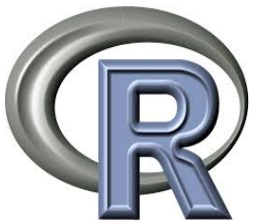
Εισαγωγή στη γλώσσα προγραμματισμού R

Αναστάσιος Κατσιλέρος

Γεωπονικό Πανεπιστήμιο Αθηνών
Εργαστήριο Βελτίωσης Φυτών και Γεωργικού Πειραματισμού

katsileros@aua.gr

Αθήνα 2020



Βασικές Στατιστικές Έννοιες

Το φαινόμενο εμφάνισης διαφορών στις παρατηρήσεις ή μετρήσεις μεταξύ μονάδων που ανήκουν στην ίδια ομάδα ή κατηγορία, ονομάζεται **παραλλακτικότητα** (variation). Η παραλλακτικότητα μπορεί να είναι έντονη σε κάποιες μονάδες, όπως για παράδειγμα μεταξύ ίδιων οργανισμών (ζώα, φυτά κ.α.) ή πάρα πολύ μικρή όταν οι μονάδες προέρχονται από μία παραγωγική διαδικασία (π.χ. προϊόντα, τρόφιμα κ.α.). Τα αίτια της παραλλακτικότητας μπορεί να είναι γνωστά (π.χ. γονότυπος, περιβάλλον, πρώτες ύλες κ.α.) ή άγνωστα στον ερευνητή.

Το χαρακτηριστικό ή ιδιότητα των μονάδων από τις οποίες λαμβάνονται αυτές οι παρατηρήσεις ή μετρήσεις αναφέρεται ως **μεταβλητή** (variable). Οι μεταβλητές διακρίνονται σε δύο γενικές κατηγορίες, τις **ποσοτικές** (quantitative) και τις **ποιοτικές** (qualitative). Οι ποσοτικές μεταβλητές μπορούν να μετρηθούν και διαχωρίζονται σε **συνεχείς** (continuous), όπου μπορούν να πάρουν αριθμητικές τιμές που καλύπτουν ένα διάστημα τιμών (π.χ. βάρος, ύψος, απόδοση κ.α.) και σε **διακριτές** (discrete), όπου μπορούν να πάρουν τιμές από ένα αριθμήσιμο σύνολο (π.χ αριθμός απογόνων, αριθμός στάχεων κ.α.).

Οι ποιοτικές μεταβλητές δεν μπορούν να μετρηθούν και διακρίνονται σε **κατηγοριοποιημένες - ονομαστικές** (nominal), όπου επιδέχονται μόνο αυθαίρετη κατάταξη (π.χ σχήμα σπόρου, χρώμα άνθεων, απουσία-παρουσία ζωνών κατά την ηλεκτροφόρηση DNA κ.α.) και οι **διατακτικές** (ordinal) οι οποίες κατατάσσονται σε ιεραρχική σειρά (π.χ. κλίμακα πλαγιάσματος σιτηρών ή ανθεκτικότητας σε μία ασθένεια κ.α.).

Το σύνολο όλων των δυνατών ή υποθετικών παρατηρήσεων μίας μεταβλητής ονομάζεται **στατιστικός πληθυσμός** (statistical population). Ο πληθυσμός μπορεί να είναι υπαρκτός ή μη και το μέγεθος του μπορεί να είναι πεπερασμένο ή άπειρο. Η μελέτη όλων των μονάδων ενός στατιστικού πληθυσμού είναι επιθυμητό, αλλά στις περισσότερες περιπτώσεις ακόμη και στους πεπερασμένους πληθυσμούς είναι αδύνατο, επομένως ο ερευνητής επιλέγει ένα τμήμα του πληθυσμού, απ' όπου τελικά εξάγει πληροφορίες για τα χαρακτηριστικά ενός συγκεκριμένου πληθυσμού.

Για να είναι το τμήμα αντιπροσωπευτικό του πληθυσμού αλλά και για να διασφαλίζεται η αμεροληψία, θα πρέπει η επιλογή των μονάδων του τμήματος να γίνεται τυχαία από τον ερευνητή.

Ο τρόπος συλλογής των παρατηρήσεων από έναν στατιστικό πληθυσμό μπορεί να γίνει μέσω μιας κατάλληλης μεθόδου **δειγματοληψίας** (sampling), οπότε το τμήμα του πληθυσμού αναφέρεται ως **δείγμα** (sample) ή μετά από μία **επέμβαση** ή **μεταχείριση** (treatment) σύμφωνα με κάποιο **πειραματικό σχέδιο** (experimental design) ή από άμεση παρατήρηση ενός φαινομένου. Οι μέθοδοι δειγματοληψίας χρησιμοποιούνται όταν ο πληθυσμός είναι υπαρκτός και πεπερασμένος (π.χ. δημοσκόπηση σε μια χώρα, το ύψος φοιτητών σε μία σχολή), ενώ οι πειραματικοί σχεδιασμοί εφαρμόζεται συνήθως όταν ο πληθυσμός είναι άπειρος και μη υπαρκτός (π.χ. αξιολόγηση ενός νέου υβριδίου αραβόσιτου ή μιας νέας θεραπευτικής αγωγής).

Όταν ένα φαινόμενο δεν μπορεί να αναπαραχθεί τεχνητά- πειραματικά (π.χ. επιδημιολογικά, μετεωρολογικά, αστρονομικά φαινόμενα), η συλλογή των παρατηρήσεων γίνεται με άμεση παρατήρηση του φαινόμενου ή μέσω προσομοιώσεων. Οι μετρήσεις ή οι παρατηρήσεις οι οποίες συλλέγονται από τις δειγματοληπτικές-πειραματικές μονάδες με κάποια μέθοδο, αναφέρονται και ως **πρωτογενή δεδομένα** (raw data).

Περιγραφή Στατιστικών Δεδομένων

Παρόλο που ο ερευνητής χρησιμοποιεί ένα τμήμα του πληθυσμού, τα πρωτογενή δεδομένα που προκύπτουν συνήθως είναι μεγάλα σε μέγεθος. Προτού προχωρήσει στην ανάλυση των δεδομένων θα πρέπει να οργανώσει, να ταξινομήσει και να παρουσιάσει τα δεδομένα απλά και κατανοητά. Για το σκοπό αυτό χρησιμοποιούνται πίνακες συχνοτήτων, γραφικοί μεθόδοι και αριθμητικά περιγραφικά μέτρα.

Όταν οι τιμές προέρχονται από μία κατηγοριοποιημένη μεταβλητή, ο πίνακας συχνοτήτων έχει την πιο απλή μορφή, όπου στην πρώτη στήλη είναι τοποθετημένες οι κατηγοριοποιημένες τιμές και στην δεύτερη στήλη είναι η **απόλυτη συχνότητα** ή **συχνότητα *fi*** (frequency) των τιμών.
Όταν οι τιμές προέρχονται από μία διακριτή μεταβλητή, στην πρώτη στήλη τοποθετούνται οι διακριτές τιμές σε αύξουσα σειρά και επίσης εκτός της στήλης με τις συχνότητες, προστίθενται στον πίνακα συχνοτήτων η **σχετική συχνότητα** p_i, η **αθροιστική συχνότητα *Fi*** και η **αθροιστική σχετική συχνότητα *Pi***.

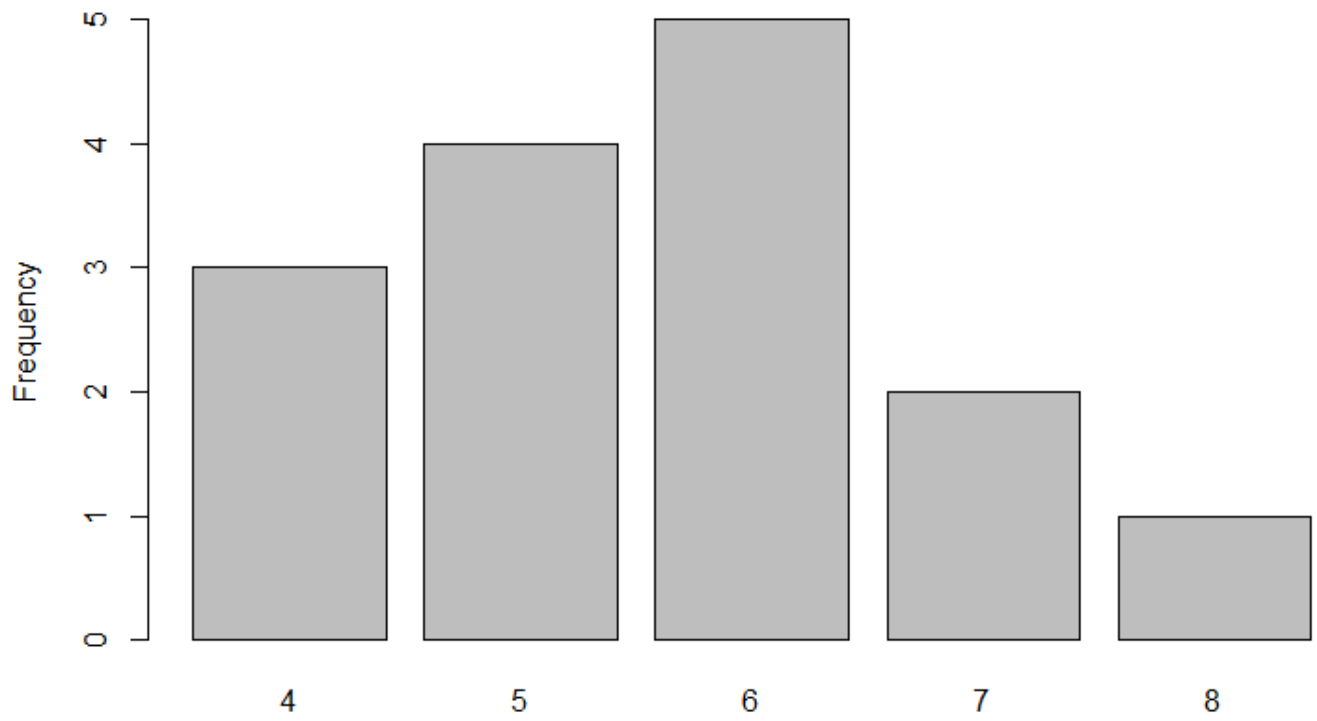
Παράδειγμα 1. Τα δεδομένα (7, 4, 5, 4, 6, 5, 7, 6, 5, 6, 8, 6, 4, 6, 5) αφορούν τη διάρκεια σπουδων δεκαπέντε φοιτητών, σε μία τετραετή σχολή.

Τιμές (Έτη)	**Απόλυτη Συχνότητα** f_i	**Σχετική Συχνότητα** p_i	**Αθροιστική Συχνότητα** F_i	**Σχετική Αθροιστική Συχνότητα** P_i
4	3	0,20	3	0,20
5	4	0,27	7	0,47
6	5	0,33	12	0,80
7	2	0,13	14	0,93
8	1	0,07	15	1
		1		

```
> y=c(7,4,5,4,6,5,7,6,5,6,8,6,4,6,5)
> Y = as.data.frame(table(y))
> Y = transform(Y, cumFreq = cumsum(Freq), relative =
prop.table(Freq))
> Y
  y Freq cumFreq relative
1 4    3       3 0.20000000
2 5    4       7 0.26666667
3 6    5      12 0.33333333
4 7    2      14 0.13333333
5 8    1      15 0.06666667
```

Το ραβδόγραμμα αποτελείται από ορθογώνιες στήλες, το πλάτος των βάσεών τους καθορίζονται αυθαίρετα και το ύψος είναι ίσο με την f_i ή p_i ή F_i.

```
> barplot(table(y), ylab = "Frequency")
```



Όταν οι τιμές προέρχονται από μία συνεχή ή διακριτή μεταβλητή με μεγάλο πλήθος τιμών, τότε δεν είναι δυνατή η εμφάνιση όλων των τιμών με τις αντίστοιχες συχνότητες στον πίνακα, οπότε είναι απαραίτητο η ομαδοποίηση των τιμών σε μια σειρά μη επικαλυπτόμενων διαστημάτων. Τα διαστήματα αυτά ονομάζονται **κλάσεις** ή **τάξεις** (classes). Ο αριθμός και το εύρος των κλάσεων επιλέγονται αυθαίρετα από τον ερευνητή, έτσι ώστε η κατανομή να περιγράφει ικανοποιητικά τα δεδομένα. Για την επιλογή του αριθμού των κλάσεων μπορούν να χρησιμοποιηθούν διάφοροι τύποι, όπως ο τύπος του Sturges στον οποίο ο αριθμός των κλάσεων υπολογίζεται ως εξής:

$$k = 1 + 3,322 \log_{10} n$$

όπου k είναι ο αριθμός των κλάσεων και n το πλήθος των παρατηρήσεων

Παράδειγμα 2. Τα δεδομένα (2,8 3,4 2,6 3,2 2,8 2,5 2,8 3,0 2,9 2,8 2,7 2,7 2,8 3,0 2,4 2,9 3,0 3,1 2,5 2,3) αφορούν τον βάρος σε γραμμάρια των σπόρων του κύριου στάχυ, ανά φυτό σιταριού.

Κλάσεις	**Κέντρο Κλάσης** m_i	**Απόλυτη Συχνότητα** f_i	**Αθροιστική Συχνότητα** F_i	**Σχετική Συχνότητα** p_i	**Σχετική Αθροιστική Συχνότητα** P_i
(2,2-2,4]	2,3	2	2	0,10	0,10
(2,4-2,6]	2,5	3	5	0,15	0,25
(2,6-2,8]	2,7	7	15	0,35	0,60
(2,8-3,0]	2,9	5	17	0,25	0,85
(3,0-3,2]	3,1	2	19	0,10	0,95
(3,2-3,4]	3,3	1	20	0,05	1
		20		1	

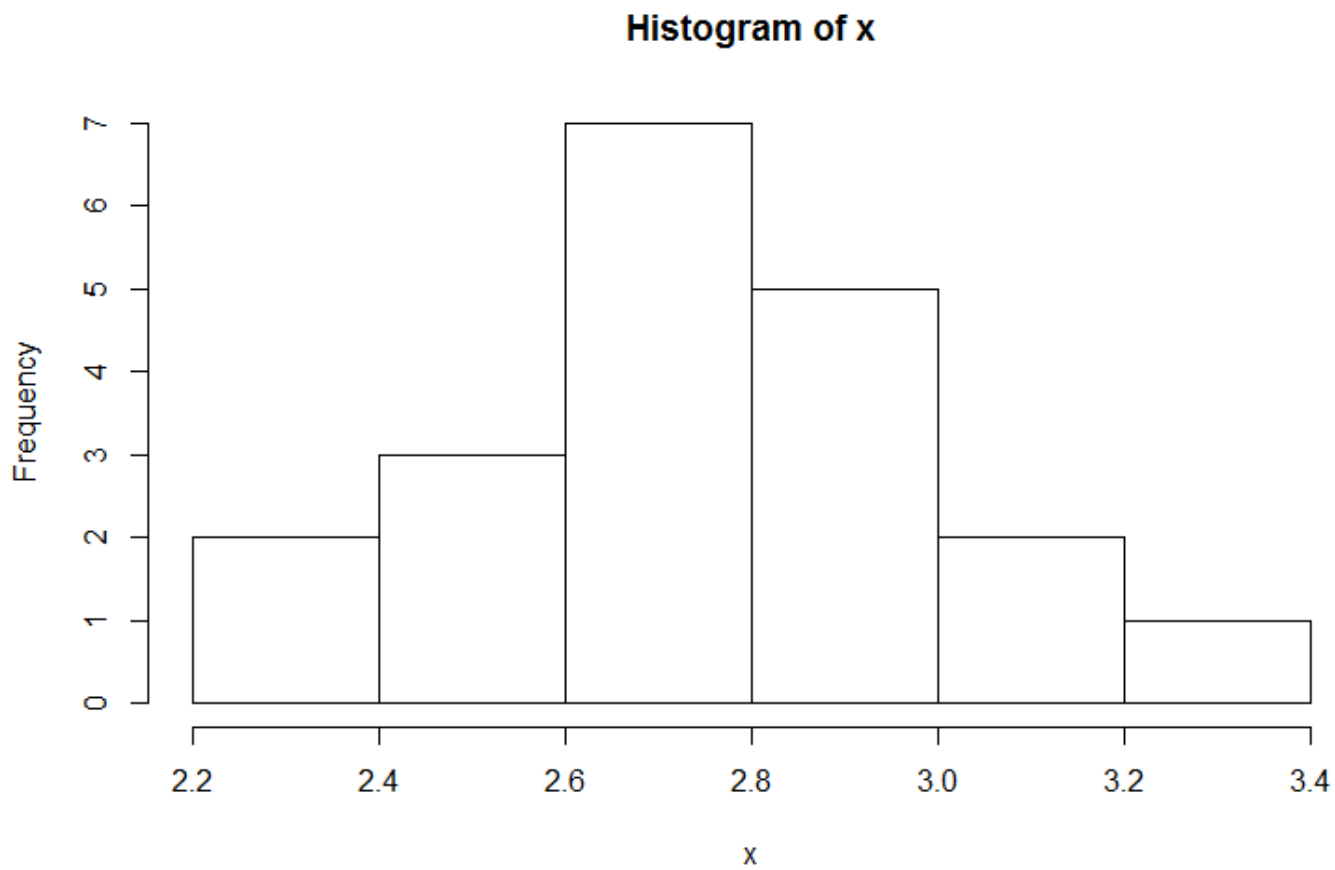
```
> x=c(2.8,3.4,2.6,3.2,2.8,2.5,2.8,3.0,2.9,2.8,2.7,2.7,2.8,3.0,
2.4,2.9,3.0,3.1,2.5,2.3)
> Sturges = 1 +3.322*log10(length(x))
> class=ceiling(Sturges)
> cut= cut(x, breaks=class) # ή breaks=nclass.Sturges(x)
> X = as.data.frame(table(cut))
> X = transform(X, cumFreq = cumsum(Freq), relative =
prop.table(Freq))
> X
```

	class	Freq	cumFreq	relative
1	(2.3,2.48]	2	2	0.10
2	(2.48,2.67]	3	5	0.15
3	(2.67,2.85]	7	12	0.35
4	(2.85,3.03]	5	17	0.25
5	(3.03,3.22]	2	19	0.10
6	(3.22,3.4]	1	20	0.05

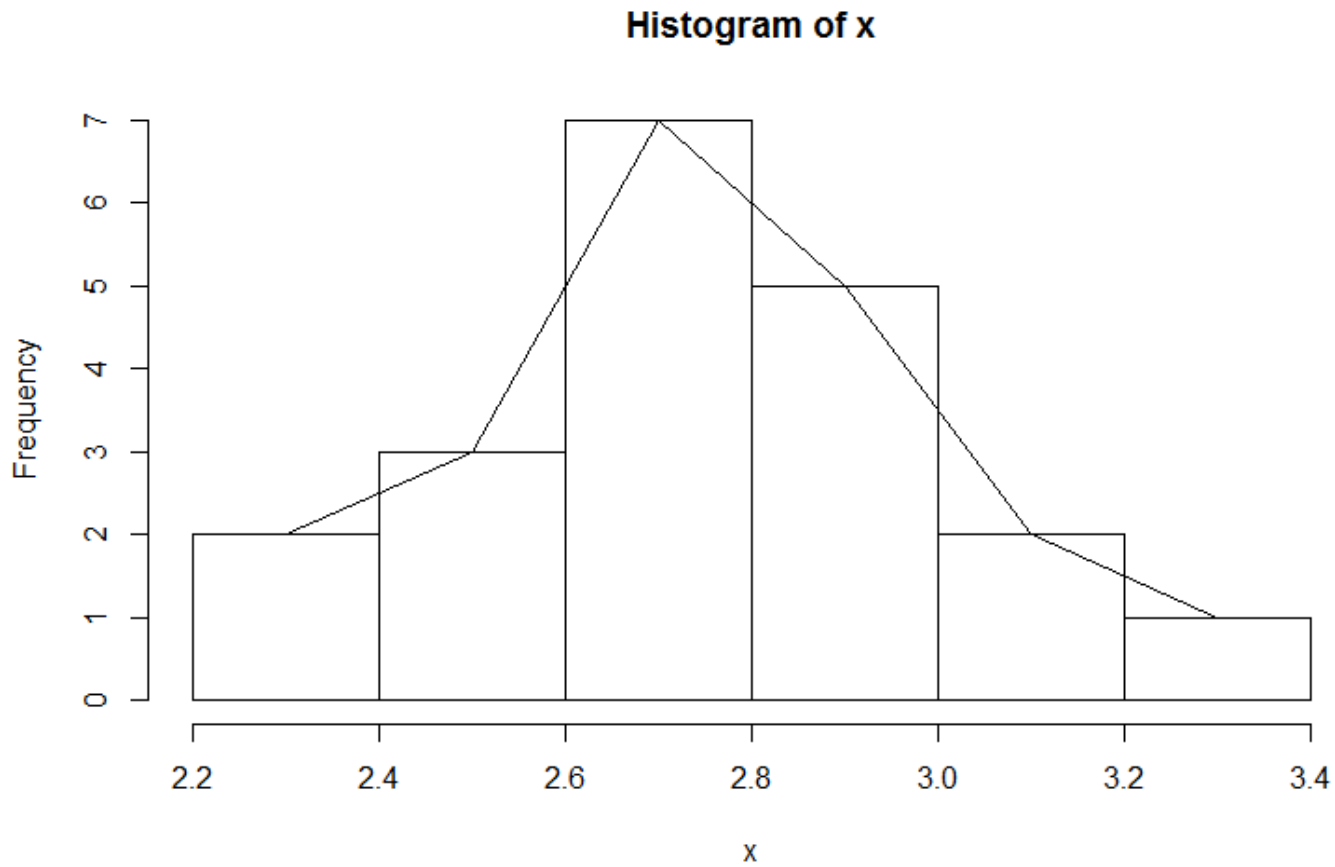
```
> h=hist(x)
> str(h)
List of 6
$ breaks : num [1:7] 2.2 2.4 2.6 2.8 3 3.2 3.4
$ counts : int [1:6] 2 3 7 5 2 1
$ density : num [1:6] 0.5 0.75 1.75 1.25 0.5 ...
$ mids : num [1:6] 2.3 2.5 2.7 2.9 3.1 3.3
$ xname : chr "x"
$ equidist: logi TRUE
- attr(*, "class")= chr "histogram"
```

Το ιστόγραμμα αποτελείται από ενωμένες μεταξύ τους ορθογώνιες στήλες, το πλάτος των βάσεων τους είναι ίσο με το εύρος της κλάσης και το ύψος είναι ίσο με yo f_i ή p_i ή F_i.

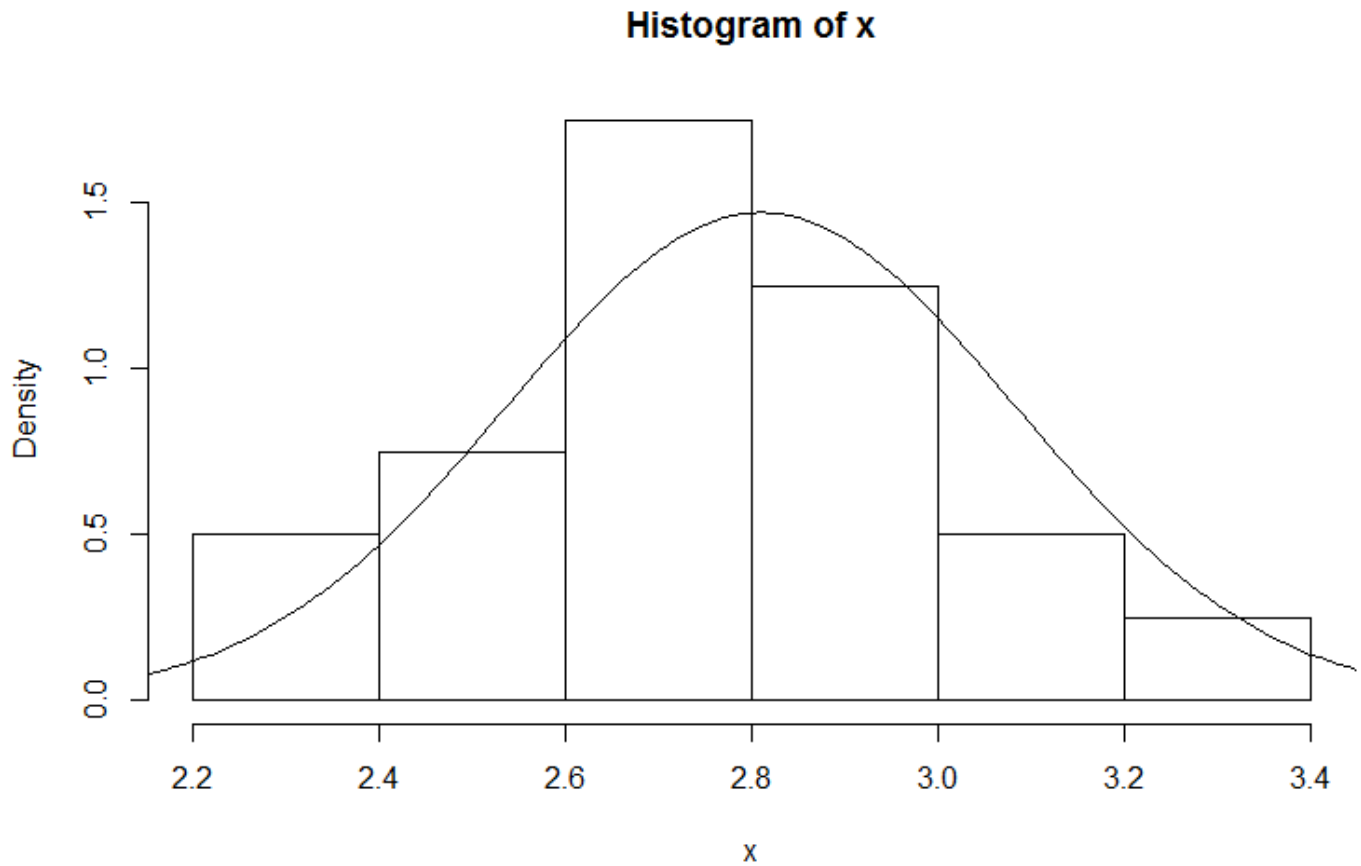
```
> hist(x, breaks = "Sturges")
```



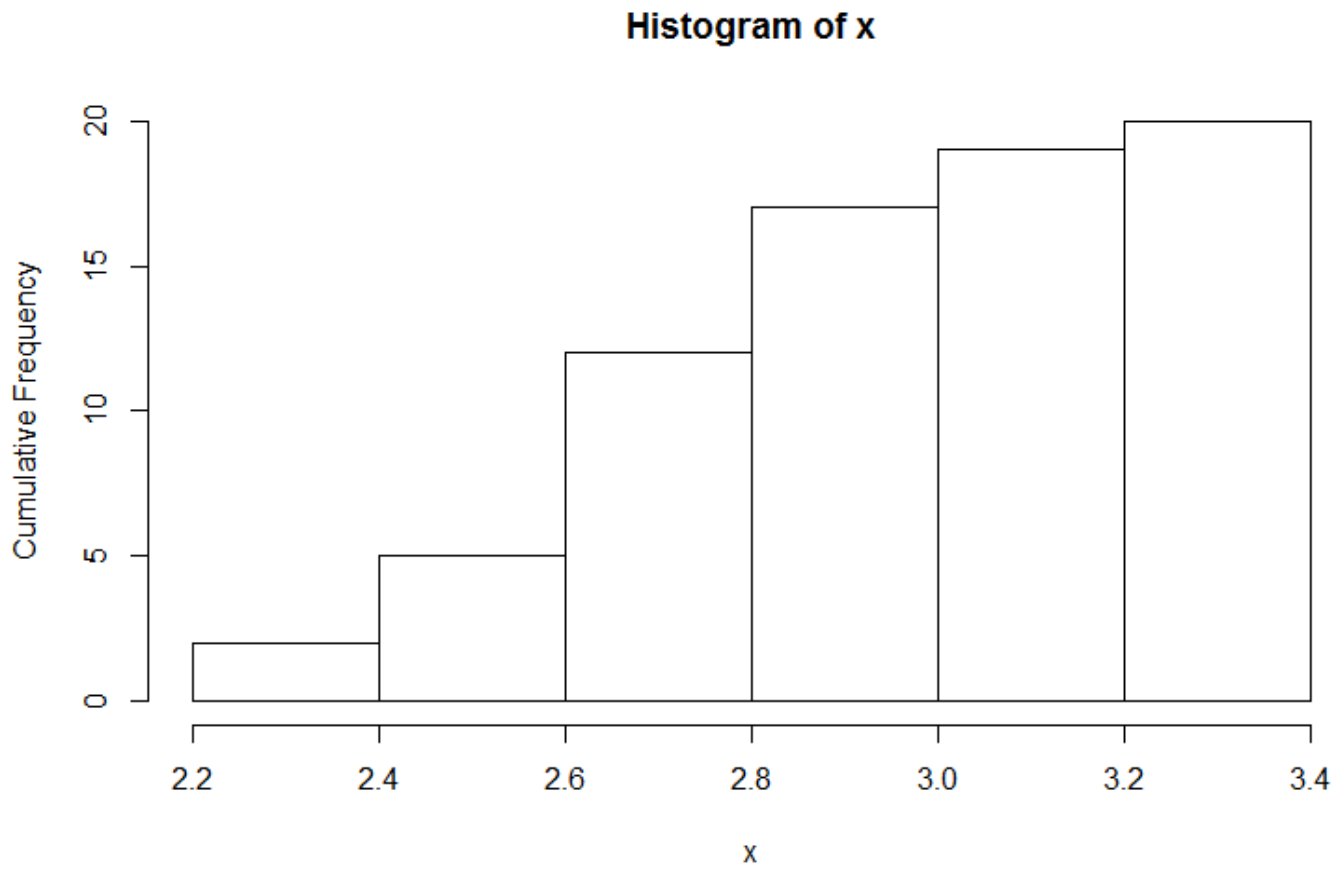
```
> lines(h$mids, h$counts)
```



```
> hist(x,freq=F)
> lines(seq(0,5,by=.01),dnorm(seq(0,5,by=0.01),mean(x),
sd(x)))
```



```
> h$counts=cumsum(h$counts)
> plot(h, ylab = "Cumulative Frequency")
```



Μέτρα Κεντρικής Τάσης

Τα μέτρα **κεντρικής τάσης** ή **θέσης** (central tendency measures) προσδιορίζουν ένα κεντρικό σημείο γύρω από το οποίο τείνουν να συγκεντρώνονται τα δεδομένα.
Τα κυριότερα μέτρα κεντρικής τάσης είναι η μέση τιμή, η διάμεσος, η επικρατούσα τιμή και τα ποσοστιαία σημεία.

Η **μέση τιμή πληθυσμού** (population mean) μεγέθους N, συμβολίζεται με το μ και ορίζεται ως:

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N} \tag{1}$$

Η **μέση τιμή δείγματος** (sample mean) αποτελεί το σημαντικότερο μέτρο θέσης και συμβολίζεται με $\bar{Y}$. Αν y_1, y_2, ..., y_n οι τιμές των παρατηρήσεων του δείγματος από έναν πληθυσμό, η μέση τιμή ορίζεται ως:

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2}$$

Η **διάμεσος** (median) ορίζεται ως η κεντρική τιμή η οποία χωρίζει ένα σύνολο διατεταγμένων σε αύξουσα σειρά παρατηρήσεων σε δύο ίσα μέρη και συμβολίζεται με $\tilde{Y}$. Αν ο αριθμός n των παρατηρήσεων είναι περιττός τότε η διάμεσος είναι η μεσαία παρατήρηση, ενώ αν είναι άρτιος είναι το ημιάθροισμα των μεσαίων παρατηρήσεων.

Η **επικρατούσα τιμή** ή **κορυφή** (mode) ορίζεται ως η τιμή ή τιμές που εμφανίζονται με τη μεγαλύτερη συχνότητα στα δεδομένα.

Το **p-ποσοστιαίο σημείο** (quantiles) ενός συνόλου παρατηρήσεων είναι η τιμή εκείνη για την οποία το $p\%$ των παρατηρήσεων είναι μικρότερες από αυτή την τιμή ($0 \leqslant p < 1$). Τα πιο χρησιμοποιούμενα ποσοστιαία σημεία είναι τα **τεταρτομόρια** (quartiles) και τα **εκατοστημόρια** (percentiles).

Το **πρώτο τεταρτημόριο** (first quartile) είναι η τιμή της παρατήρησης μέχρι την οποία περιλαμβάνεται το 25% των διατεταγμένων παρατηρήσεων του δείγματος και συμβολίζεται ως $Q1$.

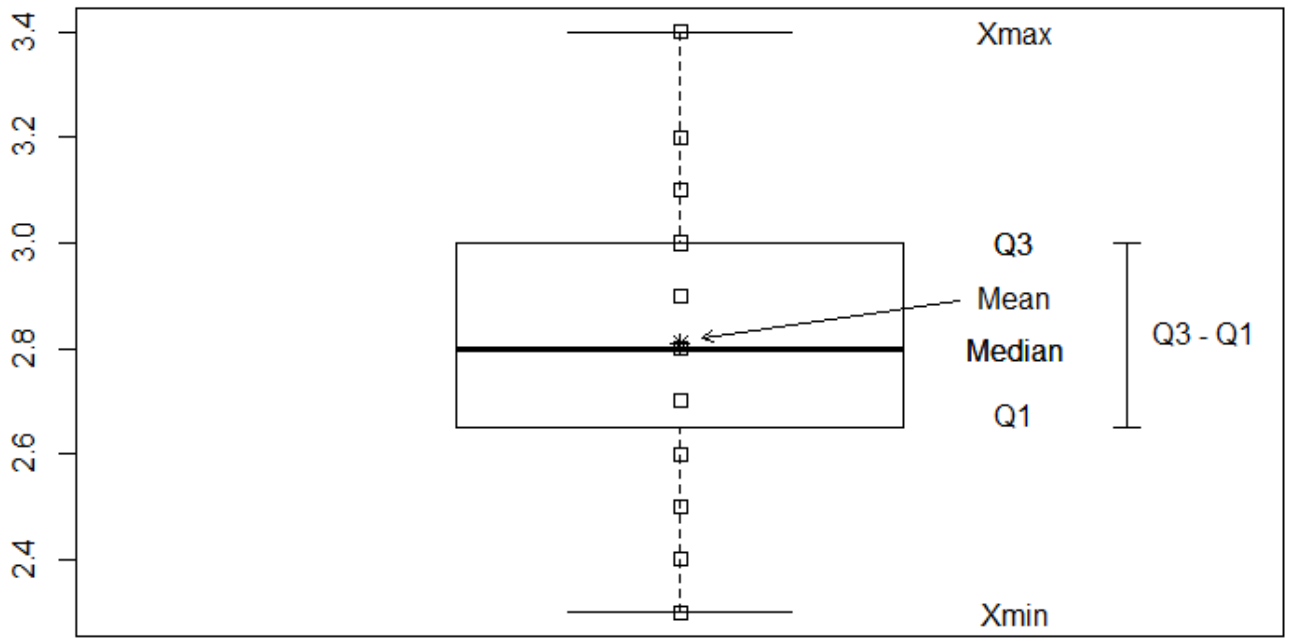
Το **δεύτερο τεταρτημόριο** (second quartile) είναι η τιμή της παρατήρησης μέχρι την οποία περιλαμβάνεται το 50% των διατεταγμένων παρατηρήσεων του δείγματος και συμβολίζεται ως $Q2$ το οποίο είναι και η διάμεσος.

Το **τρίτο τεταρτημόριο** (third quartile) είναι η τιμή της παρατήρησης μέχρι την οποία περιλαμβάνεται το 75% των διατεταγμένων παρατηρήσεων του δείγματος και συμβολίζεται ως $Q3$. Για τον υπολογισμό των τεταρτημορίων χρησιμοποιούνται διάφοροι αλγόριθμοι, στην πιο απλή περίπτωση τα τεταρτημόρια υπολογίζονται όπως η διάμεσος.

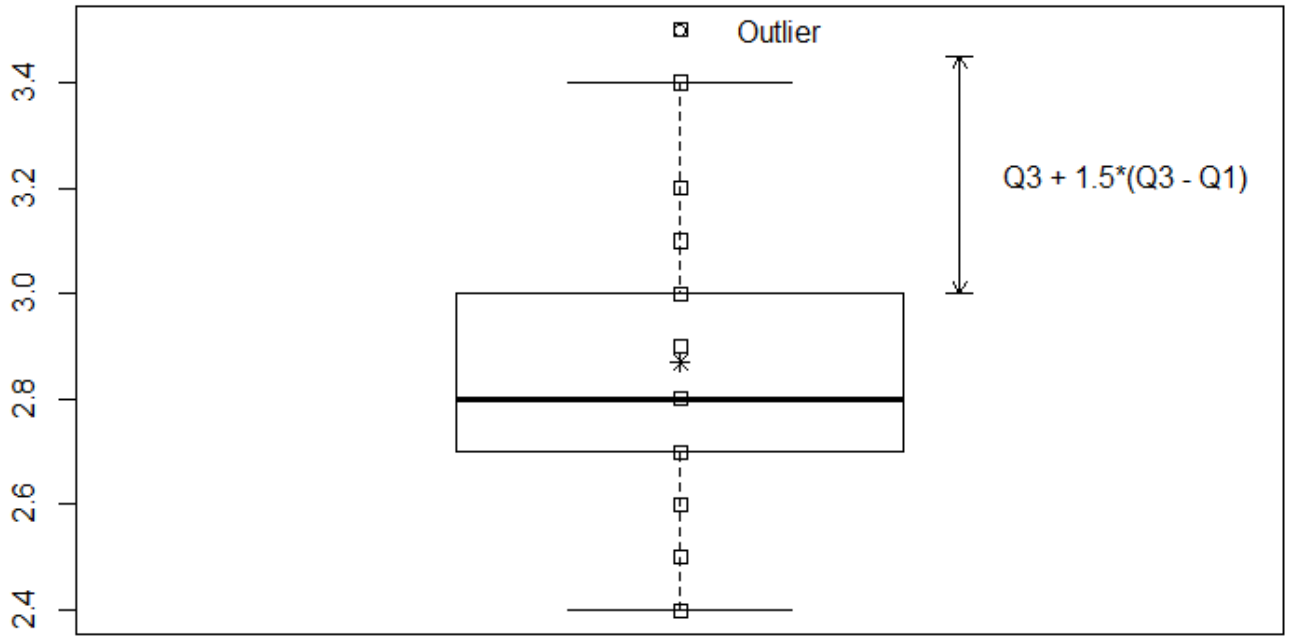
```
> mean(x, trim = 0 )
[1] 2.81
> median(x)
[1] 2.8
> quantile(x)
0% 25% 50% 75% 100%
2.300 2.675 2.800 3.000 3.400
> summary(x)
Min. 1st Qu. Median Mean 3rd Qu. Max.
2.300 2.675 2.800 2.810 3.000 3.400
```

Το θηκόγραμμα είναι πολύ χρήσιμο γραφικό εργαλείο που περιγράφει την συγκέντρωση των δεδομένων και δίνει μια εικόνα για την συμμετρικότητα της κατανομής τους. Για τη κατασκευή του θηκογράμματος δημιουργείται ένα ορθογώνιο πλαίσιο με βάσεις το πρώτο και το τρίτο τεταρτημόριο και εντός του πλαισίου τοποθετείται η διάμεσος. Από την μέση των βάσεων αναπτύσσονται γραμμές ή **κεραίες** (whiskers) που εκτείνονται μέχρι τις οριακές άνω και κάτω τιμές. Η άνω οριακή τιμή ορίζεται η μεγαλύτερη παρατήρηση η οποία είναι μικρότερη ή ίση από την τιμή $Q3 + 1.5 \cdot (Q3 - Q1)$, ενώ η κάτω οριακή τιμή, η μικρότερη παρατήρηση η οποία είναι μεγαλύτερη ή ίση από την τιμή $Q1 - 1,5 \cdot (Q3 - Q1)$.

```
> boxplot(x)
> stripchart(x, vertical=TRUE, add=TRUE); points(mean(x),
pch=8)
```



```
> boxplot(x1) # Αντικατάσταση τιμής 2,3 με την τιμή 3,5
> stripchart(x1, vertical=TRUE, add=TRUE); points(mean(x1),
pch=8)
```



Μέτρα μεταβλητότητας

Τα **μέτρα μεταβλητότητας** (variability measures) είναι αριθμητικά μεγέθη που δίνουν την **διασπορά – διασκόρπιση** (dispersion) των παρατηρήσεων γύρω από τις κεντρικές τιμές της κατανομής.
Τα κυριότερα μέτρα είναι το εύρος, το ενδοτεταρτομοριακό εύρος, η διακύμανση, η τυπική απόκλιση και ο συντελεστής παραλλακτικότητας.

Η **διακύμανση του πληθυσμού** σ^2 (population variance) μετράει τη διασπορά N παρατηρήσεων γύρω από τη μέση τιμή μ του πληθυσμού και είναι ο λόγος του αθροίσματος τετραγώνων των διαφορών ή αποκλίσεων $(Y_i - \mu)$ προς το σύνολο των παρατηρήσεων N:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(Y_i - \mu)^2}{N} \tag{3}$$

Η **διακύμανση του δείγματος** s^2 (sample variance) μετράει τη διασπορά n παρατηρήσεων γύρω από τη μέση τιμή του δείγματος ενός πληθυσμού. Η δειγματική διακύμανση ορίζεται ως ο λόγος του αθροίσματος τετραγώνων των αποκλίσεων $(Y_i - \bar{Y})$ προς τον αριθμό των παρατηρήσεων του δείγματος μείον ένα. Ο όρος του παρανομαστή ονομάζεται βαθμοί ελευθερίας.

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \tag{4}$$

Η **τυπική απόκλιση** (standard deviation) αποτελεί ένα μέτρο διασποράς των τιμών σε σχέση με την μέση τιμή, που εκφράζεται στην ίδια μονάδα μέτρησης.
Ορίζεται ως η τετραγωνική ρίζα της διακύμανσης του πληθυσμού:

$$\sigma = \sqrt{\sigma^2} \tag{5}$$

ή του δείγματος:

$$s = \sqrt{s^2} \tag{6}$$

Το **εύρος** (range) είναι η διαφορά της ελάχιστης από τη μέγιστη τιμή των δεδομένων και συμβολίζεται με R. Το εύρος αν και απλό στον υπολογισμό του δεν θεωρείται αξιόπιστο μέτρο διασποράς επειδή εξαρτάται από ακραίες τιμές.

Το **ενδοτεταρτομοριακό εύρος** (interquartile range) είναι η διαφορά του τρίτου με το πρώτο τεταρτημόριο (Q3 - Q1) η οποία περιλαμβάνει το 50% των τιμών της κατανομής και συμβολίζεται ως IQR.

Ο **συντελεστής παραλλακτικότητας** (coefficient of variation) ορίζεται ως ο λόγος της τυπικής απόκλισης s προς την μέση τιμή $\bar{Y}$ των παρατηρήσεων και συμβολίζεται ως $CV\%$.

$$CV\% = \frac{s}{\bar{Y}}\% \tag{7}$$

Ο συντελεστής παραλλακτικότητας είναι ανεξάρτητος από τις μονάδες μέτρησης και επιτρέπει τη σύγκριση της μεταβλητότητας διαφορετικών δεδομένων απαλλαγμένη από την επίδραση της μέσης τιμής. Το μέγεθος του συντελεστή παραλλακτικότητας εξαρτάται από το είδος της έρευνας.
Η γνώση του συντελεστή παραλλακτικότητας είναι πολύ χρήσιμη τόσο στην σχεδίαση των πειραμάτων όσο και στην αξιολόγηση των αποτελεσμάτων.

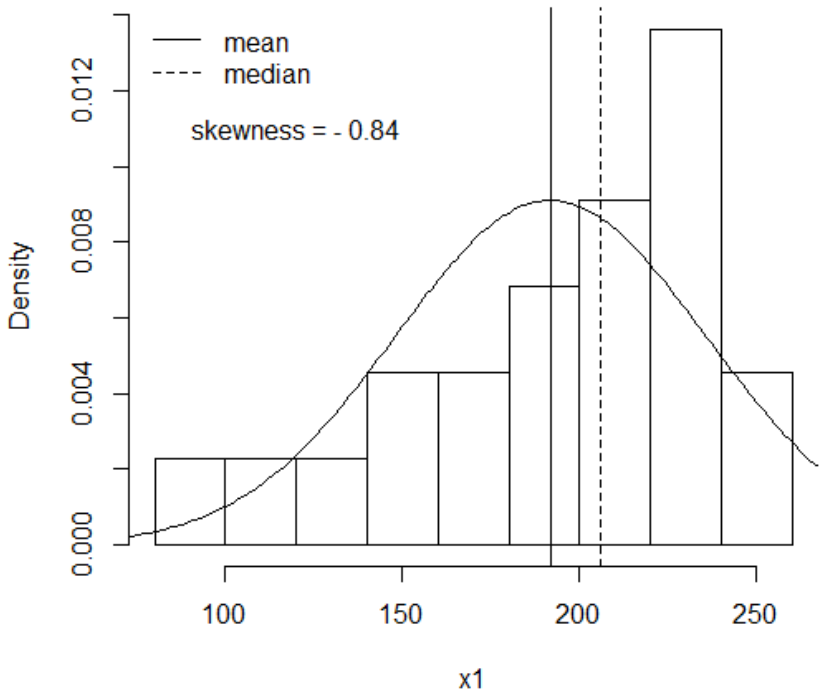
```
# Υπολογισμός ενδοτεταρτομοριακού εύρους
> IQR(x)
[1] 0.325
# Υπολογισμός διακύμανσης
> var(x)
[1] 0.07357895
# Υπολογισμός τυπικής απόκλισης
> sd(x)
[1] 0.2712544
# Υπολογισμός συντελεστή παραλλακτικότητας CV%
> CV = (sd(x)/mean(x))*100
> CV
[1] 9.653181
```

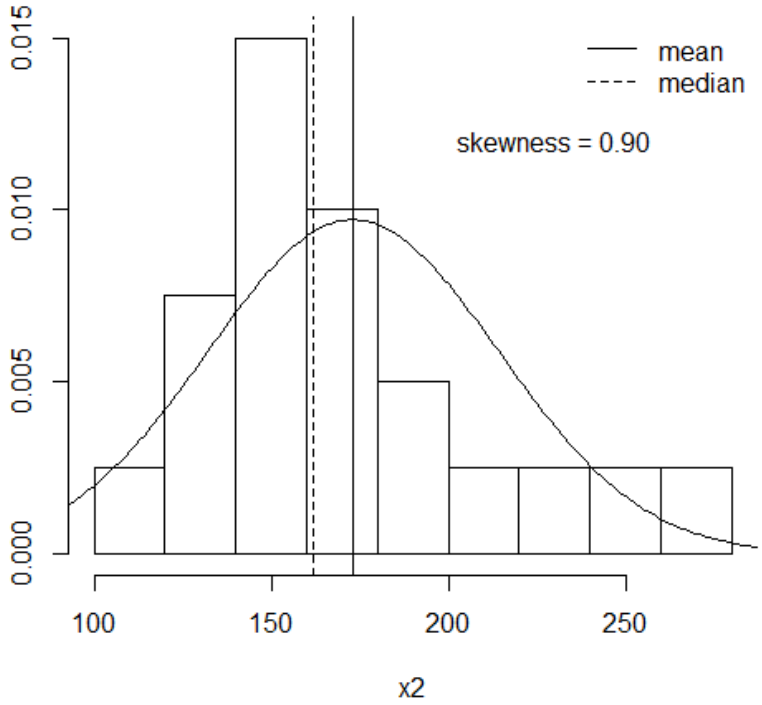
Συντελεστής ασυμμετρίας

Το **μέτρο** ή **συντελεστής ασυμμετρίας** ή **λοξότητας** (coefficient of skewness) μετράει το βαθμό της ασυμμετρίας ή απόκλισης από τη συμμετρία της κατανομής των τιμών μιας μεταβλητής. Όταν η κατανομή είναι συμμετρική, η μέση τιμή, η διάμεσος και η επικρατούσα τιμή συμπίπτουν. Αν και υπάρχουν διάφοροι συντελεστές ασυμμετρίας, από πιο διαδεδομένους είναι συντελεστής με βάση τις ροπές (moment coefficient of skewness), ο οποίος ορίζεται από τον τύπο:

$$\beta_1 = \frac{m_3}{m_2^{3/2}} \tag{8}$$

όπου $m_k = \frac{\sum (X_i - \bar{X})^k}{n}$

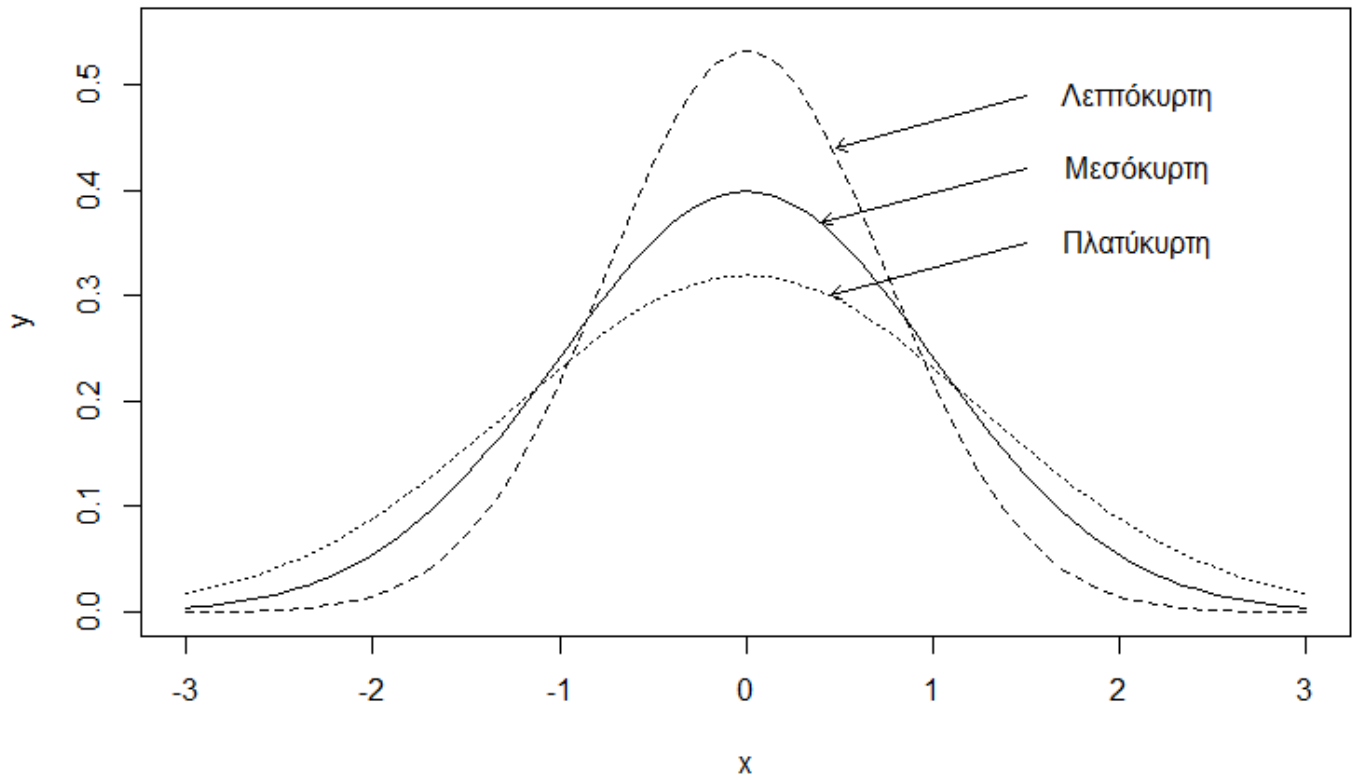

mean
median
skewness = - 0.84
Density
0.000
0.004
0.008
0.012
100
150
200
250
x1


mean
median
skewness = 0.90
0.000
0.005
0.010
0.015
100
150
200
250
x2

Συντελεστής κύρτωσης

Το **μέτρο** ή **συντελεστής κύρτωσης** (coefficient of kurtosis) μετράει το βαθμό συγκέντρωσης των τιμών γύρω από τη μέση τιμή. Μια κατανομή η οποία έχει μεγάλη συχνότητα ή συγκέντρωση τιμών γύρω από τη μέση τιμή ονομάζεται **λεπτόκυρτη** (leptokurtic), αν η συγκέντρωση τιμών είναι μικρή ονομάζεται **πλατύκυρτη** (platykurtic), ενώ οι κατανομές που προσεγγίζονται από την κανονική κατανομή λέγονται **μεσόκυρτες** (mesokurtic). Για τον υπολογισμό του συντελεστή χρησιμοποιείται ο συντελεστής του κύρτωσης με βάση τις ροπές, ο οποίος ορίζεται από τον τύπο:

$$\beta_2 = \frac{m_4}{m_2^2} \tag{9}$$


Λεπτόκυρτη
Μεσόκυρτη
Πλατύκυρτη
y
x
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3

```
# Συντελεστής ασυμμετρίας
> b1=mean((x-mean(x))^3)/mean((x-mean(x))^2)^1.5
> b1
[1] 0.1029188

# Συντελεστής κύρτωσης
> b2=mean((x-mean(x))^4)/mean((x-mean(x))^2)^2
> b2
[1] 2.78386
```

Κατανομές στην R

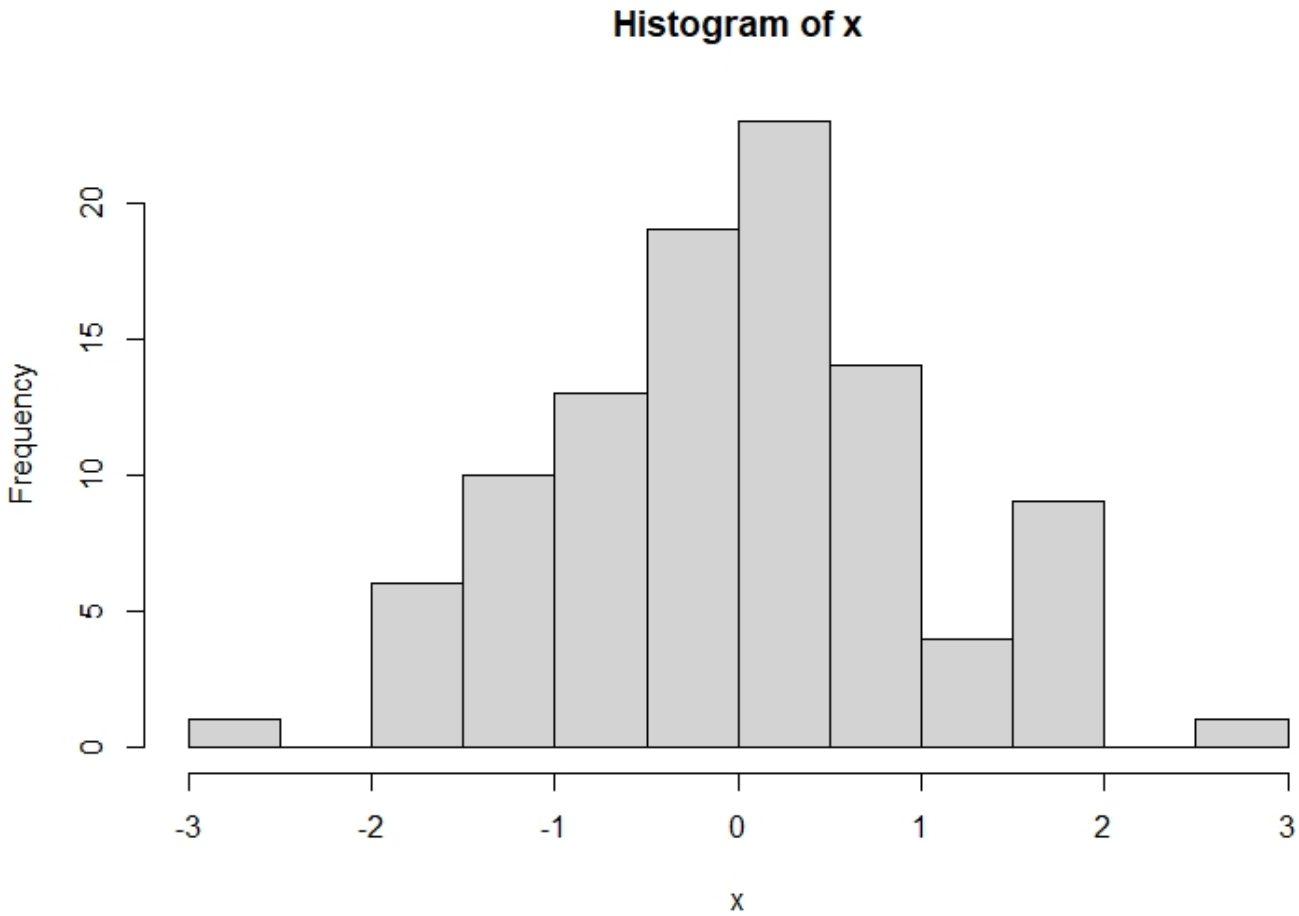
Σύμβολο	**Κατανομή**
beta	Κατανομή Βήτα
binom	Διωνυμική Κατανομή
chisq	Χ^2 Κατανομή
gamma	Γάμμα Κατανομή
exp	Εκθετική Κατανομή
F	F Κατανομή
lnorm	Lognormal Κατανομή
norm	Κανονική Κατανομή
pois	Κατανομή Poisson
t	Κατανομή Student's t
weibull	Κατανομή Weibull

r (random) Γεννήτρια τυχαίων αριθμών
d (density) Συνάρτηση πυκνότητας πιθανότητας (pdf)
p (probability) Αθροιστική συνάρτηση κατανομής (cdf)
q (quantile) Υπολογισμός ποσοστιαίου σημείου (inverse cdf)

Η συνάρτηση rnorm δημιουργεί ένα διάνυσμα τυχαίων αριθμών που ακολουθούν την κανονική κατανομή.

```
> x=rnorm(n=100, mean = 0, sd = 1)
> x
[1] -1.23887593 0.19329295 -0.24151835 0.46122473
-1.98889515 -0.67846086 -0.31029166 0.27398490
...
[97] 1.61913006 -0.06774041 -0.73915059 -1.46003503
```

```
> hist(x)
```

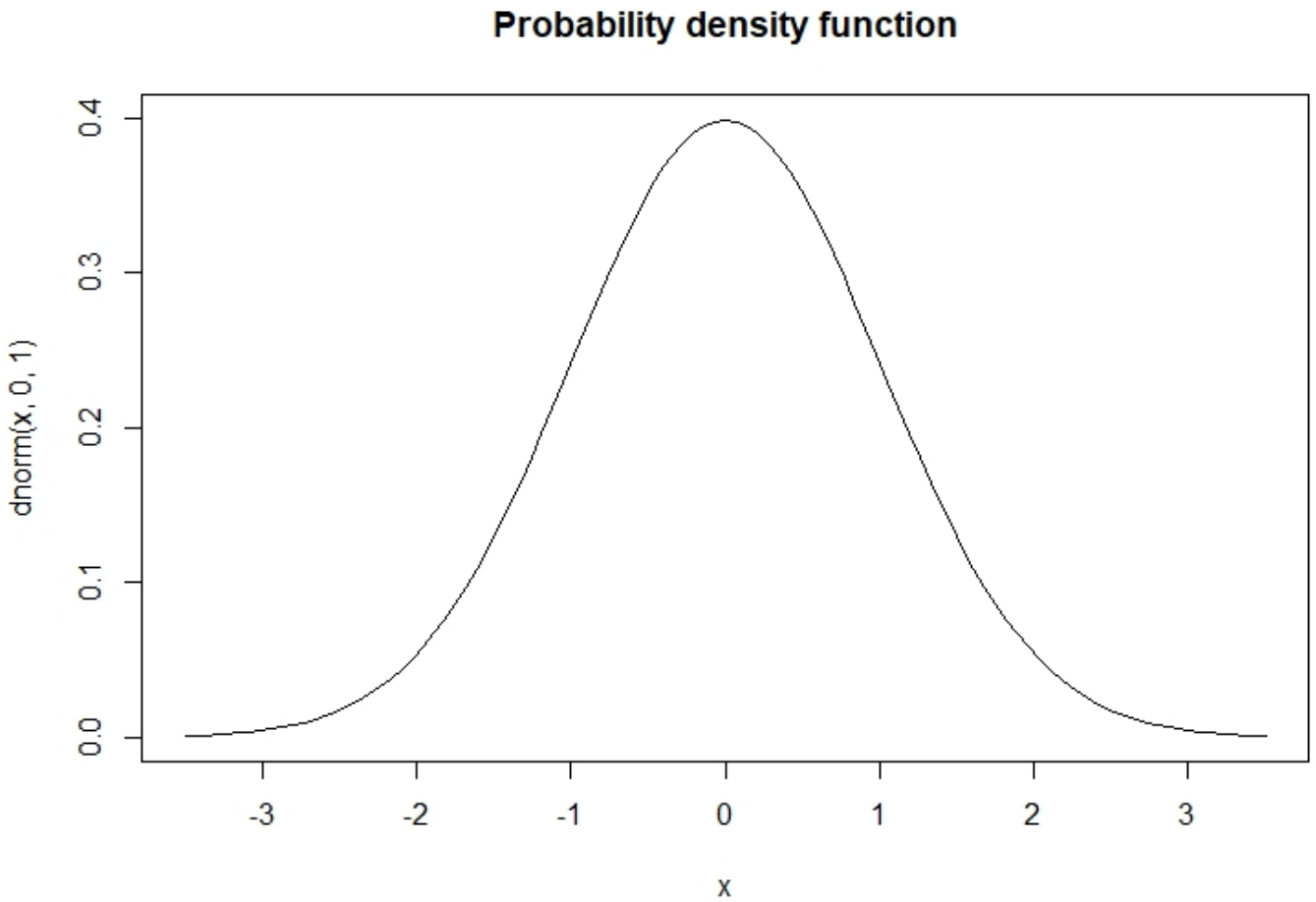


Η συνάρτηση dnorm δίνει την τιμή της συνάρτησης πυκνότητας πιθανότητας για την κανονική κατανομή.

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

```
> dnorm(0, mean = 0, sd = 1)
[1] 0.3989423
```

```
> curve(dnorm(x,0,1),xlim=c(-3.5,3.5), main="Probability density function")
```

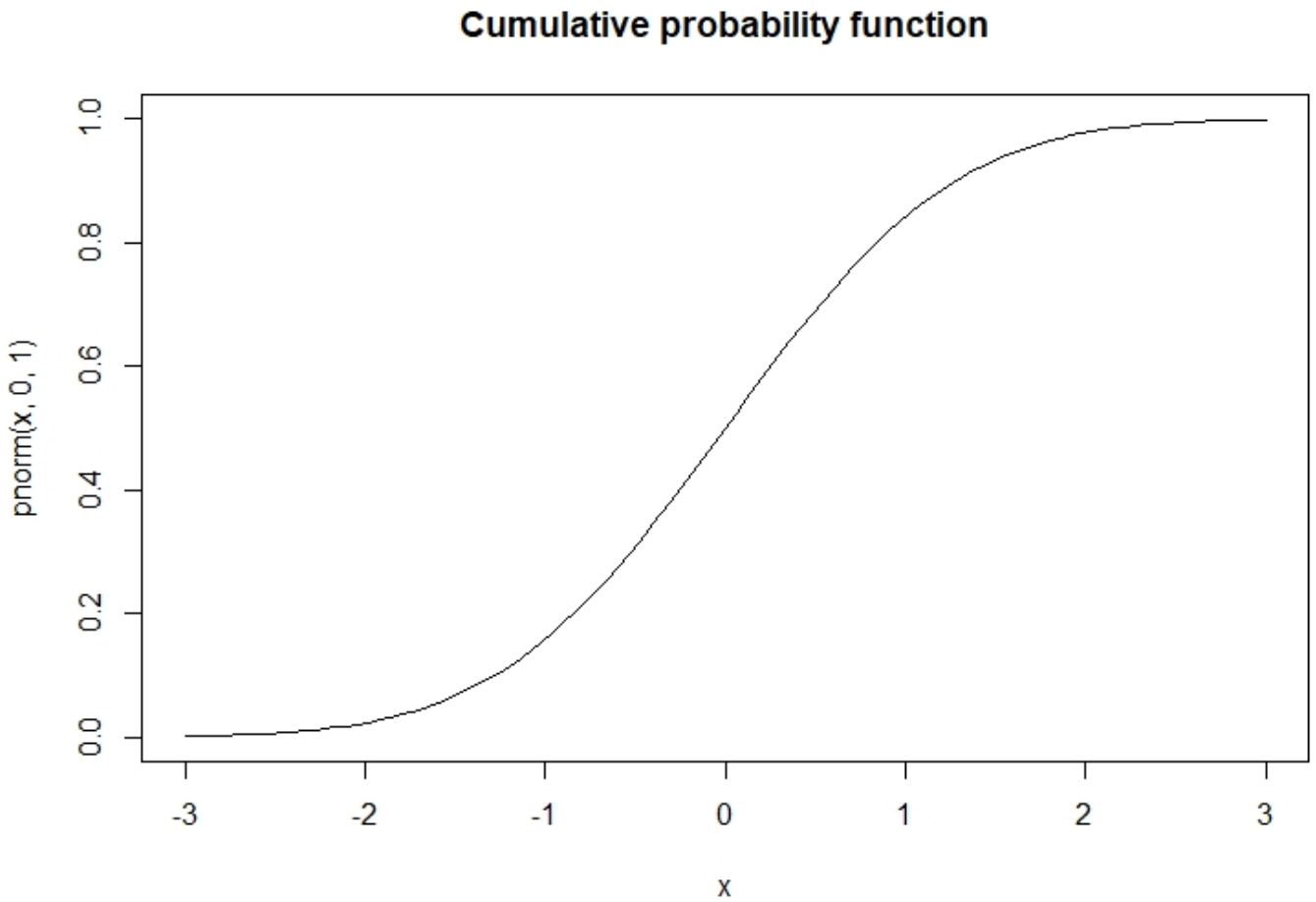


Η συνάρτηση pnorm δίνει την αθροιστική πυκνότητα της κανονικής κατανομής σε μία συγκεκριμένη τιμή.

```
> pnorm(1.21) # P(Z<1.21)
[1] 0.8868606
> pnorm(1.21, lower.tail = FALSE) # P(Z>1.21)
[1] 0.1131394
> pnorm(1.21)-pnorm(-0.81) # P(-0,81<Z<1.21)
[1] 0.6778905
> pnorm(0)-pnorm(-0.81)# P(-0,81<Z<0)
[1] 0.2910299
```

```
> pnorm(6, mean = 4, sd = 2)-pnorm(3, mean = 4, sd = 2)
#P(3 ≤ Y ≤ 6)
[1] 0.5328072
```

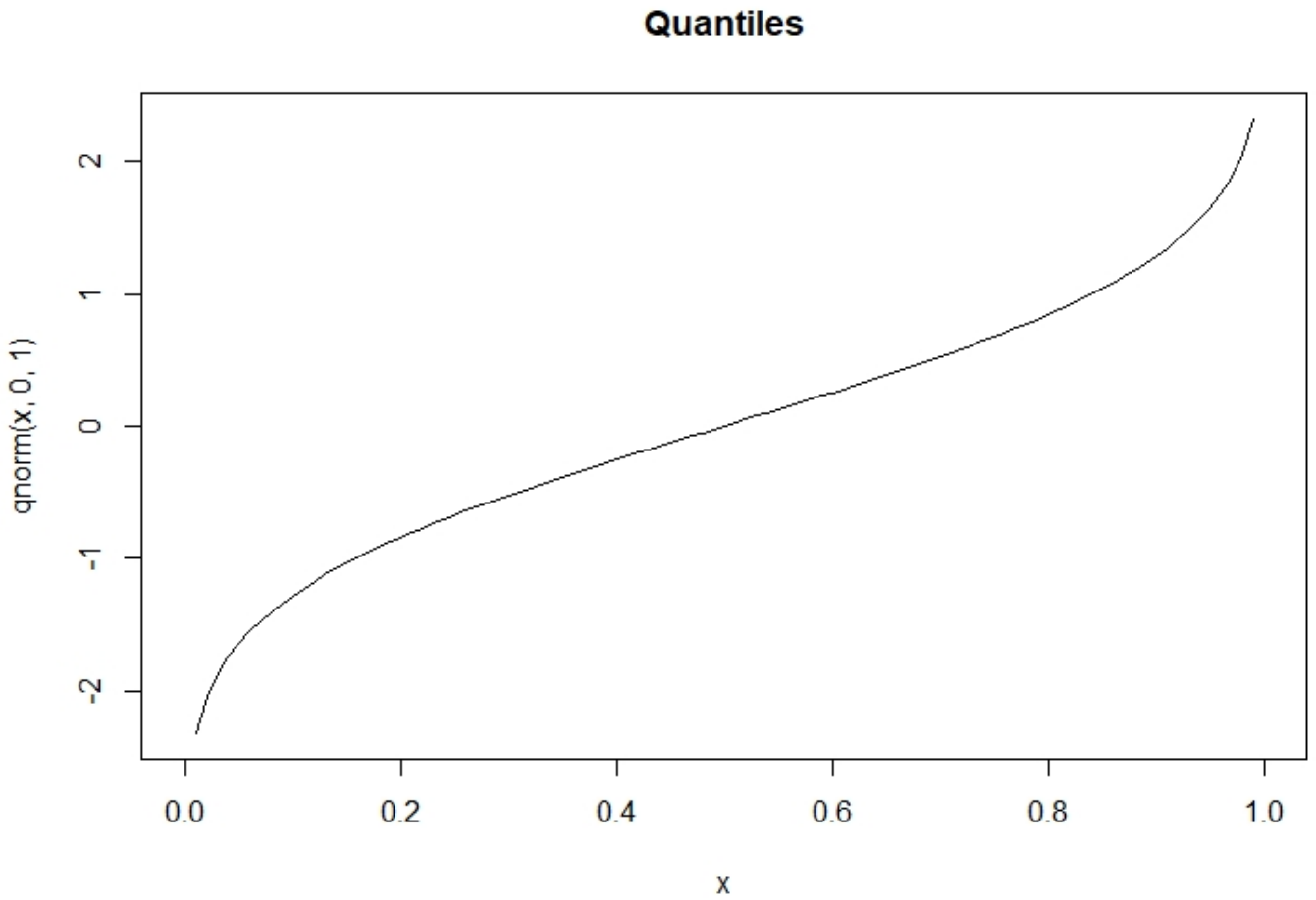
```
> curve(pnorm(x,0,1),xlim=c(-3,3), main="Cumulative probability function")
```



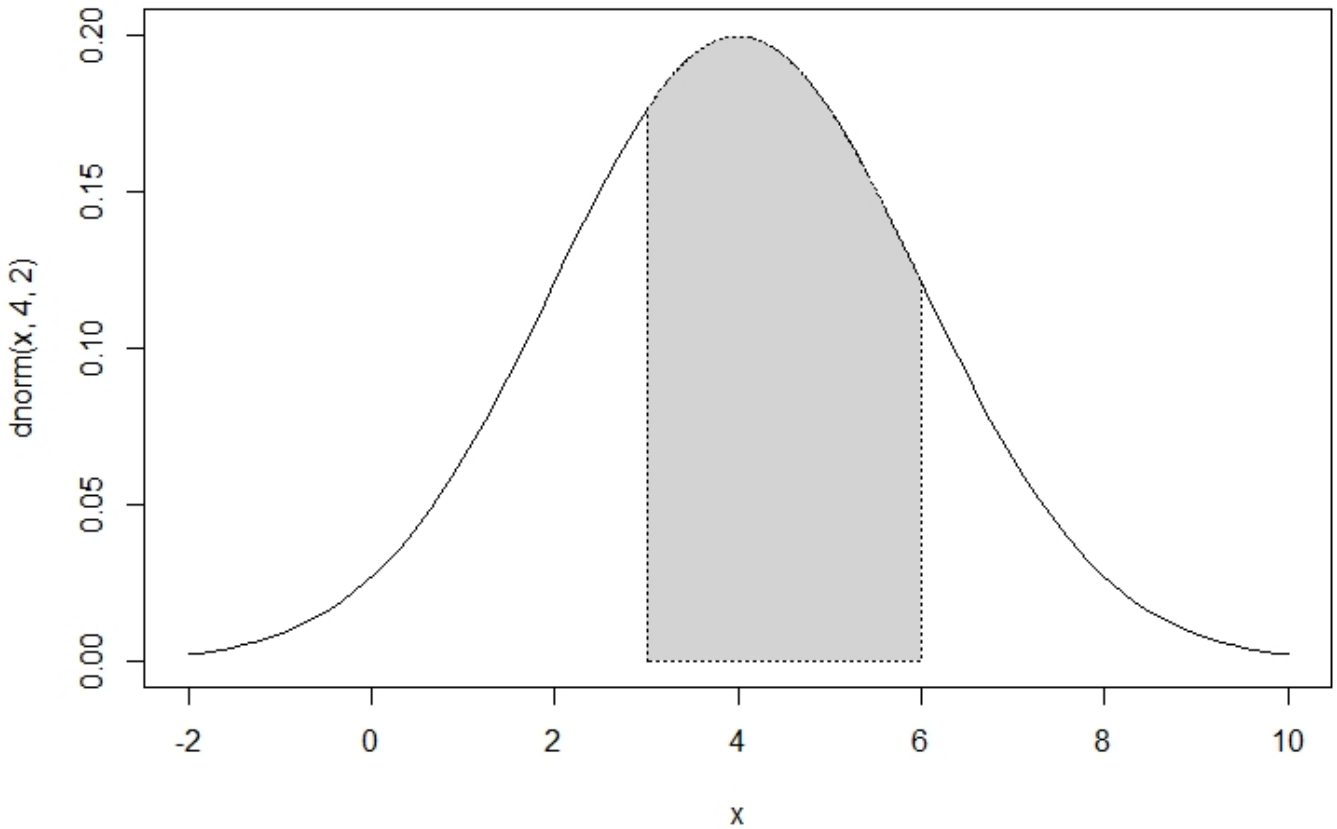
Η συνάρτηση qnorm δίνει την τιμή της κανονικής κατανομής σε μια καθορισμένη αθροιστική πυκνότητα.

```
> qnorm(0.5) # Ποια είναι η τιμή Z του 50ου ποσοστιαίου σημείου της κανονικής κατανομής;
[1] 0
> qnorm(0.95)
[1] 1.644854
> qnorm(0.05)
[1] -1.644854
```

```
> curve(qnorm(x,0,1),xlim=c(0,1), main="Quantiles")
```



```
> curve(dnorm(x,4,2), xlim=c(-2,10),ylim=c(0,0.2))
> x1=seq(3,6,length=200)
> y1=dnorm(x1,mean=4,sd=2)
> polygon(c(3,x1,6),c(0,y1,0),col="gray",border=NULL,lty=3)
```



```
> par(mfrow=c(2,2))
> hist(x)
> curve(dnorm(x,0,1),xlim=c(-3.5,3.5), main="pdf")
> curve(pnorm(x,0,1),xlim=c(-3,3), main="cdf")
> curve(qnorm(x,0,1),xlim=c(0,1), main="Quantiles")
```