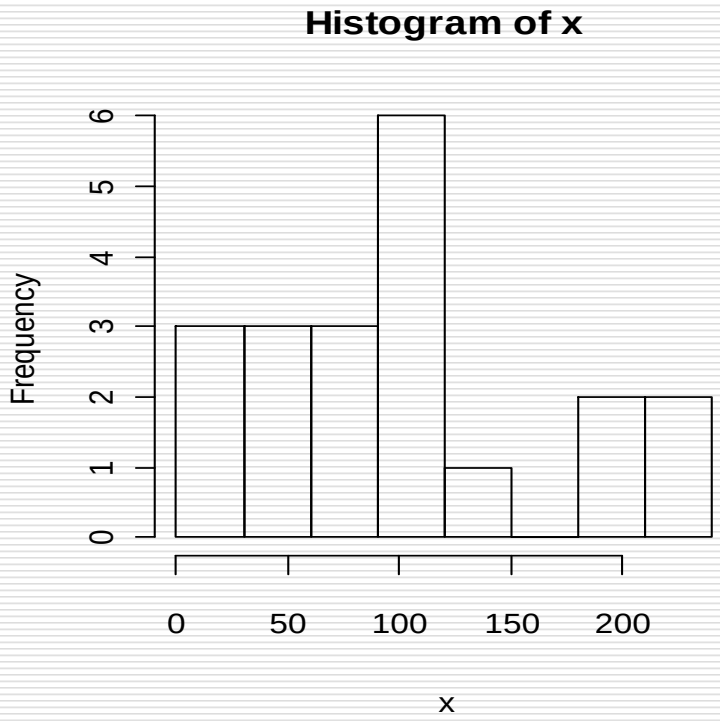
```
> hist(x, nclass=10)
```



Ποσοτικές Μεταβλητές

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

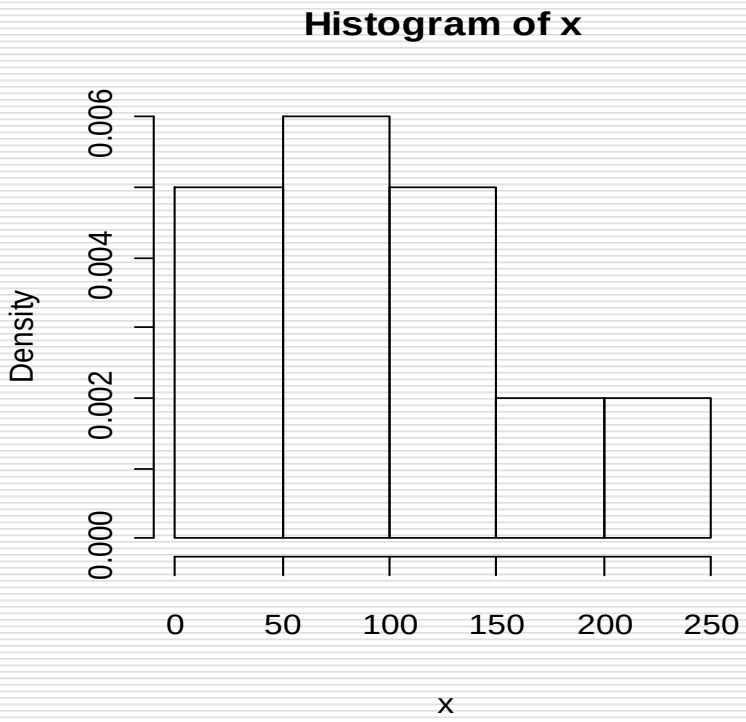
- Μπορούμε επίσης αν επιθυμούμε να ορίσουμε τα όρια των κλάσεων



Ποσοτικές Μεταβλητές

- 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.

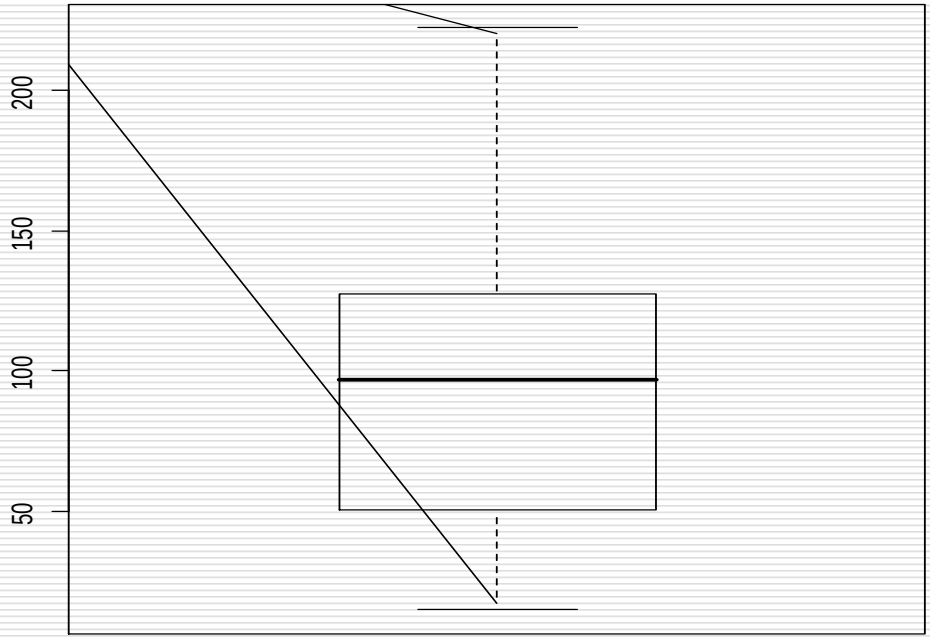
```
> hist(x, probability=T)
```



Ποσοτικές Μεταβλητές

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 3^0 (αντίστοιχα 1^o) τεταρτημόριο συν (αντίστοιχα μείον) 1.5 φορές το ενδοτεταρτημοριακό εύρος. Αν δεν υπάρχουν παρατηρήσεις τόσο απομακρυσμένες, οι γραμμές τοποθετούνται πιο κοντά στο 1^o και 3^o τεταρτημόριο. Τέλος πιο ακραίες τιμές (αν υπάρχουν) παριστάνονται με μια κουκκίδα, ενώ υπερβολικά έκτροπες τιμές παριστάνονται με αστερίσκο.

Ποσοτικές Μεταβλητές




```
> boxplot(x)
```

Ποσοτικές Μεταβλητές

- 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.

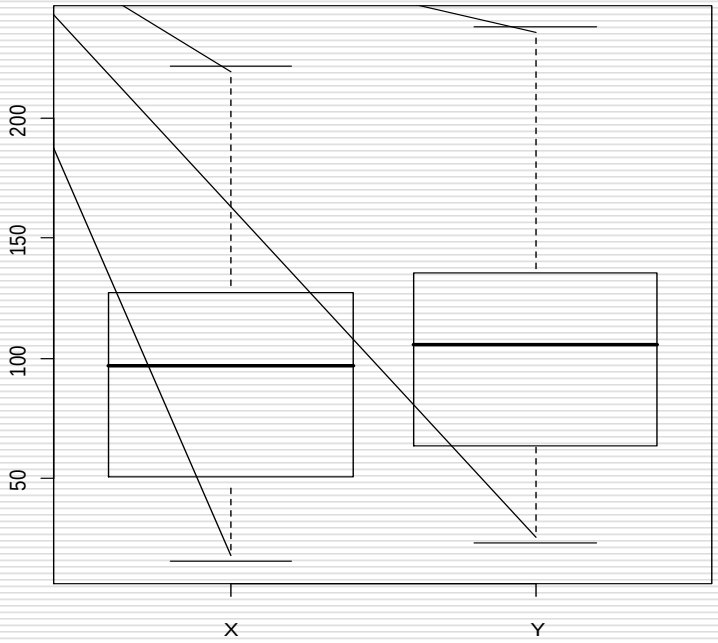
60	119	100	130	43
227	23	91	128	199
85	125	40	26	141
212	238	94	111	67

Ποσοτικές Μεταβλητές

```
> 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"))
```

όνομα για κάθε θηκογράφημα



Ποσοτικές Μεταβλητές

- 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

Α. Αριθμητικές Μέθοδοι.

Πίνακες Συχνοτήτων.

Παράδειγμα 2.
Τα παρακάτω δεδομένα 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

Κατηγορικές Μεταβλητές

- 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
```

Κατηγορικές Μεταβλητές

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

```
> Gender<-c(rep("M",10), rep("F", 10))
> Gender
 [1] "M" "M" "M" "M" "M" "M" "M" "M" "M" "M" "F" "F" "F"
     "F" "F" "F" "F" "F" "F" "F"
```

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

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

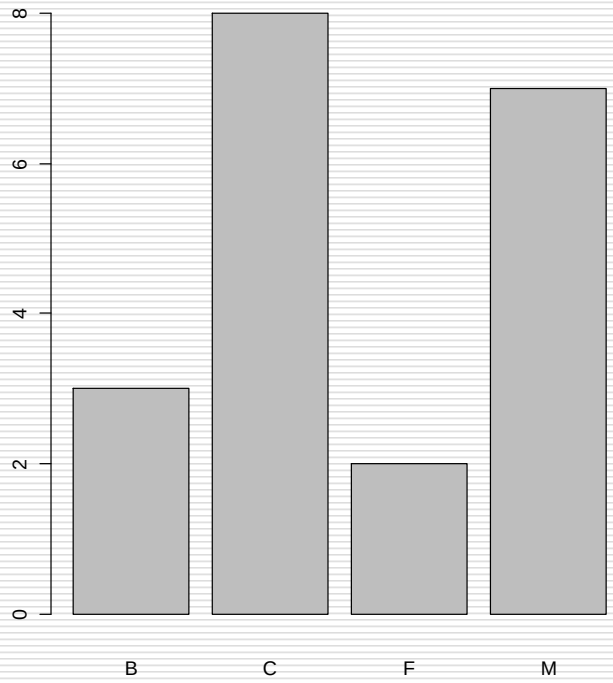
Κατηγορικές Μεταβλητές

Β. 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.

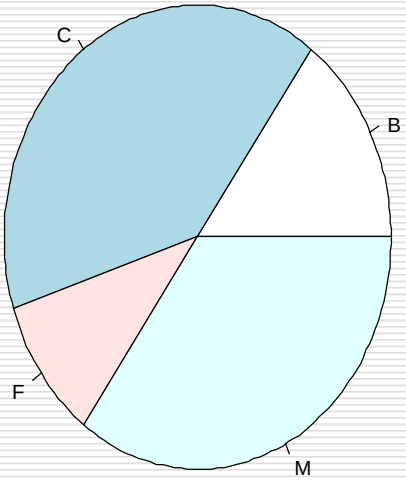
Κατηγορικές Μεταβλητές

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



Κατηγορικές Μεταβλητές

```
> pie(AA)
```

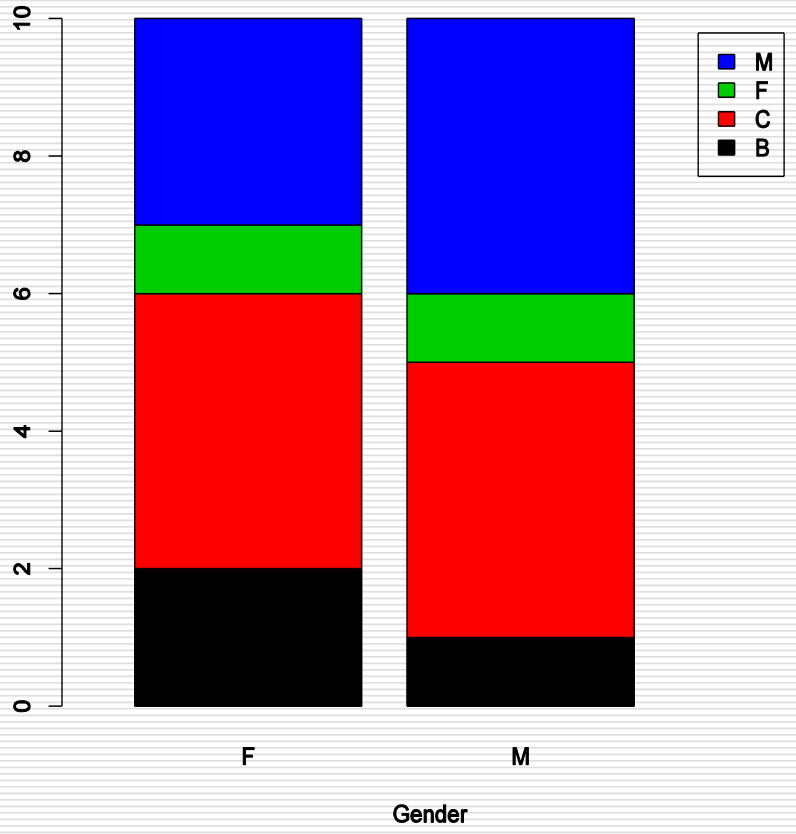


Κατηγορικές Μεταβλητές

- Στην R υπάρχει και η δυνατότητα γραφικής αναπαράστασης κατηγορικών δεδομένων προερχόμενων από δύο μεταβλητές με τη βοήθεια ενός στοιβαγμένου ραβδογράμματος (stacked barplot) ή ενός ομαδοποιημένου ραβδογράμματος (grouped barplot).

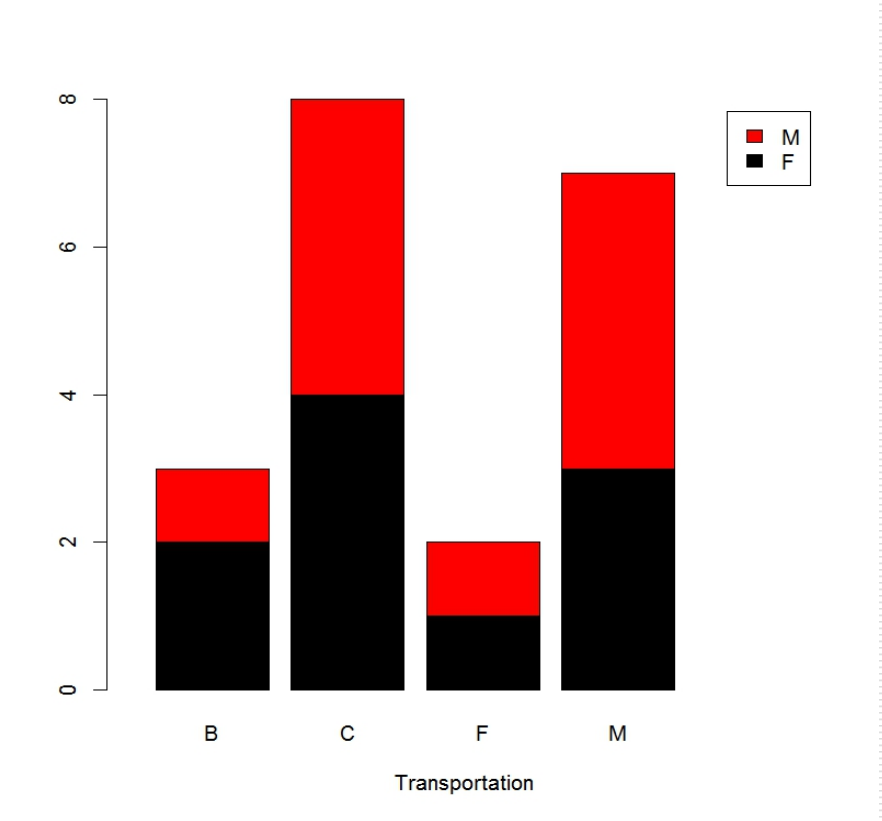
Κατηγορικές Μεταβλητές

- > freq_table<-table(A,Gender)
- > barplot(freq_table, xlim=c(0,3), xlab="Gender", legend=levels(A), col=1:4)



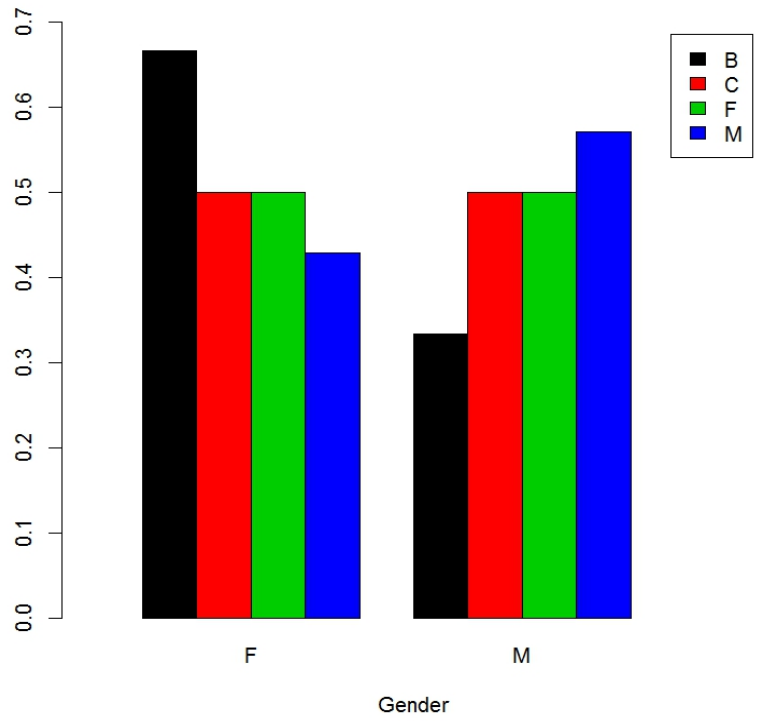
Κατηγορικές Μεταβλητές

- > freq_table<-table(Gender,A)
- > barplot(freq_table, width=0.85, xlim=c(0,5), xlab="Transportation", legend=levels(Gender), col=1:2)



Κατηγορικές Μεταβλητές

- > freq_table<-table(A,Gender)
- >barplot(prop.table(freq_table, 1), width=0.25, xlim=c(0,3), ylim=c(0,0.7), xlab="Gender", legend=levels(A), beside=TRUE, col=1:4)



Κατηγορικές Μεταβλητές

- > freq_table<-table(Gender, A)
- > barplot(prop.table(freq_table,1), width=0.25, xlim=c(0,3.6), xlab="Transportation", legend=levels(Gender), beside=TRUE, col=1:2)