ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Επαγωγή και Έλεγχος Υποθέσεων

Πέτρος Κολοβός

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ

Ελεγχος διάφορων υποθέσεων

- Ένα πρώτο βήμα πριν περάσει κανείς στον έλεγχο διάφορων υποθέσεων είναι να μελετήσει τις ιδιότητες των κατανομών που θα αναλύσει και θα συγκρίνει.
- Μεγάλη σημασία έχει έδω να γνωρίζουμε αν οι κατανομές που μελετούμε είναι κανονικές ή όχι καθώς μια σειρά από στατιστικούς ελέγχους θέτει την κανονικότητα ως προϋπόθεση για την εφαρμογή τους. Ένα πρώτο βήμα για την ανάλυση είναι έτσι συχνά ο έλεγχος της κανονικότητας.

Ελεγχος διάφορων υποθέσεων

Στη συνέχεια θα δούμε ένα παράδειγμα από την κάθε κατηγορία στα πλαίσια της ανάλυσης ενός συνόλου δεδομένων που προέρχεται από μια κλινική μελέτη. Στη συγκεκριμένη μελέτη 84 ασθενείς με Ρευματοειδή Αρθρίτιδα διαφορετικού φύλου και ηλικίας δέχτηκαν θεραπεία με ένα κανονικό φάρμακο ή με εικονική θεραπεία (placebo) και ο βαθμός βελτίωσης της υγείας τους αξιολογήθηκε με βάση μια κλίμακα κλινικών συμπτωμάτων. Για να έχουμε μια εικόνα των δεδομένων θα τα φορτώσουμε αρχικά στην R με την χρήση του πακέτου $vcd()$

Hide

```
library(vcd)
```

```
## Loading required package: grid
```

Hide

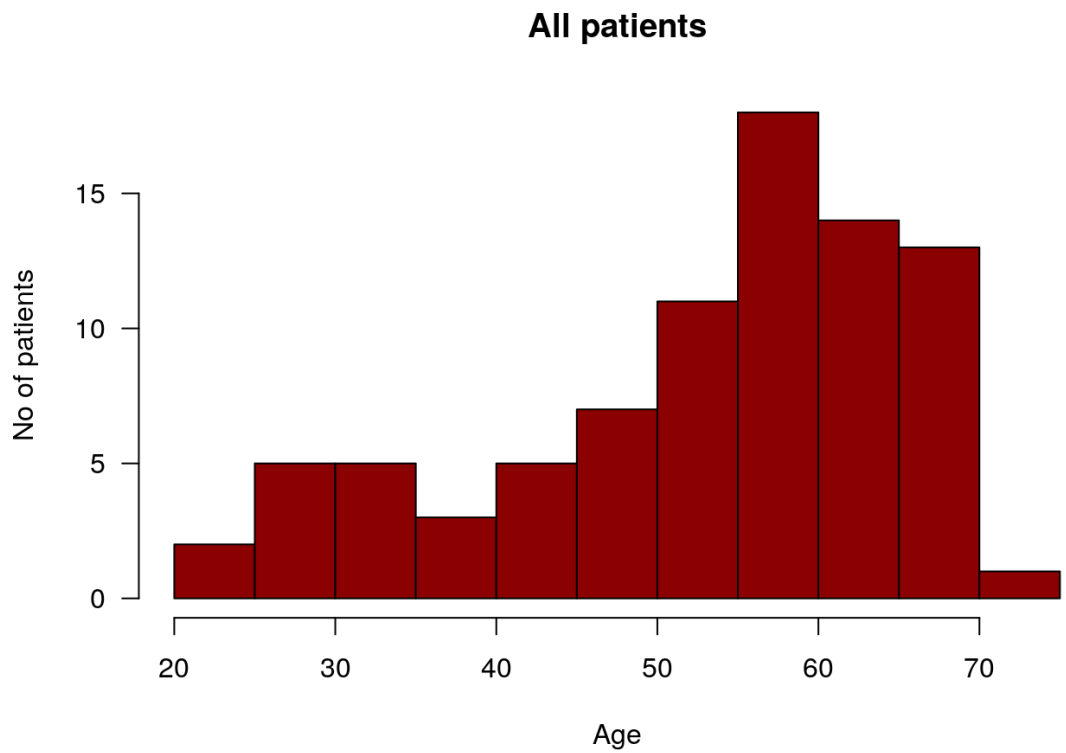
```
str(Arthritis)
```

```
## '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 ...
```

Ελεγχος διάφορων υποθέσεων

Hide

```
hist(Arthritis$Age, col="dark red", breaks=10, xlab="Age", ylab="No of patients", main="All patient
s", las=1)
```



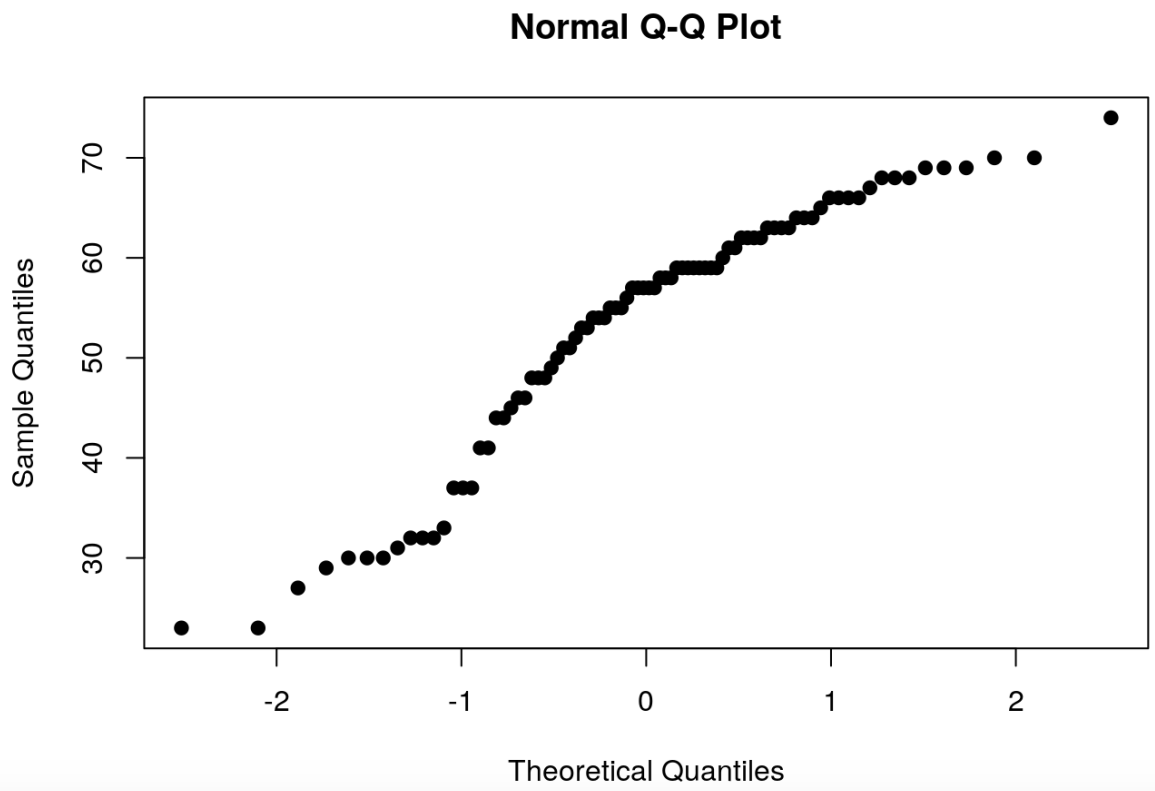
Από την εικόνα προκύπτει ότι η κατανομή δεν έχει συμμετρική τάση και απέχει από το αναμενόμενο σχήμα της κωδωνοειδούς καμπύλης.

Γραφικός έλεγχος κανονικότητας

Ένας καλύτερος τρόπος για να συγκρίνουμε την κατανομή μας με μια κανονική κατανομή με την ίδια μέση τιμή και τυπική απόκλιση είναι το διάγραμμα σκέδασης ποσοστημορίων (η πιο απλά quantile-quantile plot, και για συντομία qqplot). Σε αυτό συγκρίνονται οι τιμές ποσοστημορίων των δύο κατανομών δηλ. της πραγματικής και μιας κανονικής με ίδιες παραμέτρους. Στην R αυτό γίνεται με την απλή συνάρτηση *qqnorm*():

Hide

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



Γραφικός έλεγχος κανονικότητας

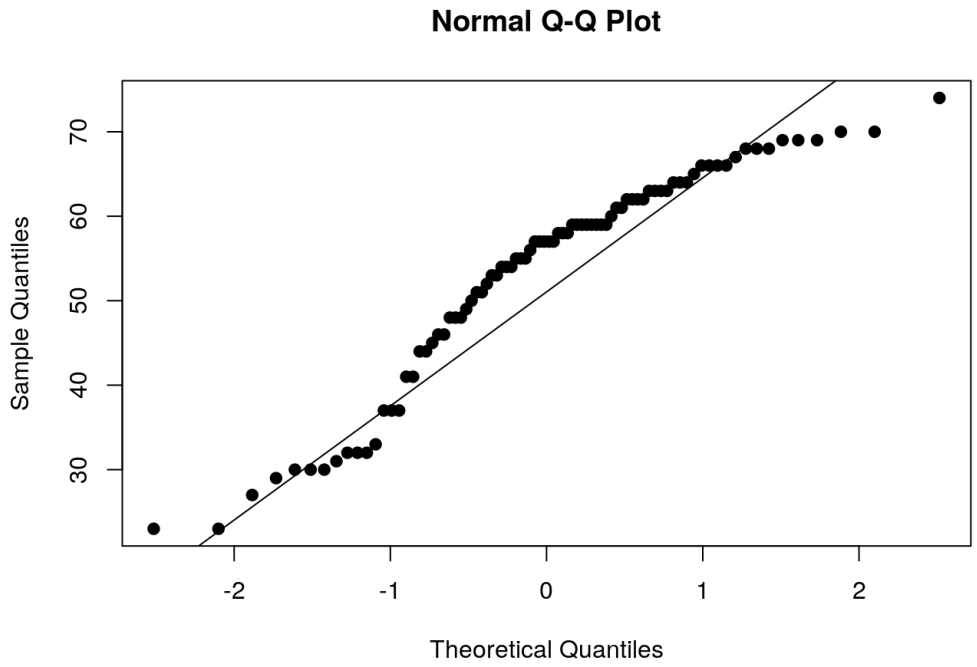
Ας δουμε για λίγο πώς θα ερμηνεύσουμε αυτό το διάγραμμα. Εφόσον ελέγχουμε για κανονικότητα, περιμένουμε ότι στην περίπτωση που τα δεδομένα μας (στον κάθετο άξονα) συμπεριφέρονται κανονικά θα τείνουν να ταυτιστούν με την θεωρητική κατανομή (στον οριζόντια άξονα) και κατά συνέπεια η κανονικότητα θα αντανακλάται με μια διαγώνια γραμμή (με γωνία κλίσης θ=45). Στην πράξη έχουμε πάντα αποκλίσεις στις χαμηλότερες και υψηλότερες τιμές αλλά ένα κεντρικό μέρος θα πρέπει να είναι διαγώνιο. Η *qqnorm*() δεν μας δίνει την δυνατότητα να έχουμε ένα μέτρο αναφοράς, μπορούμε όμως να το δημιουργήσουμε μόνοι μας με την χρήση της *rnorm*() όπως είδαμε σε προηγούμενο κεφάλαιο. Έτσι ένα σύνολο δεδομένων που ακολουθεί την κανονική κατανομή και έχει τον ίδιο αριθμό τιμών, την ίδια μέση τιμή και τυπική απόκλιση με την κατανομή των ηλικιών δίνεται από την παρακάτω εντολή:

Hide

```
nAge<-rnorm(length(Arthritis$Age), mean=mean(Arthritis$Age), sd=sd(Arthritis$Age))
```

το διάνυσμα αυτό μπορούμε τώρα να αναπαραστήσουμε γραφικά μαζί με την κανονική κατανομή με την χρήση της *qqline*():

```
qqnorm(Arthritis$Age, pch=19);qqline(nAge)
```



Γραφικός έλεγχος κανονικότητας

Hide

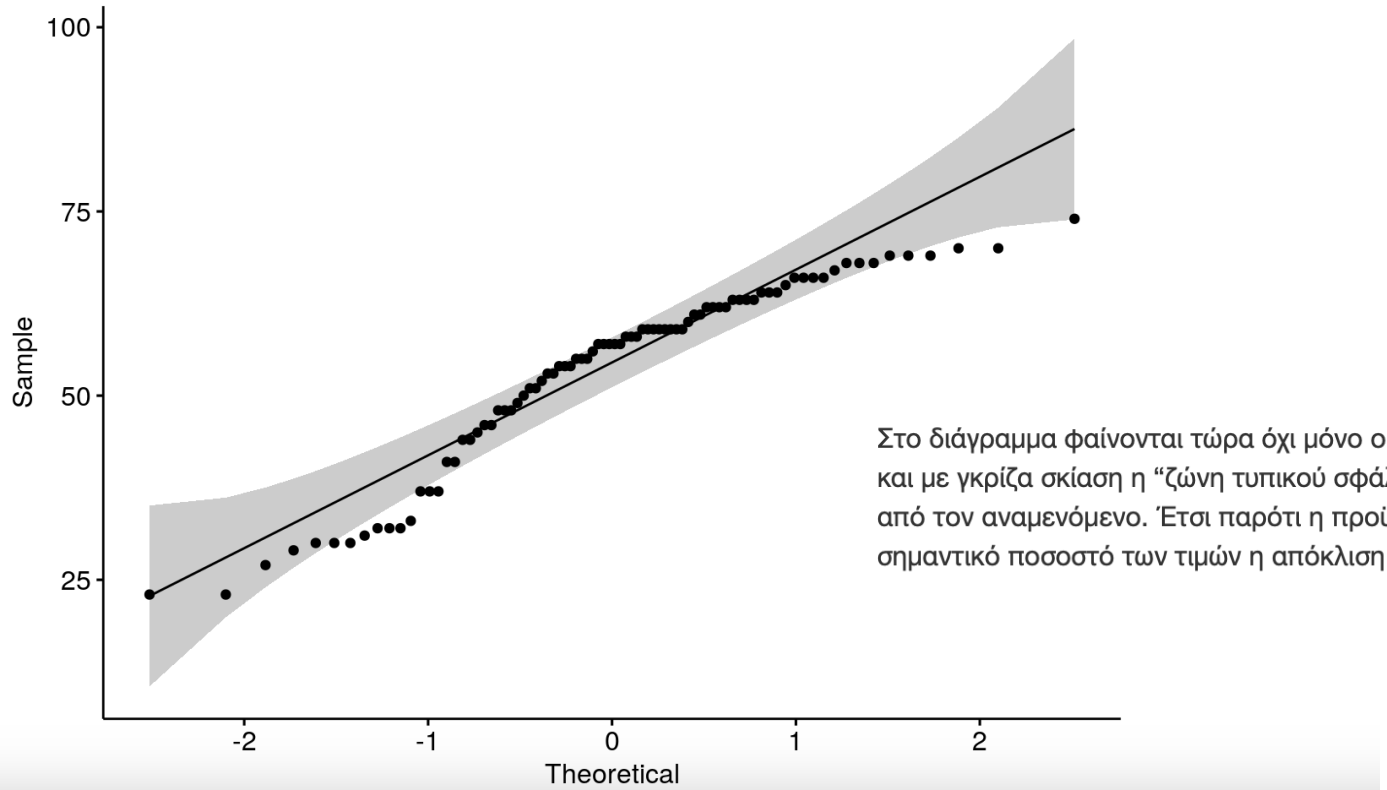
```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: magrittr
```

Hide

```
ggqqplot(Arthritis$Age);
```



Στο διάγραμμα φαίνονται τώρα όχι μόνο οι πραγματικές τιμές, αλλά και η ευθεία που αντιστοιχεί στη θεωρητική κατανομή καθώς και με γκρίζα σκίαση η "ζώνη τυπικού σφάλματος" (error band), η οποία μας δείχνει το βαθμό στον οποίο οι τιμές μας αποκλίνουν από τον αναμενόμενο. Έτσι παρότι η προϋπόθεση της κανονικότητας δεν ικανοποιείται, από το διάγραμμα βλέπουμε ότι για ένα σημαντικό ποσοστό των τιμών η απόκλιση είναι εντός ορίων.

Αριθμητικοί έλεγχοι κανονικότητας

Για έναν τόσο σημαντικό έλεγχο υπάρχει ένας μεγάλος αριθμός διαθέσιμων μεθόδων με κυριότερους τους ελέγχους Kolmogorov-Smirnov, Shapiro-Wilk, d' Agostino και Aderson-Darling. Στη συνέχεια μπορούμε να δούμε πώς εφαρμόζονται μέσω αντίστοιχων συναρτήσεων:

- Shapiro-Wilk. Ο εν λόγω έλεγχος είναι ο πιο ευρέως χρησιμοποιούμενος λόγω της ευελιξίας του και της δυνατότητάς του να εφαρμόζεται σε δεδομένα με επαναλαμβανόμενες τιμές. Η εφαρμογή του γίνεται με την συνάρτηση *shapiro.test*():

Hide

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

```
##
##  Shapiro-Wilk normality test
##
## data:  Arthritis$Age
## W = 0.91913, p-value = 5.813e-05
```


Αριθμητικοί έλεγχοι κανονικότητας

- Kolmogorov-Smirnov. Αφήνουμε τελευταίο αυτόν τον σχετικά κοινό έλεγχο λόγω της μικρότερης ευελιξίας του. Στην πραγματικότητα ο συγκεκριμένος έλεγχος δεν ελέγχει την κανονικότητα ενός δείγματος αλλά ελέγχει κατά πόσο δύο δείγματα μπορεί να έχουν προέλθει από την ίδια κατανομή. Ένα από τα προβλήματα που έχει είναι ότι δεν λειτουργεί επαρκώς ικανοποιητικά αν τα δεδομένα περιέχουν επαναλαμβανόμενες τιμές. Στο συγκεκριμένο παράδειγμα η χρήση του ελέγχου μπορεί να γίνει πάνω στις τιμές που δημιουργήσαμε με την χρήση της *rnorm*():

Hide

```
ks.test(Arthritis$Age, nAge)
```

```
## Warning in ks.test(Arthritis$Age, nAge): cannot compute exact p-value with
## ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  Arthritis$Age and nAge
## D = 0.2381, p-value = 0.0171
## alternative hypothesis: two-sided
```

Όπως βλέπουμε στο συγκεκριμένο παράδειγμα, ο έλεγχος Kolmogorov-Smirnov δίνει μια πολύ μεγαλύτερη τιμή p-value σε σχέση με τα προηγούμενα τεστ. Η διαφορά αυτή αναδεικνύει την πιο περιορισμένη ισχύ του συγκεκριμένου ελέγχου. Ένας σημαντικός παράγοντας εδώ είναι το μέγεθος του δείγματος με το οποίο συγκρίνουμε. Πράγματι, αν αντί για ένα δείγμα 84 τιμών είχαμε προσομοιώσει ένα δείγμα 10000 τιμών, αυξάνοντας έτσι την ισχύ του ελέγχου, η τιμή p αλλάζει σημαντικά:

Αριθμητικοί έλεγχοι κανονικότητας

Όπως βλέπουμε στο συγκεκριμένο παράδειγμα, ο έλεγχος Kolmogorov-Smirnov δίνει μια πολύ μεγαλύτερη τιμή p-value σε σχέση με τα προηγούμενα τεστ. Η διαφορά αυτή αναδεικνύει την πιο περιορισμένη ισχύ του συγκεκριμένου ελέγχου. Ένας σημαντικός παράγοντας εδω είναι το μέγεθος του δείγματος με το οποίο συγκρίνουμε. Πράγματι, αν αντί για ένα δείγμα 84 τιμών είχαμε προσομοιώσει ένα δείγμα 10000 τιμών, αυξάνοντας έτσι την ισχύ του ελέγχου, η τιμή p αλλάζει σημαντικά:

Hide

```
nnAge<-rnorm(10000, mean=mean(Arthritis$Age), sd=sd(Arthritis$Age))
ks.test(Arthritis$Age, nnAge)
```

```
## Warning in ks.test(Arthritis$Age, nnAge): p-value will be approximate in
## the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  Arthritis$Age and nnAge
## D = 0.14791, p-value = 0.05224
## alternative hypothesis: two-sided
```

Βλέπουμε δηλαδή ότι για σύγκριση με μια αρκετά μεγαλύτερη συλλογή (δείγμα) τιμών ενδέχεται η κανονικότητας της κατανομής μας να μην μπορεί να απορριφθεί.

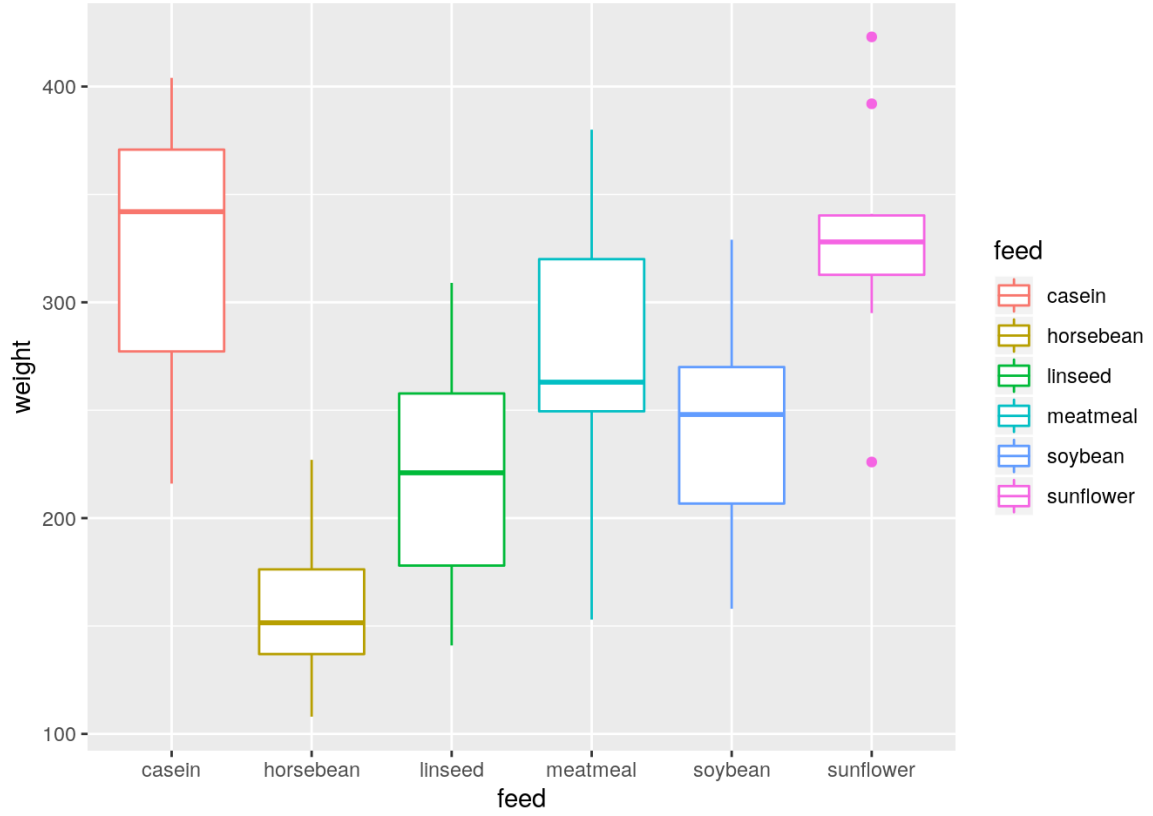
Συνοψίζοντας μπορούμε να πούμε ότι σε ό,τι αφορά τους ελέγχους κανονικότητας καλό είναι να συνδυάζονται τόσο γραφικές αναπαραστάσεις, όσο και αριθμητικοί έλεγχοι. Στη συγκεκριμένη περίπτωση παρότι μια σειρά από τεστ δεν χαρακτηρίζουν τα δεδομένα κανονικά κατανεμημένα, υπάρχει σοβαρή πιθανότητα αυτό να είναι απόρροια του περιορισμένου μεγέθους του δείγματος (finite size effect).

Σύγκριση μέσων τιμών

Το ερώτημά μας είναι αν κάποιες από τις τροφές είναι προτιμότερες από κάποιες άλλες και κατά πόσο θα πρέπει να τις προτιμήσουμε. Ουσιαστικά θέλουμε να απαντήσουμε στο ερώτημα αν η μέση τιμή των βαρών των ζώων που έχουν τραφεί με μια συγκεκριμένη τροφή είναι μεγαλύτερη από αυτήν των ζώων που έχουν τραφεί με κάποια άλλη. Μια πρώτη προσέγγιση είναι (όπως συνήθως) να αναπαραστήσουμε γραφικά τα δεδομένα. Εδώ παρουσιάζουμε έναν γρήγορο τρόπο να χρησιμοποιήσουμε την δομή του dataframe για να δημιουργήσουμε θηκογράμματα με την χρήση της *qplot*() από το πακέτο *ggplot*2:

Hide

```
library(ggplot2)
qplot(feed, weight, data=chickwts, geom="boxplot", col=feed)
```



Σύγκριση μέσων τιμών - Σε κανονικά κατανεμημένα δείγματα

Ο συγκεκριμένος έλεγχος πραγματοποίεται σε δύο κανονικά κατανεμημένα δείγματα με βάση την μηδενική υπόθεση ότι η διαφορά των μέσων τιμών τους είναι ίση με το 0 (πρακτικά δηλαδή υποθέτει ότι οι μέσες τιμές είναι ίσες). Η σύγκριση γίνεται πάντα σε δύο δείγματα, έτσι από τις παραπάνω κατηγορίες μπορούμε να την εφαρμόσουμε μόνο κατά ζεύγη. Από το θηκόγραμμα φαίνεται ότι ο λιναρόσπορος (linseed) έχει ένα πλεονέκτημα σε σχέση με τα κουκιά (horsebean). Είναι ωστόσο στατιστικά σημαντική η διαφορά τους; Για να ελέγξουμε αυτήν την υπόθεση με το t-test θα πρέπει αρχικά να εξαγάγουμε τα αντίστοιχα υποσύνολα από το σύνολο των δεδομένων, να ελέγξουμε την κανονικότητά τους και στη συνέχεια να εφαρμόσουμε το t-test με την συνάρτηση $t.test()$.

Παρακάτω βλέπουμε τις εντολές με τις οποίες πραγματοποιούμε αυτήν την ανάλυση.

Hide

```
which(chickwts$feed=="linseed")->linseed
which(chickwts$feed=="horsebean")->horsebean
shapiro.test(chickwts$weight[linseed])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chickwts$weight[linseed]
## W = 0.96931, p-value = 0.9035
```

Hide

```
shapiro.test(chickwts$weight[horsebean])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chickwts$weight[horsebean]
## W = 0.93758, p-value = 0.5264
```

Σύγκριση μέσων τιμών - Σε κανονικά κατανεμημένα δείγματα

Βλέπουμε ότι και τα δύο υποσύνολα είναι κανονικά με βάση το Shapiro-Wilk test συνεπώς μπορούμε να εφαρμόσουμε το $t.test()$ ως εξής:

Hide

```
t.test(chickwts$weight[linseed],chickwts$weight[horsebean])
```

```
##
##  Welch Two Sample t-test
##
## data:  chickwts$weight[linseed] and chickwts$weight[horsebean]
## 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
```

Ας μείνουμε λίγο στο αποτελέσμα του ελέγχου. Οι δύο πρώτες γραμμές του output μας πληροφορούν ότι πραγματοποιήθηκε ένας έλεγχος t-test σε δύο δείγματα και μας αναφέρει τα ονόματα των δειγμάτων. Στην επόμενη γραμμή έχουμε την τιμή του στατιστικού t, τους βαθμούς ελευθερίας (df) και την τιμή p που είναι αρκετά μικρή ώστε να απορρίψουμε την μηδενική υπόθεση για την εναλλακτική (alternative) που δίνεται στην αμέσως επόμενη γραμμή (και μας ενημερώνει ότι σύμφωνα με αυτήν η πραγματική διαφορά των μέσων τιμών δεν είναι ίση με το 0).

Στις επόμενες γραμμές υπάρχει μια σημαντική πτυχή του ελέγχου, το διάστημα εμπιστοσύνης 95%. Αυτό αποτελεί μια εκτίμηση του εύρους των τιμών διαφορών μέσης τιμής. Στην περίπτωση που εξετάζουμε είναι μεταξύ 18 και 99, κάτι που σημαίνει ότι αν κάναμε 100 αντίστοιχες, ανεξάρτητες δειγματοληψίες από τους ίδιους πληθυσμούς, στις 95 από αυτές η διαφορά των δύο μέσων θα ήταν μέσα σε αυτό το εύρος. Για ένα οποιοδήποτε t-test το διάστημα εμπιστοσύνης (confidence interval ή CI) είναι εξίσου σημαντικό, αν όχι σημαντικότερο από το p-value. Εφόσον δεν περιέχει την τιμή 0 και ανάλογα με την αυστηρότητά του (95%, 99%, 99.9% κλπ) μας επιτρέπει να απορρίψουμε την μηδενική υπόθεση με μεγαλύτερη ασφάλεια.

Σύγκριση μέσων τιμών - Σε κανονικά κατανεμημένα δείγματα

Ας δούμε πώς μπορούμε να υπολογίσουμε ένα πιο αυστηρό διάστημα εμπιστοσύνης, τροποποιώντας την αντίστοιχη παράμετρο:

Hide

```
t.test(chickwts$weight[linseed],chickwts$weight[horsebean], conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  chickwts$weight[linseed] and chickwts$weight[horsebean]
## 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
```

Όπως βλέπετε, δεν έχει άλλάξει τίποτα άλλο εκτός από το διάστημα εμπιστοσύνης το οποίο για ακρίβεια 99% είναι τώρα ευρύτερο (3-113) αλλά και πάλι δεν περιέχει το 0. Συνολικά λοιπόν, λαμβάνοντας υπ' όψιν το p-value και το διάστημα εμπιστοσύνης μπορούμε να πούμε ότι ο λιναρόσπορος είναι πιο αποδοτική ως τροφή. Μια σωστή διατύπωση του συμπεράσματος περιέχει όλα τα δεδομένα του στατιστικού ελέγχου και θα ήταν η εξής:

Ο λιναρόσπορος αποδίδει ζώα με μεγαλύτερο μέσο βάρος σε σχέση με τα κουκιά (t.test p-value=0.0069, df=19.8, 99%CI: (3.3, 113.8))

Σύγκριση μέσων τιμών - Σε μη κανονικά κατανεμημένα δείγματα

Όπως είδαμε παραπάνω ο έλεγχος με το t-test προϋποθέτει ότι έχουμε να κάνουμε με κανονικά κατανεμημένα δείγματα. Πώς όμως προχωράμε στη σύγκριση δειγμάτων που δεν είναι κανονικά κατανεμημένα; Στην περίπτωση αυτή θα καταφύγουμε σε κάποιο μη-παραμετρικό έλεγχο όπως αυτός που πραγματοποιείται με την μέθοδο Mann-Whitney ή Wilcoxon Rank-Sum (που είναι πρακτικά ισοδύναμες). Η Wilcoxon Rank Sum ουσιαστικά ελέγχει την μηδενική υπόθεση μια οποιαδήποτε τιμή από το ένα δείγμα να βρίσκεται ψηλότερα σε κατάταξη (εντός του δείγματος) σε σχέση με μια οποιαδήποτε άλλη από το άλλο δείγμα, είναι δηλαδή μέθοδος που βασίζεται στην κατάταξη των τιμών εντός του δείγματος και για το λόγο αυτό ανεξάρτητη από τον τύπο της κατανομής. Είναι αρκετά ισχυρή τόσο σε κανονικά όσο και σε μη κανονικά κατανεμημένα δείγματα και η εφαρμογή της είναι εξαιρετικά απλή.

Ας επιστρέψουμε στο σετ δεδομένων Arthritis όπου όπως είχαμε δει δεν είχαμε κανονικά κατανεμημένες τιμές ηλικίας. Έστω ότι θέλουμε να δούμε αν το προφίλ των ασθενών που αποκρίνονται καλύτερα στη θεραπεία έχει μια συγκεκριμένη ηλικιακή τάση. Θα χωρίσουμε τους ασθενείς με βάση το αποτέλεσμα της θεραπείας και θα προχωρήσουμε στην εφαρμογή του μη-παραμετρικού ελέγχου όπως φαίνεται παρακάτω:

Hide

```
which(Arthritis$Improved=="None")->n
which(Arthritis$Improved=="Marked")->m
shapiro.test(Arthritis$Age[n])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Arthritis$Age[n]
## W = 0.94019, p-value = 0.029
```

Hide

```
shapiro.test(Arthritis$Age[m])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Arthritis$Age[m]
## W = 0.91064, p-value = 0.02049
```

Σύγκριση μέσων τιμών - Σε μη κανονικά κατανεμημένα δείγματα

```
wilcox.test(Arthritis$Age[n], Arthritis$Age[m])
```

```
## Warning in wilcox.test.default(Arthritis$Age[n], Arthritis$Age[m]): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Arthritis$Age[n] and Arthritis$Age[m]
## W = 394.5, p-value = 0.02056
## alternative hypothesis: true location shift is not equal to 0
```

Η επισήμανση για τον μη υπολογισμό ακριβούς p-value έχει να κάνει με την ύπαρξη επαναλαμβανόμενων τιμών (ties) στα δείγματα αλλά ουσιαστικά δεν μας επηρεάζει καθώς η εκτίμηση είναι αρκετά ακριβής και όπως προκύπτει, υπάρχει μόνο ~2% πιθανότητα η τιμή W να έχει προκύψει αν οι μέσες τιμές των δύο δειγμάτων ήταν ίδιες, υπάρχει συνεπώς στατιστικά σημαντική διαφορά στις ηλικίες μεταξύ των ασθενών που έδειξαν βελτίωση και αυτών που έμειναν στάσιμοι. (Σημ. Το συμπέρασμα αυτό δεν είναι 100% ακριβές καθώς όπως θα δούμε στη συνέχεια έχουμε αγνοήσει κατά την ανάλυση ένα σημαντικό χαρακτηριστικό του πειράματος που αποτελεί όπως λέμε “συγχυτικό” παράγοντα (confounding factor). Θα επανέλθουμε σε αυτό στη συνέχεια).

Ένα σημαντικό χαρακτηριστικό των μη-παραμετρικών ελέγχων είναι ότι μπορούν να εφαρμοστούν εξίσου καλά και σε κανονικά κατανεμημένα δείγματα. Ας δούμε για παράδειγμα το αποτέλεσμα στην περίπτωση των κανονικά κατανεμημένων βαρών με βάση την τροφή που είδαμε παραπάνω:

Hide

```
wilcox.test(chickwts$weight[linseed],chickwts$weight[horsebean])
```

```
##
##  Wilcoxon rank sum test
##
## data:  chickwts$weight[linseed] and chickwts$weight[horsebean]
## W = 100, p-value = 0.007145
## alternative hypothesis: true location shift is not equal to 0
```

Στην προκειμένη περίπτωση βλέπουμε πώς η εφαρμογή του ελέγχου σε κανονικά κατανεμημένα δείγματα δίνει συγκρίσιμα αποτελέσματα με το t.test, κάτι που μας επιτρέπει να τολμήσουμε την εφαρμογή της γενικά και ανεξάρτητα από τον τύπο των κατανομών. Ένα μειονέκτημά του, ως μη παραμετρικού ελέγχου, είναι η μη αναφορά διαστημάτων εμπιστοσύνης.

Σύγκριση λόγων και αναλογιών

Ας επιστρέψουμε τώρα σε ένα άλλο ερώτημα που προκύπτει από τον τρόπο με τον οποίο προσεγγίσαμε (ελλιπώς) το πρόβλημα των ηλικιών των ασθενών με σημαντική ή καθόλου βελτίωση στην θεραπεία της Ρευματοειδούς Αρθρίτιδας. Στην ανάλυση που κάναμε παραπάνω συγκρίναμε την ηλικία τους με βάση το βαθμό βελτίωσης των συμπτωμάτων τους παραβλέποντας έναν βασικό παράγοντα που διαμορφώνει σε μεγάλο βαθμό το αποτέλεσμα και έχει να κάνει με το είδος της θεραπείας (πραγματική ή εικονική, Placebo). Καθώς το σύνολο των ατόμων που εξετάζουμε είναι ανομοιογενές δεν μπορούμε να προχωρήσουμε σε συγκρίσεις χωρίς να έχουμε λάβει υπ' όψιν το βαθμό αυτής της ανομοιογένειας.

Πριν προσπαθήσουμε έτσι να εξαγάγουμε συμπεράσματα για την ηλικία των ασθενών σε σχέση με την απόκριση στη θεραπεία ας δούμε αν υπάρχει κάποια σχέση μεταξύ της απόκρισης και του είδους της θεραπείας. Έχουμε στην ουσία να συγκρίνουμε τα δεδομένα με βάση δύο κατηγορικά δεδομένα (Treatment/Placebo και No/Marked Improvement). Αρχικά θα δημιουργήσουμε έναν πίνακα που να περιέχει μόνο αυτά αφαιρώντας τις τιμές ενδιάμεσης βελτίωσης (Some) για λογους απλότητας.

Hide

```
which(Arthritis$Improved!="Some")->i
sArthritis<-Arthritis[i,]
factor(sArthritis$Improved)->sArthritis$Improved
```

(Θυμίζουμε εδώ ότι η τελευταία εντολή εκτελείται ώστε τα επίπεδα του παράγοντα Improved για το νέο σετ δεδομένων να επανυπολογιστούν και να περιέχουν μόνο τις τιμές που αφήσαμε (None, Marked)).

Σύγκριση λόγων και αναλογιών- Ελεγχος Fisher για πίνακες σύμπτωσης 2Χ2

Αυτό που χρειαζόμαστε στη συνέχεια είναι να υπολογίσουμε τις συχνότητες μεταξύ των δύο κατηγορικών μεταβλητών θεραπείας(Treatment) και αποτελέσματος (Improved), θυμηθείτε πως αυτό επιτυγχάνεται με την συνάρτηση *table*().

Hide

```
table(sArthritis$Treatment, sArthritis$Improved)
```

```
##
##           None Marked
##   Placebo   29      7
##   Treated   13     21
```

Ο πίνακας σύμπτωσης που προκύπτει δείχνει αμέσως ότι υπάρχει μια αντίστροφη τάση μεταξύ των δύο μεταβλητών. Μεταξύ των ασθενών που έχουν πάρει Placebo η πλειοψηφία δεν έδειξε βελτίωση, ενώ το αντίθετο ισχύει για αυτούς που δέχτηκαν κανονική θεραπεία. Η διαφορά των λόγων (29/7 και 13/21) είναι αρκετά μεγάλη, ωστόσο χρειαζόμαστε έναν στατιστικό έλεγχο που θα μας πει κατά πόσο είναι, ως διαφορά, στατιστικά σημαντική. Στην περίπτωση της σύγκρισης λόγων συχνοτήτων/πιθανοτήτων ο έλεγχος που προτείνεται είναι αυτός που πρώτος περιέγραψε ο Ronald Fisher και που ελέγχει την μηδενική υπόθεση ο σύνθετος λόγος των δύο λόγων πιθανοτήτων (odds ratio) να είναι ίσος με την μονάδα. Πράγματι αν δεν υπάρχει σύνδεση μεταξύ των ενδεχομένων λήψης κανονικού φαρμάκου ή placebo και βελτίωσης τότε οι λόγοι του παραπάνω πίνακα θα έτειναν να είναι ίσοι καθώς τα άτομα δε θα είχαν προτίμηση και θα κατανέμονταν με τον ίδιο τρόπο μεταξύ των διαφόρων κατηγοριών.

Σύγκριση λόγων και αναλογιών- Ελεγχος Fisher για πίνακες σύμπτωσης 2Χ2

Ο έλεγχος του Fisher εξετάζει σε ποιο βαθμό τέτοιες “προτιμήσεις” οδηγούν σε ανισοκατανομές που αποκλίνουν σημαντικά από το αναμενόμενο και υπολογίζει μια τιμή p-value που προκύπτει ακριβώς από τον πλήθος των πιθανών διατάξεων των ατόμων σε έναν 2Χ2 πίνακα σύμπτωσης (ανήκει δηλαδή στην κατηγορία των ακριβών (exact) ελέγχων). Η εφαρμογή του είναι απλή και γίνεται με την συνάρτηση $fisher.test()$ πάνω στα δεδομένα που έχουμε χρησιμοποιήσει για τον παραπάνω πίνακα:

Hide

```
fisher.test(sArthritis$Treatment, sArthritis$Improved)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  sArthritis$Treatment and sArthritis$Improved
## p-value = 0.0005272
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   2.04206 23.01700
## sample estimates:
## odds ratio
##   6.487789
```

Όπως βλέπετε, το output είναι παρόμοιο με αυτό του ελέγχου του t.test που είδαμε παραπάνω και περιγράφει τον έλεγχο και τα δεδομένα πάνω στα οποία εφαρμόστηκε, την τιμή p-value, διαστήματα εμπιστοσύνης και το στατιστικό που υπολογίζει που σε αυτήν την περίπτωση είναι ο λόγος των λόγων πιθανοτήτων (odds ratio). Αν ανατρέξετε στον πίνακα σύμπτωσης που προκύπτει από τα ίδια δεδομένα θα δείτε ότι είναι η τιμή που προκύπτει από το σύνθετο κλάσμα των συχνοτήτων. Το διάστημα εμπιστοσύνης αναφέρεται σε αυτήν την τιμή odds ratio (για συντομία OR) και στην περίπτωση του ελέγχου Fisher θέλουμε τα όριά του να βρίσκονται πάνω ή κάτω από την μονάδα που είναι η τιμή βάσης.

Σε αντιστοιχία με το t.test κατά την διενέργεια ενός ελέγχου Fisher αναφέρουμε την τιμή OR συνοδευόμενη από την τιμή p-value και τα διαστήματα εμπιστοσύνης:

Ασθενείς που έλαβαν θεραπεία με το πραγματικό φάρμακο έδειξαν σημαντική βελτίωση των συμπτωμάτων σε σχέση με αυτούς που έλαβαν εικονική θεραπεία placebo (Fisher’s exact test, OR=6.48, p-value=0.00052, 95%CI:(2.04, 23.02))

Σύγκριση λόγων και αναλογιών-Ελεγχος Fisher πίνακες σύμπτωσης >2X2

Σε πολλές περιπτώσεις ο αριθμός των κατηγοριών είναι μεγαλύτερος από 2 κάτι που οδηγεί σε πίνακες σύμπτωσης μεγαλύτερων διαστάσεων, για τους οποίους ο ακριβής υπολογισμός όλων των πιθανών μεταθέσεων καθίσταται εξαιρετικά δύσκολος λόγω του εκρηκτικά αυξανόμενου αριθμού τους (θυμηθείτε από την συνδυαστική ότι στον υπολογισμό υπεισέρχονται παραγοντικά γινόμενα). Για το λόγο αυτό, σε πίνακες με διαστάσεις >2X2 καταφεύγουμε σε προσεγγιστικές μεθόδους ανάλογες του ελέγχου Fisher, με χαρακτηριστικότερο παράδειγμα τον έλεγχο χ^2 του Pearson (Pearson's Chi-square) που βασίζεται σε προσεγγιστικούς υπολογισμούς των παρατηρούμενων προς αναμενόμενων λόγων πιθανοτήτων.

Θυμηθείτε ότι στο αρχικό σύνολο δεδομένων *Arthritis* υπήρχε μια ακόμα κατηγορία βελτίωσης συμπτωμάτων που αντιστοιχούσε σε μερική βελτίωση (Some). Αν επανέλθουμε στο αρχικό dataframe και επαναλάβουμε την διαδικασία που είδαμε παραπάνω θα έχουμε καταρχάς έναν 2X3 πίνακα σύμπτωσης:

Hide

```
table(Arthritis$Treatment, Arthritis$Improved)
```

```
##
##           None Some Marked
##   Placebo   29    7      7
##   Treated   13    7     21
```

τη στατιστική εκτίμηση του οποίου θα πραγματοποιήσουμε τώρα με τον έλεγχο χ^2 όπως πραγματοποιείται από την συνάρτηση *chisq. test*():

Hide

```
chisq.test(Arthritis$Treatment, Arthritis$Improved)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Arthritis$Treatment and Arthritis$Improved
## X-squared = 13.055, df = 2, p-value = 0.001463
```

Το ouput είναι πιο απλό σε αυτήν την περίπτωση αναφέροντας μόνο το στατιστικό χ^2 , την τιμή p, και τους βαθμούς ελευθερίας (df). Σε αντιστοιχία με τις παραπάνω περιπτώσεις, η σωστή πρακτική κατά την περιγραφή της ανάλυσής μας είναι να αναφέρουμε και τα τρία αυτά χαρακτηριστικά της ανάλυσης όπως στο παρακάτω παράδειγμα:

Ασθενείς που έλαβαν θεραπεία με το πραγματικό φάρμακο έδειξαν σημαντική βελτίωση των συμπτωμάτων σε σχέση με αυτούς που έλαβαν εικονική θεραπεία placebo (Pearson's χ^2 test, p-value=0.00146, df=2, X-squared=13.055)

Έλεγχος υπερ-εκπροσωπήσεων με την υπεργεωμετρική κατανομή

Σε αυτές τις περιπτώσεις οι έλεγχοι Fisher ή χ^2 μπορούν να αποτελέσουν ικανοποιητικές προσέγγισεις, ωστόσο για μεγάλο αριθμό συνδυασμών με γνωστά τα συνολικά πλήθη των ατόμων/αντικειμένων είναι προτιμότερο να χρησιμοποιείται ο έλεγχος μέσω της υπεργεωμετρικής κατανομής. Από αυστηρά μαθηματική άποψη, η υπεργεωμετρική κατανομή αποδίδει την στατιστική σημασία της υπερεκπροσώπησης k στοιχείων επιλεγμένων από έναν γνωστό πληθυσμό συνολικών K αντικειμένων μιας συγκεκριμένης ιδιότητας από ένα ευρύτερο δείγμα μεγέθους N.

Θα καταλάβουμε καλύτερα τόσο την έννοια όσο και την λειτουργία της υπεργεωμετρικής κατανομής μέσα από ένα (λίγο μακάβριο) παράδειγμα, χρησιμοποιώντας το σύνολο δεδομένων $Titanic$ της R. Το Titanic είναι ένας πολυδιάστατος πίνακας (array, βλ. Κεφάλαιο 3) που περιέχει συγκεντρωμένα τα στοιχεία των επιβατών του Τιτανικού με βάση την θέση που ταξίδευαν, το φύλο, την ηλικία τους και το αν τελικά επέζησαν ή όχι.

Hide

```
str(Titanic)
```

```
##  'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

Έλεγχος υπερ-εκπροσωπήσεων με την υπεργεωμετρική κατανομή

Μπορούμε να θυμηθούμε για λίγο πώς αποκτούμε πρόσβαση στα στοιχεία ενός πολυδιάστατου πίνακα στην R μέσω δεικτών σε αγκύλες. Ο αριθμός των δεικτών πρέπει να είναι ίσος με τις διαστασείς της λίστας των μεταβλητών (που εδώ είναι 4) και με την σειρά που υποδεικνύεται από την δομή του αντικειμένου. Για παράδειγμα αν κανείς ενδιαφέρεται να δει την μοίρα των παιδιών στον Τιτανικό θα πρέπει να επιλέξει με βάση την τιμή 1 για τον τρίτο δείκτη στις αγκύλες (μεταβλητή Age).

Hide

```
Titanic[,,1,]
```

```
## , , Survived = No
##
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
##
## , , Survived = Yes
##
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
```

απ' όπου προκύπτει ότι όλα τα παιδιά στην πρώτη και δεύτερη θέση επέζησαν, ωστόσο 52 (35+17) συνολικά παιδιά-επιβάτες της τρίτης θέσης χάθηκαν στο ναυάγιο.

Έλεγχος υπερ-εκπροσωπήσεων με την υπεργεωμετρική κατανομή

```
str(Titanic)
```

```
##  'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

Hide

```
N<-sum(Titanic[,,,]) # total sum of passengers
K<-sum(Titanic[,2,,]) # sum of female passengers
n<-sum(Titanic[,,,2]) # sum of passengers who survived
k<-sum(Titanic[,2,,2]) # sum of female passengers who survived
c(k, n, K, N)
```

```
## [1]  344  711  470 2201
```

Απ' όπου βλέπουμε ότι 344 από τους 711 επιζώντες ήταν γυναίκες, την στιγμή που οι γυναίκες ήταν συνολικά 470 σε σύνολο 2201 επιβατών. Υπολογίζοντας απλώς τα ποσοστά γυναικών επιζώντων (344/711=48.3%) και γυναικών συνολικά (470/2201=21.3%) φαίνεται ότι υπάρχει μια προφανής τάση υπερ-εκπροσώπησης, την οποία μπορούμε να υπολογίσουμε μέσω του ελέγχου Fisher σε έναν πίνακα σύμπτωσης 2X2:

Έλεγχος υπερ-εκπροσωπήσεων με την υπεργεωμετρική κατανομή

Απ' όπου βλέπουμε ότι 344 από τους 711 επιζώντες ήταν γυναίκες, την στιγμή που οι γυναίκες ήταν συνολικά 470 σε σύνολο 2201 επιβατών. Υπολογίζοντας απλώς τα ποσοστά γυναικών επιζώντων (344/711=48.3%) και γυναικών συνολικά (470/2201=21.3%) φαίνεται ότι υπάρχει μια προφανής τάση υπερ-εκπροσώπησης, την οποία μπορούμε να υπολογίσουμε μέσω του ελέγχου Fisher σε έναν πίνακα σύμπτωσης 2X2:

Hide

```
N<-sum(Titanic[,,,]) # total sum of passengers
K<-sum(Titanic[,2,,]) # sum of female passengers
n<-sum(Titanic[,,,2]) # sum of passengers who survived
k<-sum(Titanic[,2,,2]) # sum of female passengers who survived
#
m<-matrix(c(k, n, K, N), nrow=2, ncol=2)
colnames(m)<-c("Survived", "Not Survived")
rownames(m)<-c("Female", "Male")
ft<-fisher.test(m)
ft$p.value
```

```
## [1] 2.530126e-22
```

Όπου βλέπουμε ότι για ένα OR=2.26 η τιμή p-value είναι εξαιρετικά μικρή (της τάξης του 10^{-22}).

Έλεγχος υπερ-εκπροσωπήσεων με την υπεργεωμετρική κατανομή

Η υπεργεωμετρική κατανομή θα υπολογίσει *ακριβώς* την πιθανότητα να υπάρχουν k ή περισσότερες γυναίκες μεταξύ n επιζώντων από ένα σύνολο K γυναικών σε N επιβάτες. Η εφαρμογή της γίνεται με την συνάρτηση $phyper(k, K, N-K, n)$, όπου με βάση τους ορισμούς που έχουμε δώσει, N-K είναι ο αριθμός των επιβατών που δεν είναι γυναίκες. Στην R θα γράψουμε:

Hide

```
N<-sum(Titanic[,,,]) # total sum of passengers
K<-sum(Titanic[,2,,]) # sum of female passengers
n<-sum(Titanic[,,,2]) # sum of passengers who survived
k<-sum(Titanic[,2,,2]) # sum of female passengers who survived
phyper(k, K, N-K, n, lower.tail = F)
```

```
## [1] 2.638026e-97
```

Στην εφαρμογή της συνάρτησης $phyper$ έχει σημασία η σειρά με την οποία εισάγουμε τα δεδομένα τα οποία κατά σειρά είναι:
1. k=επιτυχίες στο δείγμα,
2. K="επιτυχίες" στον πληθυσμό,
3. $N-K$="αποτυχίες" στον πληθυσμό,
4. n=μέγεθος του δείγματος.

Οι "επιτυχίες" και "αποτυχίες" ορίζονται με βάση την υπόθεση που θέλουμε να εξετάσουμε. Στην προκειμένη περίπτωση "επιτυχία" συνιστά το να είναι κάποια γυναίκα (που είναι επιτυχία και σε πολλές άλλες περιπτώσεις).

Η παράμετρος lower.tail=F λέει στην R ότι μας ενδιαφέρει να υπολογίσουμε την πιθανότητα του ανώτερου μέρους της κατανομής δηλαδή την πιθανότητα για τιμές επιτυχιών στο δείγμα $>= k$.

Το αποτέλεσμα του υπερ-γεωμετρικού ελέγχου είναι μια μοναδική τιμή p-value που αντιστοιχεί στην πιθανότητα να υπήρχαν 344 (επιτυχίες στο δείγμα) ή περισσότερες γυναίκες μεταξύ των 711 επιζώντων (δείγμα) από ένα σύνολο 2201 επιβατών (πληθυσμός) στο οποίο οι γυναίκες ήταν 470 (επιτυχίες στον πληθυσμό).

Βλέπουμε ότι η τιμή είναι πολλές τάξεις μεγέθους μικρότερη από την αντίστοιχη του ελέγχου Fisher καθώς εκμεταλλευόμαστε την μεγαλύτερη ισχύ που έχει ο υπεργεωμετρικός έλεγχος. Στην περιγραφή της ανάλυσής μας θα γράψουμε:

Οι γυναίκες ήταν υπερ-εκπροσωπούμενες μεταξύ των επιζώντων του Τιτανικού με OR=2.26 και p-value=$2.26\mathrm{X}10^{-97}$ με βάση έναν υπεργεωμετρικό έλεγχο.

Ύλη

Ανάλυση δεδομένων με την R

Κεφάλαιο 11