Τύποι, Ποιότητα και Προεπεξεργασία Δεδομένων

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Συνοπτική Στατιστική και Οπτικοποίηση

Συνοπτική Στατιστική και Οπτικοποίηση

Ο βασικός στόχος αυτού του μαθήματος είναι να εξετάσουμε στις διάφορες τεχνικές που διατίθενται για την εξερεύνηση των δεδομένων, έτσι ώστε να είμαστε σε θέση να τις γνωρίσουμε καλύτερα, για την πιο επιτυχή κατάληξη της εργασίας της Εξόρυξης Δεδομένων.

Πιο συγκεκριμένα, θα εξετάσουμε τεχνικές

- της Συνοπτικής Στατιστικής και
- της Οπτικοποίησης.

Συνοπτική Στατιστική και Οπτικοποίηση

Ο βασικός στόχος αυτού του μαθήματος είναι να εξετάσουμε στις διάφορες τεχνικές που διατίθενται για την εξερεύνηση των δεδομένων, έτσι ώστε να είμαστε σε θέση να τις γνωρίσουμε καλύτερα, για την πιο επιτυχή κατάληξη της εργασίας της Εξόρυξης Δεδομένων.

Πιο συγκεκριμένα, θα εξετάσουμε τεχνικές

- της Συνοπτικής Στατιστικής και
- της Οπτικοποίησης.

Συνοπτική Στατιστική και Οπτικοποίηση

Θα μελετήσουμε και θα εφαρμόσουμε μέτρα

- θέσης
- διασποράς
- συσχέτισης, και
- τεχνικές οπτικοποίησης, όπως
  - ιστογράμματα
  - θηκογράμματα, και
  - διαγράμματα διασποράς

Ο φοιτητής/τρια θα μάθει να υπολογίζει τα διάφορα

- μέτρα θέσης
- διασποράς και συσχέτισης, αλλά και
- να δημιουργεί ιστογράμματα
- θηκογράμματα και
- διαγράμματα διασποράς

με χρήση της γλώσσας R.

Συνοπτική Στατιστική και Οπτικοποίηση

Η Συνοπτική ή Περιγραφική Στατιστική αποτελεί εκείνη την περιοχή της επιστήμης της Στατιστικής, που ασχολείται με τη συνοπτική και αποτελεσματική παρουσίαση των στατιστικών δεδομένων.

Ανάλογα με την περιοχή εφαρμογής, τα στατιστικά δεδομένα μπορούν να παρουσιαστούν συνοπτικά είτε μέσω συγκεκριμένων **αριθμητικών μέτρων, γνωστών ως μέτρων θέσης και διασποράς, είτε μέσω κατάλληλων διαγραμμάτων.**

Στην πιο αναλυτική, αλλά όχι τόσο κατάλληλη μορφή για την απόδοση συμπερασμάτων, τα στατιστικά δεδομένα μπορούν να παρουσιαστούν μέσω **διανυσμάτων ή και πινάκων**.

Μέτρα Θέσης

Τα μέτρα θέσης (ή μέτρα κεντρικής τάσης) περιγράφουν περιληπτικά τη θέση των δεδομένων πάνω στην ευθεία των πραγματικών αριθμών.

Προσδιορίζουν ένα κεντρικό σημείο γύρω από το οποίο έχουν την τάση να συγκεντρώνονται τα δεδομένων.

Τα κυριότερα μέτρα θέσης είναι η **μέση τιμή** (mean value) και η **διάμεσος** (median).

Μέση τιμή

Η μέση τιμή (mean value) ή αριθμητικός μέσος είναι το συνηθέστερο **μέτρο κεντρικής θέσης**.

Αν n είναι το πλήθος των παρατηρήσεων $x_i, i = 1, \ldots, n$, η μέση τιμή ορίζεται ως

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

Στην **R**, ο υπολογισμός της μέσης τιμής γίνεται με χρήση της συνάρτησης **"mean()"**.

Μέση τιμή

Ας θεωρήσουμε, για παράδειγμα, τις παρακάτω τιμές, που αφορούν το πλήθος των ωρών χρήσης του διαδικτύου τον τελευταίο μήνα, ενός δείγματος 10 εφήβων: 22, 0, 7, 12, 5, 33, 14, 8, 0, 9.

Εισάγουμε τα δεδομένα αυτά στην R, σε μια μεταβλητή με το όνομα ***internet_usage***, και υπολογίζουμε τη μέση τιμή τους ως εξής

```
> internet_usage = c(22, 0, 7, 12, 5, 33, 14, 8, 0, 9)
> internet_usage
[1] 22  0  7 12  5 33 14  8  0  9
> mean(internet_usage)
[1] 11
```

Μέση τιμή

Σε περίπτωση που τα διαθέσιμα δεδομένα έχουν ελλιπείς τιμές, τότε για τον υπολογισμό των μέτρων προσθέτουμε το όρισμα **na.rm = TRUE**.

Για παράδειγμα:

```
> internet_usage = c(22, 0, 7, 12, 5, NA, 14, NA, 0, 9)
> mean(internet_usage, na.rm = FALSE)
[1] NA
> mean(internet_usage, na.rm = TRUE)
[1] 8.625
```

Διάμεσος

Η **διάμεσος** (median) είναι η τιμή της μεσαίας παρατήρησης, όταν οι παρατηρήσεις ταξινομηθούν με αύξουσα ή φθίνουσα διάταξη.

Στην περίπτωση που το πλήθος n των παρατηρήσεων είναι περιττός αριθμός, τότε η μεσαία παρατήρηση είναι η (n+1)/2 – οστή, ενώ, όταν το πλήθος των παρατηρήσεων είναι άρτιος αριθμός, έχουμε δύο «μεσαίες» παρατηρήσεις, στις θέσεις n/2 και n/2+1, οπότε η διάμεσος είναι το ημι-άθροισμα αυτών.

Για παράδειγμα, για τις ταξινομημένες παρατηρήσεις (περιττού πληθους):

2 8 16 17 21 33 33 35 37

η διάμεσος είναι η 5η παρατήρηση, δηλαδή είναι ίση με 21,

ενώ για τις παρατηρήσεις (άρτιου πλήθους): 100 150 170 220 230 380

η διάμεσος είναι ίση με (170 + 220)/2 = 195.

Διάμεσος

Στην R, ο υπολογισμός της διαμέσου γίνεται με χρήση της συνάρτησης **median()**.

Η εφαρμογή της συνάντησης στο προηγούμενο παράδειγμα με τις ώρες χρήσης διαδικτύου θα μας δώσει:

```
> median(internet_usage)
[1] 8.5
```

Ως μεσαία παρατήρηση, η διάμεσος έχει το χαρακτηριστικό να είναι μεγαλύτερη ή ίση από το 50% των παρατηρήσεων του δείγματος.

Το χαρακτηριστικό αυτό θα το δούμε στη συνέχεια και σε άλλα αριθμητικά μέτρα.

Μέτρα Διασποράς

Τα μέτρα διασποράς ή μέτρα μεταβλητότητας περιγράφουν περιληπτικά τη διασπορά των δεδομένων πάνω στην ευθεία των πραγματικών αριθμών.

Με άλλα λόγια, τα μέτρα διασποράς φανερώνουν τη μεταβλητότητα των παρατηρήσεων.

Η μεταβλητότητα των παρατηρήσεων δεν είναι πάντα φανερή από τα μέτρα θέσης, για παράδειγμα, τη μέση τιμή.

Αν τα δεδομένα είναι συγκεντρωμένα γύρω από τη μέση τιμή, δηλαδή αν η διασπορά είναι μικρή, τότε πράγματι η μέση τιμή αντιπροσωπεύει ικανοποιητικά τα δεδομένα.

Σε διαφορετική περίπτωση, τα μέτρα θέσης δεν δίνουν καλή συνοπτική περιγραφή των παρατηρήσεων.

Μέτρα Διασποράς

Είναι δυνατό, επίσης, διαφορετικά δείγματα παρατηρήσεων από τον ίδιο πληθυσμό να έχουν το ίδιο μέτρο θέσης.

Αυτό μπορεί να γίνει εύκολα κατανοητό, αν θεωρήσουμε τα σύνολα παρατηρήσεων A και B, όπου
A = 33, 37, 48, 49, 52, 54, 62, 63, 64, 68, 71 και
B = 1, 37, 38, 41, 45, 47, 48, 51, 56, 90, 147.

Η μέση τιμή και των δύο συνόλων είναι ίση με 54.636, ωστόσο, οι τιμές των δύο συνόλων έχουν διαφορετική μεταβλητότητα (διασπορά στην ευθεία των πραγματικών αριθμών).

Μέτρα Διασποράς

Τα κυριότερα μέτρα διασποράς, τα οποία και θα περιγράψουμε στη συνέχεια, είναι

- το εύρος,
- η διακύμανση,
- η τυπική απόκλιση,
- ο συντελεστής μεταβλητότητας και
- τα ποσοστιαία σημεία.

Ελάχιστη τιμή, Μέγιστη τιμή, Εύρος

Ας θεωρήσουμε το σύνολο παρατηρήσεων

A = 49, 33, 37, 63, 48, 54, 62, 52, 64, 71, 68.

Η **ελάχιστη (min)** και η μέγιστη (max) παρατήρηση μπορεί να υπολογιστεί μέσω των συναρτήσεων **min()** και **max()** , αντίστοιχα.

```
> A = c(49, 33, 37, 63, 48, 54, 62, 52, 64, 71, 68)
> min(A)
[1] 33
> max(A)
[1] 71
```

Οι παρακάτω εντολές προσδιορίζουν τη θέση, όπου εμφανίζεται η ελάχιστη και η μέγιστη τιμή του A, αντίστοιχα.

```
> which.min(A)
[1] 2
> which.max(A)
[1] 10
```

Ελάχιστη τιμή, Μέγιστη τιμή, Εύρος

Το **εύρος** (range) ορίζεται ως η διαφορά της μικρότερης παρατήρησης (min) από τη μεγαλύτερη παρατήρηση (max).

Στην R μπορούμε εύκολα να υπολογίσουμε το εύρος με χρήση των αντίστοιχων συναρτήσεων.

```
> print(max(A) - min (A))
[1] 38
```

Η συνάρτηση **range()** επιστρέφει ένα διάνυσμα με τη μικρότερη και την ελάχιστη παρατήρηση του διανύσματος x.

```
> range(A)
[1] 33 71
```

Οπότε, μέσω της **range()** έχουμε έναν εναλλακτικό τρόπο υπολογισμού του εύρους εντός συνόλου παρατηρήσεων.

```
> print(range(A)[2] - range(A)[1])
[1] 38
```

Ποσοστιαία σημεία

Το p-ποσοστιαίο σημείο ενός δείγματος με n παρατηρήσεις, ορίζεται ως η παρατήρηση, για την οποία το πολύ $p\%$ των παρατηρήσεων είναι μικρότερες από αυτήν και το πολύ $(1-p)\%$ των παρατηρήσεων μεγαλύτερες από την αυτήν.

Για την εύρεση του p-ποσοστιαίου σημείου, $1 \leqslant p \leqslant 99$, οι παρατηρήσεις θα πρέπει να είναι διατεταγμένες σε αύξουσα διάταξη και τότε η παρατήρηση βρίσκεται στη θέση $\frac{n+1}{100}p$

Στην R, για τον υπολογισμό του p-ποσοστιαίου σημείου χρησιμοποιούμε τη συνάρτηση **quantile(x,p)**, όπου x είναι το διάνυσμα με τις παρατηρήσεις μας και p είναι το p-ποσοστιαίο σημείο, $0.1 \leqslant p \leqslant 0.99$

Ποσοστιαία σημεία

Ας θεωρήσουμε τις παρακάτω n=20 παρατηρήσεις, οι οποίες για διευκόλυνση είναι ήδη διατεταγμένες σε αύξουσα διάταξη: 3, 4, 5, 6, 7, 8, 10, 10, 11, 12, 14, 14, 14, 15, 16, 17, 21, 25, 27, 32.

Για παράδειγμα, το 80-ποσοστιαίο σημείο βρίσκεται στη θέση $\frac{20+1}{100}80 = 16.8$, άρα μεταξύ της 16ης και της 17ης παρατήρησης και συγκεκριμένα θα βρίσκεται δεξιότερα της 16ης παρατήρησης κατά 0.8 της διαφοράς μεταξύ των δύο παρατηρήσεων.

Η 16η παρατήρηση είναι ίση με 17 και η 17η παρατήρησης είναι ίση με 21.

Οπότε, η παρατήρηση που ψάχνουμε θα είναι ίση με $x_{16} + 0.8(x_{17} - x_{16}) = 17 + 0.8(21 - 17) = 20.2$.

Ποσοστιαία σημεία

Ο τρόπος που χρησιμοποιήσαμε αντιστοιχεί στον υπ' αριθμό 7 αλγόριθμο υπολογισμού, ο οποίος είναι ενσωματωμένος στην R, οπότε, μέσω της εντολής **quantile()**, προκύπτει ότι:

```
> x= c(3, 4,5,6,7,8,10,10,11,12,14,14,14,15,16,17,21,25,27,32)
> quantile(x, 0.80)
80%
17.8
```

Ιδιαίτερο ενδιαφέρον παρουσιάζει το 25-ποσοστιαίο σημείο (η παρατήρηση που είναι μεγαλύτερη ή ίση από το 25% των παρατηρήσεων), που καλείται πρώτο τεταρτημόριο, **το 50-ποσοστιαίο σημείο (η παρατήρηση που είναι μεγαλύτερη ή ίση από το 50% των παρατηρήσεων), που καλείται δεύτερο τεταρτημόριο και δεν είναι άλλο από τη διάμεσο**, καθώς και το 75-ποσοστιαίο σημείο (η παρατήρηση που είναι μεγαλύτερη ή ίση από το 75% των παρατηρήσεων), που καλείται τρίτο τεταρτημόριο.

Ποσοστιαία σημεία

Για το προηγούμενο παράδειγμα, και με χρήση της R, έχουμε ότι:

```
> quantile(x, 0.80, type = 7)
80%
17.8
> quantile(x, 0.25, type = 7)
25%
7.75
> quantile(x, 0.50, type = 7)
50%
13
> quantile(x, 0.75, type = 7)
75%
16.25
>quantile(x,  probs = c(0.25, 0.5, 0.75, 0.85))
 25%   50%   75%   85%
7.75 13.00 16.25 21.60
```

Ποσοστιαία σημεία

Για περισσότερες πληροφορίες σχετικά με τη συνάρτηση quantile(), πληκτρολογήστε **help(«quantile»)**.

Η εντολή **summary()** συνοψίζει κάποια από τα αριθμητικά μέτρα, τα οποία έχουμε περιγράψει μέχρι τώρα.

```
> summary(x)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3.00    7.75   13.00   13.55   16.25   32.00
```

Ενδοτεταρτημοριακό εύρος

Το **ενδοτεταρτημοριακό** εύρος (interquartile range, IRQ) ορίζεται ως η διαφορά τρίτου και πρώτου τεταρτημορίου.

Το **ενδοτεταρτημοριακό** εύρος χρησιμοποιείται συχνά για την εύρεση ακραίων τιμών στα δεδομένα. Οι ακραίες τιμές εδώ ορίζονται ως παρατηρήσεις που πέφτουν κάτω από το Q1 − 1,5 IQR ή πάνω από το Q3 + 1,5 IQR.

Στην R δεν υπάρχει ενσωματωμένη συνάρτηση που να το υπολογίζει, ωστόσο μπορούμε να δημιουργήσουμε τη δική μας:

```
> irq = function(x){quantile(x,0.75) - quantile(x,0.25)}
> irq(x)
75%
8.5
```

Διασπορά

Η διασπορά (variance) ή διακύμανση s^2 ενός δείγματος n παρατηρήσεων δίνεται από τον τύπο

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Η διασπορά μπορεί να υπολογιστεί απλούστερα ως

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - \frac{n}{n-1}\overline{x}^2$$

Όταν τα δεδομένα αποτελούν το σύνολο του πληθυσμού και όχι δείγμα αυτού, τότε η διακύμανση συμβολίζεται με σ^2 και δίνεται από τον τύπο

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x - m)^2$$

όπου N είναι το μέγεθος και m η μέση τιμή του πληθυσμού.

Διασπορά

Ο υπολογισμός της **διασποράς** στην **R** γίνεται μέσω της συνάρτησης **var()**.

Για παράδειγμα, οι τιμές που ακολουθούν αντιστοιχούν στο πλήθος των μαθημάτων, που «οφείλει» ένα δείγμα 20 φοιτητών «επί πτυχίω»:

6, 2, 1, 9, 17, 4, 3, 2, 1, 5, 11, 4, 3, 1, 2, 2, 5, 4, 3, 6.

```
> courses = c(6,2,1,9,17,4,3,2,1,5,11,4,3,1,2,2,5,4,3,6)
> var(courses)
[1] 15.41842
```

Διασπορά

Μπορούμε να υπολογίσουμε τη διασπορά χωρίς χρήση της συνάρτησης, αλλά μέσω του μαθηματικού ορισμού της.

```
> sum((courses - mean(courses))^2 / (length(courses)-1))
[1] 15.41842
```

Η συνάρτηση **sum()** υπολογίζει το άθροισμα της τετραγωνικής διαφοράς κάθε τιμής του διανύσματος courses από τη μέση τιμή mean(courses) αυτού, το οποίο στη συνέχεια διαιρείται με το πλήθος των παρατηρήσεων length(courses) μείον 1.

Τυπική Απόκλιση

Η τυπική απόκλιση (standard deviation) s ενός δείγματος παρατηρήσεων ορίζεται ως η τετραγωνική ρίζα της διακύμανσης των παρατηρήσεων.

Στην R, η τυπική απόκλιση υπολογίζεται μέσω της συνάρτησης **sd()**.

```
> sd(courses)
[1] 3.92663
```

Ωστόσο, ο υπολογισμός της τυπικής απόκλισης μπορεί να γίνει εύκολα με χρήση των συναρτήσεων **var()** για τον υπολογισμό της διασποράς και της **sqrt()** για τον υπολογισμό της τετραγωνικής ρίζας. Για παράδειγμα:

```
> sqrt(var(courses))
[1] 3.92663
```

Μπορούμε, επίσης, να ορίσουμε εμείς τη συνάρτηση υπολογισμού της τυπικής απόκλισης, για παράδειγμα ως:

```
> std = function(x) sqrt(var(x))
> std(courses)
```

Συντελεστής μεταβλητότητας

Ο συντελεστής μεταβλητότητας (coefficient of variation, cv) ενός δείγματος παρατηρήσεων ορίζεται ως το πηλίκο της τυπικής απόκλισης προς τη μέση τιμή.

Εκφράζει την τυπική απόκλιση ως ποσοστό της μέσης τιμής.

Μπορούμε να δημιουργήσουμε την κατάλληλη συνάρτηση στην R για τον υπολογισμό του συντελεστή μεταβλητότητας ως εξής:

```
> cv = function(x){sd(x)/mean(x)}
> cv(courses)
[1] 0.8629955
```

Οπτικοποίηση Ποιοτικών Δεδομένων

Στην ενότητα που ακολουθεί θα παρουσιάσουμε τρόπους, με τους οποίους μπορούμε να οπτικοποιήσουμε τις παρατηρήσεις μας.

Η οπτικοποίηση μπορεί να γίνει είτε μέσω διανυσμάτων και πινάκων αλλά είτε και μέσω διαγραμμάτων, όπως ιστογράμματα, ραβδογράμματα, διαγράμματα πίτας κ.ά. Σε κάθε περίπτωση, σημαντική είναι η διάκριση των παρατηρήσεων μας σε ποσοτικά ή ποιοτικά δεδομένα.

Ξεκινάμε με τα ποιοτικά ή κατηγορικά δεδομένα, τα οποία συνήθως παρουσιάζουμε με πίνακες, ραβδογράμματα και διαγράμματα πίτας.

Οπτικοποίηση Ποιοτικών Δεδομένων

Ως ένα παράδειγμα εργασίας, ας υποθέσουμε ότι έχουμε στη διάθεση μας τις απαντήσεις που έδωσαν 20 άτομα αναφορικά με τη χρήση μέσων μεταφοράς σε καθημερινή βάση, προκειμένου να μεταβούν στον χώρο εργασίας τους.

Οι ερωτηθέντες είχαν να επιλέξουν τη μετακίνηση τους με αυτοκίνητο (car), με λεωφορείο (bus), με μετρό (metro) και με τα πόδια (foot). Οι απαντήσεις που δόθηκαν ήταν οι εξής:

car, car, bus, metro, metro, car, metro, metro, foot, cat, foot, bus. bus, metro, metro, car, car, car, metro, car

Οπτικοποίηση Ποιοτικών Δεδομένων

Αρχικά εισάγουμε τα δεδομένα, δημιουργώντας το διάνυσμα m. Τα δεδομένα μας είναι κατηγορικά, οπότε θα πρέπει να συμπεριλάβουμε τις τιμές τους σε διπλά εισαγωγικά.

```
> m = c("car", "car", "bus", "metro", "metro", "car", "metro", "met
+               ro", "foot", "car", "foot", "bus", "bus", "metro", "metr
+               "car", "car", "metro", "car")
```

Μπορούμε να απεικονίσουμε τις παρατηρήσεις μας σε μορφή πίνακα συχνοτήτων με χρήση της εντολής **table()**.

```
> table(m)
m
bus      car     foot  metn      ro    metro
3        8       2                 1     6
```

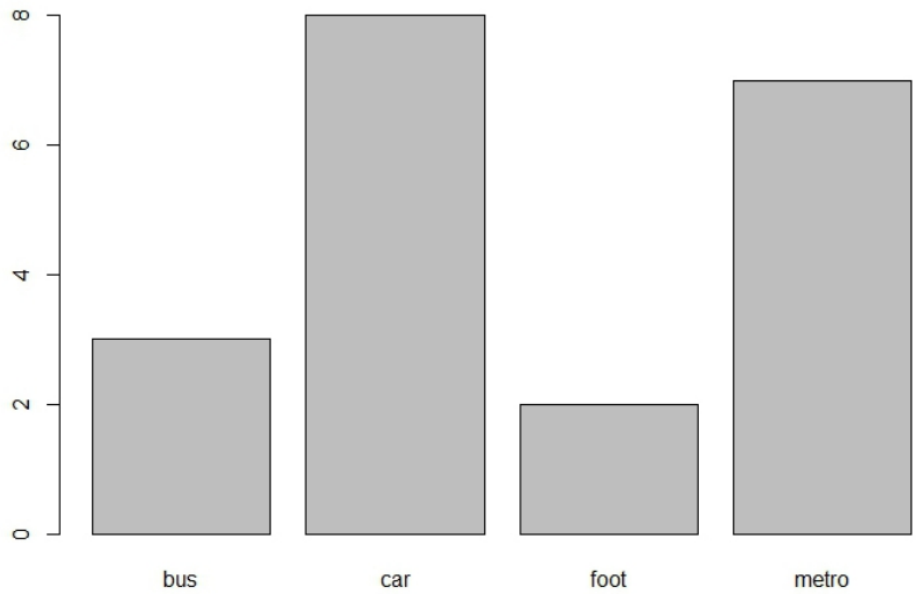
Ο παραπάνω πίνακας στην πρώτη γραμμή εμφανίζει τις διακεκριμένες τιμές των παρατηρήσεων και στη δεύτερη γραμμή τη συχνότητα εμφάνισης της κάθε τιμής.

Μπορείτε εύκολα να επαληθεύσετε ότι, για παράδειγμα, 8 απάντησαν πως μετακινούνται στην εργασία τους με αυτοκίνητο (car).

Αν επιθυμούμε να δούμε τις σχετικές συχνότητας (την αναλογία συχνότητας, δηλαδή, κάθε παρατήρησης προς το πλήθος των παρατηρήσεων), τότε χρησιμοποιούμε την εντολή **prop.table(table())**.

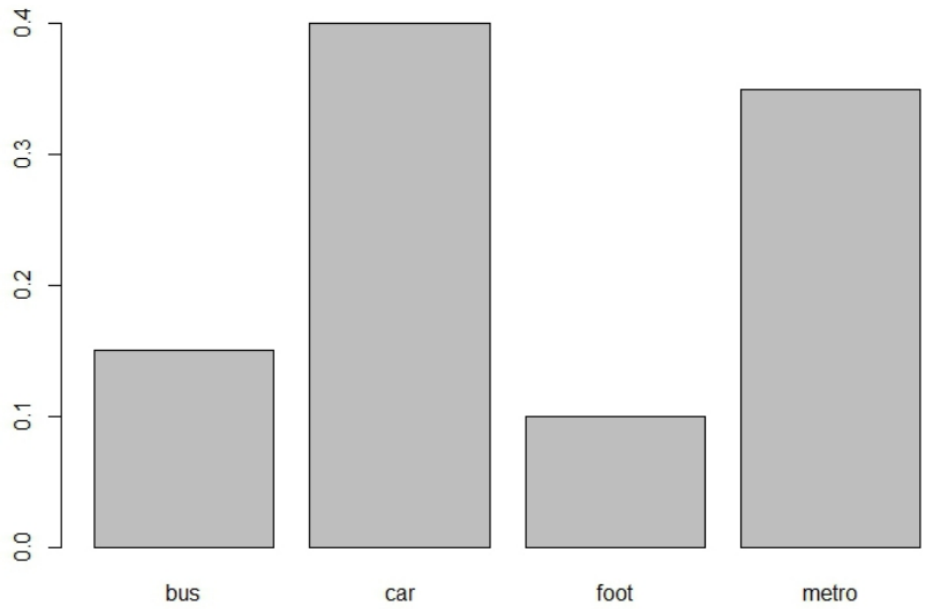
```
> prop.table(table(m))
m
bus     car    foot   metro   metro
0.15    0.40   0.10   0.05    0.30
```

Το ραβδόγραμμα (bar chart) που ακολουθεί αφορά στις συχνότητες του δείγματος και προκύπτει μέσω της εντολής: **barplot(table(m))**



Σχήμα: Παράδειγμα σχεδίασης ραβδογράμματος με χρήση συχνοτήτων.

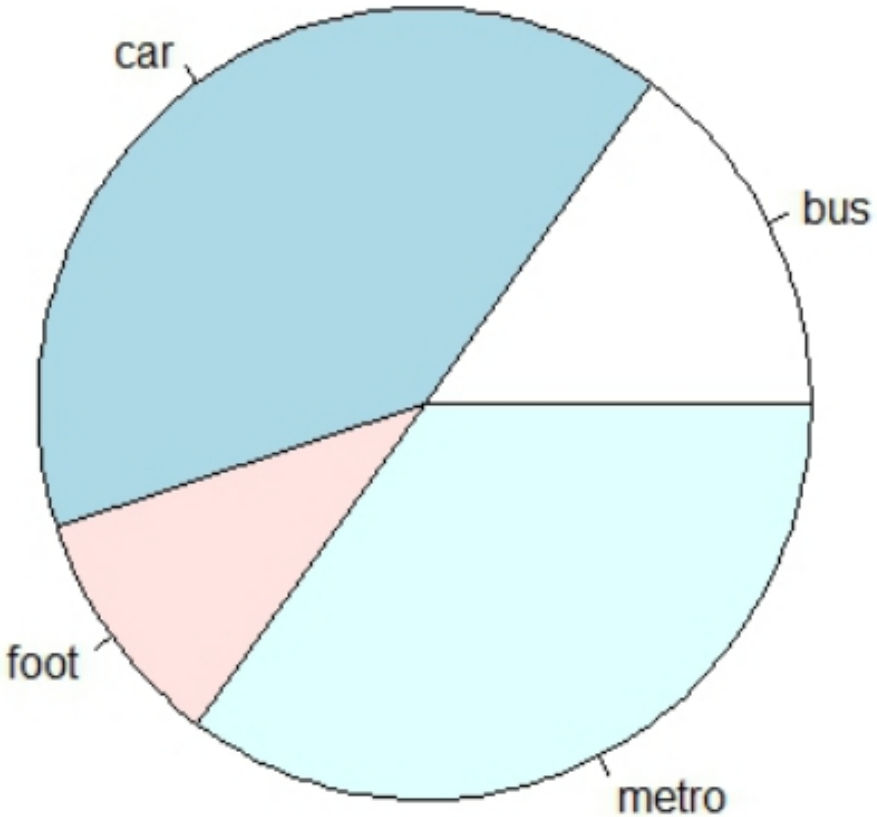
Αντίστοιχα, το ραβδόγραμμα σχετικών συχνοτήτων προκύπτει με την εντολή:
barplot(prop.table(table(m)))



Σχήμα: Παράδειγμα σχεδίασης ραβδοβράμματος με χρήση σχετικών συχνοτήτων.

Διαγράμματα Πίτας

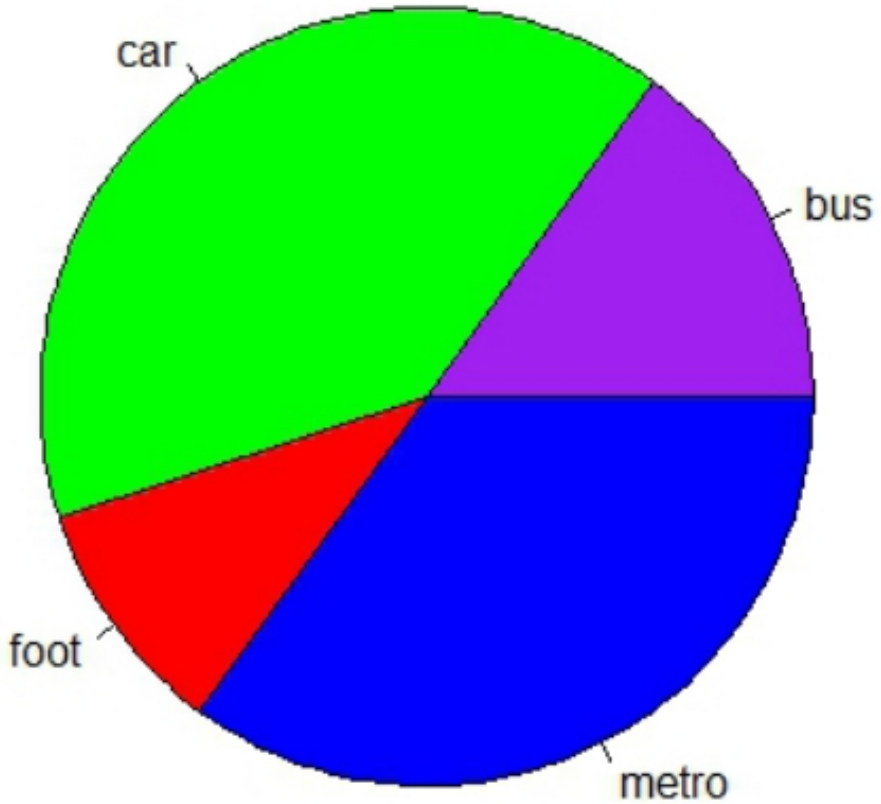
Μπορούμε να απεικονίσουμε τα δεδομένα μας σε μορφή διαγράμματος πίτας ή τομεογράφημα (pie chart), μέσω της εντολής: **barplot(pie(table(m))**



Σχήμα: Παράδειγμα διαγράμματος πίτας.

Διαγράμματα Πίτας

ή ακόμα και να αποδώσουμε διαφορετικό χρωματισμό στους κυκλικούς τομείς του διαγράμματος πίτας: **pie(prop.table(table(m)), col=c("purple", "green", "red", "blue"))**



Σχήμα: Παράδειγμα διαγράμματος πίτας.

Πίνακες Συνάφειας

Ένας πίνακας συνάφειας (contingency matrix) αφορά δύο κατηγορηματικές μεταβλητές και απεικονίζει την κατανομή των συχνοτήτων τους.

Η χρήση του θα γίνει κατανοητή με το ακόλουθο παράδειγμα.

Ας υποθέσουμε ότι στα δεδομένα της προηγούμενης έρευνας έχουμε συμπεριλάβει και το φύλο κάθε ατόμου.

Έστω ότι, για απλότητα, τα 8 πρώτα άτομα ήταν άνδρες (M) και τα υπόλοιπα 12 ήταν γυναίκες (F).

Δημιουργούμε το διάνυσμα g.

```
g = c(rep("M",8),rep("F",12))
> g
 [1] "M" "M" "M" "M" "M" "M" "M" "M" "F" "F" "F" "F" "F" "F" "F" "F"
```

Πίνακες Συνάφειας

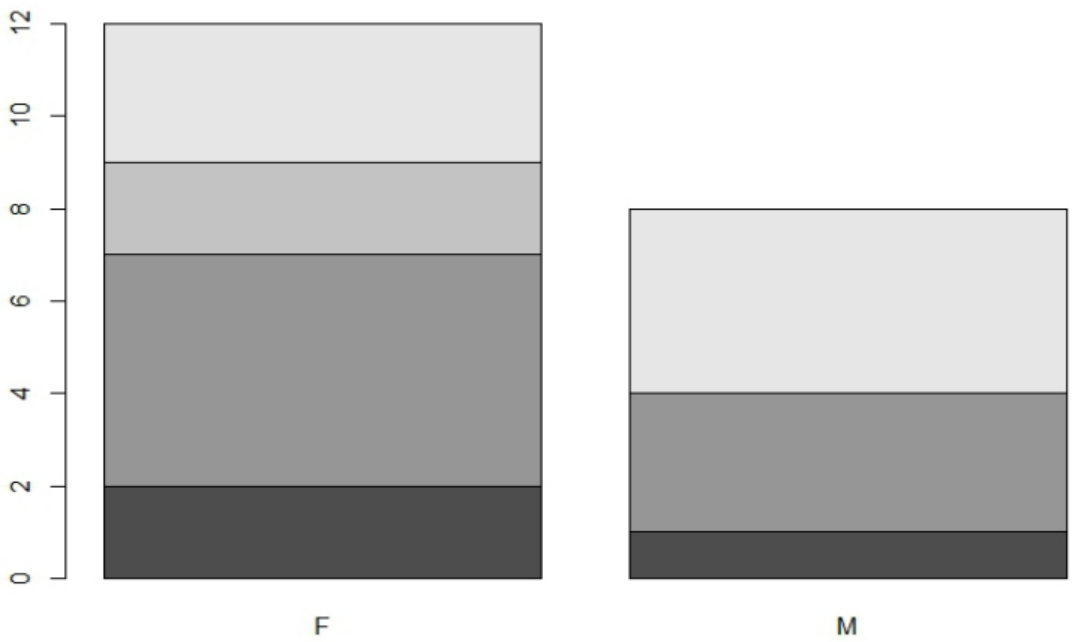
Παρατηρήστε τον γρήγορο τρόπο αρχικοποίησης του διανύσματος g μέσω της συνάρτησης rep(), λόγω της υπόθεσης που κάναμε παραπάνω. Δημιουργούμε τον πίνακα συχνοτήτων διπλής εισόδου mg ως εξής:

```
> m = c("car", "car", "bus", "metro", "metro", "car", "metro", "met
+           "car", "car", "metro", "car")
> m
 [1] "car"   "car"   "bus"   "metro" "metro" "car"   "metro" "metro"
"car"   "foot"  "bus"
[13] "bus"   "metro" "metro" "car"   "car"   "car"   "metro" "car"
> mg = table(m,g)
> mg
g
m       F M
  bus   2 1
  car   5 3
  foot  2 0
  metro 3 4
```

Στοιβαγμένα Ραβδογράμματα και Ομαδοποιημένα Ραβδογράμματα

Μπορούμε να αναπαραστήσουμε ποιοτικά δεδομένα, τα οποία προέρχονται από τις τιμές δύο μεταβλητών με χρήση ενός στοιβαγμένου ραβδογράμματος (stacked barplot) ή ενός ομαδοποιημένου ραβδογράμματος (grouped barplot).

```
barplot(mg)
```

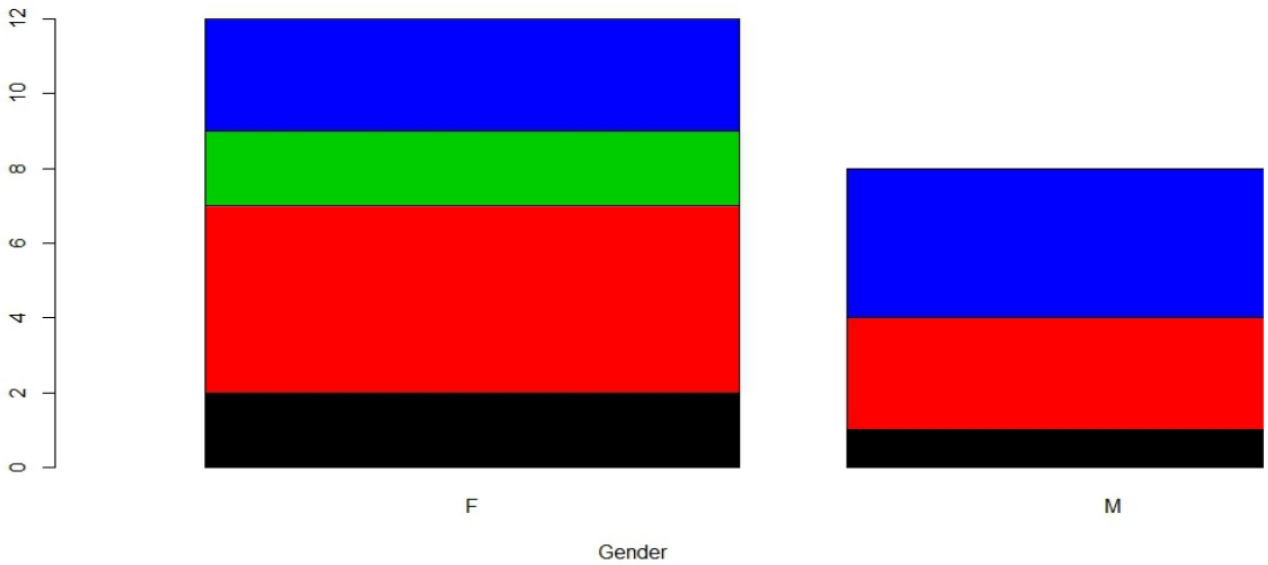


Σχήμα: Παράδειγμα στοιβαγμένου ραβδογράμματος.

Στοιβαγμένα Ραβδογράμματα και Ομαδοποιημένα Ραβδογράμματα

Χρησιμοποιώντας τις παραμέτρους της εντολής, το αποτέλεσμα είναι πολύ καλύτερο:

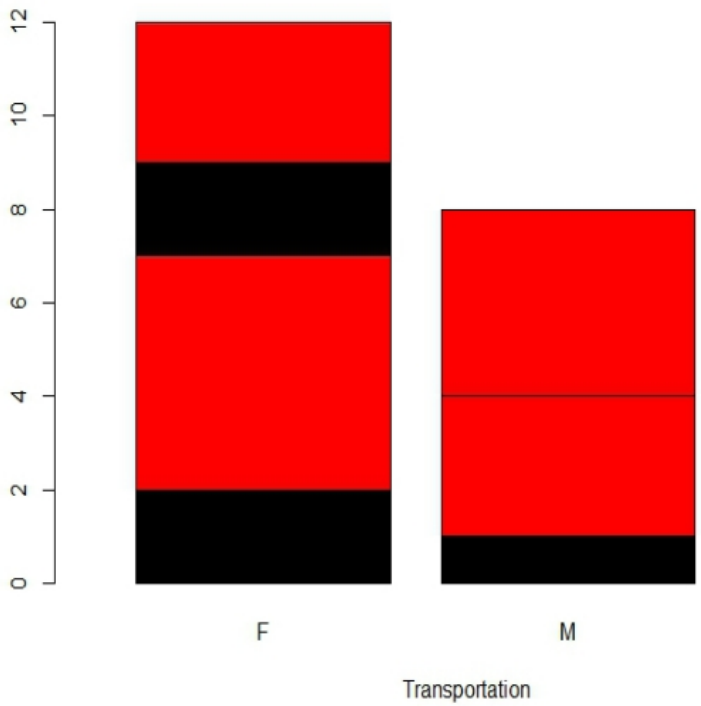
```
barplot(mg, xlim = c(0,2), xlab"="Gender, legent = levels(m),
col = 1:4)
```



Σχήμα: Παράδειγμα σχεδίασης στοιβαγμένου ραβδογράμματος.

Στοιβαγμένα Ραβδογράμματα και Ομαδοποιημένα Ραβδογράμματα

```
barplot(mg, xlim = c(0,2), xlab="Transportation",
legent =levels(g), col=1:2)
```

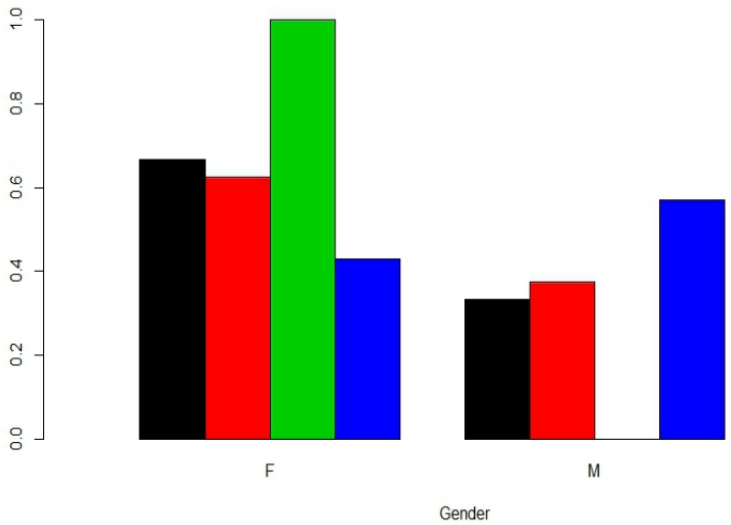


Σχήμα: Παράδειγμα σχεδίασης στοιβαγμένου ραβδογράμματος

Στοιβαγμένα Ραβδογράμματα και Ομαδοποιημένα Ραβδογράμματα

Οι εντολές που ακολουθούν δημιουργούν ομαδοποιημένα ραβδογράμματα.

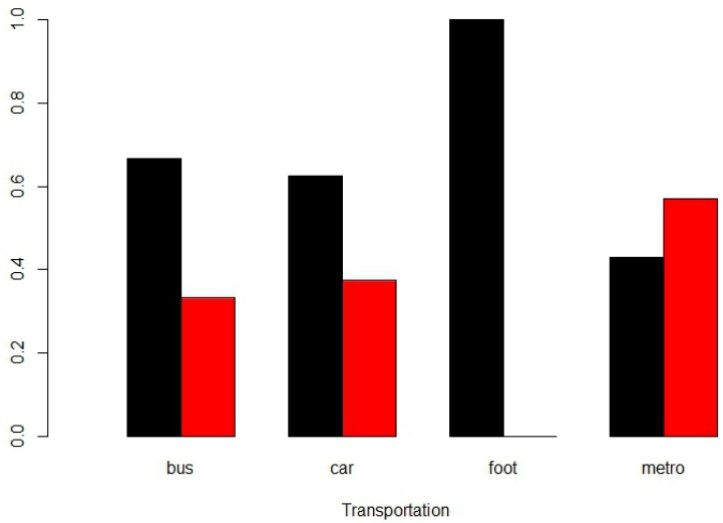
```
barplot(prop.table(mg,1), width=0.25, xlim = c(0,3), ylim =
c(0,1), xlab="Gender", legent = levels(m), beside=T, col = 1:4)
```



Σχήμα: Παράδειγμα σχεδίασης ομαδοποιημένου ραβδογράμματος

Στοιβαγμένα Ραβδογράμματα και Ομαδοποιημένα Ραβδογράμματα

```
> mg = table(g,m)
> barplot(prop.table(mg,2), width=0.25, xlim = c(0,3), ylim =
c(0,1), xlab"="Transportation, legent = levels(a), beside=T,
col = 1:2)
```



Σχήμα: Παράδειγμα σχεδίασης ομαδοποιημένου ραβδογράμματος.

Οπτικοποίηση Ποσοτικών Δεδομένων

Ας υποθέσουμε ότι οι παρακάτω τιμές αφορούν τη βαθμολογία ενός δείγματος 30 φοιτητών στο μάθημα «Εξόρυξη Δεδομένων»:

10, 10, 5, 9, 7, 6, 8, 6, 5, 8, 10, 7, 7, 8, 5, 6, 4, 7, 9, 7, 4, 8, 10, 10, 7, 4, 9, 5, 8, 9

Μπορούμε να παρουσιάσουμε τα δεδομένα μέσω ενός πίνακα συχνοτήτων, ως εξής:

```
> x = c(10, 10, 5, 9, 7, 6, 8, 6, 5, 8, 10, 7, 7, 8, 5, 6, 4,
+       7, 9, 7, 4, 8, 10, 10, 7, 4, 9, 5, 8, 9)
> table(x)
x
4  5  6  7  8  9 10
3  4  3  6  5  4  5
```

Ο παραπάνω πίνακας στην πρώτη γραμμή εμφανίζει τις διακεκριμένες τιμές των παρατηρήσεων και στη δεύτερη γραμμή τη συχνότητα εμφάνισης της κάθε τιμής.

Μπορείτε εύκολα να επαληθεύσετε ότι για παράδειγμα η τιμή 7 εμφανίζεται 6 φορές στο δείγμα.

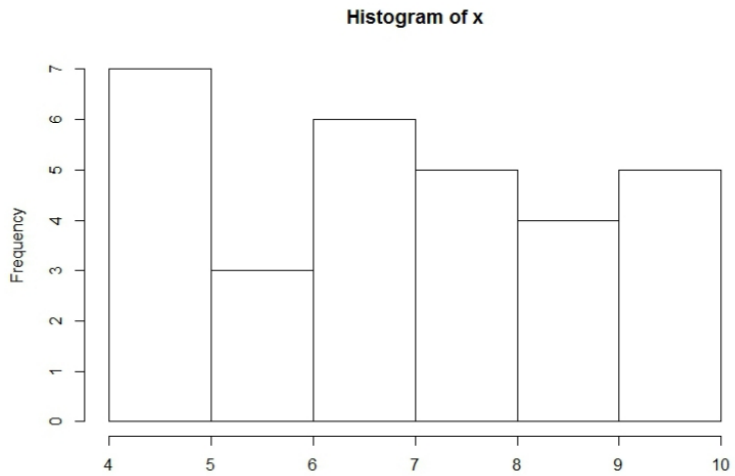
Μπορούμε αντίστοιχα να εμφανίζουμε και τον πίνακα σχετικών συχνοτήτων:

```
> prop.table(table(x))
x
        4         5         6         7         8         9        10
0.1000000 0.1333333 0.1000000 0.2000000 0.1666667 0.1333333 0.16666
```

Για την παρουσίαση ποσοτικών δεδομένων χρησιμοποιούμε τα ιστογράμματα (histograms), μέσω της εντολής hist().

Για το διάνυσμα x, το ιστόγραμμα συχνοτήτων προκύπτει με την εντολή:

```
hist(x)
```



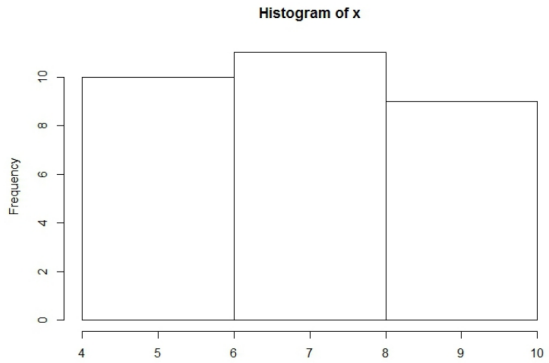
Σχήμα: Παράδειγμα ιστογράμματος.

Τα δεδομένα ομαδοποιούνται σε κλάσεις, οι οποίες αναπαριστούνται με διαδοχικά ορθογώνια.

Η βάση κάθε ορθογωνίου αντιστοιχεί στο εύρος της αντίστοιχης κλάσης, ενώ το ύψος αντιστοιχεί στη συχνότητα της αντίστοιχης παρατήρησης.

Συνήθως δημιουργούμε κλάσεις ιδίου εύρους. Μπορούμε, αν θέλουμε, να προσδιορίσουμε εμείς το πλήθος των κλάσεων, ως εξής:

```
hist(x, nclass=3)
```



Σχήμα: Παράδειγμα σχεδίασης ιστογράμματος με προκαθορισμένο πλήθος κλάσεων.

Η χρησιμότητα του ιστογράμματος θα γίνει πιο κατανοητή σε ένα παράδειγμα με μεγαλύτερο πλήθος παρατηρήσεων.

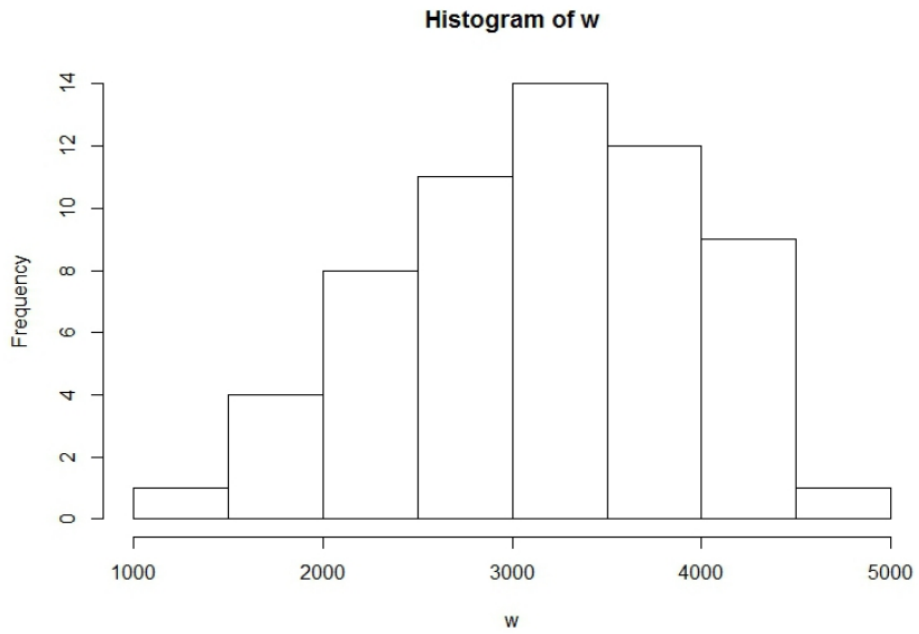
Ας υποθέσουμε ότι έχουμε διαθέσιμα τα αποτελέσματα της μέτρησης του βάρους (σε γραμμάρια) γέννησης 60 βρεφών:

1950, 2090, 2700, 3350, 4200, 3720, 4400, 2980, 3850, 4550 3050, 2350, 1850, 2820, 3670, 2950, 3750, 1850, 2420, 3150 3000, 3470, 3920, 3100, 2400, 2900, 2650, 3450, 3650, 4020 4450, 3120, 3660, 3070, 3550, 2020, 3500, 2500, 3780, 3940 3540, 2800, 2850, 4450, 1950, 3020, 2800, 3500, 1480, 4495 2850, 3100, 2250, 3300, 4100, 3220, 3600, 2130, 4020, 4075

Αρχικά εισάγουμε τις παρατηρήσεις μας μέσω του διανύσματος w:

```
> w = c(1950, 2090, 2700, 3350, 4200, 3720, 4400, 2980, 3850, 4550,
+       3050, 2350, 1850, 2820, 3670, 2950, 3750, 1850, 2420, 3150,
+       4450, 3120, 3660, 3070, 3550, 2020, 3500, 2500, 3780, 3940,
+       3000, 3470, 3920, 3100, 2400, 2900, 2650, 3450, 3650, 4020,
+       3540, 2800, 2850, 4450, 1950, 3020, 2800, 3500, 1480, 4495,
+       2850, 3100, 2250, 3300, 4100, 3220, 3600, 2130, 4020, 4075)
```

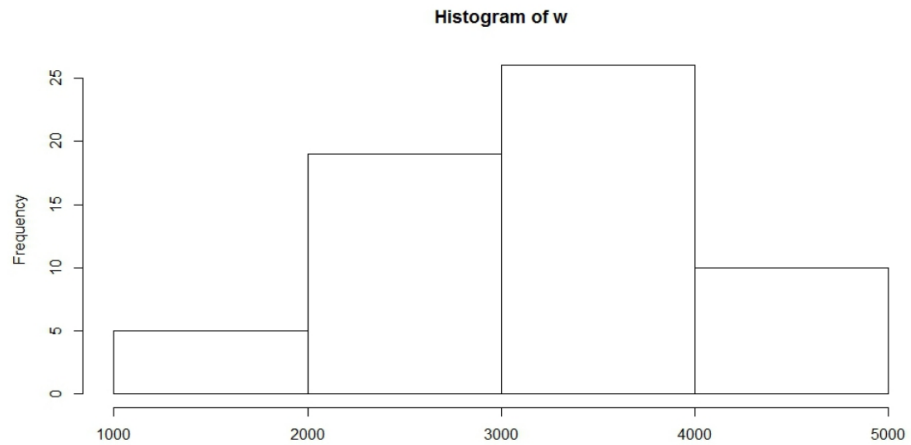
Η χρήση της εντολής: **hist(w)** δίνει το ιστογράφημα που ακολουθεί:



Σχήμα: Παράδειγμα σχεδίασης ιστογράμματος.

Η R υπολογίζει αυτόματα το πλήθος και το εύρος των κλάσεων.

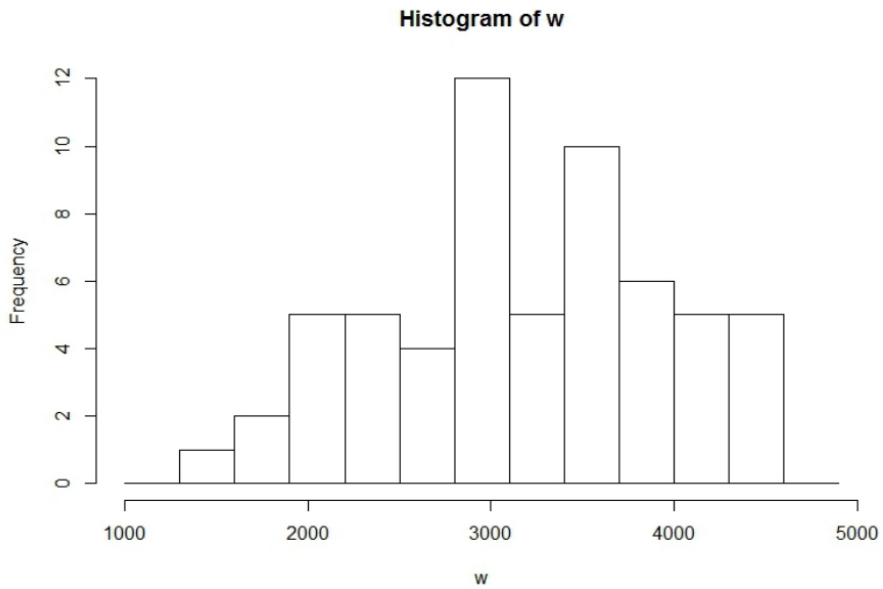
Αν θέλουμε να δημιουργήσουμε ένα ιστόγραμμα, όπου οι παρατηρήσεις μας να είναι ομαδοποιημένες, για παράδειγμα, σε 4 κλάσεις, αυτό γίνεται μέσω της εντολής: **hist(w, nclass = 4)**



Σχήμα: Σχεδίαση ιστογράμματος με προκαθορισμένο πλήθος κλάσεων.

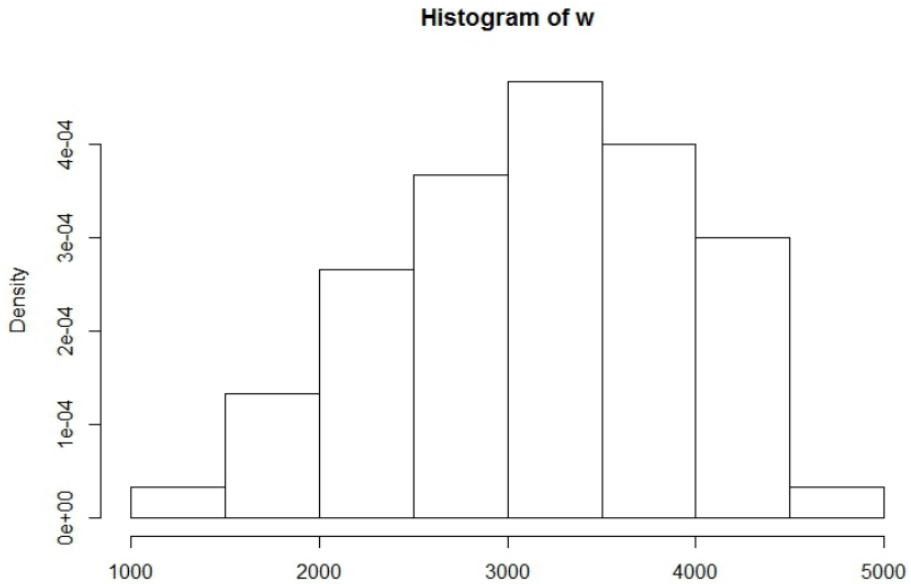
Μπορούμε, επίσης, να ορίσουμε την αρχή και το τέλος των κλάσεων μέσω μιας επαναληπτικής διαδικασίας.

> hist(w, breaks = seq(from = 1000, to=5000, by=300))



Σχήμα: Σχεδίαση ιστογράμματος με προκαθορισμένη αρχική και τελική τιμή και εύρος κλάσεων.

Επιπλέον, μπορούμε στον κάθετο άξονα, αντί για συχνότητα, να έχουμε τη σχετική συχνότητα, δηλαδή την πυκνότητα πιθανότητας, ως εξής: **hist(w, probability=T)**



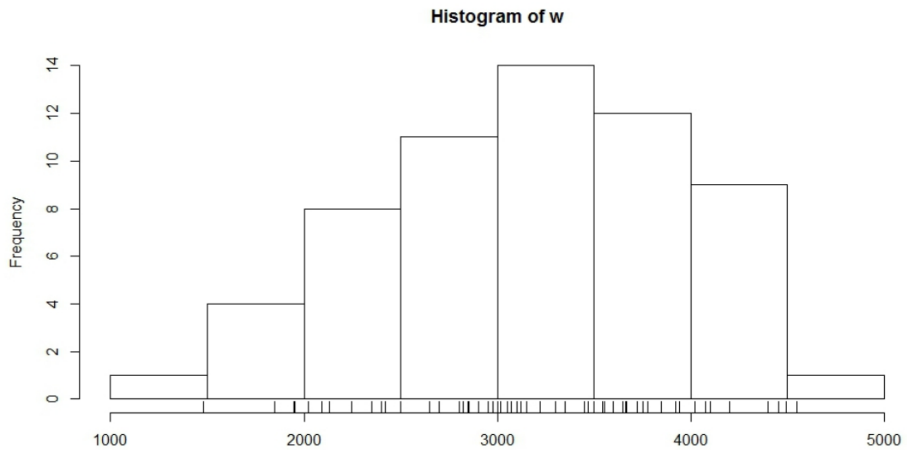
Σχήμα: Σχεδίαση ιστογράμματος μέσω σχετικών συχνοτήτων.

Οι εντολές:

> hist(w)

> rug(jitter(w))

εμφανίζουν, εντός από το ιστόγραμμα, και τις τιμές των παρατηρήσεων.



Σχήμα: Εμφάνιση ιστογράμματος και των τιμών των παρατηρήσεων.

Το θηκόγραμμα (boxplot) είναι ένας κατάλληλος τρόπος, για να παρουσιάσουμε τα κυριότερα χαρακτηριστικά της κατανομής των παρατηρήσεων του δείγματος.

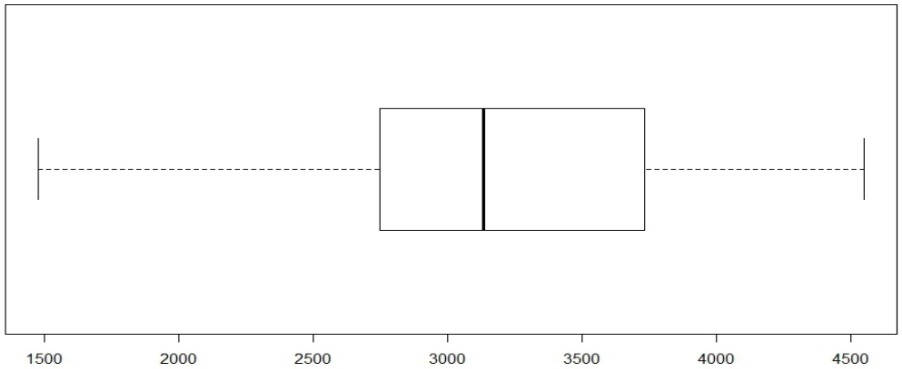
Πρόκειται για ένα ορθογώνιο που βασίζεται στις τιμές του 1ου, του 2ου (της διαμέσου) και του 3ου τεταρτημορίου, ενώ οι μύστακες εκτείνονται από τη μικρότερη μέχρι τη μεγαλύτερη τιμή των παρατηρήσεων.

Το θηκόγραμμα (boxplot) είναι ένας κατάλληλος τρόπος, για να παρουσιάσουμε τα κυριότερα χαρακτηριστικά της κατανομής των παρατηρήσεων του δείγματος.

Πρόκειται για ένα ορθογώνιο που βασίζεται στις τιμές του 1ου, του 2ου (της διαμέσου) και του 3ου τεταρτημορίου, ενώ οι μύστακες εκτείνονται από τη μικρότερη μέχρι τη μεγαλύτερη τιμή των παρατηρήσεων.

Το θηκόγραμμα, για το παράδειγμα των βαρών γέννησης των βρεφών, μπορεί να σχεδιαστεί μέσω της εντολής:

```
boxplot(w, horizontal = T )
```



Σχήμα: Παράδειγμα οριζόντιας σχεδίασης θηκογράμματος.

Αν θέλετε, μπορείτε να καταργήσετε την οριζόντια σχεδίαση του θηκογράμματος (θα προκύψει το ίδιο θηκόγραμμα στραμμένο 90 μοίρες δεξιά).

Τις τιμές των πέντε στατιστικών μεγεθών, που χρησιμοποιούνται για την κατασκευή του θηκογράμματος, μπορούμε να τις δούμε μέσω της εντολής:

```
> fivenum(w)
[1] 1480 2750 3135 3735 4550
```

Τα θηκογράμματα είναι ένας ωραίος τρόπος, για να συγκρίνουμε δυο δείγματα μεταξύ τους.

Ας υποθέσουμε, για παράδειγμα, πως το δείγμα w1 αποτελείται από τις τιμές:

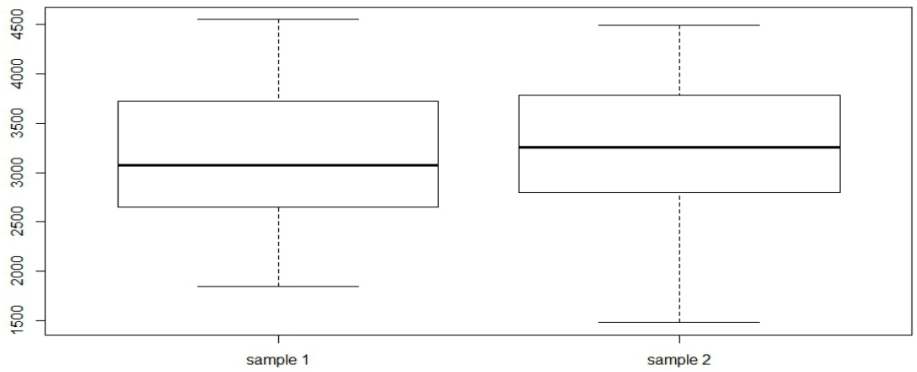
1950, 2090, 2700, 3350, 4200, 3720, 4400, 2980, 3850, 4550
3050, 2350, 1850, 2820, 3670, 2950, 3750, 1850, 2420, 3150
3000, 3470, 3920, 3100, 2400, 2900, 2650, 3450, 3650, 4020

και το δείγμα w2 από τις τιμές:
4450, 3120, 3660, 3070, 3550, 2020, 3500, 2500, 3780, 3940
3540, 2800, 2850, 4450, 1950, 3020, 2800, 3500, 1480, 4495
2850, 3100, 2250, 3300, 4100, 3220, 3600, 2130, 4020, 4075

την κατασκευή του θηκογράμματος, μπορούμε να τις δούμε μέσω της εντολής:

```
> fivenum(w)
[1] 1480 2750 3135 3735 4550
> w1 = c(1950, 2090, 2700, 3350, 4200, 3720, 4400, 2980, 3850, 4550
+        3050, 2350, 1850, 2820, 3670, 2950, 3750, 1850, 2420, 3150
+        3000, 3470, 3920, 3100, 2400, 2900, 2650, 3450, 3650, 4020
> fivenum(w1)
[1] 1850 2650 3075 3720 4550
> w2 = c(4450, 3120, 3660, 3070, 3550, 2020, 3500, 2500, 3780, 3940
+        3540, 2800, 2850, 4450, 1950, 3020, 2800, 3500, 1480, 4495
+        2850, 3100, 2250, 3300, 4100, 3220, 3600, 2130, 4020, 4075
> fivenum(w2)
[1] 1480 2800 3260 3780 4495
```

boxplot(w1, w2, names = c(“sample 1”,”sample 2”))



Σχήμα: Σύγκριση διαφορετικών δειγμάτων μέσω θηκογραμμάτων.