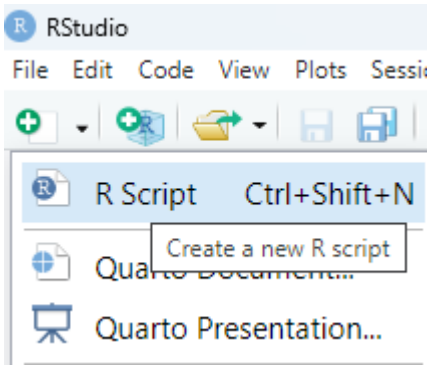
ΧΩΡΙΚΗ ΑΝΑΛΥΣΗ ΟΙΚΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

ΔΙΑΛΕΞΗ 2: ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ – ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

Δουλεύοντας με την R

Αρχικά δημιουργούμε ένα νέο Script στο Rstudio (File > R Script)

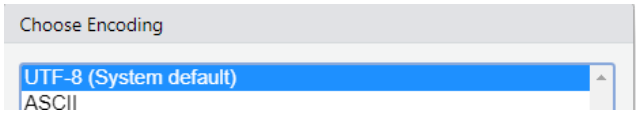


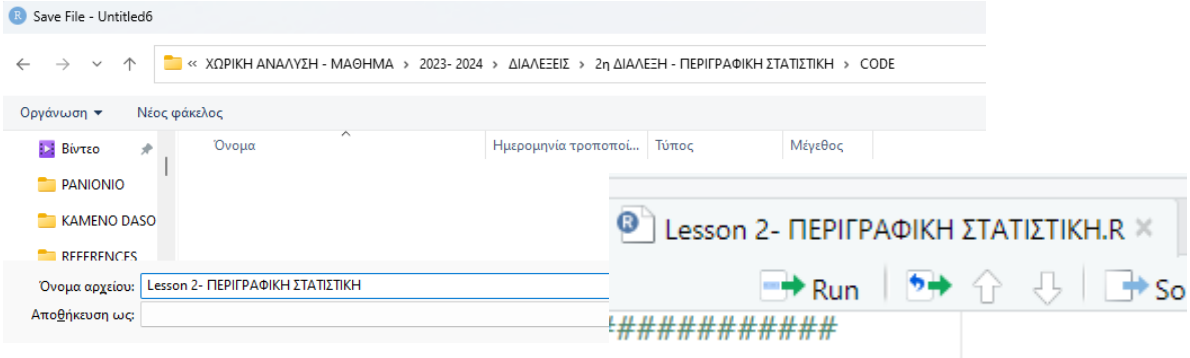
και το ονομάζουμε ως **Lesson 2_Descriptives.**

Για να το σώσουμε (αφού το έχουμε δημουργήσει) πηγαίνουμε ξανά στο μενού File και επιλέγουμε το Save as…

Αφού επιλέξουμε το default Encoding, στο φάκελο μας σώζουμε το script με το παραπάνω όνομα.

Επειδή το script είναι ουσιαστικά ο κώδικας που θα γράψουμε ως ένα αρχείο κειμένου, μπορούμε μετά αυτό να το ανοίγουμε και από ένα notebook για να δούμε το περιεχόμενο του.





Στην αρχή του κώδικα γράφουμε μερικά επεξηγηματικά κείμενα για να μας βοηθάνε στην κατανόηση των ενεργειών μας.

Για να κατανοεί το πρόγραμμα ότι αυτά δεν είναι εντολές πρέπει στην αρχή να εισάγουμε το σύμβολο #

Μπορούμε να βάλουμε περισσότερα ## για να κατηγοριοποιούμε τις ενότητες

```
1  ###############    LESSON 2 - ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ  ######################
2
3  ### ΔΗΜΙΟΥΡΓΙΑ ΑΝΤΙΚΕΙΜΕΝΩΝ
4
5  ## Απλές αριθμητικές σειρές
```

☑

☑ ΕΝΕΡΓΕΙΑ 1: Δημιουργία αντικειμένου/ -ων αριθμοδεδομένων

⇨ **Δημιουργία τυχαίων αριθμών** με την εντολή sample

Γράφοντας τον κώδικα, το πρόγραμμα μας λέει τη πληροφορίες θα πρέπει να δώσουμε

```
# καθορισμός απλού αντικειμένου συνεχόμενων τιμών

x1 = sam
```

```
sample          {base}
sample.int      {base}
```

```
sample(x, size, replace = FALSE, prob = NULL)
Random Samples and Permutations

sample takes a sample of the specified size from the elements of x using
either with or without replacement.
```

Για το sample, π.χ. για το x = 1:20 το εύρος του αριθμητικού δείγματος θα είναι από 1 ως 20, για x= 1:53 από 1 ως 53

Το size είναι ο αριθμός των τυχαίων αριθμών, π.χ. 30 = 30 διαφορετικοί αριθμοί μέσα στο καθορισμένο εύρος (που έχει καθοριστεί από το x)

Το replace παίρνει τιμές λογικές TRUE ή FALSE. Με το TRUE ή ίδια τιμή μπορεί να υπολογιστεί ξανά, σε αντίθεση με το FALSE.

```
# καθορισμός απλού αντικειμένου συνεχόμενων τιμών

x1 = sample(x=1:40, size = 50, replace=TRUE )

x1
```

Η παραπάνω εντολή δημιουργεί το ακόλουθο διάνυσμα τιμών, 50 τιμές σε εύρος 1 ως 40 με δυνατότητα επανάληψης της ίδιας τιμής (π.χ. το 10 εμφανίζεται πέντε φορές)

```
> x1
 [1] 34 18 34 16 31 15 28 10  3 16 25 38  8 20  4 19 32 14 16 26 31 32 25 26 36 40  8 40  4 27 10 10
[33] 30 17  1 28 39 29  2  6 36 20 21 19 28  5  2 36 10 10
>
```

Ο ίδιος κώδικας θα μπορούσε να γραφτεί και πιο απλά

```
x2 = sample (1:40, size=50, replace=TRUE)
x2
```

```
> x2
 [1]  2  2 32 14  8 18 40 17 28 12 11  1 16 10 35 13  9 27 16  7  5 24  9 36 34 35 30 12 38 16 15 37
[33] 33 21 18 36 19 39 40  5 20 26 14 15  1 21  8  3 16 33
```

Τα νούμερα τώρα είναι διαφορετικά από τα προηγούμενα, γιατί η αρχή της τυχαιοποίησης είναι κάθε φορά διαφορετική.

Αν θέλουμε η αρχική τυχαιοποίηση να παραμένει η ίδια σε επαναλαμβανόμενα επόμενα στάδια θα πρέπει αρχικά να καθορίσουμε το αρχικό τυχαίο σημείο με την εντολή set.seed

Π.χ. οι δύο παρακάτω εντολές έχουν δώσει τα ίδια τυχαία 60 νούμερα,

```
> set.seed (15)
> x3 = sample (1:100, 60, replace=TRUE)
> x3
 [1] 37 34 38 49  5 89 76 84 65 12 37 87 66 25 10 90 85 53 99  2 34 58 95 19 72 30 21 59 37 79 27 76
[33] 86  2 95 14 58 10 73 45 81 61 33 39 60 49 92  6 98 23 35 49 75 49 56 65 39 88 34 81
> set.seed(15)
> x4 = sample(x=1:100, size = 60, replace=TRUE )
> x4
 [1] 37 34 38 49  5 89 76 84 65 12 37 87 66 25 10 90 85 53 99  2 34 58 95 19 72 30 21 59 37 79 27 76
[33] 86  2 95 14 58 10 73 45 81 61 33 39 60 49 92  6 98 23 35 49 75 49 56 65 39 88 34 81
```

Ας κρατήσουμε το αντικείμενο x1 και να δούμε την οπτικοποίηση του μέσα από τη δημιουργία ιστογράμματος.

Θα το δούμε με δύο διαφορετικούς τρόπους

1. Με βάση την core της R
2. Με βάση τα ggplot

⇨ **Δημιουργία ιστογράμματος (barplot)** από την core R, με την εντολή hist

Ενεργοποιώντας τη βοήθεια της εντολής (πατώντας το F1) ή με το ?hist, θα μας παρουσιάσει τι ενσωματώνουμε στην εντολή

```
hist(x, ...)

## Default S3 method:
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE, fuzz = 1e-7,
     density = NULL, angle = 45, col = "lightgray", border = NULL,
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, warn.unused = TRUE, ...)
```

Αρχικά γράφουμε το αντικείμενο που θέλουμε να δημιουργήσουμε το ιστόγραμμα (πρέπει να είναι αριθμητικό). Στη συνέχεια δηλώνουμε σε πόσες κλάσεις θέλουμε να χωριστεί το δείγμα μας. Αν απλά βάλουμε ένα νούμερο, θα υπολογιστεί με βάση τη φόρμουλα Sturges (δες πλαίσιο). *Αλλιώς υπάρχουν και άλλοι δύο τρόποι (με βάση το standard error ή με βάση το inter-quartile range).*

The default for breaks is "Sturges"

`nclass.Sturges` *uses Sturges' formula, implicitly basing bin sizes on the range of the data.*

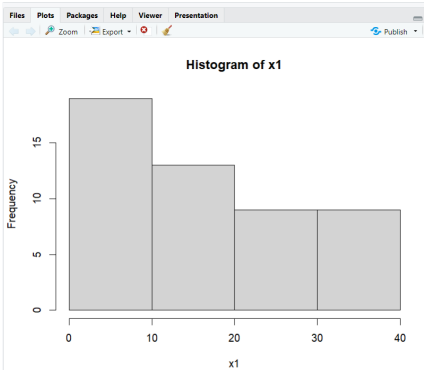
`nclass.scott` *uses Scott's choice for a normal distribution based on the estimate of the standard error, unless that is zero where it returns* `1`*.*

Ας αρχίσουμε από τα απλά…

Θέλουμε στο αντικείμενο x1 να κάνουμε δύο ιστογράμματα με 5 και 10 κλάσεις αντίστοιχα. Ο κώδικας μας θα ήταν για τις 5 κλάσεις

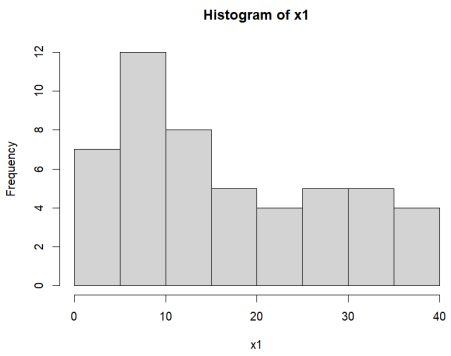
```
# διαμόρφωση ιστογράμματος με την coreR
hist (x1, breaks = 5)
```

και στην καρτέλα Plots εμφανίζεται το διάγραμμα



Ο κώδικας μας για τις 10 κλάσεις

```
hist (x1, breaks = 10)
```



Αν θέλαμε να θέλουμε να δούμε και τις λεπτομέρειες του διαγράμματος θα το σώσουμε ως αντικείμενο και θα «τρέξουμε» το αντικείμενο.

```
histog1 <- hist(x1, breaks = 5)
histog1
```

```
> histog1
$breaks
[1]  0 10 20 30 40

$counts
[1] 19 13  9  9

$density
[1] 0.038 0.026 0.018 0.018

$mids
[1]  5 15 25 35

$xname
[1] "x1"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"

>
```

```
histog2 <- hist(x1, breaks = 10)
histog2
> histog2
$breaks
[1]  0  5 10 15 20 25 30 35 40

$counts
[1]  7 12  8  5  4  5  5  4

$density
[1] 0.028 0.048 0.032 0.020 0.016 0.020 0.020 0.016

$mids
[1]  2.5  7.5 12.5 17.5 22.5 27.5 32.5 37.5

$xname
[1] "x1"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

Στην περιγραφή της πληροφορίας των αντικειμένων βλέπουμε και την ανάλυση

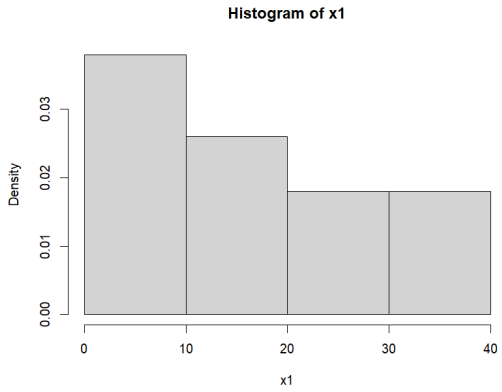
```
$counts
[1]  7 12  8  5  4  5  5  4

$density
[1] 0.028 0.048 0.032 0.020 0.016 0.020 0.020 0.016
```

Το πρώτο (counts) αντιστοιχεί στη συχνότητα n_k των μετρήσεων σε κάθε διάστημα k (και είναι το default, αν δεν προσδιορίσουμε κάτι άλλο). Αν θέλουμε στο διάγραμμα να φαίνεται το density (σχετικές συχνότητες ως προς το σύνολο του δείγματος), θα πρέπει να γράψουμε στον κώδικα επιπλέον το freq = FALSE

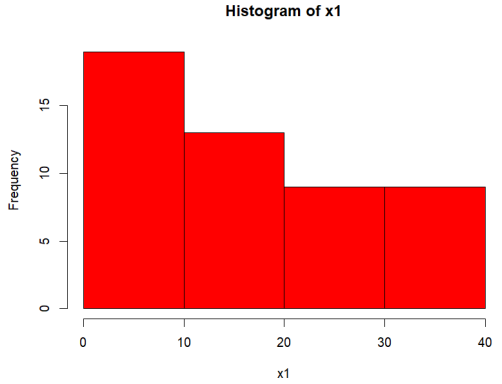
```
hist (x1, breaks = 5)
hist (x1, breaks=5, freq = FALSE)
```

`freq` logical; if `TRUE`, the histogram graphic is a representation of frequencies, the `counts` component of the result; if `FALSE`, probability densities, component `density`, are plotted (so that the histogram has a total area of one).



Ας ομορφύνουμε λίγο περισσότερο το ιστόγραμμα. Να δώσουμε κόκκινο χρώμα στις κλάσεις, να αλλάξουμε το Histogram of x1 σε κάτι δικό μας, καθώς και τις ονομασίες στους άξονες x και y.

```
hist (x1, breaks = 5, col="red")
```



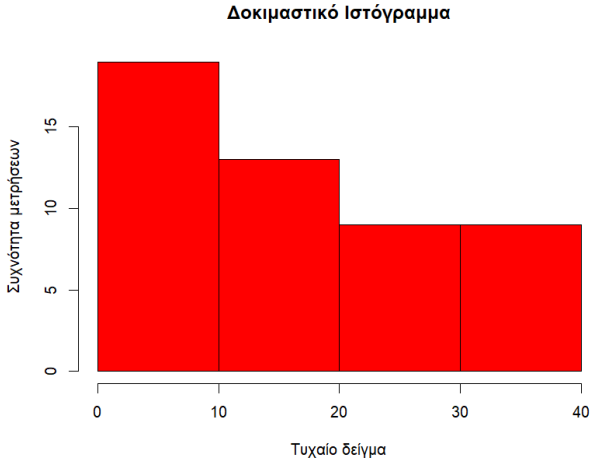
Η R έχει πολλούς τρόπους να χρωματίσει κάτι. Με όνομα, νούμερο, κωδικό κ.α.

Π.χ. υπάρχουν 657 ονόματα για χρώματα. Γράφοντας την εντολή colors() θα εμφανιστούν αυτά τα ονόματα στην R.

Για να βάλουμε τίτλους γράφουμε επιπλέον τον κώδικα. Οι ονομασίες θέλουν “ ” στο κείμενο

- xlab = "Τυχαίο δείγμα" για τον άξονα Χ
- ylab = "Συχνότητα μετρήσεων" για τον άξονα Ψ και
- main = "Δοκιμαστικό Ιστόγραμμα " για τον τίτλο

```
hist (x1, breaks = 5, col="red", xlab = "Τυχαίο δείγμα",
      ylab = "Συχνότητα μετρήσεων", main = "Δοκιμαστικό Ιστόγραμμα")
```

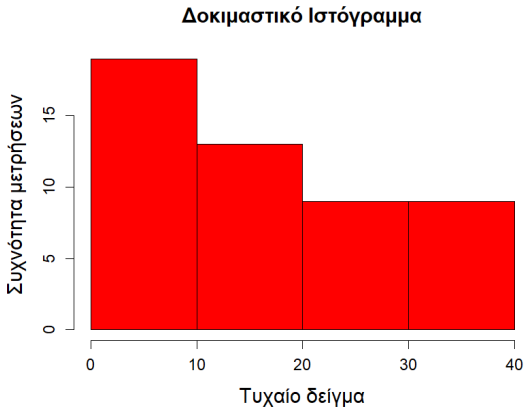


Τέλος θέλουμε να αλλάξουμε τα μεγέθη των τίτλων και των αριθμών στο ιστόγραμμα. Χρησιμοποιούμε το όρισμα cex. Το default = 1.

There are actually five **cex** arguments you can use to change the size of specific plot elements:

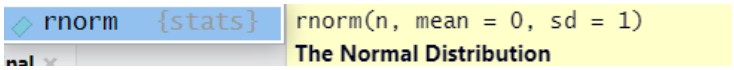
- **cex**: Changes the size of symbols
- **cex.axis**: Changes the size of axis tick mark annotations
- **cex.lab**: Changes the size of x-axis and y-axis labels
- **cex.main**: Changes the size of the plot title
- **cex.sub**: Changes the size of the plot subtitle

```
hist (x1, breaks = 5, col="red", xlab = "Τυχαίο δείγμα",
      ylab = "Συχνότητα μετρήσεων", main = "Δοκιμαστικό Ιστόγραμμα", cex.lab=1.5,
      cex.axis=1.2, cex.main=1.5)
```



⇨ Δημιουργία τυχαίων αριθμών που ακολουθούν την κανονική κατανομή.

Εδώ μας ενδιαφέρουν και πρέπει να ορίσουμε α) τον αριθμό των τυχαίων αριθμών, τη μέση τιμή και την τυπική απόκλιση. Χρησιμοποιούμε την εντολή rnorm



Ας φτιάξουμε μια κατανομή 1000 αριθμών με μέση τιμή = 150 και τυπική απόκλιση 30

Ο κώδικας θα είναι ο παρακάτω:

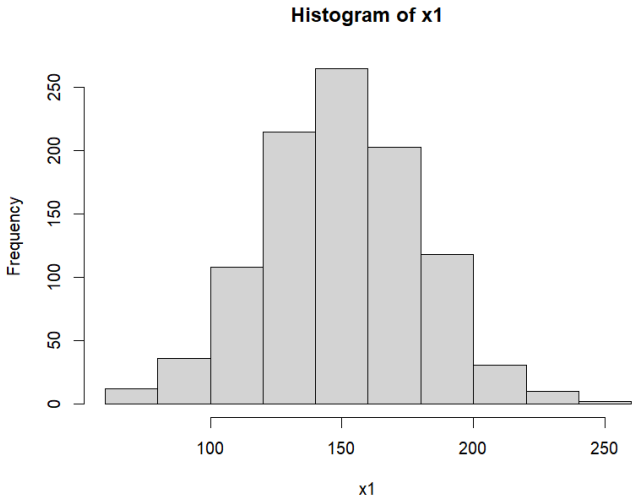
```
## διαμόρφωση διανύσματος κανονικής κατανομής
# καθορισμός δείγματος
n1 = 1000
#δημιουργία διαγράμματος
x1 <- rnorm(n1, mean =150, sd=30)
```

```
Τρέχουμε την εντολή
summary (x1)
```

Τι παρατηρούμε;

Στη συνέχεια παράγουμε ιστόγραμμα με τις 1000 τυχαίες τιμές και αν χρειαστεί το τροποποιούμε όπως παραπάνω.

```
hist(x1)
```

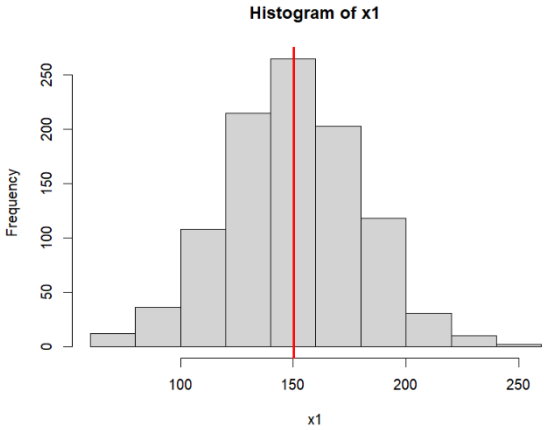


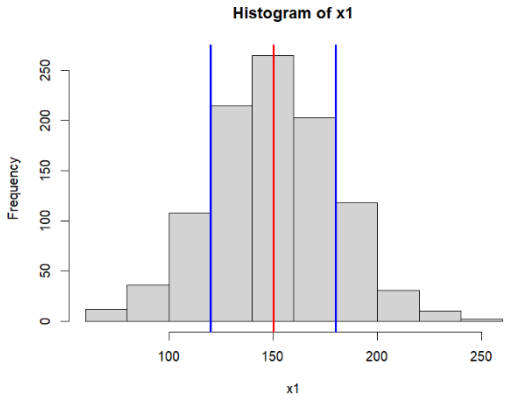
Με το όρισμα abline, μπορούμε να προσθέσουμε τις τιμές για μέση τιμή κλπ.

```
hist(x1)
abline(v = mean(x1),col = "red", lwd = 3) # Add line for mean
abline (v=median(x1), col = "blue", lwd = 3) # Add line for median
abline (v=quantile(x1), col = "black", lwd = 3) # Add line for quantile
```

Για την τυπική απόκλιση θα πρέπει να υπολογίσουμε την τιμή με βάση τη μέση τιμή ± 1 τυπική απόκλιση και να την ενσωματώσουμε σε ένα νέο abline

```
abline (v=120, col = "blue", lwd = 3) # Add line for MO-1sd
abline (v=180, col = "blue", lwd = 3) # Add line for MO+1sd
```

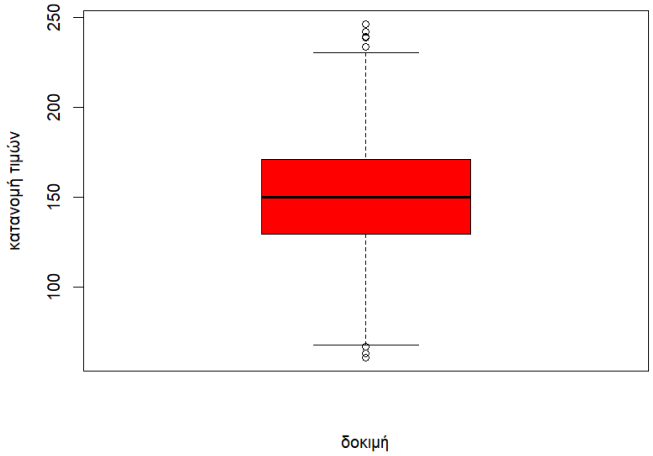




ΘΗΚΟΓΡΑΜΜΑΤΑ

Με βάση τις προηγούμενες κατανομές, πειραματιστείτε με τα θηκογράμματα, χρησιμοποιώντας την εντολή boxplot από την βασική R

```
boxplot(x1,col="red", xlab = "δοκιμή", ylab = "κατανομή τιμών")
```



Στη συνέχεια κάντε αθροιστική καμπύλη με δύο νέες εντολές την ecdf (Empirical Cumulative Distribution Function) και πλοτάρεται το αντικείμενο που θα δημιουργήσετε με την εντολή plot. Το plot είναι μια εξαιρετική εντολή, που προσαρμόζεται αντίστοιχα σε σχέση με το αντικείμενο που καλείται να αναπαραστήσει γραφικά!!!

```
x11 = x1[order(x1)]
p = ecdf(x11)
plot(p, col="red", lwd = 3, main="Αθροιστική Σχετική Συχνότητα")
```