# Κεφάλαιο 7

# Στατιστική Συμπερασματολογία

Στο κεφάλαιο αυτό γίνεται μια εισαγωγή σε μερικές απλές στατιστικές μεθόδους σε προβλήματα συμπερασματολογίας σε ένα και δύο δείγματα. Το πρώτο μέρος αναφέρεται στην Περιγραφική Στατιστική η οποία είναι μια συλλογή από γραφικές μεθόδους για εξερεύνηση και κατανόηση των δεδομένων. Στη συνέχεια θα γίνει αναφορά σε στατιστικούς ελέγχους που χρησιμοποιούνται κατά γενικό κανόνα, όπως τους ελέγχους t και Wilcoxon για να εξαχθούν στατιστικά συμπεράσματα για τις υπό διερεύνηση παραμέτρους. Προχωρώντας, παρουσιάζονται περαιτέρω έλεγχοι υποθέσεων όπως ο έλεγχος καλής προσαρμογής και ο έλεγχος για ποσοστά. Στο τέλος, εξετάζεται ο τρόπος χειρισμού πινάκων συνάφειας για εξαγωγή συμπεράσματος για τη συσχέτιση μεταξύ δυο μεταβλητών όταν ταξινομηθούν σε σχέση μιας τρίτης.

## 7.1 Περιγραφική Στατιστική

Η Περιγραφική Στατιστική χρησιμοποιεί γραφήματα τα οποία βοηθούν στο να εξερευνηθεί αν ισχύουν οι υποθέσεις των στατιστικών μοντέλων. Μερικές ερωτήσεις που ενδιαφέρουν σε αυτές τις περιπτώσεις είναι:

- Τα δεδομένα ακολουθούν την κανονική κατανομή;
- Υπάρχουν απομακρυσμένες τιμές στα δεδομένα;

- Αν τα δεδομένα συγκεντρώθηκαν διαδοχικά στον χρόνο (χρονοσειρές), υπάρχει ένδειξη σειριακή συσχέτισης;

Τα κύρια γραφήματα τα οποία βοηθούν στην εποπτική εξερεύνηση των δεδομένων είναι τα ακόλουθα:

- ιστογράμματα (histograms),
- κυτιογραφήματα (boxplots),
- γράφημα πυκνότητας (density plots),
- QQ plots (quantile-quantile plots).

Ακολουθεί μια απλή συνάρτηση στην R η οποία κατασκευάζει αυτά τα τέσσερα γραφήματα:

```
eda.shape <- function(x)
 {
   par(mfrow = c(2, 2))
   hist(x)
   boxplot(x)
   iqd <- summary(x)[5] - summary(x)[2]
   plot(density(x,width=2*iqd), xlab = "x",ylab = "", type = "l")
   qqnorm(x)
   qqline(x)
 }
```

Με την εντολή `hist()` κατασκευάζεται το ιστόγραμμα και με την εντολή `boxplot` το κυτιόγραμμα. Η εντολή `summary()` δίνει την περίληψη πέντε αριθμών και άρα το `iqd` στη συνάρτηση υπολογίζει το ενδοτεταρτημοριακό εύρος (interquartile range). Η εντολή `plot(density())` κατασκευάζει το γράφημα μη παραμετρικής εκτιμήτριας συνάρτησης πυκνότητας πιθανότητας με την μέθοδο των πυρήνων και οι εντολές `qqnorn` και `qqline` δημιουργούν το QQ plot μαζί με την γραμμή για απόκλιση από την κανονική κατανομή.

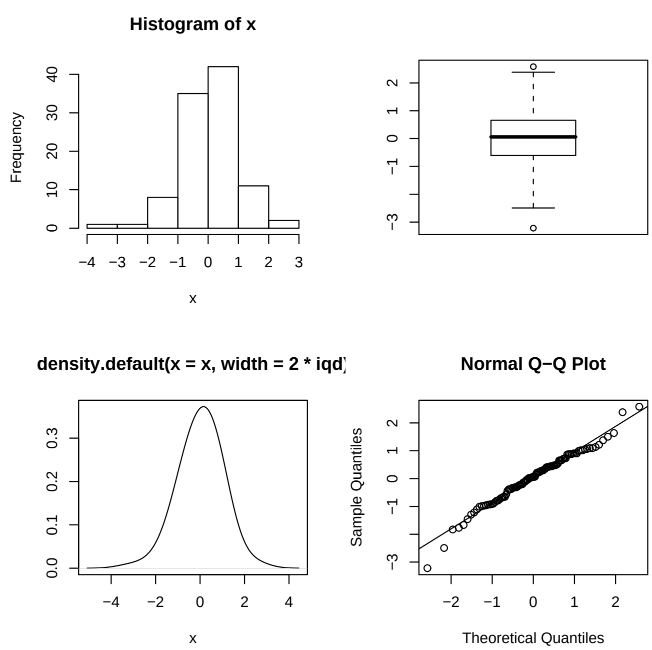
Για να εξερευνηθεί εποπτικά αν υπάρχει σειριακή συσχέτιση είναι χρήσιμο να γίνει στην αρχή η γραφική παράσταση χρονοσειρών των δεδομένων. Η ακόλουθη συνάρτηση χειρίζεται αυτήν την περίπτωση:

```
eda.ts <- function(x)
```

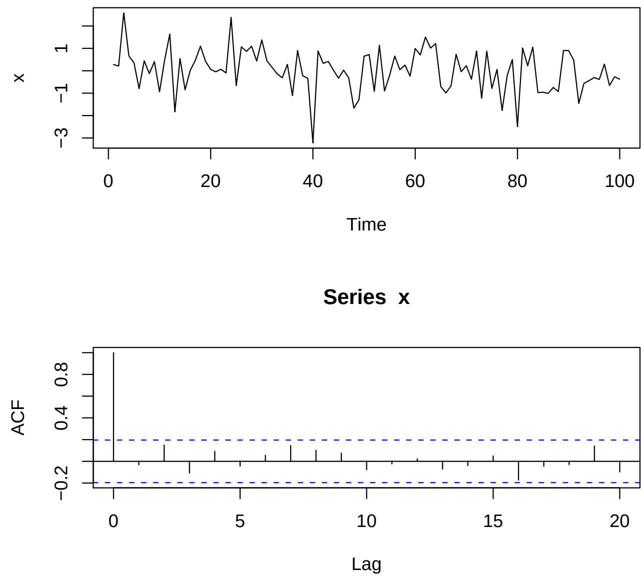
```
{
   par(mfrow=c(2,1))
   ts.plot(x)
   acf(x)
   invisible()
}
```

Η εντολή ts.plot παράγει γράφημα των διαδοχικών παρατηρήσεων ενώ η εντολή acf δίνει την δειγματική συνάρτηση αυτοσυσχέτισης. Οι δυο πιο πάνω συναρτήσεις εφαρμόζονται σε 100 τιμές από την κανονική και τα αποτελέσματα παρουσιάζονται στα σχήματα 7.1 και 7.2, αντίστοιχα.

```
x <- rnorm(100)
eda.shape(x)
eda.ts(x)
```




Σχήμα 7.1: Γραφήματα για τη μορφή των δεδομένων.

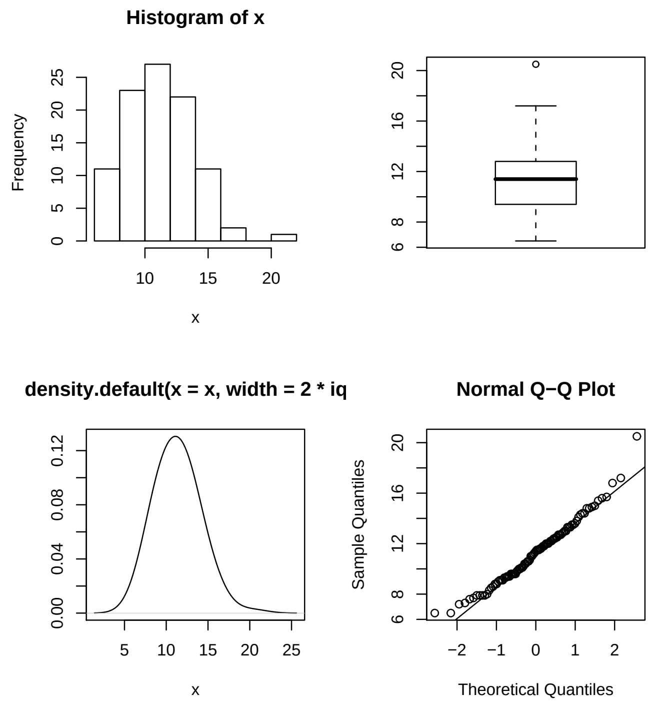



Σχήμα 7.2: Γραφήματα για σειριακή συσχέτιση.

## 7.2 Συμπερασματολογία για Ένα Δείγμα

Για το κομμάτι αυτό του κεφαλαίου θα χρησιμοποιηθεί το πλαίσιο δεδομένων `cats`, το οποίο αποτελείται από παρατηρήσεις του βάρους του σώματος και της καρδιάς 140 αρσενικών και θηλυκών γάτων οι οποίες χρησιμοποιήθηκαν για ένα πείραμα. Το συγκεκριμένο πλαίσιο βρίσκεται στη βιβλιοθήκη της R `MASS` (κατασκευάστηκε από τους Venables και Ripley)

```
> library(MASS)
> cats
> male   <-  cats[cats$Sex=="M",]$Hwt
> female <-  cats[cats$Sex=="F",]$Hwt
> eda.shape(male)
```




Σχήμα 7.3: Γραφήματα για τη μορφή των δεδομένων για το βάρος της καρδιάς των αρσενικών.

```
> t.test(male, mu=10)

        One Sample t-test

data:  male
t = 5.1241, df = 96, p-value = 1.544e-06
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
 10.81030 11.83506
sample estimates:
mean of x
 11.32268

> t.test(male, mu=10, conf.level=0.99)
```

```
        One Sample t-test

data:  male
t = 5.1241, df = 96, p-value = 1.544e-06
alternative hypothesis: true mean is not equal to 10
99 percent confidence interval:
 10.64431 12.00105
sample estimates:
mean of x
 11.32268
```

Οι εντολές `cats[cats$Sex=="M",]$Hwt` και `cats[cats$Sex=="F",]$Hwt` χρησιμοποιήθηκαν για να εξάγουν το βάρος της καρδιάς των αρσενικών και θηλυκών γάτων, αντίστοιχα. Από τα γραφήματα για τη μορφή των παρατηρήσεων για το βάρος της καρδιάς των αρσενικών γάτων συμπεραίνεται ότι είναι λογικό να γίνει η υπόθεση ότι ακολουθούν την κανονική κατανομή με μέση τιμή περίπου ίση με 11 (βλ. Σχήμα 7.3). Η συνάρτηση `t.test` μπορεί να χρησιμοποιηθεί για να γίνει ο έλεγχος t με μηδενική υπόθεση $\mu = \mu_0$ και στατιστική συνάρτηση

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

Το αποτέλεσμα δείχνει ότι για $\mu_0 = 10$, η μηδενική υπόθεση απορρίπτεται (το `p-value` είναι πολύ κοντά στο 0 και άρα μικρότερο από 0.05). Πρέπει να σημειωθεί ότι εξ ορισμού το αποτέλεσμα της συνάρτησης δίνει το 95% διαστήματα εμπιστοσύνης. Είναι δυνατόν όμως να δοθεί το διάστημα εμπιστοσύνης για οποιοδήποτε επίπεδο εμπιστοσύνης $(1 - \alpha)$% δίνοντας τιμή στο όρισμα `conf.level`. Για να γίνει ο απαραμετρικός έλεγχος Wilcoxon για τη μέση τιμή χρησιμοποιείται η συνάρτηση `wilcox.test` όπως πιο κάτω:

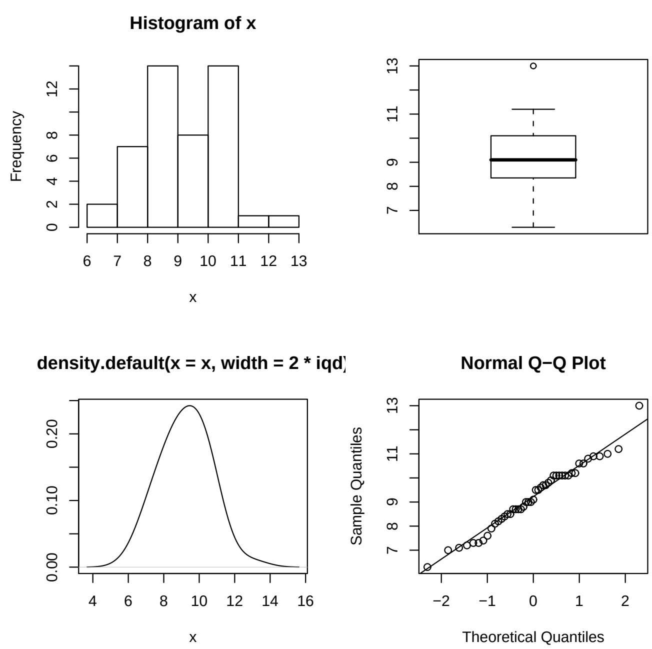
```
> wilcox.test(male, mu=10)

        Wilcoxon signed rank test with continuity correction

data:  male
V = 3491.5, p-value = 6.934e-06
alternative hypothesis: true location is not equal to 10
```

Η μηδενική υπόθεση απορρίπτεται και σε αυτήν την περίπτωση και συνεπώς η μέση τιμή $\mu$ διαφέρει στατιστικώς σημαντικά από την τιμή 10.

## 7.3 Συμερασματολογία για Δυο Δείγματα

Ας εξεταστεί τώρα ο έλεγχος της διαφοράς δύο μέσων τιμών βασισμένος σε πρόβλημα δύο δειγμάτων. Για παράδειγμα, έστω ότι είναι ανάγκη να εξεταστεί η διαφορά μεταξύ της μέσης τιμής του βάρους της καρδιάς των αρσενικών και των θηλυκών γάτων από το πλαίσιο δεδομένων που χρησιμοποιήθηκε πιο πάνω.

```
> eda.shape(female)
```




Σχήμα 7.4: Γραφήματα για τη μορφή των δεδομένων για το βάρος της καρδιάς των θηλυκών.

```
> var.test(male,female)

        F test to compare two variances

data:  male and female
F = 3.5064, num df = 96, denom df = 46, p-value = 8.159e-06
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
 2.075412 5.664645
sample estimates:
ratio of variances
           3.50642


> t.test(male,female,var.equal=T)

        Two Sample t-test

data:  male and female
t = 5.3539, df = 142, p-value = 3.38e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.337588 2.903517
sample estimates:
mean of x mean of y
11.322680  9.202128

> t.test(male,female,var.equal=F)

        Welch Two Sample t-test

data:  male and female
t = 6.5179, df = 140.608, p-value = 1.186e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.477352 2.763753
sample estimates:
mean of x mean of y
11.322680  9.202128

> wilcox.test(male,female)

        Wilcoxon rank sum test with continuity correction
```

```
data:  male and female
W = 3460.5, p-value = 4.882e-07
alternative hypothesis: true location shift is not equal to 0
```

Από το γράφημα στο Σχήμα 7.3 συμπεραίνεται ότι τα δεδομένα για το βάρος της καρδιάς των θηλυκών δεν εφαρμόζουν στην κανονική κατανομή. Ο πρώτος έλεγχος που γίνεται στο πιο πάνω ελέγχει την ισότητα των διακυμάνσεων των δύο δειγμάτων (μηδενική υπόθεση) με την στατιστική συνάρτηση

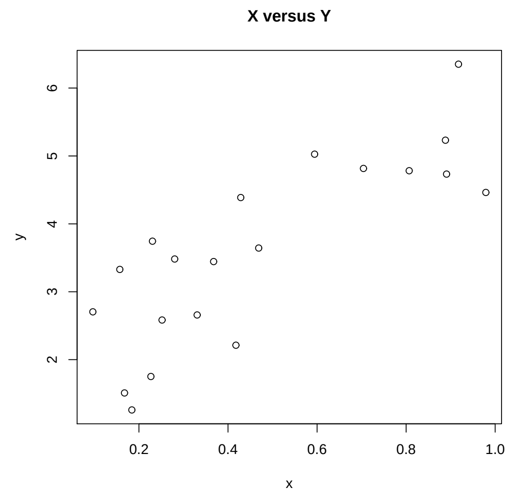
$$F = \frac{S_x^2}{S_y^2}$$

και δείχνει ότι η ισότητα απορρίπτεται. Συνεπώς, ο συνηθισμένος έλεγχος t για τη διαφορά των μέσων τιμών δεν μπορεί να εφαρμοστεί σε αυτήν την περίπτωση. Ωστόσο, στη συνάρτηση `t.test` μπορεί να προστεθεί το όρισμα `var.equal` στο οποίο δηλώνεται αν ισχύει η ισότητα των διακυμάνσεων με τις τιμές ¨TRUE¨ και ¨FALSE¨. Με αυτόν το τρόπο είναι δυνατό να γίνει ο έλεγχος t και στο πιο πάνω παράδειγμα συμπεραίνεται ότι η υπόθεση ισότητας των μέσων τιμών των δυο δειγμάτων απορρίπτεται. Το ίδιο συμπέρασμα βγαίνει και από τον απαραμετρικό έλεγχο Wilcoxon.

## 7.4 Συμπερασματολογία για Δείγματα Κατά Ζεύγη

Πολλές φορές τα δεδομένα παρουσιάζονται σε ζεύγη της μορφής $(X_i, Y_i)$. Για να γίνει ο έλεγχος t σε αυτήν την περίπτωση, είναι χρήσιμο να οριστεί μια καινούργια μεταβλητή $Z$ με τιμές $Z_i = X_i - Y_i$ και μετά να εργαστούμε όπως πιο πάνω. Για να μελετηθεί η συσχέτιση μεταξύ των δυο μεταβλητών είναι πάντα χρήσιμο να γίνει το γράφημα διασποράς τους για να εξεταστεί γραφικά η σχέση τους. Στη συνέχεια γίνεται ο έλεγχος συσχέτισης με την εντολή `cor.test()` για να ελεγχθεί αν ο συντελεστής συσχέτισης είναι μεγαλύτερος, μικρότερος ή διαφορετικός από 0. Ακολουθεί ένα απλό παράδειγμα.

```
> x <- runif(20)
> y <- 2+3*x+rnorm(20)
> plot(x,y,main="X versus Y")


> cor.test(x,y)
```




Σχήμα 7.5: Διάγραμμα διασποράς του $y$ συναρτήσει του $x$.

```
        Pearson's product-moment correlation

data:  x and y
t = 5.9907, df = 18, p-value = 1.149e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5847064 0.9246687
sample estimates:
      cor
0.8160727

> cor.test(x,y, alt="l")

        Pearson's product-moment correlation

data:  x and y
t = 5.9907, df = 18, p-value = 1
alternative hypothesis: true correlation is less than 0
95 percent confidence interval:
 -1.0000000  0.9127702
```

```
sample estimates:
      cor
0.8160727
```

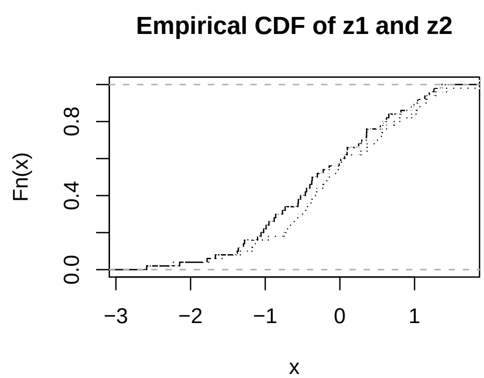
Όπως φαίνεται από το Σχήμα 7.5, υπάρχει μεγάλη υψηλή συσχέτιση μεταξύ του $x$ και του $y$ και άρα αναμένεται ο συντελεστής συσχέτισης να είναι κοντά στο 1. Για τον λόγο αυτό, ο πρώτος έλεγχος, με εναλλακτική υπόθεση η συσχέτιση να είναι διαφορετική από 0, δεν απορρίπτεται ενώ ο δεύτερος έλεγχος, με εναλλακτική υπόθεση η συσχέτιση να είναι αρνητική, απορρίπτεται.

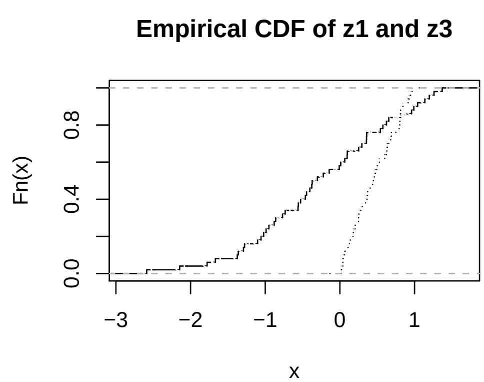
## 7.5 Έλεγχος Καλής Προσαρμογής

Ο έλεγχος καλής προσαρμογής (goodness of fit test) θεωρείται ως ακόμη ένας τρόπος για να εξακριβωθούν οι υποθέσεις της κατανομής των δεδομένων. Η R αξιολογεί την ¨καλή προσαρμογή¨ των δεδομένων με τον έλεγχο Kolmogorov-Smirnov. Ο έλεγχος Kolmogorov-Smirnov χρησιμοποιείται στην περίπτωση σύγκρισης ενός δείγματος με μια κατανομή, αλλά και στην περίπτωση σύγκρισης της κατανομής δυο δειγμάτων μεταξύ τους. Στην πρώτη περίπτωση, η μηδενική υπόθεση $H_0$ δηλώνει ότι τα δεδομένα ακολουθούν την συγκεκριμένη κατανομή, ενώ στη δεύτερη δηλώνει ότι τα δυο δείγματα ακολουθούν την ίδια κατανομή. Η σύγκριση των κατανομών δυο δειγμάτων μπορεί να γίνει και γραφικά με τη βοήθεια της γραφικής παράστασης της εμπειρικής κατανομής τους. Πιο κάτω παρουσιάζεται ένα παράδειγμα στο οποίο συγκρίνεται γραφικά η κατανομή δύο δειγμάτων από την κανονική και ένα από την ομοιόμορφη κατανομή (Σχήμα 7.6). Στο πάνω γράφημα συγκρίνονται τα δύο δείγματα από την ίδια κατανομή (κανονική), ενώ στο δεύτερο δυο δείγματα από διαφορετική κατανομή (κανονική και ομοιόμορφη) και η διαφορά τους είναι εμφανής.

```
> z1<-rnorm(50)
> z2<-rnorm(50)
> z3<-runif(50)
> par(mfrow=c(2,1))
> plot.ecdf(z1,verticals=TRUE, col.p="white",col.v="black",
+ main="Empirical CDF of z1 and z2")
> plot(ecdf(z2),add=TRUE,verticals=TRUE,col.p="white",col.v="black",
+ lty="dotted")
> plot.ecdf(z1,verticals=TRUE, col.p="white",col.v="black",
```

```
+ main="Empirical CDF of z1 and z3")
> plot(ecdf(z3),add=TRUE,verticals=TRUE,col.p="white",col.v="black",
+ lty="dotted")
```




Σχήμα 7.6: Σύγκριση γραφήματος εμπειρικής συνάρτησης κατανομής δύο δειγμάτων.

Ακολουθούν τα αποτελέσματα από τον έλεγχο Kolmogorov–Smirnov σε ένα και δύο δείγματα για το πιο πάνω παράδειγμα. Ο πρώτος έλεγχος συγκρίνει ένα δείγμα με την κανονική κατανομή, ο δεύτερος δυο όμοια δείγματα μεταξύ τους και ο τρίτος δυο διαφορετικά δείγματα μεταξύ τους. Στον πρώτο και δεύτερο έλεγχο η μηδενική υπόθεση δεν απορρίπτεται ενώ στον τρίτο απορρίπτεται, όπως αναμένεται.

```
> ks.test(z1,"pnorm")
```

```
        One-sample Kolmogorov-Smirnov test

data:  z1
D = 0.0973, p-value = 0.6946
alternative hypothesis: two-sided

> ks.test(z1,z2)

        Two-sample Kolmogorov-Smirnov test

data:  z1 and z2
D = 0.16, p-value = 0.5487
alternative hypothesis: two-sided

> ks.test(z1,z3)

        Two-sample Kolmogorov-Smirnov test

data:  z1 and z3
D = 0.46, p-value = 3.801e-05
alternative hypothesis: two-sided
```

Μια παραλλαγή του ελέγχου Kolmogorov–Smirnov είναι ο έλεγχος Anderson-Darling ο οποίος βρίσκεται στη βιβλιοθήκη `nortest` της R.

## 7.6 Έλεγχος Υποθέσεων για Ποσοστά

Η R παρέχει την εντολή `binom.test` για έλεγχο υποθέσεων για ποσοστά. Για παράδειγμα, έστω το πείραμα ρίψεως ενός νομίσματος 500 φορές με αποτέλεσμα εμφάνισης *κεφαλή* 226 φορές. Θα ελεγχθεί η υπόθεση αν η πιθανότητα να έρθει *κεφαλή* είναι ίση με $p = 0.5$.

```
> binom.test(226,500,p=0.5)

        Exact binomial test
```

```
data:  226 and 500
number of successes = 226, number of trials = 500, p-value = 0.03546
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4077723 0.4967984
sample estimates:
probability of success
                 0.452
```

Όπως φαίνεται, ο έλεγχος απορρίπτεται σε επίπεδο σημαντικότητας $\alpha = 0.05$, αλλά δεν απορρίπτεται για $\alpha = 0.01$. Επίσης, το διάστημα στο διάστημα εμπιστοσύνης δεν περιέχει την τιμή 0.5. Συνεπώς, η πιθανότητα να έρθει *κεφαλή* είναι στατιστικά μικρότερη από 0.5. Παρόμοια αποτελέσματα εξάγονται και για $p = 0.4$, συμπεραίνοντας ότι το $p$ είναι μεγαλύτερο.

```
> binom.test(226,500,p=0.4)

        Exact binomial test

data:  226 and 500
number of successes = 226, number of trials = 500, p-value = 0.01983
alternative hypothesis: true probability of success is not equal to 0.4
95 percent confidence interval:
 0.4077723 0.4967984
sample estimates:
probability of success
                 0.452
```

Με την εντολή prop.test γίνεται ο έλεγχος για ποσοστό $p = 0.5$ λαμβάνοντας υπόψη τη διόρθωση συνέχειας. Στο επόμενο παράδειγμα υποθέστε ότι εμφανίστηκε *κεφαλή* 266 φορές. Η μηδενική υπόθεση δεν απορρίπτεται με αποτέλεσμα η πιθανότητα να έρθει *κεφαλή* δεν διαφέρει στατιστικώς σημαντικά από 0.5.

```
> prop.test(266,500)

        1-sample proportions test with continuity correction

data:  266 out of 500, null probability 0.5
```

```
X-squared = 1.922, df = 1, p-value = 0.1656
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4871883 0.5763127
sample estimates:
    p
0.532
```

Στη συνέχεια παρουσιάζεται πως μπορεί να γίνει ο προσημικός έλεγχος για ένα ή δυο δείγματα με τη βοήθεια της διαμέσου. Για ένα δείγμα ελέγχεται η υπόθεση αν η διάμεσος παίρνει μια συγκεκριμένη τιμή, ενώ για δυο δείγματα αν η διάμεσος της διαφοράς τους είναι διαφορετική από 0.

```
> x<-rnorm(100)
> y<-sum(x>0)
> binom.test(y,100)

        Exact binomial test

data:  y and 100
number of successes = 54, number of trials = 100, p-value = 0.4841
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4374116 0.6401566
sample estimates:
probability of success
                  0.54

> z<-rnorm(100)
> d<-x-z
> binom.test(sum(d>0),length(d))

        Exact binomial test

data:  sum(d > 0) and length(d)
number of successes = 47, number of trials = 100, p-value = 0.6173
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
```

```
 0.3694052 0.5724185
sample estimates:
probability of success
                  0.47
```

Έστω, τώρα, ο έλεγχος σύγκρισης δυο κερμάτων και έστω ότι σε 200 ρίψεις του πρώτου κέρματος εμφανίστηκε *κεφαλή* 80 φορές, ενώ σε 150 ρίψεις του δεύτερου κέρματος εμφανίστηκε 100 φορές *κεφαλή*. Κάνοντας το έλεγχο στην R όπως πιο κάτω, συμπεραίνεται ότι υπάρχει στατιστική διαφορά μεταξύ των δυο κερμάτων.

```
> x<-c(80,100)
> n<-c(200,150)
> prop.test(x,n)

        2-sample test for equality of proportions with continuity correction

data:  x out of n
X-squared = 23.345, df = 1, p-value = 1.354e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.3739929 -0.1593405
sample estimates:
   prop 1    prop 2
0.4000000 0.6666667
```

## 7.7 Πίνακες Συνάφειας

Για τα πιο κάτω θα χρησιμοποιηθεί το πλαίσιο δεδομένων `solder` το οποίο περιέχεται στη βιβλιοθήκη `faraway`. Οι μεταβλητές προς εξέταση είναι η `Solder` (παράγοντας με 5 επίπεδα) και η `Mask` (με 2 επίπεδα). Για να κατασκευαστεί ο $2x5$ πίνακας συνάφειας (contigency table) χρησιμοποιείται η εντολή `table`. Έπειτα, με την εντολή `summary` δίνεται ο Pearson $X^2$ έλεγχος ανεξαρτησίας. Επιπρόσθετα, η σταυρωτή ταξινόμηση (cross classification) είναι δυνατή με την εντολή `xtabs`. Το όρισμα στ' αριστερά δίνει την μεταβλητή η οποία θα ταξινομηθεί, ενώ τα ορίσματα στα δεξιά δίνουν τις αντίστοιχες κατηγορίες. Με την εντολή `CrossTable`, η οποία βρίσκεται στη βιβλιοθήκη `gmodels` παρουσιάζεται η ταξινόμηση σε πίνακα μαζί με την συνεισφορά στον $X^2$ έλεγχο κάθε συνδυασμού. Η εντολή `summary` χρησιμοποιείται και σ' αυτήν την περίπτωση για τον $X^2$ έλεγχο ανεξαρτησίας αφού γίνει η ταξινόμηση.

```
> library("gmodels")
> library("faraway")
> attach(solder)
> names(solder)
[1] "Opening" "Solder"  "Mask"    "PadType" "Panel"   "skips"
> X<-table(Solder,Mask)
> X
       Mask
Solder  A1.5  A3  A6  B3  B6
  Thick   90 150  30  90  90
  Thin    90 120  60  90  90
> summary(X)
Number of cases in table: 900
Number of factors: 2
Test for independence of all factors:
        Chisq = 13.333, df = 4, p-value = 0.009757
> cross<-xtabs(.~Solder+Mask)
Error in eval(expr, envir, enclos) : object "." not found
> cross<-xtabs(~Solder+Mask)
> cross
       Mask
Solder  A1.5  A3  A6  B3  B6
  Thick   90 150  30  90  90
  Thin    90 120  60  90  90
```

```
> CrossTable(cross)


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  900


             | Mask
      Solder |      A1.5 |        A3 |        A6 |        B3 |        B6 | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
       Thick |        90 |       150 |        30 |        90 |        90 |       450 |
             |     0.000 |     1.667 |     5.000 |     0.000 |     0.000 |           |
             |     0.200 |     0.333 |     0.067 |     0.200 |     0.200 |     0.500 |
             |     0.500 |     0.556 |     0.333 |     0.500 |     0.500 |           |
             |     0.100 |     0.167 |     0.033 |     0.100 |     0.100 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
        Thin |        90 |       120 |        60 |        90 |        90 |       450 |
             |     0.000 |     1.667 |     5.000 |     0.000 |     0.000 |           |
             |     0.200 |     0.267 |     0.133 |     0.200 |     0.200 |     0.500 |
             |     0.500 |     0.444 |     0.667 |     0.500 |     0.500 |           |
             |     0.100 |     0.133 |     0.067 |     0.100 |     0.100 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
Column Total |       180 |       270 |        90 |       180 |       180 |       900 |
             |     0.200 |     0.300 |     0.100 |     0.200 |     0.200 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
```

## 7.8 Παράδειγμα

Τα δεδομένα αυτά αποτελούνται από το δείκτη νοημοσύνης (IQ) παιδιών ηλικίας 5 χρονών, τα οποία τα ξεχωρίζουμε βάση του κριτηρίου ότι οι μητέρες τους έχουν υποφέρει από επεισόδιο επιλόχειας κατάθλιψης. Επικεντρωνόμαστε στο να απαντήσουμε το ερώτημα αν τα παιδιά από τις δύο κατηγορίες έχουν διαφορετικό δείκτη νοημοσύνης. Τα δεδομένα βρίσκονται στο παράρτημα αυτού του κεφαλαίου.

Συνήθως, χρησιμοποιείται ο έλεγχος $t$ για τον έλεγχο με μηδενική υπόθεση ότι οι δύο κατηγορίες έχουν ίσες πληθυσμιακές μέσες τιμές και εναλλακτική ότι δεν έχουν. Ο έλεγχος υποθέτει ότι οι παρατηρήσεις :

1. είναι ανεξάρτητες μεταξύ τους,
2. προέρχονται από πληθυσμό από την κανονική κατανομή,
3. προέρχονται από πληθυσμούς με ίσες διασπορές.

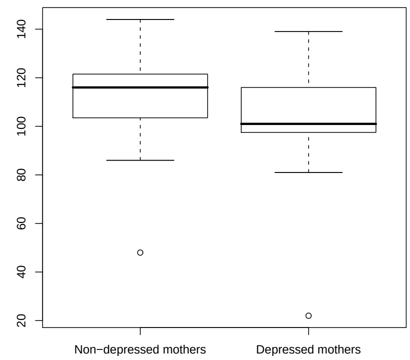
Εκτός από το p–value, το οποίο λαμβάνεται από τον έλεγχο, συνήθως είναι χρήσιμο ένα διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών.

Πριν την εφαρμογή του ελέγχου υποθέσεων, πρέπει να εξεταστεί αν τα δεδομένα ικανοποιούν τις υποθέσεις στις οποίες βασίζεται ο έλεγχος. Η αρχική εξέταση των δεδομένων γίνεται εποπτικά με τη βοήθεια γραφημάτων, όπως τα ιστογράμματα, τα κυτιογραφήματα και τα QQ-γραφήματα. Με τα γραφήματα αυτά μπορεί να αναγνωριστεί η απόκλιση από την κανονική κατανομή, η παρουσία απομακρυσμένων τιμών κ.τ.λ. Τα γραφήματα αυτά και για τις δύο κατηγορίες παρουσιάζονται στα Σχήματα 7.7-7.9.

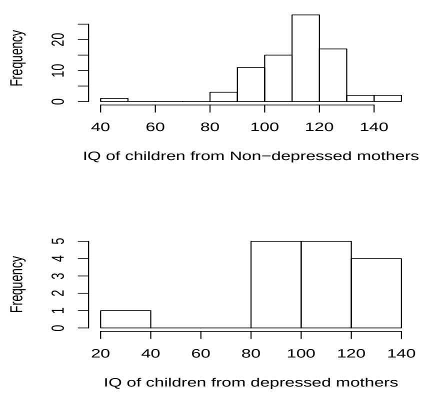
Για να κατασκευαστούν τα γραφήματα και η ανάλυση πρέπει πρώτα να διαχωριστούν οι δύο κατηγορίες των δεδομένων, `scorend` για τα παιδιά με μη καταθλιπτικές μητέρες και `scored` για τα παιδιά με καταθλιπτικές μητέρες.

```
> iqdata<-read.table("iqdata.txt")
> attach(iqdata)
> scorend <- iqdata$V2[V1=="nd"]
> scored <- iqdata$V2[V1=="d"]
> boxplot(scorend, scored, names=c("Non-depressed mothers",
+ "Depressed mothers"))
> par(mfrow=c(2,1))
> hist(scorend, xlab="IQ of children from Non-depressed mothers")
> hist(scored, xlab="IQ of children from depressed mothers")
> qqnorm(scorend,main="")
> qqline(scorend)
> title(main="Normal Probability plot for IQ of children
+ from Non-depressed mothers")
> qqnorm(scored,main="")
> qqline(scored)
```

```
> title(main="Normal Probability plot for IQ of children
+ from depressed mothers")
```

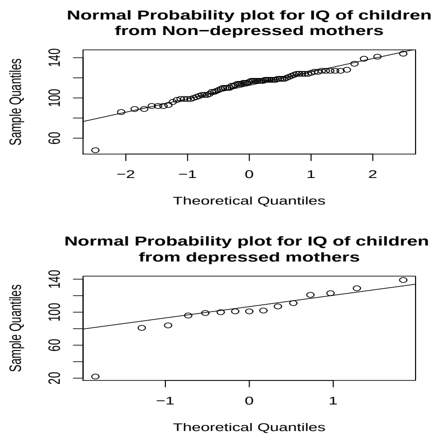



Σχήμα 7.7: Κυτιογραφήματα των δεικτών νοημοσύνης των παιδιών.




Σχήμα 7.8: Ιστογράμματα των δεικτών νοημοσύνης των παιδιών.

Το πιο σημαντικό στοιχείο από όλα τα γραφήματα είναι ότι και στις δύο κατηγορίες παρουσιάζεται από μια απομακρυσμένη παρατήρηση για το δείκτη νοημοσύνης, η οποία και στις δυο περιπτώσεις ανήκει σε παιδί με πολύ χαμηλό δείκτη νοημοσύνης. Τέτοιες παρατηρήσεις μπορούν να επηρεάσουν σημαντικά περιγραφικά μέτρα όπως τη μέση τιμή και τη διακύμανση, αλλά μπορούν να επηρεάσουν




Σχήμα 7.9: QQ-γράφημα των δεικτών νοημοσύνης των παιδιών.

επίσης τους στατιστικούς ελέγχους υποθέσεων οι οποίοι βασίζονται στην κανονικότητα. Προς το παρών, έστω ότι δε λαμβάνεται κανένα διορθωτικό μέτρο για αυτές τις παρατηρήσεις, και εφαρμόζεται ο έλεγχος $t$ για τον οποίο θεωρείται ότι τα δεδομένα των δυο κατηγοριών έχουν την ίδια διακύμανση. Τα αποτελέσματα δίνονται πιο κάτω:

```
> t.test(scorend,scored,var.equal=T)

        Two Sample t-test

data:  scorend and scored
t = 2.4637, df = 92, p-value = 0.01561
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.27152 21.16477
sample estimates:
mean of x mean of y
 112.7848  101.0667
```

Η διαφορά στη μέση τιμή του δείκτη νοημοσύνης των δύο κατηγοριών είναι στατιστικώς σημαντική, γεγονός που φαίνεται από τη μικρή τιμή για το p–value, αλλά και από το ότι το μηδέν δεν ανήκει στο διάστημα εμπιστοσύνης. Ωστόσο, η παρουσία των απομακρυσμένων τιμών μπορεί να επηρεάζει αυτό το αποτέλεσμα. Η διακύμανση σε κάθε κατηγορία, για παράδειγμα, είναι 205.50 (παιδιά από μη καταθλιπτικές μητέρες) και 729.21 (παιδιά από καταθλιπτικές μητέρες). Συνεπώς, υπάρχει μια σημαντική διαφορά στις διακυμάνσεις των δύο κατηγοριών, με αποτέλεσμα να παραβιάζεται μία από τις υποθέσεις του ελέγχου πιο πάνω. Αυτό μπορεί να αποδειχθεί και στατιστικά με τον έλεγχο $F$ για την ισότητα των διακυμάνσεων.

```
> var.test(scorend,scored)

        F test to compare two variances

data:  scorend and scored
F = 0.2818, num df = 78, denom df = 14, p-value = 0.0003276
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1089903 0.5746361
sample estimates:
ratio of variances
          0.281818
```

Η υπόθεση ισότητας των διακυμάνσεων απορρίπτεται ξεκάθαρα, αφού η σχετική p–value είναι πολύ μικρή. Τι συμβαίνει όμως όταν ο έλεγχος επαναληφθεί αφού αφαιρεθούν οι απομακρυσμένες τιμές από τα δεδομένα; Εφαρμόζεται ξανά ο έλεγχος ισότητας των διακυμάνσεων των δύο κατηγοριών αφού αφαιρεθούν οι τιμές των δύο παιδιών με χαμηλό δείκτη νοημοσύνης, δηλαδή δείκτη νοημοσύνης μικρότερο ή ίσο του 50.

```
> var.test(scorend[scorend>50],scored[scored>50])

        F test to compare two variances

data:  scorend[scorend > 50] and scored[scored > 50]
F = 0.5664, num df = 77, denom df = 13, p-value = 0.1283
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2102836 1.1765891
sample estimates:
ratio of variances
```

```
       0.5664062
```

Ο έλεγχος τώρα δίνει ότι ο λόγος των διακυμάνσεων είναι ίσος με 0.57, και επισημαίνει ότι η διαφορά ανάμεσα στις διακυμάνσεις δεν είναι στατιστικά σημαντική αφού το p–value είναι μεγαλύτερο από 0.05 και το 1 ανήκει στο σχετικό 95% διάστημα εμπιστοσύνης. Άρα, τα δεδομένα με την αφαίρεση των απομακρυσμένων τιμών έγιναν καταλληλότερα για την εφαρμογή του έλεγχου $t$. Επίσης, βάση της μελέτης των δεδομένων αυτών είναι πιο λογικό να αφαιρεθούν αφού η μία παρατήρηση ανήκει σε αυτιστικό παιδί ενώ η άλλη σε παιδί το οποίο έπαθε ζημιά στον εγκέφαλο κατά τον τοκετό. Επαναλαμβάνεται, λοιπόν, ο έλεγχος $t$ αφού έχουν αφαιρεθεί οι απομακρυσμένες τιμές :

```
> t.test(scorend[scorend>50],scored[scored>50],var.equal=T)

        Two Sample t-test

data:  scorend[scorend > 50] and scored[scored > 50]
t = 1.8242, df = 90, p-value = 0.07145
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6148208 14.4170186
sample estimates:
mean of x mean of y
 113.6154  106.7143
```

Είναι φανερό ότι αφού αφαιρεθούν οι δύο απομακρυσμένες τιμές, δεν υπάρχει πλέον η ένδειξη της διαφοράς των μέσων τιμών των δεικτών νοημοσύνης των παιδιών από μη καταθλιπτικές και καταθλιπτικές μητέρες.

## Παράρτημα

Τα δεδομένα που χρησιμοποιούνται σε αυτό το κεφάλαιο για απλή στατιστική συμπερασματολογία.

```
> iqdata
   V1  V2
1  nd 103
2  nd 124
```

```
3   nd 124
4   nd 104
5    d  96
6   nd  92
7   nd 124
8   nd  99
9   nd  92
10  nd 116
11  nd  99
12   d  22
13   d  81
14  nd 117
15   d 100
16  nd  89
17  nd 125
18  nd 127
19  nd 112
20  nd  48
21  nd 139
22  nd 118
23   d 107
24  nd 106
25   d 129
26  nd 117
27  nd 123
28  nd 118
29   d  84
30  nd 117
31   d 101
32  nd 141
33  nd 124
34  nd 110
35  nd  98
36  nd 109
37  nd 120
38  nd 127
39  nd 103
```

```
40 nd 118
41 nd 117
42 nd 115
43 nd 119
44 nd 117
45 nd  92
46 nd 101
47 nd 119
48 nd 144
49 nd 119
50 nd 127
51 nd 113
52 nd 127
53 nd 103
54 nd 128
55 nd  86
56 nd 112
57 nd 115
58 nd 117
59 nd  99
60 nd 110
61  d 139
62 nd 117
63 nd  96
64  d 111
65 nd 118
66 nd 126
67 nd 126
68 nd  89
69 nd 102
70 nd 134
71 nd  93
72 nd 115
73  d  99
74 nd  99
75 nd 122
76 nd 106
```

```
77 nd 124
78 nd 100
79 nd 114
80 nd 121
81 nd 119
82 nd 108
83 nd 110
84 nd 127
85 nd 118
86 nd 107
87  d 123
88  d 102
89 nd 110
90 nd 114
91 nd 118
92  d 101
93  d 121
94 nd 114
```