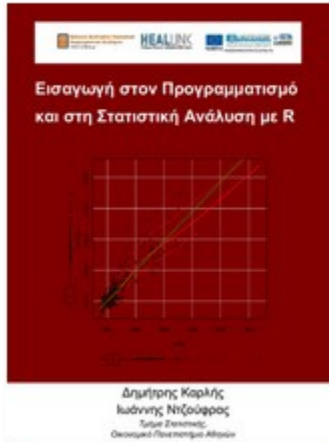


Βιοστατιστική

Εφαρμογές και Παραδείγματα Ελέγχου Υποθέσεων

Πηγές υλικού

- Διαφάνειες και ασκήσεις του Dr Theophanis Tsandilas (National Institute for Research in Digital Science and Technology, INRIA, Γαλλία)
- Διαφάνειες και ασκήσεις του καθηγητή Σπύρου Γαλατσίδα (Τμ. Δασολογίας & Διαχείρισης Περιβάλλοντος & Φυσικών Πόρων, ΔΠΘ)



Εισαγωγή στον προγραμματισμό και στη στατιστική ανάλυση με R

Author(s) :*Ntzoufras, Ioannis; Karlis, Dimitrios*

Type :Undergraduate textbook



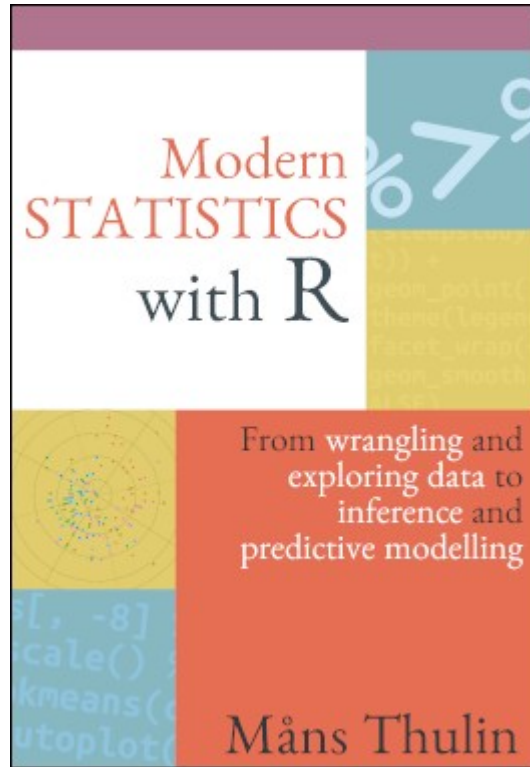
Μεθοδολογία της έρευνας στις επιστήμες υγείας

Author(s) :*Lagoumintzis, Georgios; Vlachopoulos, Georgios; Koutsogiannis, Konstantinos*

Type :Undergraduate textbook



- <https://repository.kallipos.gr/>



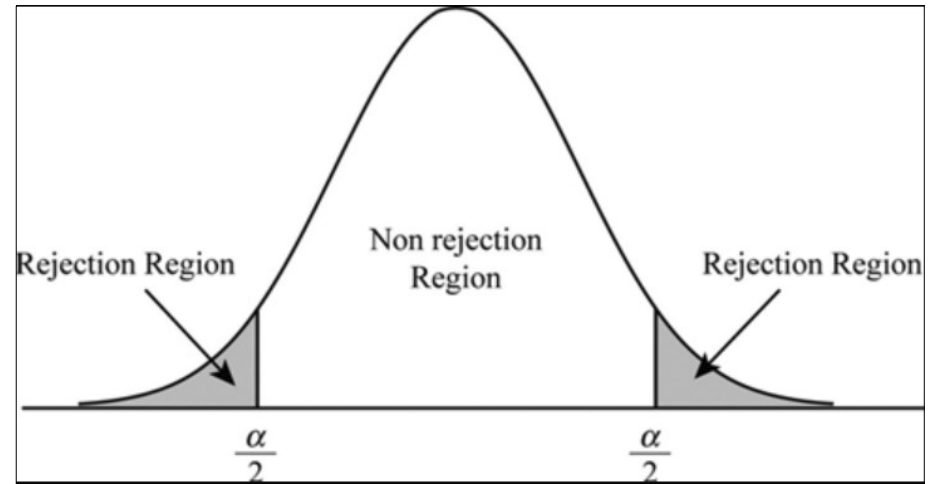
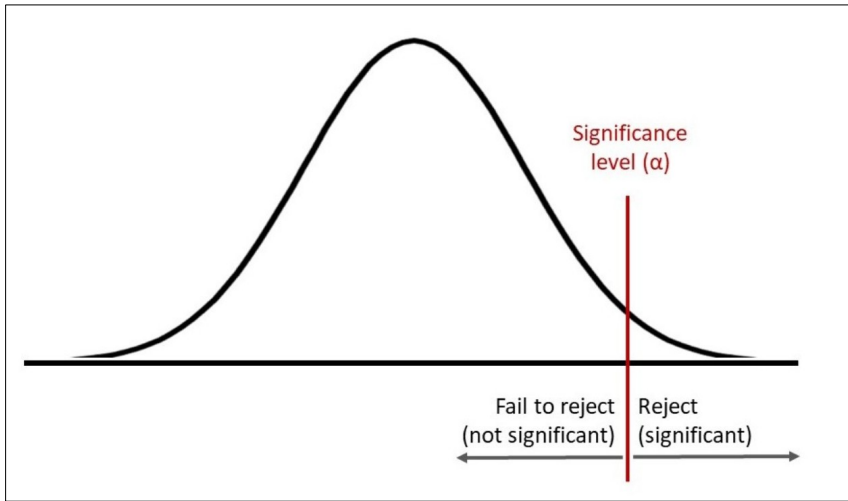
Και άλλα παραδείγματα προέρχονται από το αυτό το e-book
<https://modernstatisticswithr.com/index.html>

This is the online version of the book Modern Statistics with R. It is free to use, and always will be.

Στατιστικό συμπέρασμα

- Η διαδικασία της εξαγωγής των παραμέτρων μιας υποκείμενης κατανομής πιθανοτήτων από ένα δείγμα
- Τέσσερις κατηγορίες:
 - Σημειακές εκτιμήσεις
 - Εκτιμήσεις διαστήματος
 - Έλεγχος υποθέσεων
 - Πρόβλεψη

Επίπεδο σημαντικότητας α



Μεθοδολογία ελέγχου υποθέσεων



STEP 1: State the hypotheses.
(Population)



α

STEP 2: Set the level of Significance
(Criterion)



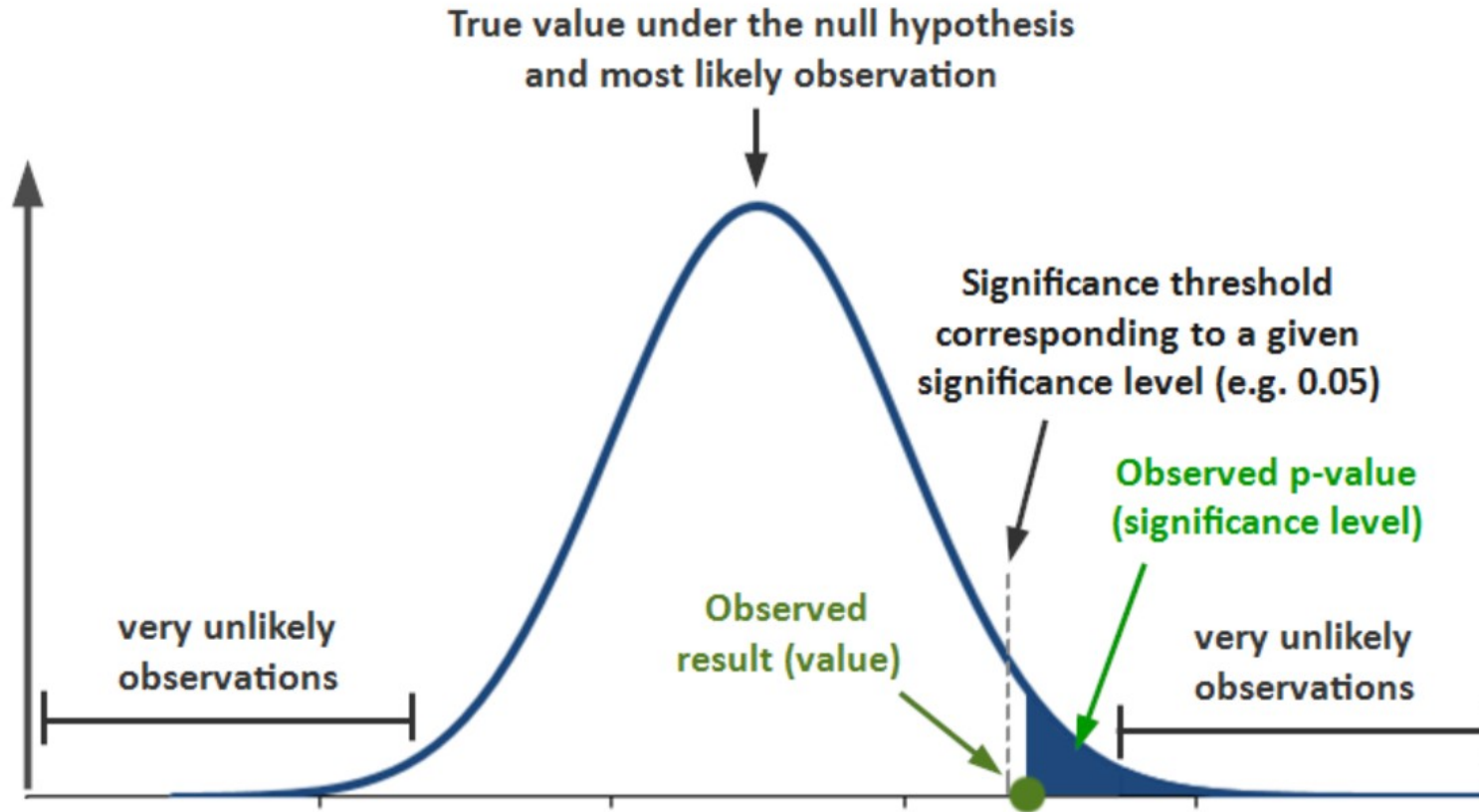
STEP 3: Compute test Statistics
(Sample)



p

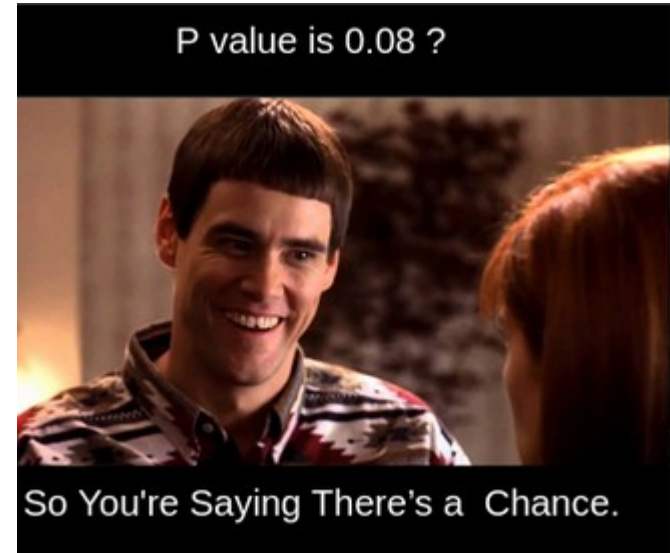
STEP 4: Make a decision based on p value

p-value και α



p-value και α

- Όταν η p-value είναι μικρότερη από το επίπεδο σημαντικότητας α , τότε η μηδενική υπόθεση H_0 απορρίπτεται
- Το α είναι η μέγιστη p-value για την οποία απορρίπτεται η H_0
 - Είναι η μέγιστη πιθανότητα να κάνουμε σφάλμα τύπου I

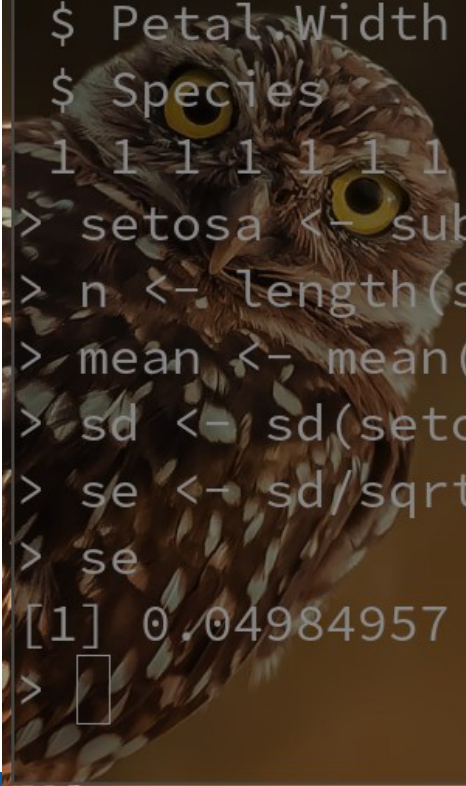


Εφαρμογές και παραδείγματα ελέγχου
υποθέσεων

Εύρεση διαστημάτων εμπιστοσύνης

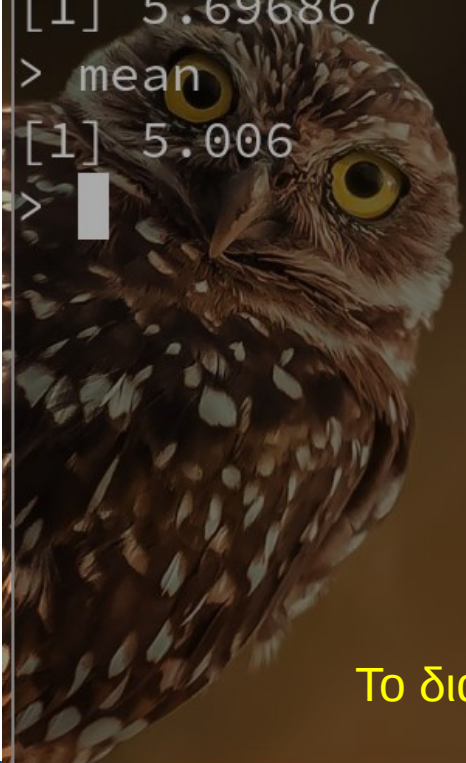
Κανονική κατανομή

- Η γενική προσέγγιση υπολογισμού του CI είναι:
 - $CI = \text{σημειακή εκτίμηση (π.χ. μέσος)} \pm (\text{critical value}) * (\text{τυπικό σφάλμα})$
- Η κρίσιμη τιμή για την κανονική κατανομή βρίσκεται με την εντολή `qnorm()` στην R
 - Ας χρησιμοποιήσουμε το σετ δεδομένων **iris** για να βρούμε τα διαστήματα εμπιστοσύνης για τον μέσο του είδους *setosa* για τη μεταβλητή `Sepal.Length`
 - `qnorm(p, mean = mean, sd = sd, lower.tail = TRUE)`



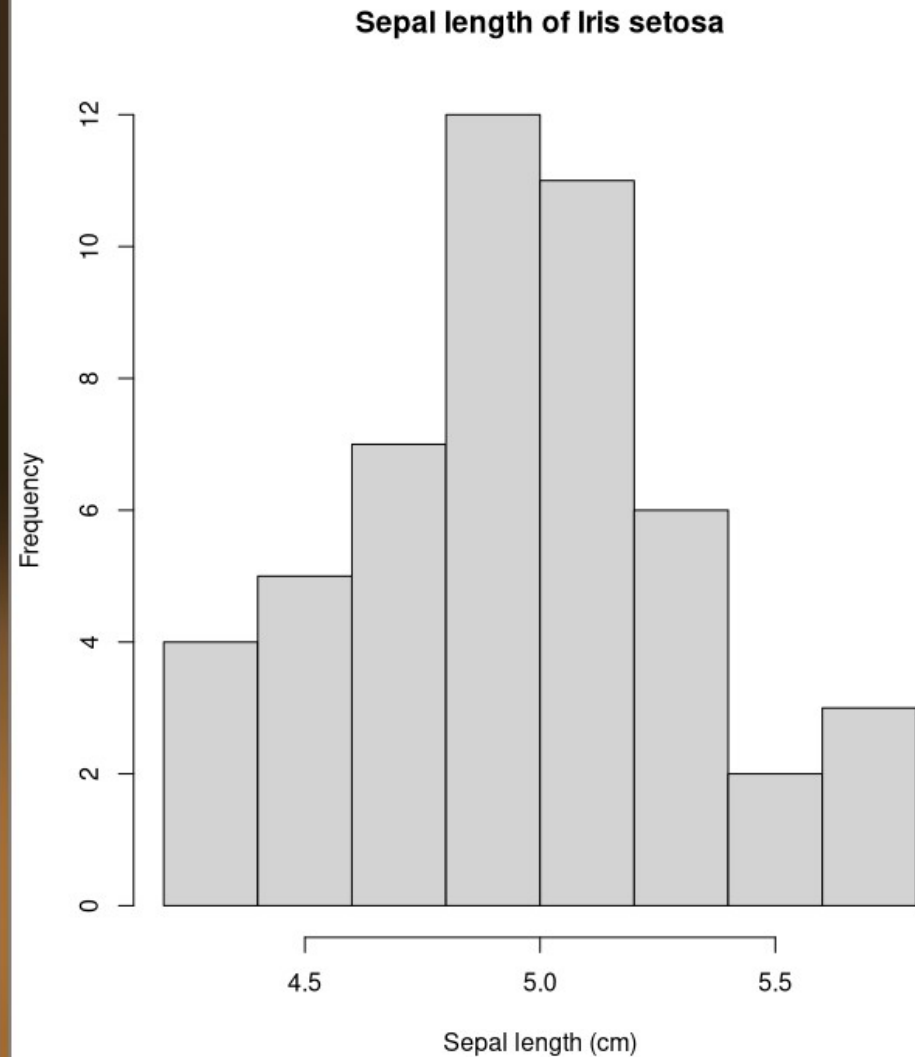
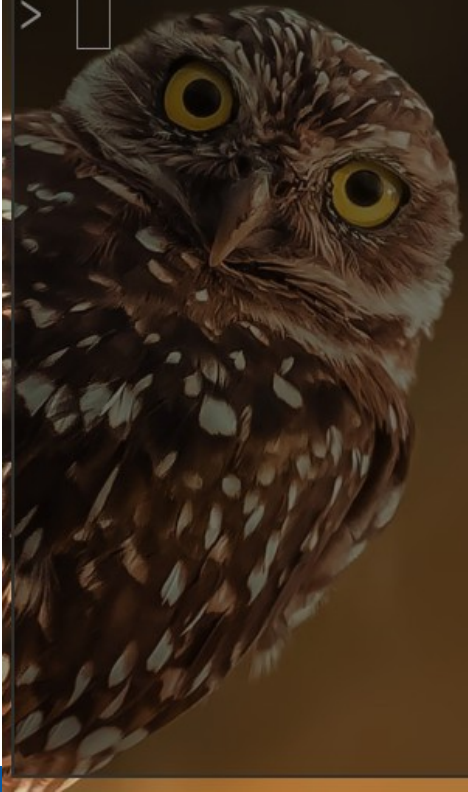
```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width  : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width  : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1
1 1 1 1 1 1 1 ...
> setosa <- subset(iris, Species == "setosa")
> n <- length(setosa$Sepal.Length)
> mean <- mean(setosa$Sepal.Length)
> sd <- sd(setosa$Sepal.Length)
> se <- sd/sqrt(n)
> se
[1] 0.04984957
> 
```

```
> ci1<- qnorm(0.025, mean = mean, sd = sd, lower.tail = TRUE)
> ci1
[1] 4.315133
> ci2<- qnorm(0.975, mean = mean, sd = sd, lower.tail = FALSE)
> ci2
[1] 5.696867
> mean
[1] 5.006
> █
```

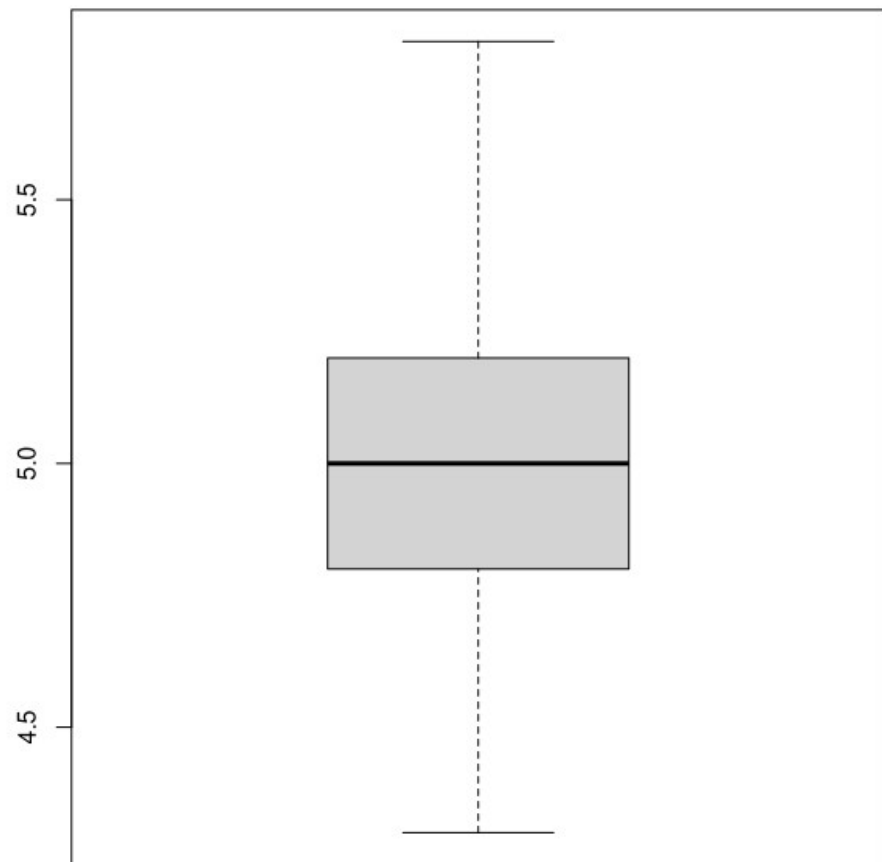


Το διάστημα εμπιστοσύνης είναι 4.31 ως 5.7 cm με μέση τιμή 5 cm

```
main = "Sepal length of Iris setosa",
xlab = "Sepal length (cm)",
ylab = "Frequency"
)
> 
```

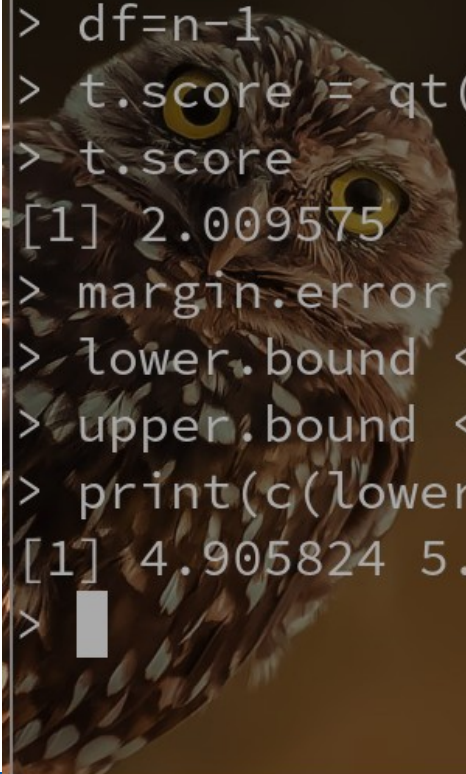


```
main = "Sepal length of Iris setosa",
xlab = "Sepal length (cm)",
ylab = "Frequency"
)
> boxplot(setosa$Sepal.Length)
> 
```



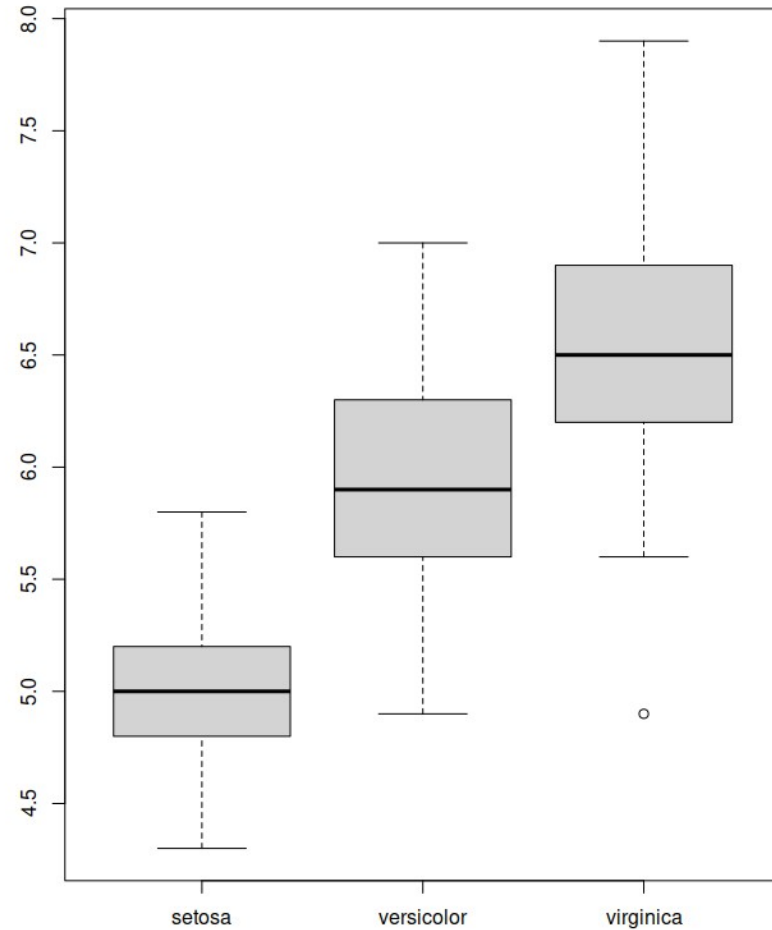
Κατανομή t

- Επειδή δεν γνωρίζουμε την πραγματική τυπική απόκλιση του πληθυσμού και χρησιμοποιούμε το τυπικό σφάλμα από την τυπική απόκλιση του δείγματος, πιο σωστό είναι να χρησιμοποιήσουμε την κατανομή t - Student
 - $margin.error \leftarrow t.score * se$
 - $t.score = qt(p=alpha/2, df=(n-1), lower.tail=F)$

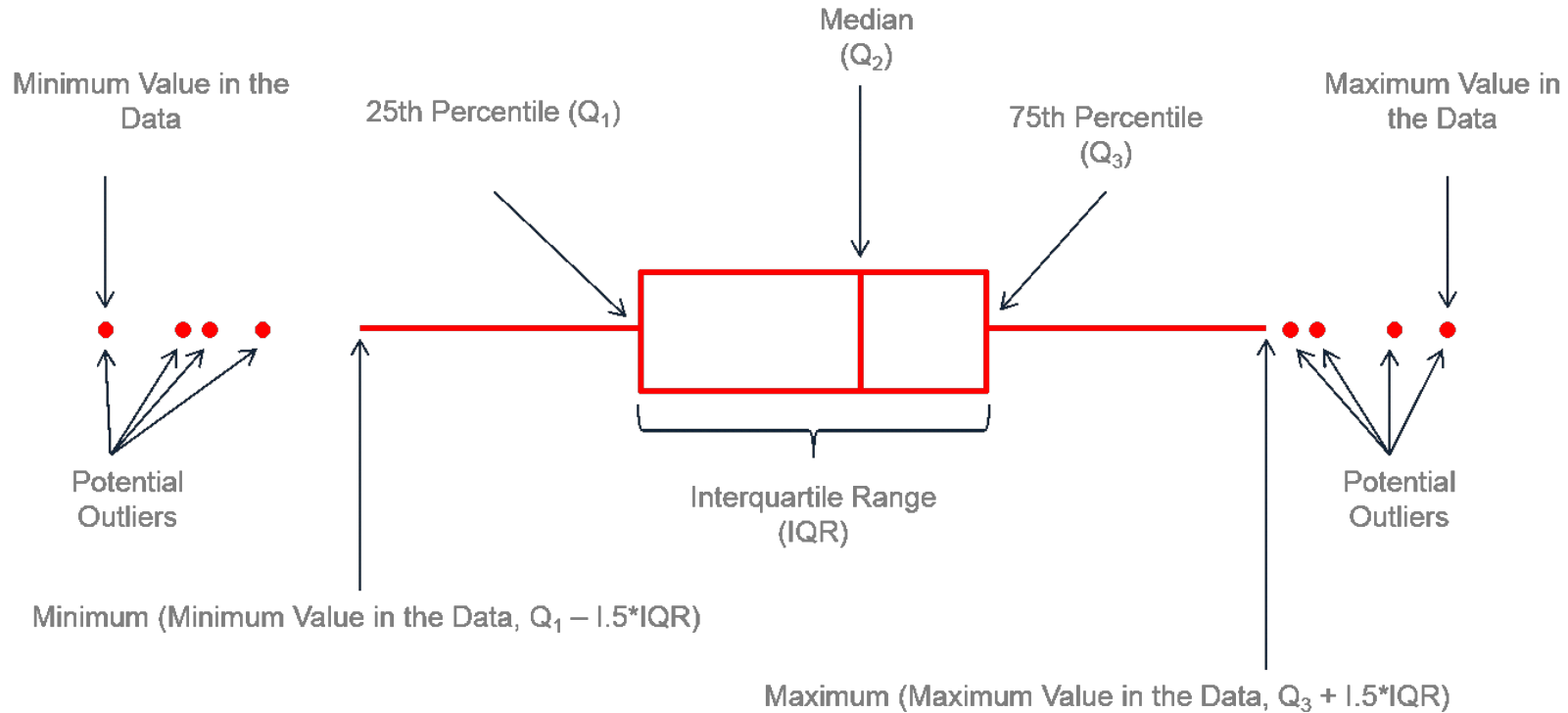


```
> alpha=0.05
> df=n-1
> t.score = qt((alpha/2), df,lower.tail=F)
> t.score
[1] 2.009575
> margin.error <- t.score * se
> lower.bound <- mean - margin.error
> upper.bound <- mean + margin.error
> print(c(lower.bound,upper.bound))
[1] 4.905824 5.106176
> █
```

```
> virginica <- subset(iris, Species == "virginica")
> boxplot(setosa$Sepal.Length, versicolor$Sepal.Length, virginica$Sepal.Length, names=c("setosa", "versicolor", "virginica"))
```



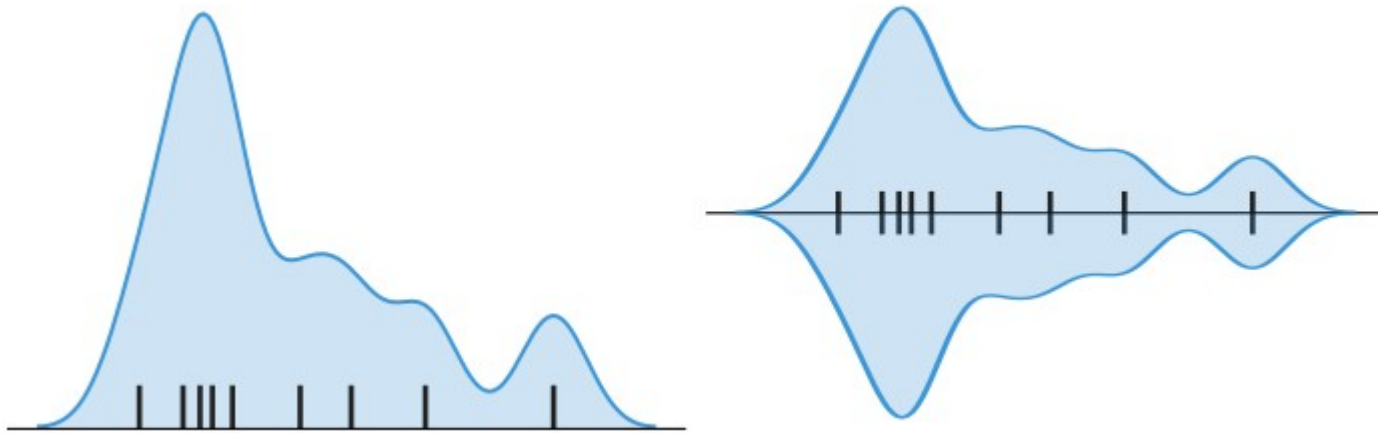
Ανατομία ενός boxplot



Violin plots

- Τα boxplot είναι πολύ χρήσιμα για να συγκρίνουν συνοπτικά διάφορα σετ δεδομένων, αλλά είναι υπερβολικά λιτά και απλοποιούν την παρεχόμενη πληροφορία
 - Για το λόγο αυτό, μπορούμε να κατασκευάσουμε violin plots
- Ένα **violin plot** (γράφημα βιολιού) απεικονίζει κατανομές αριθμητικών δεδομένων για μία ή περισσότερες ομάδες χρησιμοποιώντας καμπύλες πυκνότητας (density plots)
 - Το πλάτος κάθε καμπύλης αντιστοιχεί με την κατά προσέγγιση συχνότητα των σημείων δεδομένων σε κάθε περιοχή

Density plot και violin plot

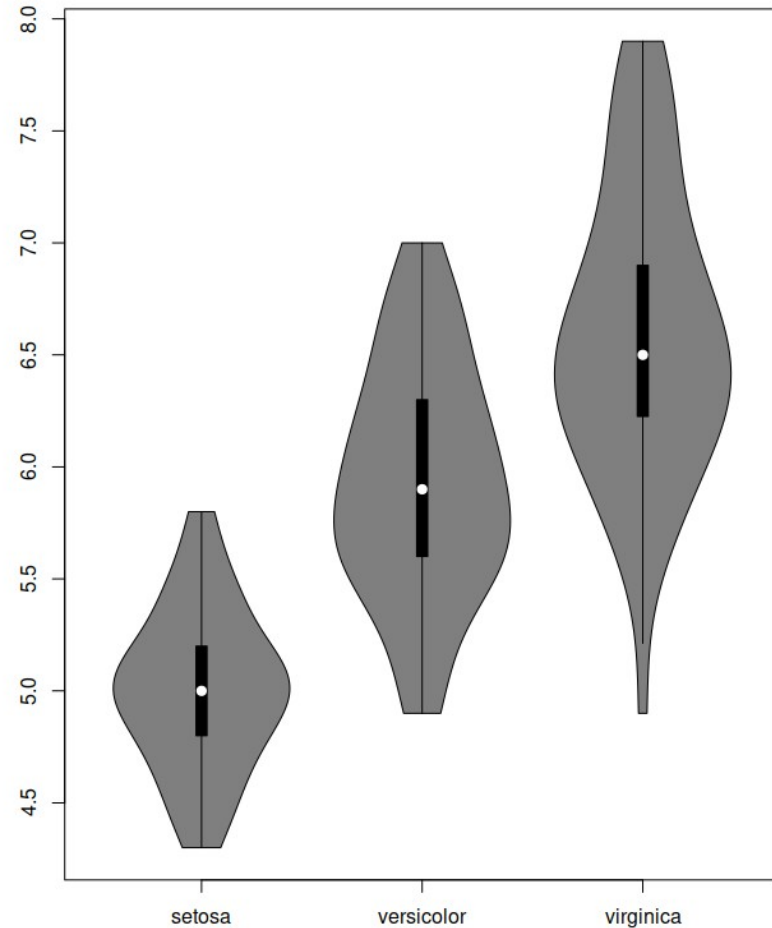


```
> virginica <- subset(iris, Species == "virginica")
> boxplot(setosa$Sepal.Length, versicolor$Sepal.Length, virginica$Sepal.Length, names=c("setosa", "versicolor", "virginica"))
> library(vioplot)
Loading required package: sm
Package 'sm', version 2.2-5.7: type help(sm) for summary information
Loading required package: zoo

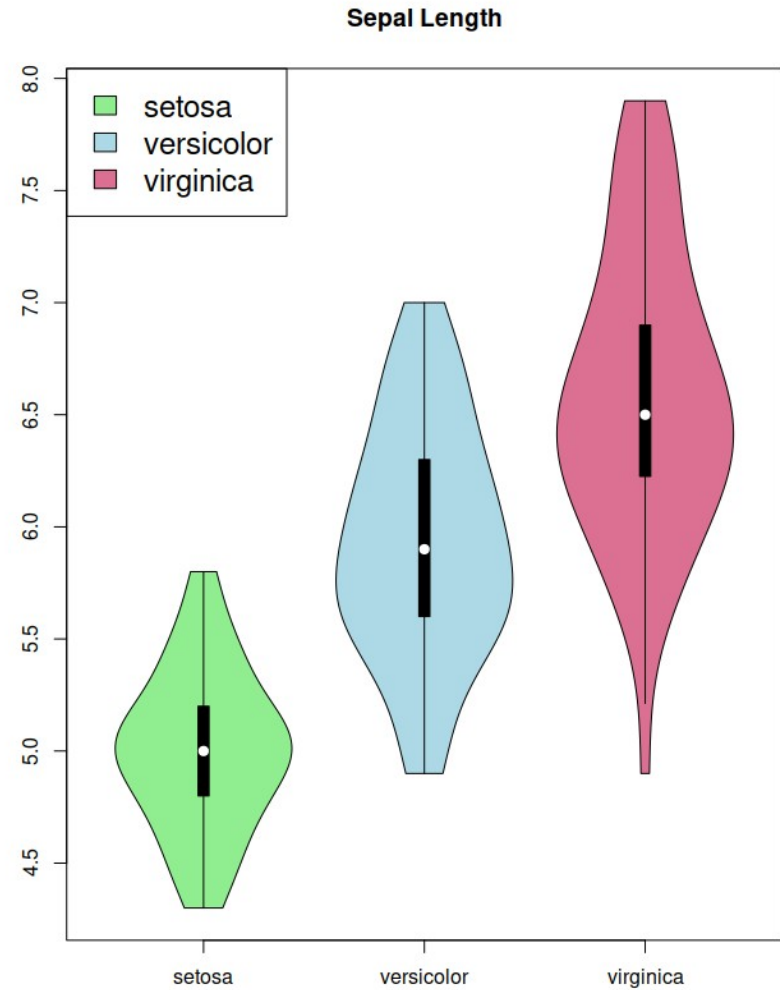
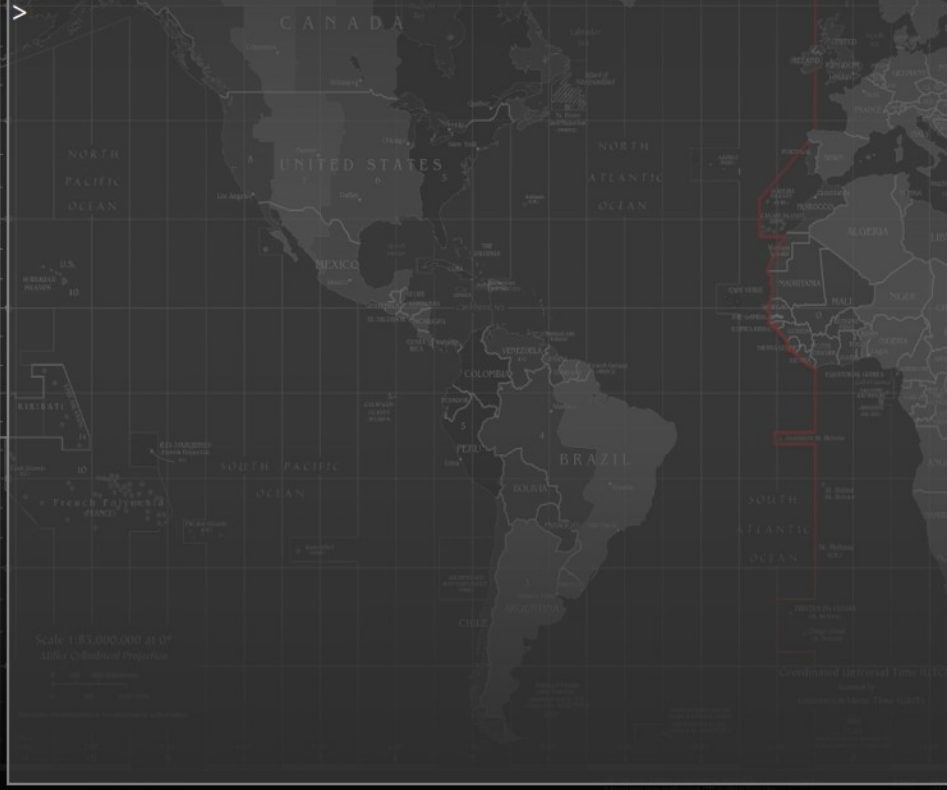
Attaching package: 'zoo'

The following objects are masked from 'package:base':
:
  as.Date, as.Date.numeric

Warning message:
package 'zoo' was built under R version 4.3.0
> vioplot(setosa$Sepal.Length, versicolor$Sepal.Length, virginica$Sepal.Length, names=c("setosa", "versicolor", "virginica"))
>
```



```
> violplot(setosa$Sepal.Length, versicolor$Sepal.Length, virginica$Sepal.Length, names=c("setosa", "versicolor", "virginica"), main = "Sepal Length", col=c("lightgreen", "lightblue", "palevioletred"))
legend("topleft", legend=c("setosa", "versicolor", "virginica"), fill=c("lightgreen", "lightblue", "palevioletred"), cex = 1.5)
```



Άσκηση 1

- Κατασκευάστε boxplot και violin plot από το σετ δεδομένων “iris” για τις υπόλοιπες τρεις μεταβλητές
 - Sepal.Width
 - Petal.Length
 - Petal.Width

Εφαρμογές και παραδείγματα ελέγχου υποθέσεων

Έλεγχος κανονικότητας

Έλεγχος κανονικότητας

- Παράδειγμα με το σετ δεδομένων “**Arthritis**” και το πακέτο *vcd*
 - Κλινική μελέτη για 84 ασθενείς με ρευματοειδή αρθρίτιδα διαφορετικού φύλου και ηλικίας, που δέχτηκαν θεραπεία με ένα κανονικό φάρμακο ή με εικονική θεραπεία (placebo)
 - Ο βαθμός βελτίωσης της υγείας τους αξιολογήθηκε με βάση μια κλίμακα κλινικών συμπτωμάτων
- Κάνουμε μια αρχική οπτική εκτίμηση με ένα γράφημα της κατανομής των τιμών

```
[sudo] password for telis:
```

```
R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"  
Copyright (C) 2022 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

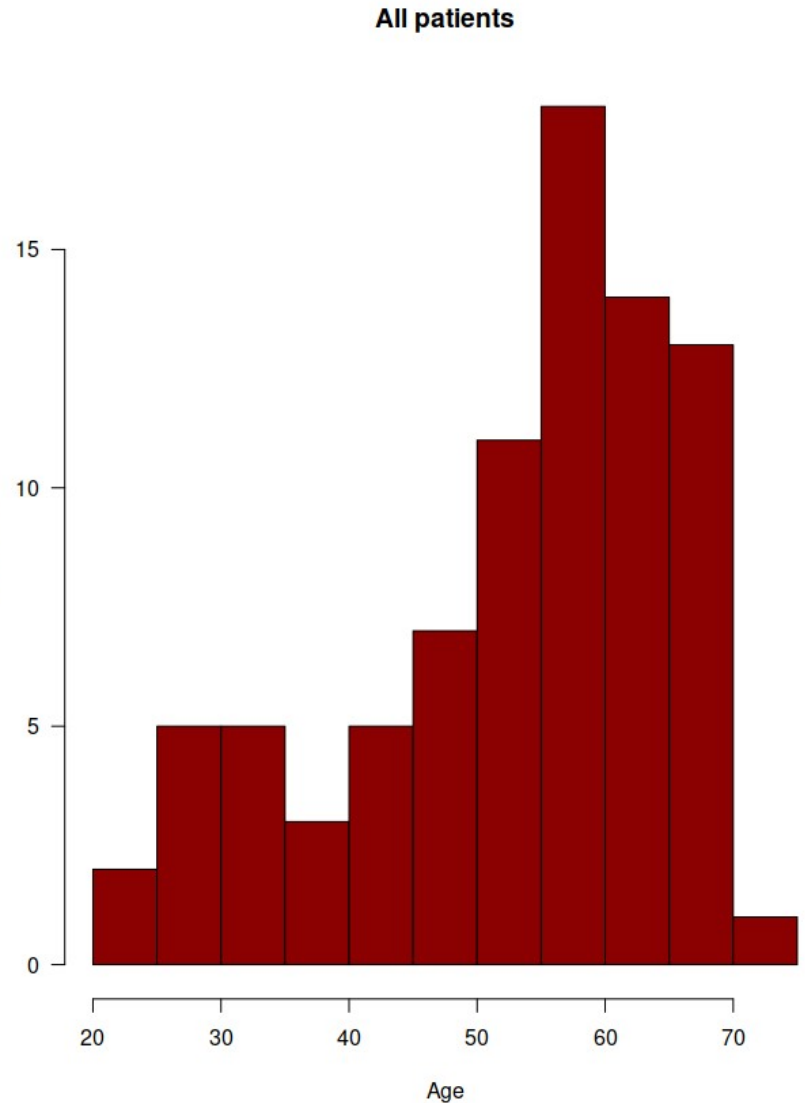
```
Natural language support but running in an English locale
```

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```

```
> library(vcd)  
str(Arthritis)  
hist(Arthritis$Age, col="dark red", breaks=10, xlab="Age", ylab="No of patients",  
main="All patients", las=1)  
Loading required package: grid  
'data.frame': 84 obs. of 5 variables:  
 $ ID : int 57 46 77 17 36 23 75 39 33 55 ...  
 $ Treatment: Factor w/ 2 levels "Placebo","Treated": 2 2 2 2 2 2 2 2 2 2 ...  
 $ Sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...  
 $ Age : int 27 29 30 32 46 58 59 59 63 63 ...  
 $ Improved : Ord.factor w/ 3 levels "None"<"Some"<: 2 1 1 3 3 3 1 3 1 1 ...  
> _
```

No of patients

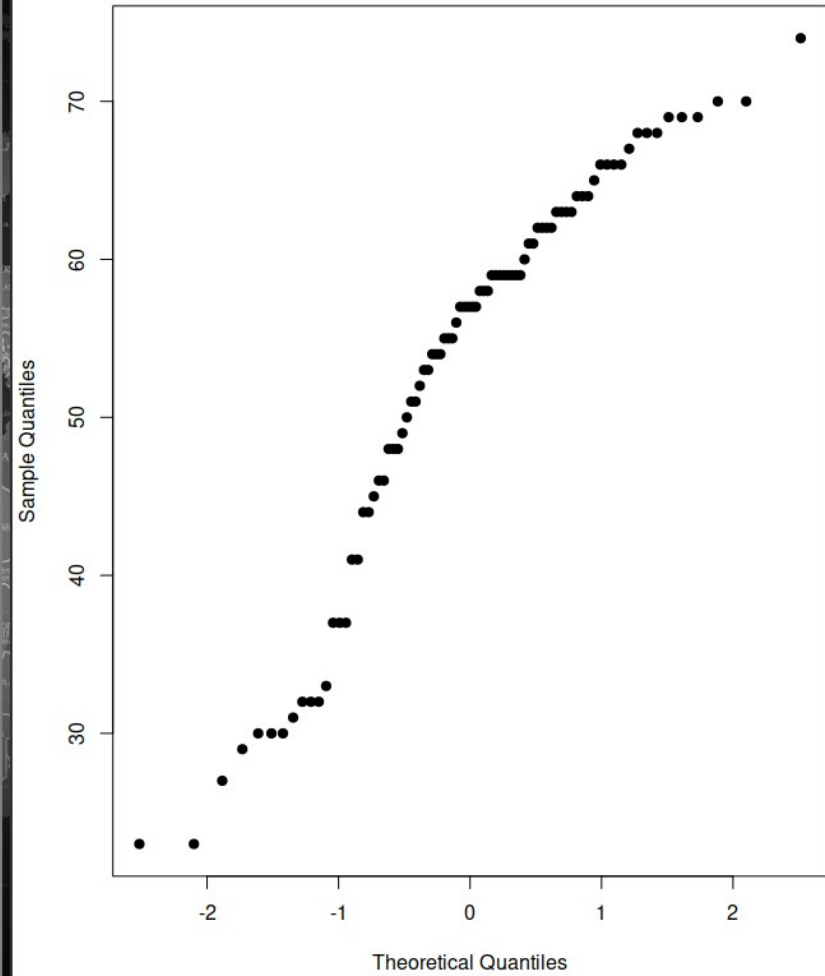


```
> qqnorm(Arthritis$Age, pch=19)
```

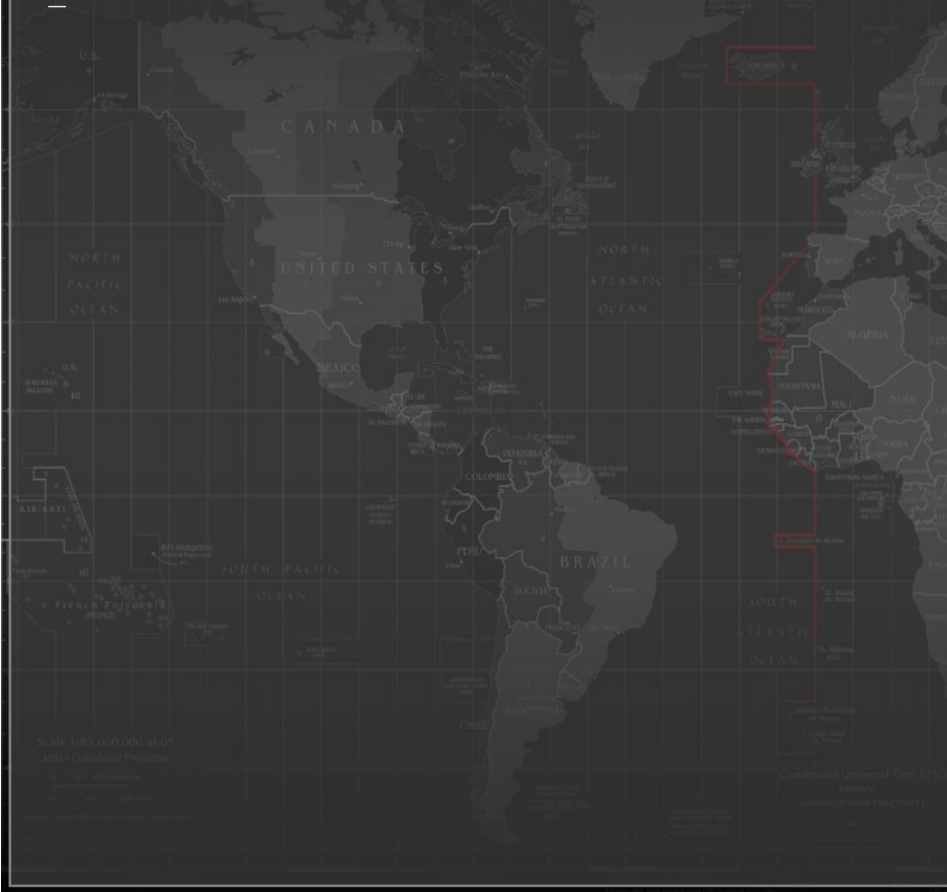
```
>
```



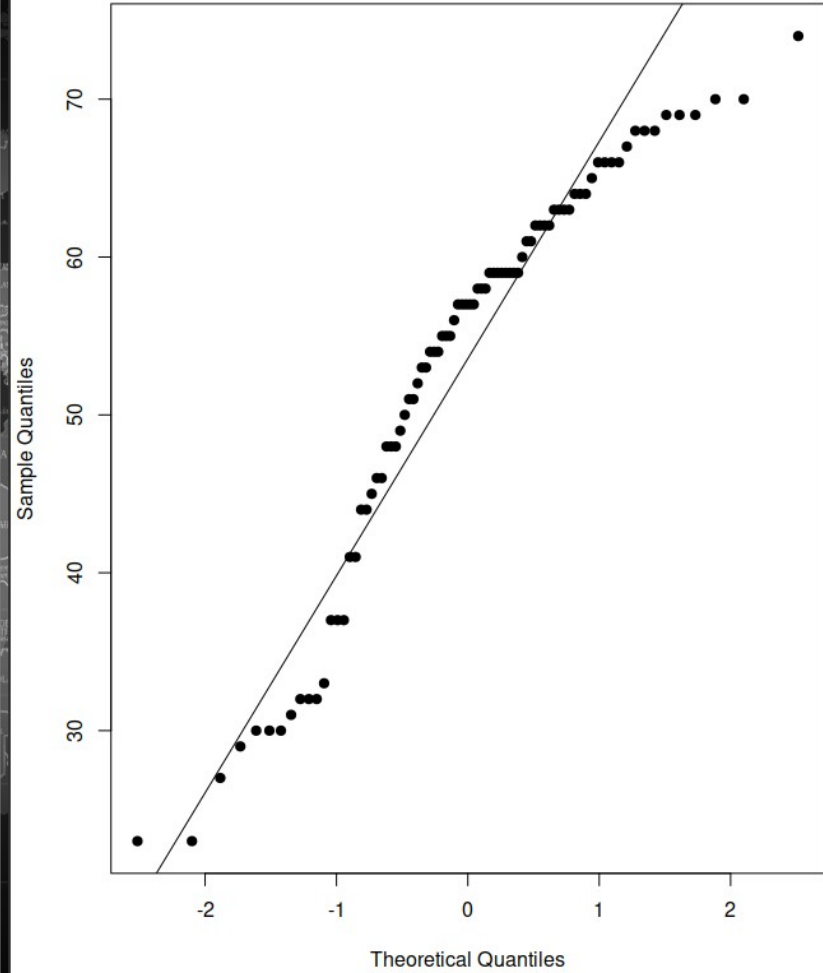
Normal Q-Q Plot



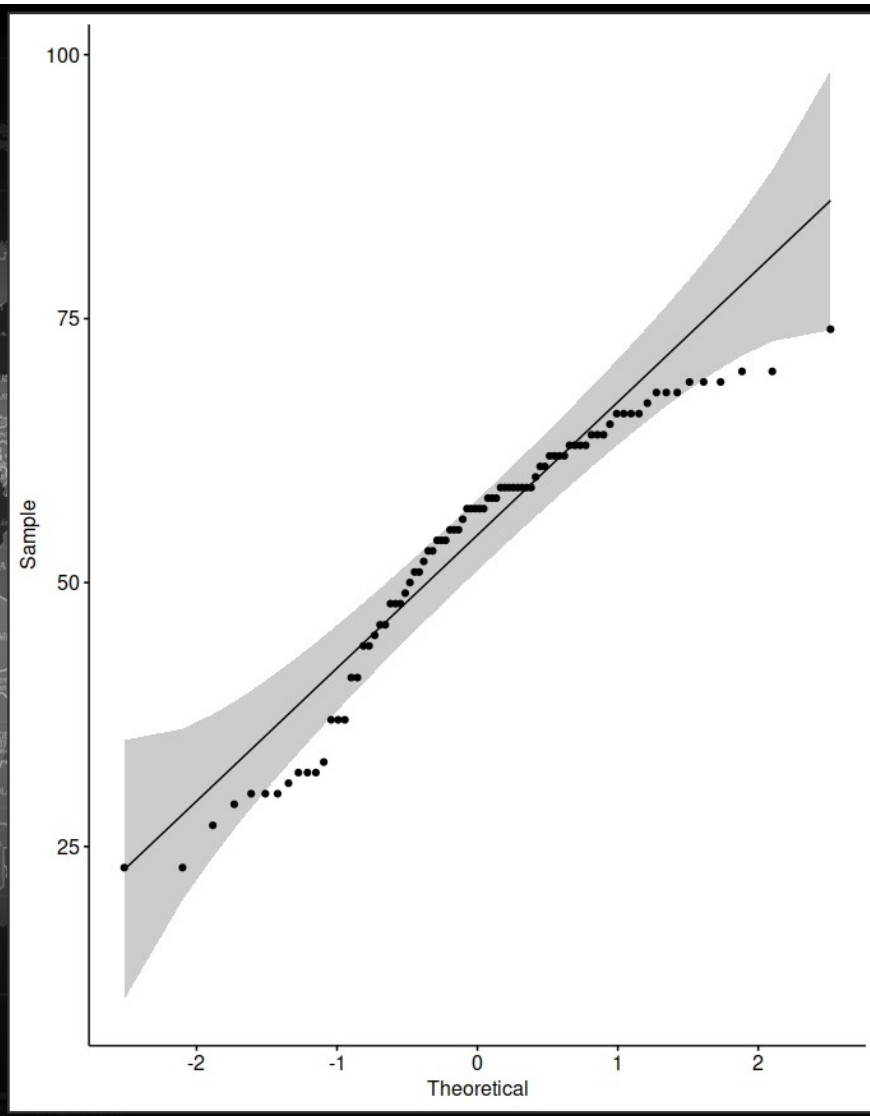
```
> nAge<-rnorm(length(Arthritis$Age), mean=mean(Arthritis$Age), sd=sd(Arthritis$Age))
> qqnorm(Arthritis$Age, pch=19)
> qqline(nAge)
>
```



Normal Q-Q Plot



```
> library(ggpubr)
> ggqqplot(Arthritis$Age)
>
```



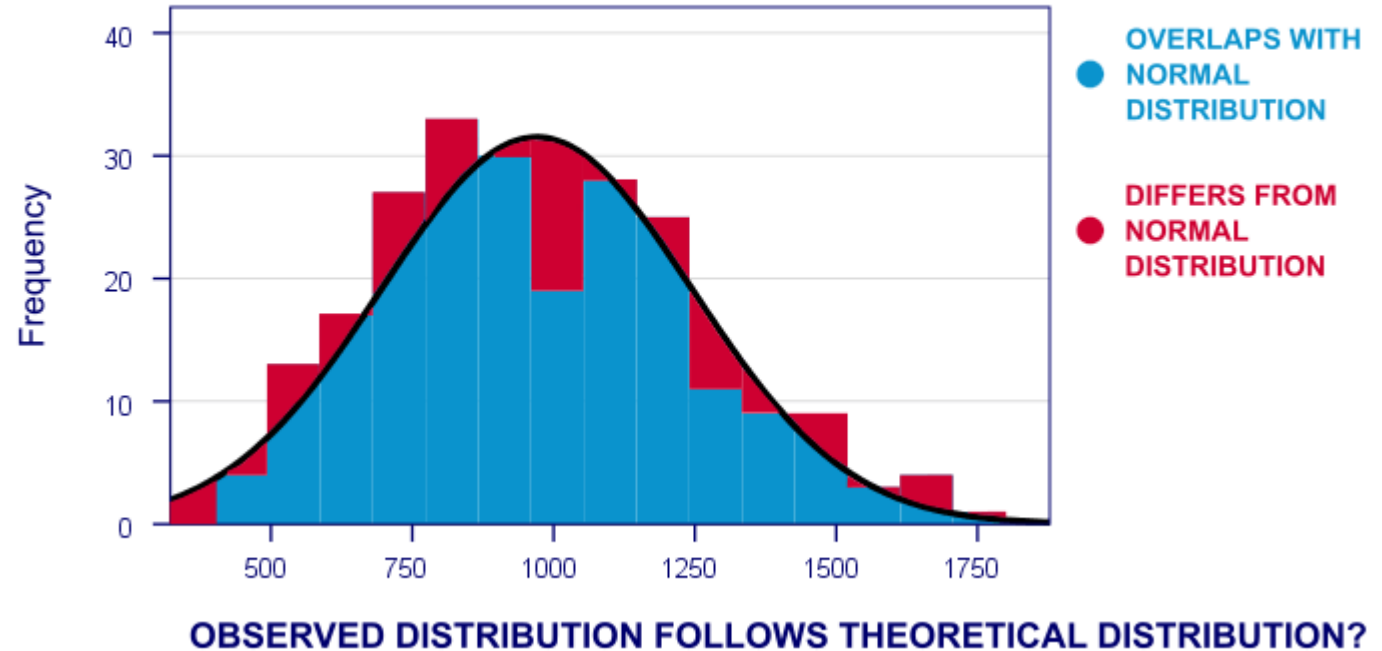
Αριθμητικός έλεγχος κανονικότητας

© www.spss-tutorials.com

SHAPIRO-WILK NORMALITY TEST

- Έλεγχος Shapiro-Wilk
- 1965: Samuel Sanford Shapiro & Martin Wilk
- Έχει αποδειχτεί ότι είναι το πιο ισχυρό τεστ
- Υπολογίζεται η τιμή W

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$



Αριθμητικός έλεγχος κανονικότητας

- Shapiro-Wilk

- Ο πιο συχνά χρησιμοποιούμενος λόγω της ευελιξίας του
- Η εφαρμογή του γίνεται με την συνάρτηση `shapiro.test()`
- Αποδίδει δύο τιμές, την τιμή **W** που είναι το στατιστικό και την p-value

```
> shapiro.test(Arthritis$Age)

      Shapiro-Wilk normality test

data:  Arthritis$Age
W = 0.91913, p-value = 5.813e-05
```

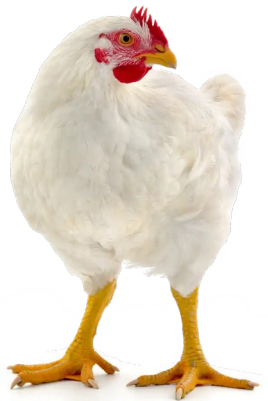
- Η πιθανότητα να προκύψει μια τιμή **W=0.91913**, αν τα δεδομένα μας κατανέμονται κανονικά, είναι μικρότερη από 0.000058
- Απορρίπτουμε τη μηδενική υπόθεση
- Τα δεδομένα μας **δεν** είναι κανονικά κατανομημένα

Άσκηση 2

- Ακολουθήστε τη διαδικασία του οπτικού και του αριθμητικού ελέγχου κανονικότητας, ακριβώς όπως παρουσιάζεται στις προηγούμενες διαφάνειες, για **όλο** το σετ δεδομένων “iris” για τις μεταβλητές
 - Petal.Length
 - Petal.Width
- Επαναλάβετε τη διαδικασία, για τις ίδιες μεταβλητές, αλλά **μόνο** για το είδος *setosa*
 - Υπάρχει κάποια διαφορά;

Παραδείγματα ελέγχου υποθέσεων

Σύγκριση μέσων τιμών



Σύγκριση μέσω τιμών

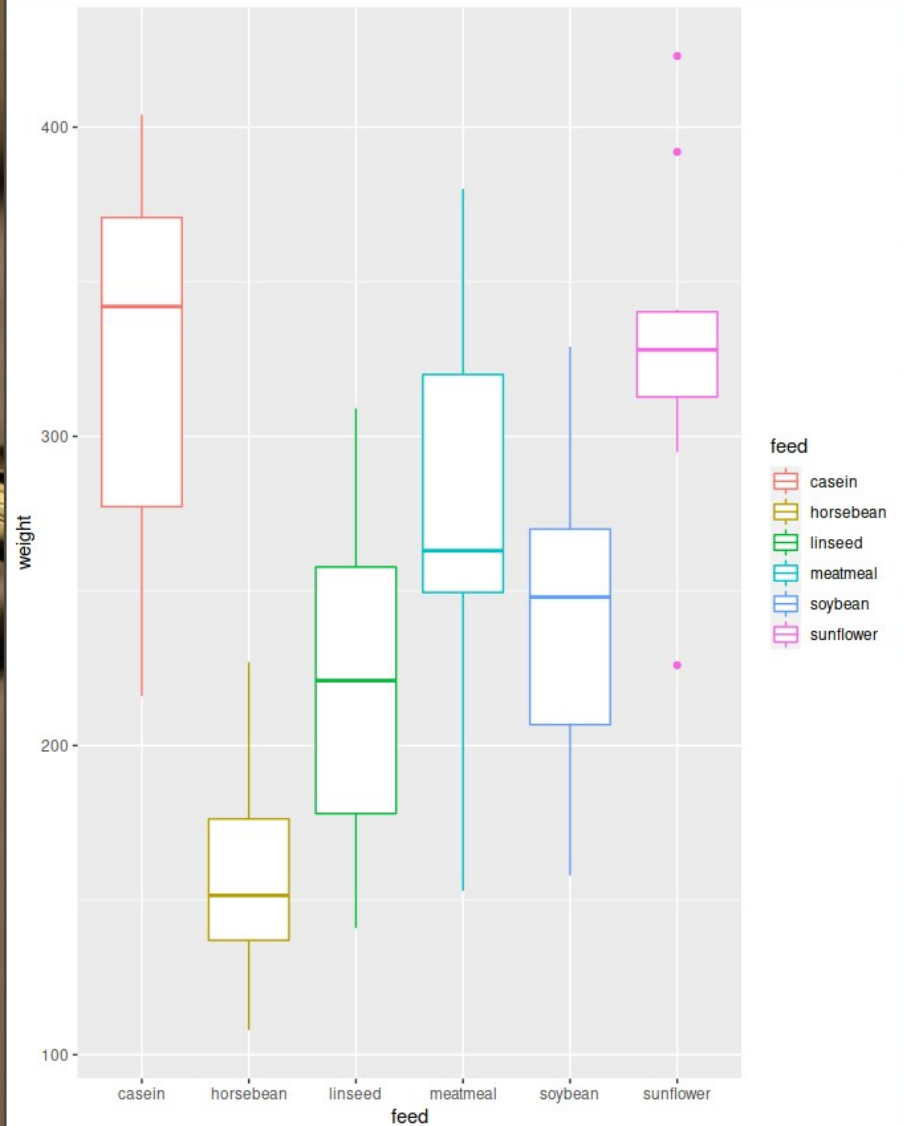
- Παράδειγμα: μελετάμε το βάρος που αποκτούν τα κοτόπουλα μιας κτηνοτροφικής μονάδας με βάση την κύρια πηγή πρωτεΐνης στην τροφή τους
 - Το σετ δεδομένων ονομάζεται **chickwts** και περιέχει τιμές βάρους για έξι διαφορετικά είδη τροφών:

```
> str(chickwts)
levels(chickwts$feed)
'data.frame': 71 obs. of 2 variables:
 $ weight: num 179 160 136 227 217 168 108 124 143 140 ...
 $ feed : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...
[1] "casein" "horsebean" "linseed" "meatmeal" "soybean" "sunflower"
> _
```

Οπτική εκτίμηση

- Το ερώτημά είναι αν η μέση τιμή του βάρους των ζώων που έχουν τραφεί με μια συγκεκριμένη τροφή είναι μεγαλύτερη από αυτήν των ζώων που έχουν τραφεί με κάποια άλλη
- Θα αναπαραστήσουμε γραφικά τα δεδομένα
 - Χρησιμοποιούμε τη δομή του dataframe για να δημιουργήσουμε θηκογράμματα (boxplots) με την χρήση της `qplot()` από το πακέτο **ggplot2**

```
> str(chickwts)
levels(chickwts$feed)
'data.frame': 71 obs. of 2 variables:
 $ weight: num 179 160 136 227 217 168 108 124 143 140 ...
 $ feed : Factor w/ 6 levels "casein","horsebean",...: 2 2 2 2 2 2 2 2 2 2 ...
[1] "casein" "horsebean" "linseed" "meatmeal" "soybean" "sunflower"
> library(ggplot2)
> qplot(feed, weight, data=chickwts, geom="boxplot", col=feed)
> _
```



Student's t-test



- Το 1908 ο William Sealy Gosset δημοσίευσε το τεστ αυτό με το ψευδώνυμο *Student*
- Ο εργοδότης του (ζυθοποιία Guinness, Δουβλίνο) απαγόρευε τη δημοσίευση επιστημονικών μεθόδων
- Χρησιμοποιούνταν για να ελεγχθεί η ποιότητα της μπύρας από διαφορετικούς τύπους
- Η μέθοδος έγινε γνωστή από τον Fisher

- Μπορεί να χρησιμοποιηθεί με ένα ή δύο δείγματα
- Προϋποθέτει κανονική κατανομή
- Για σύγκριση μέσων όρων δύο δειγμάτων:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1) s_{X_1}^2 + (n_2 - 1) s_{X_2}^2}{n_1 + n_2 - 2}}$$

Σύγκριση μέσων όρων δύο δειγμάτων

- Αρχικά εξάγουμε τα υποσύνολα των δύο δειγμάτων από το σύνολο των δεδομένων
- Ελέγχουμε την κανονικότητα
 - OK
- Εφαρμόσουμε το t-test με την συνάρτηση *t.test()*
 - *Οι δύο μέσοι όροι διαφέρουν σημαντικά*
- Επαναλαμβάνουμε για 99% διάστημα εμπιστοσύνης

```
> str(chickwts)
'data.frame':  71 obs. of  2 variables:
 $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
 $ feed  : Factor w/ 6 levels "casein","horsebean",...: 2 2 2 2 2 2 2 2 2 2 ...
> levels(chickwts$feed)
[1] "casein" "horsebean" "linseed" "meatmeal" "soybean" "sunflower"
> shapiro.test(linseed$weight)
```

Shapiro-Wilk normality test

```
data:  linseed$weight
W = 0.96931, p-value = 0.9035
```

```
> shapiro.test(horsebean$weight)
```

Shapiro-Wilk normality test

```
data:  horsebean$weight
W = 0.93758, p-value = 0.5264
```

```
>
```



```
> t.test(linseed$weight, horsebean$weight)
```

Welch Two Sample t-test

data: linseed\$weight and horsebean\$weight

t = 3.0172, df = 19.769, p-value = 0.006869

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

18.0403 99.0597

sample estimates:

mean of x	mean of y
218.75	160.20

```
>
```

Τι σημαίνει αυτό
το p-value?

Default επίπεδο
εμπιστοσύνης 95%

```
> t.test(linseed$weight, horsebean$weight, conf.level = 0.99)
```

Welch Two Sample t-test

data: linseed\$weight and horsebean\$weight

t = 3.0172, df = 19.769, p-value = 0.006869

alternative hypothesis: true difference in means is not equal to 0

99 percent confidence interval:

3.267538 113.832462

sample estimates:

mean of x	mean of y
218.75	160.20

```
>
```

Το p-value δεν αλλάζει, τι
μας λέει αυτό;

Επίπεδο εμπιστοσύνης
99%

```
> student<-t.test(linseed$weight,horsebean$weight)
> student

      Welch Two Sample t-test

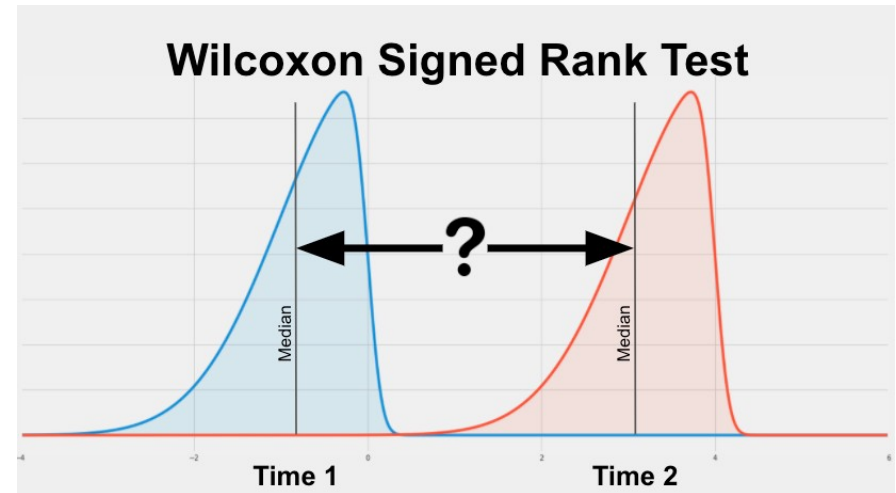
data: linseed$weight and horsebean$weight
t = 3.0172, df = 19.769, p-value = 0.006869
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 18.0403 99.0597
sample estimates:
mean of x mean of y
 218.75   160.20

> str(student)
List of 10
 $ statistic  : Named num 3.02
 ..- attr(*, "names")= chr "t"
 $ parameter  : Named num 19.8
 ..- attr(*, "names")= chr "df"
 $ p.value    : num 0.00687
 $ conf.int   : num [1:2] 18 99.1
 ..- attr(*, "conf.level")= num 0.95
 $ estimate   : Named num [1:2] 219 160
 ..- attr(*, "names")= chr [1:2] "mean of x" "mean of y"
 $ null.value : Named num 0
 ..- attr(*, "names")= chr "difference in means"
 $ stderr     : num 19.4
 $ alternative: chr "two.sided"
 $ method     : chr "Welch Two Sample t-test"
 $ data.name  : chr "linseed$weight and horsebean$weight"
 - attr(*, "class")= chr "htest"
> student$p.value
[1] 0.006869064
> student$conf.int
[1] 18.0403 99.0597
attr(,"conf.level")
[1] 0.95
> _
```

Μπορούμε να ορίσουμε το `t.test` σαν **αντικείμενο** της R (εδώ `student`) και να καλέσουμε τα στοιχεία του (π.χ. `p.value`) για ευκολία

Σύγκριση μέσων τιμών σε μη κανονικά κατανεμημένα δείγματα

- Πώς προχωράμε στη σύγκριση δειγμάτων που δεν είναι κανονικά κατανεμημένα;
- Θα καταφύγουμε σε κάποιο **μη-παραμετρικό έλεγχο** όπως αυτός που πραγματοποιείται με την μέθοδο Mann-Whitney ή Wilcoxon Rank-Sum (που είναι πρακτικά ισοδύναμες)



Wilcoxon Rank-Sum


- Στο σετ δεδομένων Arthritis δεν είχαμε κανονικά κατανομημένες τιμές ηλικίας
- Θέλουμε να δούμε αν το προφίλ των ασθενών που αποκρίνονται καλύτερα στη θεραπεία έχει μια συγκεκριμένη ηλικιακή τάση
- Χωρίζουμε τους ασθενείς με βάση το αποτέλεσμα της θεραπείας

```
> shapiro.test(Arthritis$Age)

      Shapiro-Wilk normality test

data:  Arthritis$Age
W = 0.91913, p-value = 5.813e-05

>
```



```
> None <- subset(Arthritis, Improved == "None")
> Marked <- subset(Arthritis, Improved == "Marked")
> shapiro.test(None$Age)
```

Shapiro-Wilk normality test

```
data: None$Age
W = 0.94019, p-value = 0.029
```

```
> shapiro.test(Marked$Age)
```

Shapiro-Wilk normality test

```
data: Marked$Age
W = 0.91064, p-value = 0.02049
```

```
> _
```

Χωρίζουμε το σετ σε δύο νέα, ανάλογα με την ανταπόκριση στη θεραπεία.
Και οι δύο νέες ομάδες δεδομένων δεν ακολουθούν την κανονική κατανομή

```
> None <- subset(Arthritis, Improved == "None")
> Marked <- subset(Arthritis, Improved == "Marked")
> shapiro.test(None$Age)

      Shapiro-Wilk normality test

data:  None$Age
W = 0.94019, p-value = 0.029

> shapiro.test(Marked$Age)

      Shapiro-Wilk normality test

data:  Marked$Age
W = 0.91064, p-value = 0.02049

> wilcox.test(None$Age, Marked$Age)

      Wilcoxon rank sum test with continuity correction

data:  None$Age and Marked$Age
W = 394.5, p-value = 0.02056
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(None$Age, Marked$Age) :
  cannot compute exact p-value with ties
> _
```

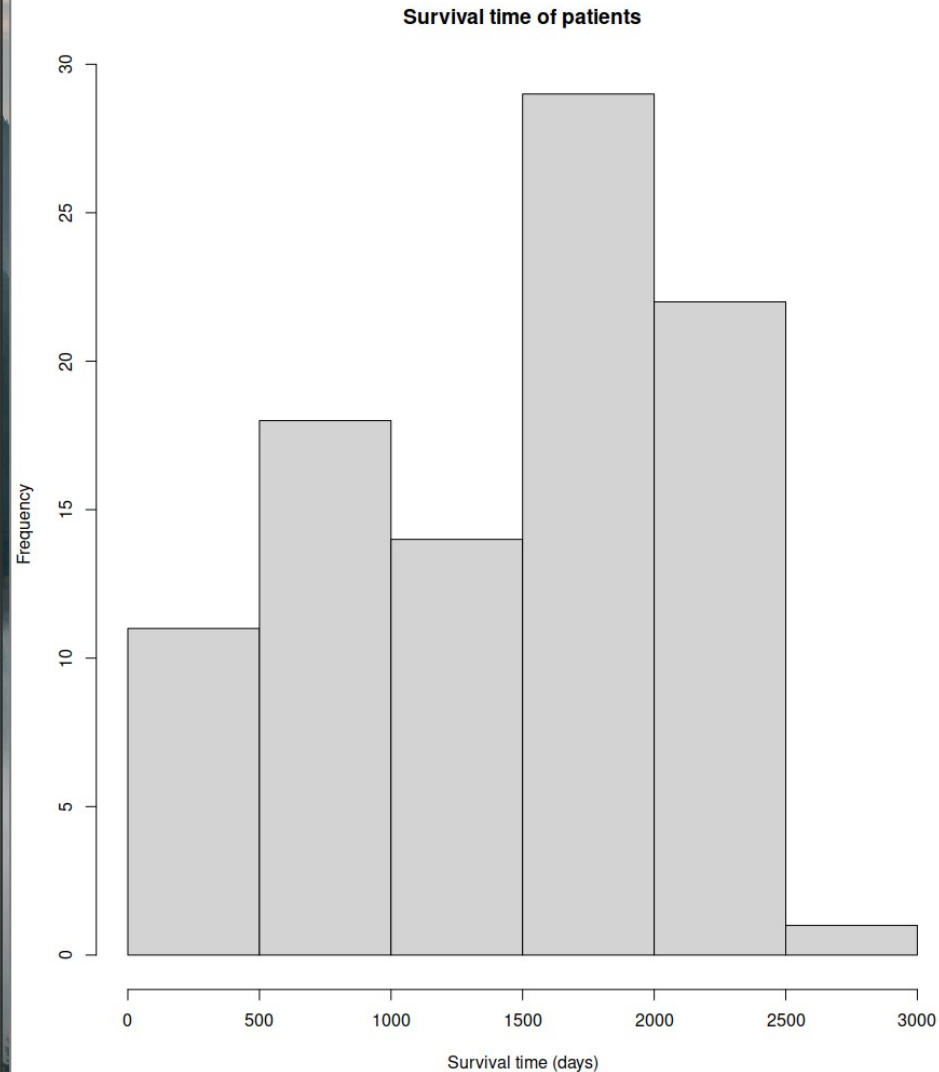
Εφαρμόζουμε τον μη-παραμετρικό έλεγχο Wilcoxon Rank Sum. Όπως βλέπουμε από το p-value που είναι μικρότερο του 0.05, υπάρχουν διαφορές μεταξύ των δύο ομάδων.

```
> data<-read.csv("cancer1.csv")
> str(data)
'data.frame': 95 obs. of 14 variables:
 $ Patient.ID : chr "P109342" "P124450" "P131833" "P131888" ...
 $ Survival_time: int 2329 2532 2271 2193 2387 2225 2240 2314 299 2295 ...
 $ death : int 1 0 0 0 0 0 0 0 1 0 ...
 $ Tumor_size : num 2.6 3.5 2 3 4 2.5 2.2 1.2 2.8 2 ...
 $ Grade : int 3 3 2 2 2 2 3 3 1 2 ...
 $ Stage : chr "IB" "IB" "IA2" "IA3" ...
 $ Age : int 67 68 80 63 88 66 69 52 67 59 ...
 $ Sex : chr "Male" "Female" "Female" "Male" ...
 $ Cigarette : chr "Former" "Former" "Never" "Former" ...
 $ Pack_per_year: num 50 52.5 0 47 0 ...
 $ Therapy : chr "None" "None" "None" "None" ...
 $ batch : int 1 1 1 1 1 1 2 2 1 ...
 $ EGFR : chr NA "Negative" "Negative" "Negative" ...
 $ KRAS : chr NA "Negative" "Negative" "G12C" ...
>
```

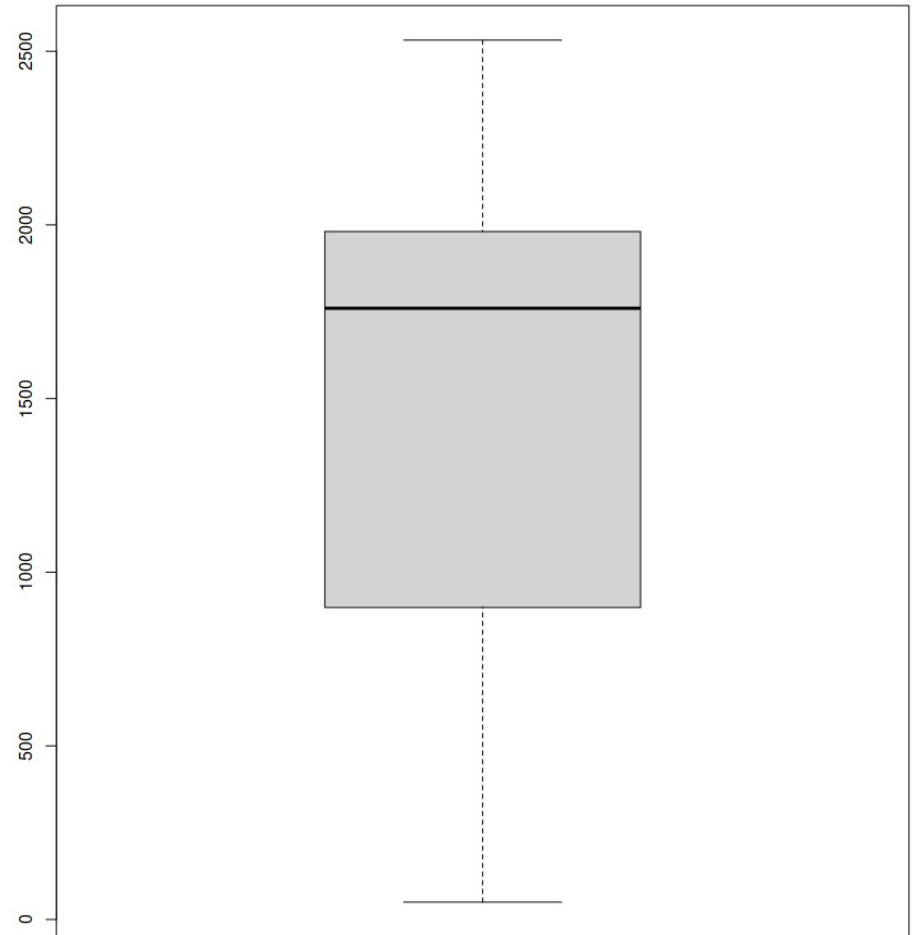
Θα κάνουμε το ίδιο και με πραγματικά δεδομένα από κλινικές έρευνες που σχετίζονται με διάφορους παράγοντες και υπάρχουν στο αρχείο “cancer1.csv” που αποκτήθηκε από το kaggle

<https://www.kaggle.com/datasets/imtkaggleteam/clinical-dataset/data>

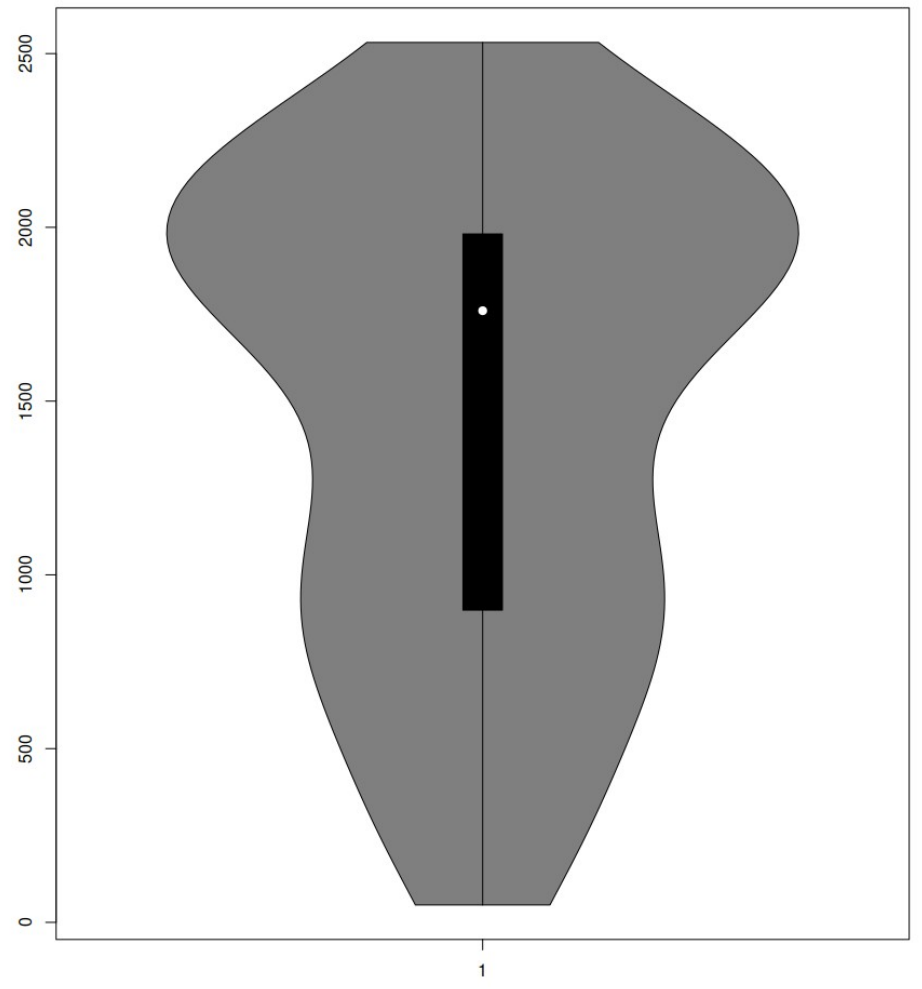

```
> data<-read.csv("cancer1.csv")
> str(data)
'data.frame': 95 obs. of 14 variables:
 $ Patient.ID : chr "P109342" "P124450" "P131833" "P131888" ...
 $ Survival_time: int 2329 2532 2271 2193 2387 2225 2240 2314 299 2295 ...
 $ death : int 1 0 0 0 0 0 0 1 0 ...
 $ Tumor_size : num 2.6 3.5 2 3 4 2.5 2.2 1.2 2.8 2 ...
 $ Grade : int 3 3 2 2 2 2 3 3 1 2 ...
 $ Stage : chr "IB" "IB" "IA2" "IA3" ...
 $ Age : int 67 68 80 63 88 66 69 52 67 59 ...
 $ Sex : chr "Male" "Female" "Female" "Male" ...
 $ Cigarette : chr "Former" "Former" "Never" "Former" ...
 $ Pack_per_year: num 50 52.5 0 47 0 ...
 $ Therapy : chr "None" "None" "None" "None" ...
 $ batch : int 1 11 1 1 1 2 2 2 1 ...
 $ EGFR : chr NA "Negative" "Negative" "Negative" ...
 $ KRAS : chr NA "Negative" "Negative" "G12C" ...
> hist(data$Survival_time,
main = "Survival time of patients",
xlab = "Survival time (days)",
ylab = "Frequency"
)
> _
```



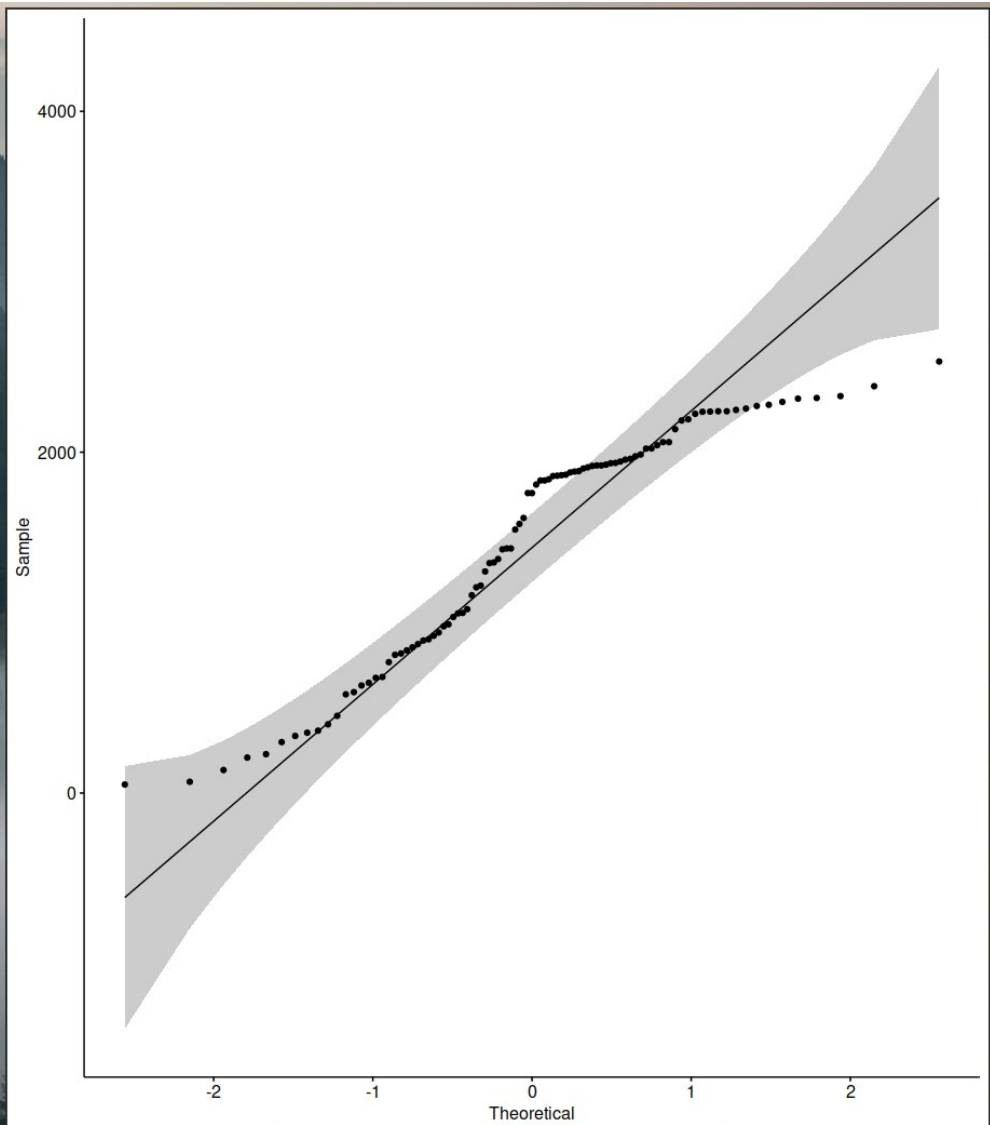
```
> data<-read.csv("cancer1.csv")
> str(data)
'data.frame': 95 obs. of 14 variables:
 $ Patient.ID : chr "P109342" "P124450" "P131833" "P131888" ...
 $ Survival_time: int 2329 2532 2271 2193 2387 2225 2240 2314 299 2295 ...
 $ death : int 1 0 0 0 0 0 0 1 0 ...
 $ Tumor_size : num 2.6 3.5 2 3 4 2.5 2.2 1.2 2.8 2 ...
 $ Grade : int 3 3 2 2 2 2 3 3 1 2 ...
 $ Stage : chr "IB" "IB" "IA2" "IA3" ...
 $ Age : int 67 68 80 63 88 66 69 52 67 59 ...
 $ Sex : chr "Male" "Female" "Female" "Male" ...
 $ Cigarette : chr "Former" "Former" "Never" "Former" ...
 $ Pack_per_year: num 50 52.5 0 47 0 ...
 $ Therapy : chr "None" "None" "None" "None" ...
 $ batch : int 1 1 1 1 1 1 2 2 2 1 ...
 $ EGFR : chr NA "Negative" "Negative" "Negative" ...
 $ KRAS : chr NA "Negative" "Negative" "G12C" ...
> hist(data$Survival_time,
main = "Survival time of patients",
xlab = "Survival time (days)",
ylab = "Frequency"
)
> boxplot(data$Survival_time)
> _
```



```
> vioplot(data$Survival_time)
>
> _
```



```
> vioplot(data$Survival_time)
> ggqqplot(data$Survival_time)
>
```



```
> vioplot(data$Survival_time)
> ggqqplot(data$Survival_time)
> shapiro.test(data$Survival_time)
```

Shapiro-Wilk normality test

```
data: data$Survival_time
W = 0.92043, p-value = 2.353e-05
```

```
> _
```

Όπως φάνηκε και από τα γραφήματα, τα δεδομένα της επιβίωσης των ασθενών δεν ακολουθούν κανονική κατανομή

Ερώτημα

- Κάνουμε το ερώτημα, αν η επιβίωση των ασθενών σχετίζεται με το κάπνισμα
 - μεταβλητή Cigarette με τρεις κατηγορίες: "current", "former" and "never"
- Χωρίζουμε τα δεδομένα στις τρεις κατηγορίες και ελέγχουμε ξανά την κανονικότητα
- Στη συνέχεια και ανά δύο κατηγορίες, θα κάνουμε t.test αν τα δεδομένα ακολουθούν κανονική κατανομή, ή τον μη-παραμετρικό έλεγχο Wilcoxon Rank Sum, αν δεν έχουμε κανονικές κατανομές

```
> never <- subset(data, Cigarette == "Never")
> current <- subset(data, Cigarette == "Current")
> former <- subset(data, Cigarette == "Former")
> shapiro.test(never$Survival_time)
```

Shapiro-Wilk normality test

```
data: never$Survival_time
W = 0.8959, p-value = 0.02464
```

```
> shapiro.test(current$Survival_time)
```

Shapiro-Wilk normality test

```
data: current$Survival_time
W = 0.92154, p-value = 0.3317
```

```
> shapiro.test(former$Survival_time)
```

Shapiro-Wilk normality test

```
data: former$Survival_time
W = 0.92176, p-value = 0.0007286
```

```
> _
```

H_0 : η κατανομή είναι κανονική

Απορρίπτεται η H_0 για επίπεδο εμπιστοσύνης 95%

Δεν απορρίπτεται η H_0 για επίπεδο εμπιστοσύνης 95%

Απορρίπτεται η H_0 για επίπεδο εμπιστοσύνης 95%

```
> wilcox.test(never$Survival_time,current$Survival_time)

    Wilcoxon rank sum exact test

data: never$Survival_time and current$Survival_time
W = 118, p-value = 0.9252
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(never$Survival_time,former$Survival_time)

    Wilcoxon rank sum test with continuity correction

data: never$Survival_time and former$Survival_time
W = 610.5, p-value = 0.4701
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(current$Survival_time,former$Survival_time)

    Wilcoxon rank sum test with continuity correction

data: current$Survival_time and former$Survival_time
W = 284, p-value = 0.3836
alternative hypothesis: true location shift is not equal to 0

>
```

H_0 : Οι μέσοι όροι για κάθε σύγκριση δεν διαφέρουν

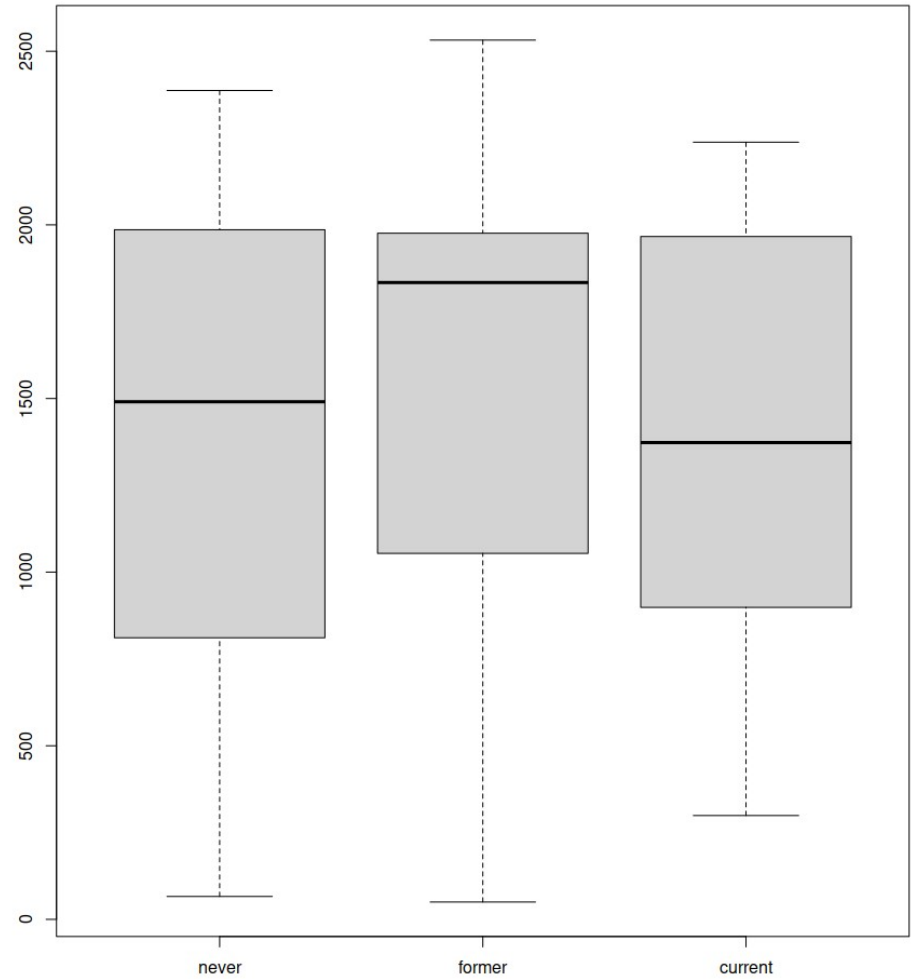
Καμία από τις συγκρίσεις δεν έδωσε $p\text{-value} < 0,05$ άρα σε καμία από τις τρεις συγκρίσεις δεν μπορέσαμε να απορρίψουμε τη μηδενική υπόθεση

Τα επίπεδα καπνίσματος δεν φαίνονται να επηρεάζουν τον χρόνο επιβίωσης των ασθενών

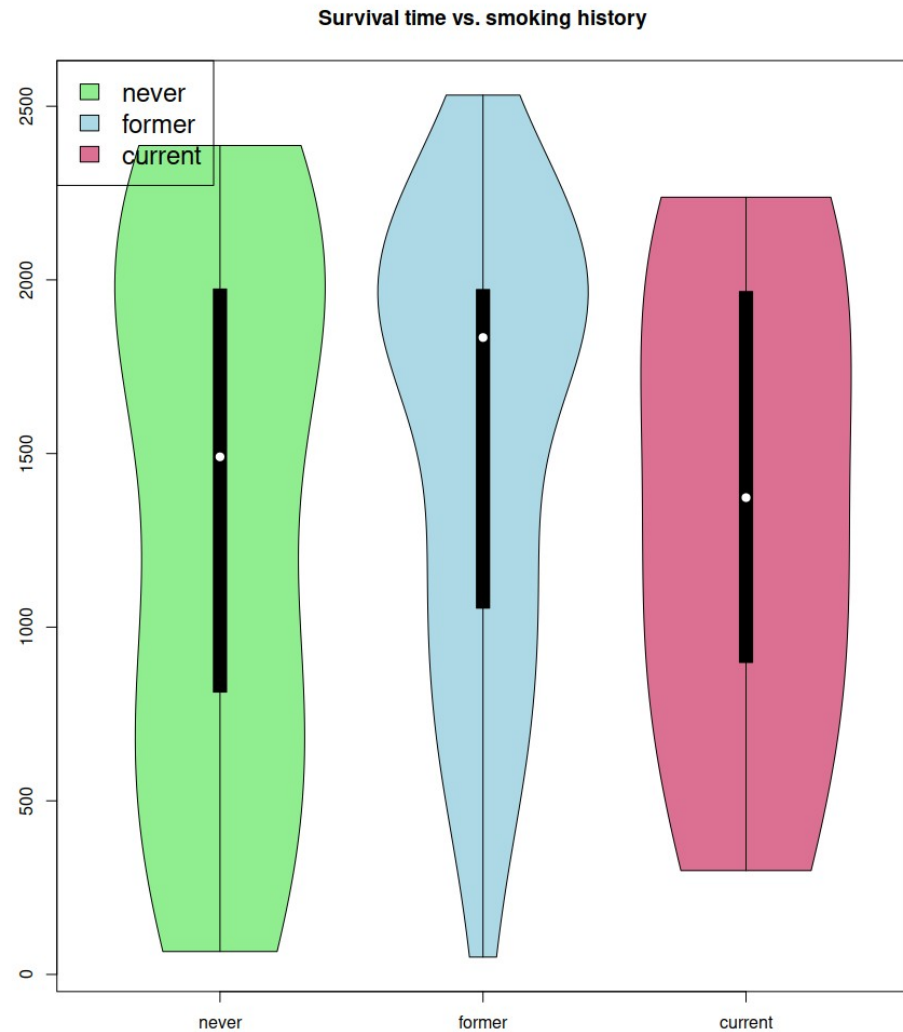
Μπορούμε να συμπεράνουμε ότι το κάπνισμα δεν κάνει κακό στην υγεία;


```
> boxplot(never$Survival_time, former$Survival_time, current$Survival_time,  
names=c("never", "former", "current"))
```

```
> _
```



```
> vioplot(never$Survival_time, former$Survival_time, current$Survival_time,
names=c("never", "former", "current"), main = "Survival time vs. smoking h
istory", col=c("lightgreen", "lightblue", "palevioletred"))
> legend("topleft", legend=c("never", "former", "current"), fill=c("lightgr
een", "lightblue", "palevioletred"), cex = 1.5)
> _
```



Άσκηση 3

- Κατεβάστε τα δεδομένα `carcer1.csv` από το `eclass` και βρείτε αν το φύλο παίζει ρόλο στον χρόνο επιβίωσης των ασθενών
 - Ελέγξτε την κανονικότητα των επιμέρους σετ δεδομένων
 - Κάντε τη σύγκριση μέσω όρων με τον σωστό τρόπο
 - Κατασκευάστε `boxplot` και `violin plot` συγκριτικά για τα δύο φύλα

Έλεγχος υποθέσεων

False positive

False negative

Type I Error



Type II Error



Αξιοπιστία των στατιστικών ελέγχων

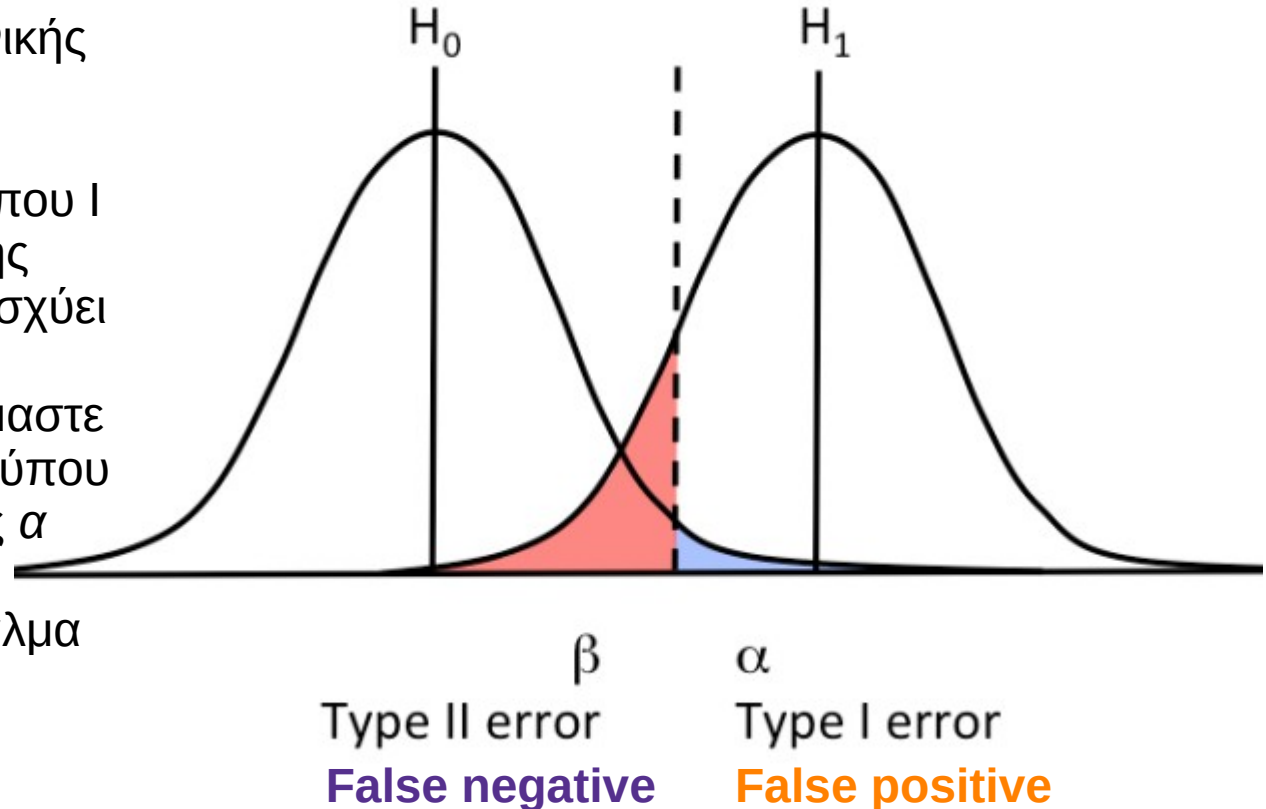
Σφάλμα τύπου I (Type I error)

Εσφαλμένη απόρριψη της μηδενικής υπόθεσης (false positive)

Η πιθανότητα του σφάλματος τύπου I είναι η πιθανότητα απόρριψης της μηδενικής υπόθεσης, ενώ αυτή ισχύει

Σε έναν στατιστικό έλεγχο, δεχόμαστε την πιθανότητα του σφάλματος τύπου I σαν το επίπεδο σημαντικότητας α

Για $\alpha = 0.05$, δεχόμαστε ένα σφάλμα τύπου I ίσο με 5%

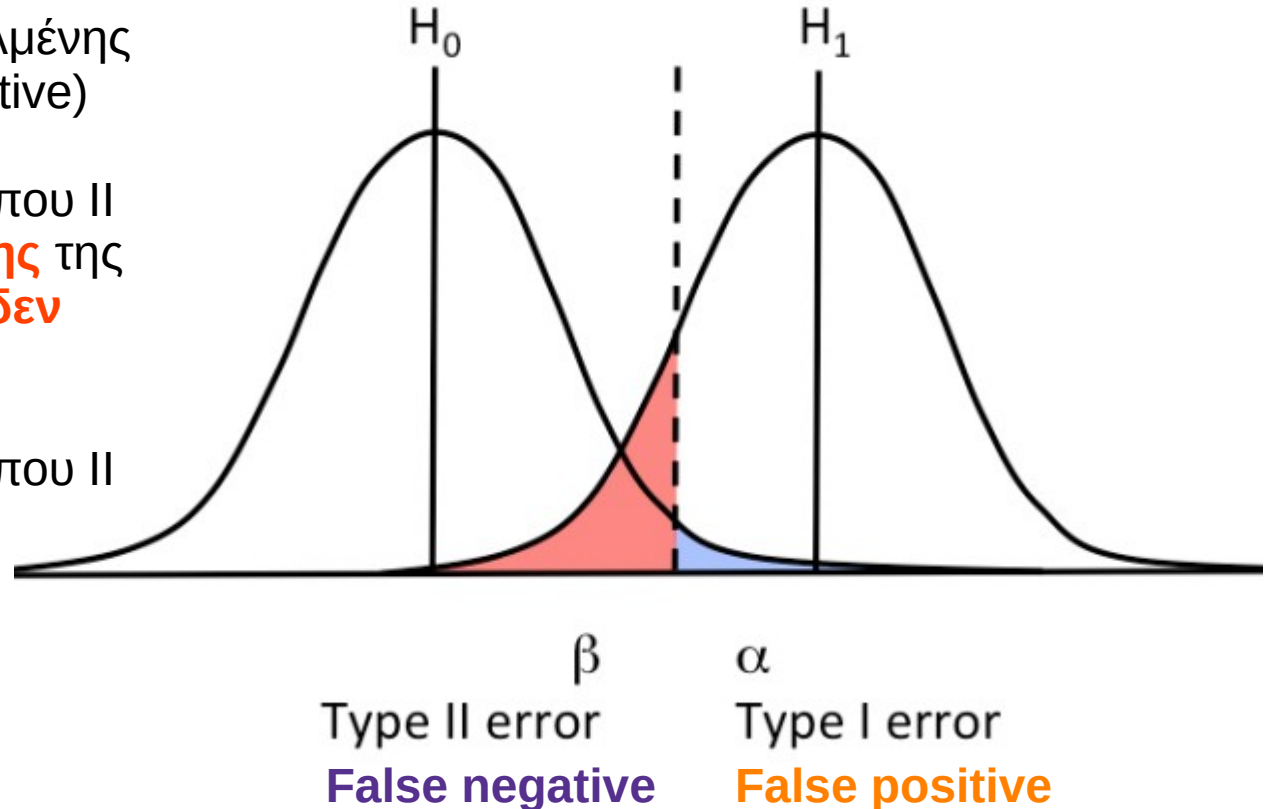


Σφάλμα τύπου II (Type II error)

Αποτυχία απόρριψης μιας εσφαλμένης μηδενικής υπόθεσης (false negative)

Η πιθανότητα του σφάλματος τύπου II είναι η πιθανότητα **μη απόρριψης** της μηδενικής υπόθεσης, ενώ αυτή **δεν** ισχύει

Η πιθανότητα του σφάλματος τύπου II αναφέρεται σαν β



Έλεγχος του σφάλματος τύπου II

- Οι έλεγχοι σημαντικότητας δεν παρέχουν καμία εγγύηση για το ποσοστό σφάλματος τύπου II
 - Ελέγχουν μόνο ως προς την πιθανότητα του σφάλματος τύπου I
- Πώς μπορούμε να μειώσουμε την πιθανότητα του σφάλματος τύπου II;
 - Όσο μεγαλύτερο είναι το **μέγεθος του δείγματος**, τόσο χαμηλότερο αναμένεται να είναι το ποσοστό σφάλματος τύπου II
 - Εξαρτάται επίσης από το πόσο μεγάλη ή μικρή είναι η επίδραση της υπόθεσης που ελέγχεται (π.χ. η διαφορά μεταξύ δύο μέσων)

Στατιστική ισχύς

- Η στατιστική ισχύς (statistical power) είναι η πιθανότητα να απορρίψουμε (σωστά) τη μηδενική υπόθεση, όταν ισχύει εναλλακτική υπόθεση
 - $power = Pr(\text{reject } H_0 \mid H_1 \text{ is true}) = \mathbf{1 - \beta}$
- Μία έρευνα έχει μεγάλη στατιστική ισχύ όταν η πιθανότητα απόρριψης της μηδενικής υπόθεσης (H_0) είναι μεγάλη, ενώ ισχύει η εναλλακτική υπόθεση (H_1)
- Ικανοποιητική στατιστική ισχύς θεωρείται π.χ. το 80%, δηλαδή $\beta=0.20$

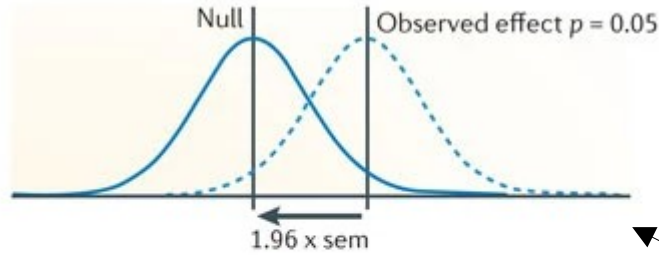
Πρόβλεψη της ισχύος ενός πειράματος

- Εκτιμούμε την ελάχιστη αναμενόμενη επίδραση του αντικειμένου της έρευνας
 - Πόσο μικρές είναι οι διαφορές που θέλουμε να εντοπίσουμε;
- Εκτιμούμε τη διακύμανση στον πληθυσμό
 - (Αυτό μπορεί να μην είναι εύκολο)
- Με βάση τις παραπάνω παραμέτρους, υπολογίζουμε το ελάχιστο μέγεθος δείγματος που απαιτείται
 - Οι μικρές αναμενόμενες διαφορές και η μεγάλη διακύμανση απαιτούν μεγαλύτερα μεγέθη δείγματος

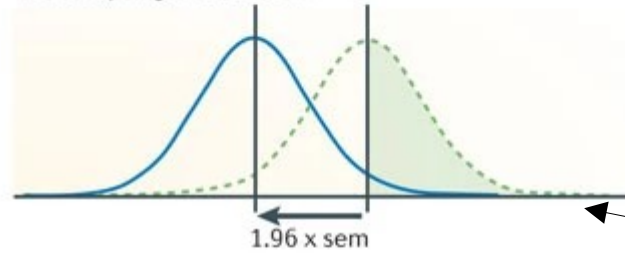
Power failure: why small sample size undermines the reliability of neuroscience

[Katherine S. Button](#), [John P. A. Ioannidis](#), [Claire Mokrysz](#), [Brian A. Nosek](#), [Jonathan Flint](#), [Emma S. J. Robinson](#) & [Marcus R. Munafò](#) 

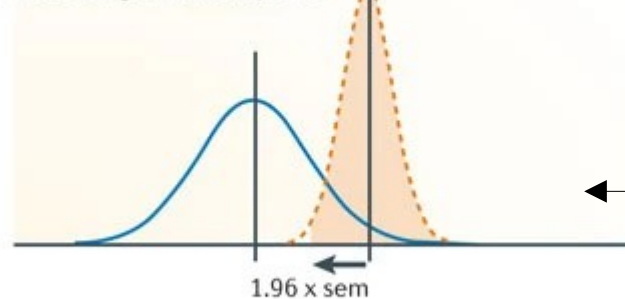
a Rejecting the null hypothesis



b The sampling distribution



c Increasing statistical power



Σε έναν στατιστικό έλεγχο βρίσκουμε $p=0.05$ και απορρίπτουμε την H_0

Επαναλαμβάνοντας τον έλεγχο με το ίδιο μέγεθος δείγματος, έχουμε πολύ μεγάλη πιθανότητα να κάνουμε σφάλμα τύπου II

Μεγαλώνοντας το δείγμα μειώνουμε την πιθανότητα να κάνουμε σφάλμα τύπου II, οπότε αυξάνουμε την ισχύ του ελέγχου

Έλεγχοι με χαμηλή στατιστική ισχύ

- Πώς προκύπτουν έλεγχοι με χαμηλή στατιστική ισχύ;
 - Μικρό δείγμα
 - **Αυξάνουμε το μέγεθος του δείγματος**
 - Αποτυχία αναγνώρισης πιθανών παραγόντων που προκαλούν ποικιλομορφία στις μετρήσεις
 - **Σχεδιάζουμε σωστά το πείραμα**
 - Λάθος στις μετρήσεις
 - **Βελτιώνουμε την ακρίβεια των μετρήσεων και επιλέγουμε τις σωστές τεχνικές, με πολλές επαναλήψεις**

ΠΡΟΣΟΧΗ

- Μπορούμε να μειώσουμε το σφάλμα τύπου II (και να αυξήσουμε τη στατιστική ισχύ ενός ελέγχου) αν αυξήσουμε το επίπεδο σημαντικότητας α
 - Κάνουμε τον έλεγχο πιο “χαλαρό” ως προς την πιθανότητα του σφάλματος τύπου I
 - **Δεν το θέλουμε αυτό!**
- Είναι προτιμότερο σε έναν στατιστικό έλεγχο να κάνουμε σφάλμα τύπου II από το να κάνουμε σφάλμα τύπου I
 - Μπορείτε να εξηγήσετε γιατί;

Πλάνες στατιστικών ελέγχων

Τι πρέπει να αποφεύγουμε!

Η πλάνη της δημοσίευσης

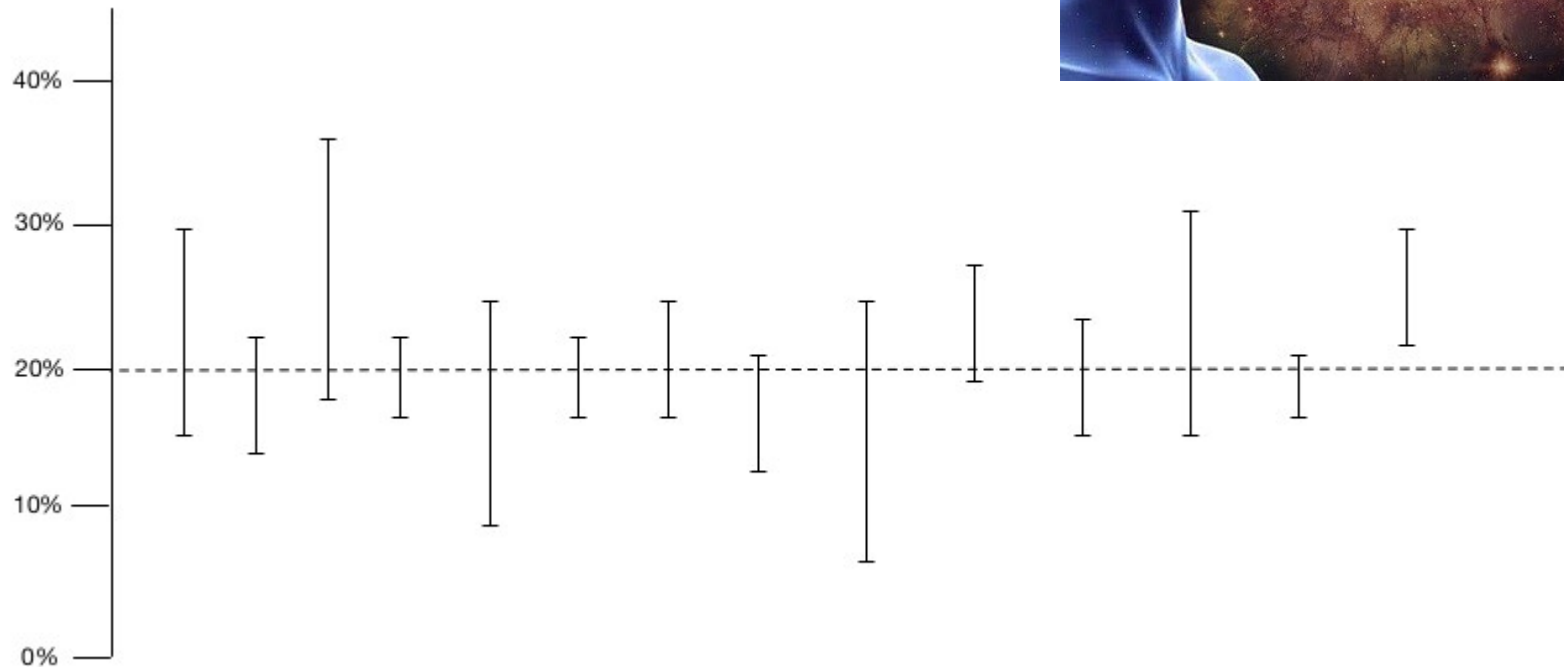
- Οι δημοσιευμένες εργασίες περιέχουν τα λεγόμενα **θετικά** αποτελέσματα, αυτά που απορρίπτουν τη μηδενική υπόθεση
 - Τα **αρνητικά** αποτελέσματα είναι βαρετά και μπορούν εύκολα να προκύψουν από ακατάλληλα πειράματα
- Όμως η μη δημοσίευση αρνητικών αποτελεσμάτων μειώνει την επίδραση των σφαλμάτων τύπου II
 - Και μοιραία αυξάνει την εμφάνιση σφαλμάτων τύπου I

Παράδειγμα: πείραμα τηλεπάθειας

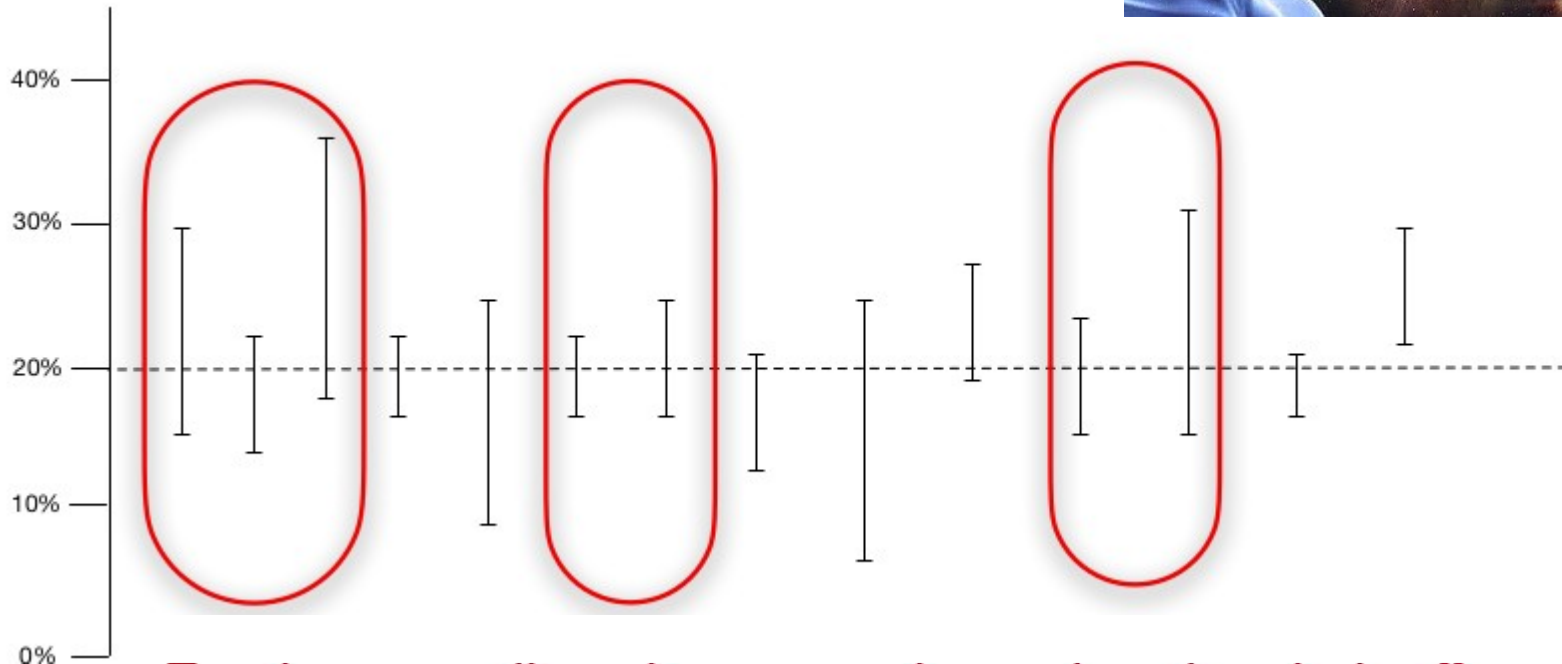
- Ένα πείραμα μελετά εάν δύο απομακρυσμένα άτομα (ένας αποστολέας και ένας παραλήπτης) μπορούν να επικοινωνήσουν πληροφορίες χωρίς καμία φυσική αλληλεπίδραση
- Ο αποστολέας επιλέγει μια κάρτα με έναν τυχαίο αριθμό από το 1 έως το 5 και προσπαθεί να την κοινοποιήσει στον παραλήπτη και ο παραλήπτης προσπαθεί να μαντέψει τον αριθμό που κοινοποίησε ο αποστολέας
 - H_0 : Ο δέκτης έχει 20% πιθανότητα να μαντέψει τον σωστό αριθμό
 - H_1 : Ο δέκτης μαντεύει τον σωστό αριθμό με ποσοστό μεγαλύτερο από την πιθανότητα (20%)

Διάφοροι παραψυχολόγοι διεξάγουν το ίδιο πείραμα πολλές φορές

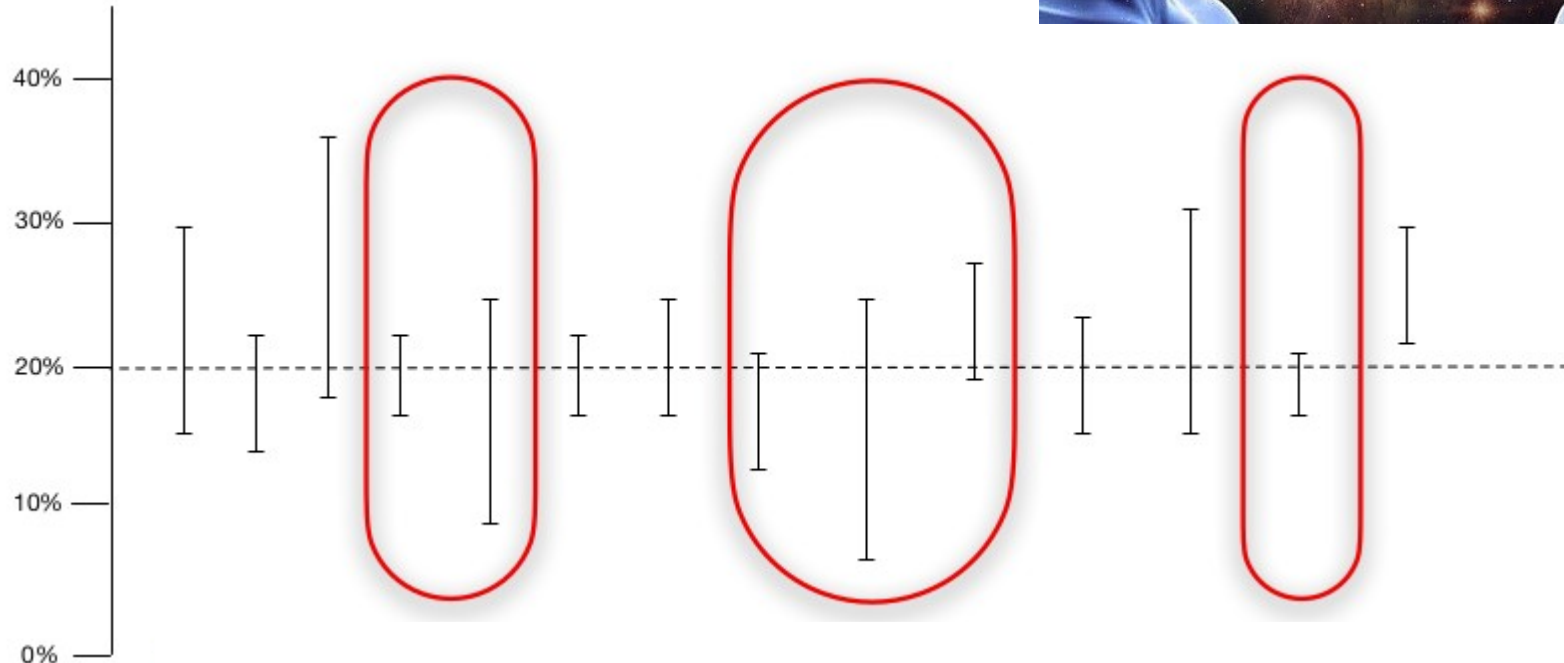
Το γράφημα δείχνει το 95% CI 14 διαφορετικών πειραμάτων



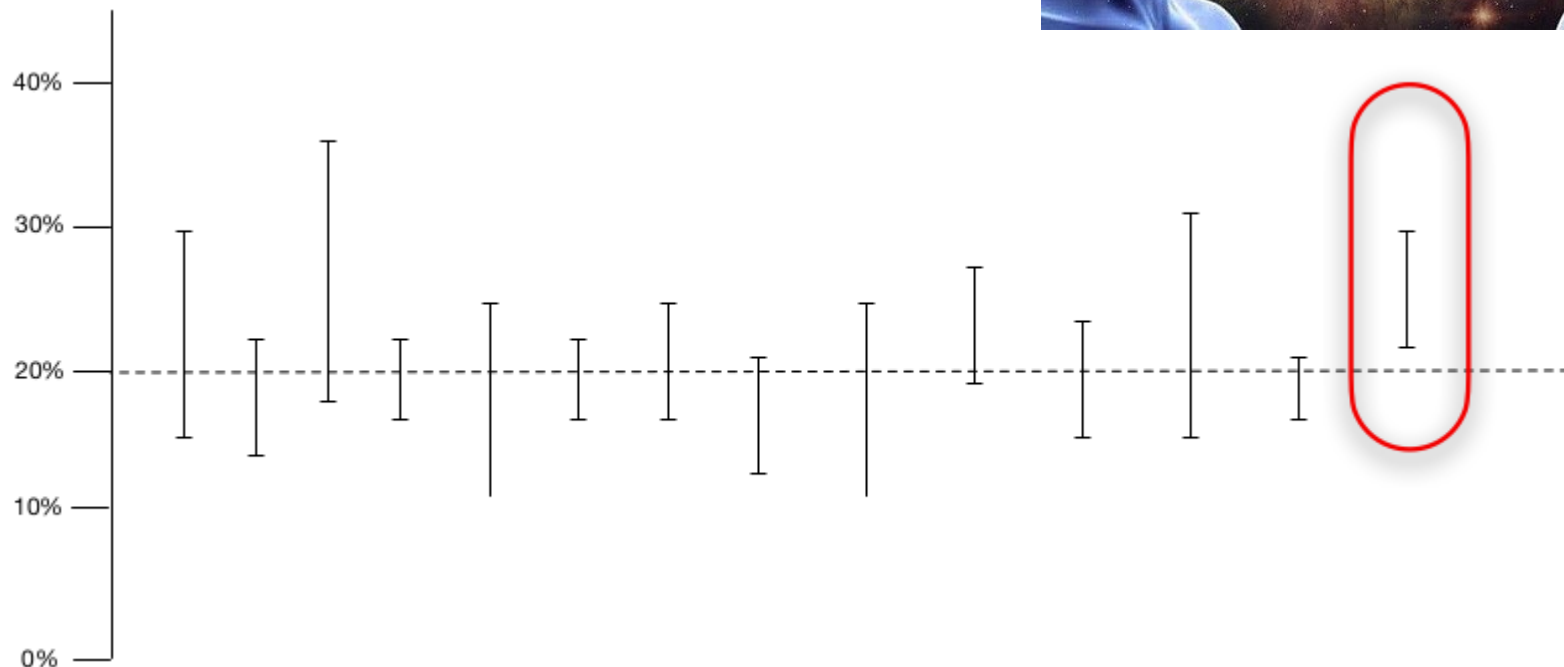
Κάποιοι ερευνητές απογοητεύονται και δεν δημοσιεύουν ποτέ τα αποτελέσματά τους



Κάποιοι ερευνητές προσπαθούν να δημοσιεύσουν τα αποτελέσματά αυτά αλλά απορρίπτονται από τα περιοδικά



Αυτοί οι ερευνητές γράφουν ιστορία: είναι οι πρώτοι που αποδεικνύουν πειραματικά την ύπαρξη της τηλεπάθειας. Η εργασία τους γίνεται δεκτή από κάποιο περιοδικό και σύντομα γίνονται viral στα social media...



Επανάληψη ελέγχων



The screenshot shows the top portion of a news article on the Nature website. The header includes the 'nature' logo and navigation links. The article title is 'Over half of psychology studies fail reproducibility test' by Monya Baker, dated 27 August 2015. A sub-headline reads 'Largest replication study to date casts doubt on many published positive results.' A 'Rights & Permissions' button is visible. The main text begins with 'Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.' and continues with 'In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.' A photo of Brian Nosek is partially visible on the right side of the article.

- Η επανάληψη των πειραμάτων και των στατιστικών ελέγχων είναι σημαντική πρακτική ενάντια στην πλάνη της δημοσίευσης
 - Δυστυχώς δεν γίνεται πάντα και σε όλα τα επιστημονικά πεδία

Η πλάνη της p-value

- Η εμφάνιση στατιστικής σημαντικότητας $p < 0.05$ έχει γίνει σκοπός σε πολλές ερευνητικές προσπάθειες

What is p-hacking?

P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	SIGNIFICANT
0.04	
0.049	OH CRAP. REDO CALCULATIONS.
0.050	
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

https://imgs.xkcd.com/comics/p_values.png



Αυτό παρακινεί πολλούς ερευνητές να παραποιούν τους ελέγχους μέχρι να βρουν την επιθυμητή σημαντικότητα

p-hacking

<https://www.youtube.com/watch?v=i60wwZDA1CI>

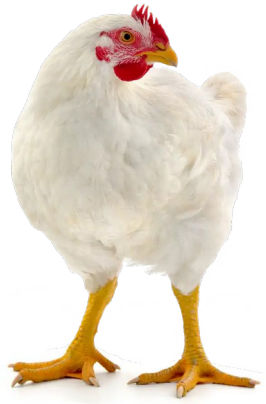
Μην το κάνετε

Συνηθισμένες τεχνικές p-hacking

- Σταδιακή αύξηση του μεγέθους του δείγματος μέχρι να επιτευχθεί στατιστική σημαντικότητα
- Επιλεκτική κατάργηση «ακραίων τιμών» που δεν συμμορφώνονται με τις προσδοκίες των ερευνητών
- Αλλαγή των στόχων της υπόθεσης με βάση τα αποτελέσματα.
 - Π.χ. η εστίαση σε τυχαία «στατιστικά σημαντικά» αποτελέσματα, ακόμη και αν αυτά τα αποτελέσματα δεν ήταν μέρος των αρχικών πειραματικών στόχων
- Εφαρμογή διαφορετικών ελέγχων σημαντικότητας και επιλογή μόνο αυτού που οδηγεί σε σημαντικότητα, συχνά παραβιάζοντας τις προϋποθέσεις

Οδηγίες για τις ασκήσεις αυτού του κεφαλαίου

- Οι ασκήσεις δεν είναι υποχρεωτικές, όμως δίνουν **bonus** για τον τελικό βαθμό του μαθήματος.
- Μπορείτε να κάνετε τις ασκήσεις σε ομάδες των 2 ή 3 ατόμων.
 - Όχι ατομικά και όχι πάνω από τρία άτομα στην ομάδα!
- Οι τρεις αυτές ασκήσεις θα μπουν σε ένα αρχείο **.pdf** που θα περιέχει τον κώδικα, τα γραφήματα και μια περιγραφή της διαδικασίας. Το αρχείο θα έχει τίτλο τα επώνυμα των μελών της ομάδας.
- Κάθε ομάδα θα ανεβάσει μία φορά το αρχείο με τις λυμένες ασκήσεις, από ένα μέλος της. Το αρχείο θα ανέβει στις “Εργασίες” του eclass του μαθήματος, στην **Εργασία Β**.
 - Δεν θα δεχτούμε ασκήσεις με email
- *Αυστηρή προθεσμία: Κυριακή 12 Μαΐου!*



Thank you

