ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Ανάλυση Διακύμανσης και Έλεγχοι Πολλαπλών Υποθέσεων

Πέτρος Κολοβός

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ

Εισαγωγή

- Τι συμβαίνει στην περίπτωση που τα δείγματα είναι περισσότερα από δύο; Οι μέθοδοι που έχουμε ως τώρα εξετάσει μας επιτρέπουν προφανώς να συγκρίνουμε τα δείγματα σε ζεύγη ωστόσο αυτό δημιουργεί προβλήματα σε δύο επίπεδα.
- Το πρώτο είναι κάπως τεχνικό και σχετίζεται με τον αριθμό των συγκρίσεων που κάνουμε. Καθώς αυξάνεται το πλήθος των δειγμάτων οι έλεγχοι που κάνουμε αυξάνονται με ακόμα μεγαλύτερο ρυθμό. Αυτό έχει ως αποτέλεσμα να ελέγχουμε έναν αυξανόμενο αριθμό υποθέσεων που με την σειρά τους δημιουργούν μια σειρά από τιμές p-value που επίσης πρέπει να ελεγχθούν. Η έννοια του ελέγχου των “πολλαπλών υποθέσεων” αποκτά έτσι ιδιαίτερη σημασία
- Το δεύτερο και ίσως πιο ουσιαστικό επίπεδο είναι ότι πολλές φορές υπάρχει πληροφορία μεταξύ των δειγμάτων που χάνεται σε ζευγαρωτές συγκρίσεις. Μπορείτε να φανταστείτε περιπτώσεις όπου η σχέση δύο δειγμάτων με ένα τρίτο μπορεί να μας δώσει πληροφορίες και για την μεταξύ τους σχέση. Συνεπώς χρειαζόμαστε τρόπους για να εξετάσουμε τέτοιες πολύπλοκες σχέσεις που βασίζονται τόσο στις κεντρικές τάσεις όσο και στη διασπορά δειγμάτων. Χαρακτηριστικότερη μεθοδολογία για κάτι τέτοιο αποτελεί η Ανάλυση Διακύμανσης

Πολλαπλοί ζευγαρωτοί έλεγχοι (pairwise tests)

Hide

```
table(chickwts$feed)
```

```
##
##    casein horsebean   linseed  meatmeal   soybean sunflower
##        12        10        12        11        14        12
```

ενώ με την χρήση συναρτήσεων όπως η *aggregate*() μπορούμε να εξετάσουμε κεντρικές τάσεις:

Hide

```
aggregate(chickwts$weight, by=list(chickwts$feed), FUN=mean)
```

```
##     Group.1        x
## 1    casein 323.5833
## 2 horsebean 160.2000
## 3   linseed 218.7500
## 4  meatmeal 276.9091
## 5   soybean 246.4286
## 6 sunflower 328.9167
```

και μέτρα διασποράς ανα κατηγορία:

Hide

```
aggregate(chickwts$weight, by=list(chickwts$feed), FUN=sd)
```

```
##     Group.1        x
## 1    casein 64.43384
## 2 horsebean 38.62584
## 3   linseed 52.23570
## 4  meatmeal 64.90062
## 5   soybean 54.12907
## 6 sunflower 48.83638
```

Πολλαπλοί ζευγαρωτοί έλεγχοι (pairwise tests)

Η βασική μας επιδίωξη παραμένει να συγκρίνουμε τις κατηγορίες σε επίπεδο (καταρχάς) μέσων τιμών. Μπορούμε να κάνουμε κάτι τέτοιο απευθείας με την χρήση συναρτήσεων του τύπου $pairwise.X.test()$ από το πακέτο $stats$. Βασική προϋπόθεση είναι τα δεδομένα να είναι οργανωμένα σε dataframe με την κατηγορική μεταβλητή ως ένα από τα διανύσματα. Το X εδώ μπορεί να είναι ένα εκ των t (t-test) ή wilcox (Wilcoxon Rank Sum test). Εφόσον το κριτήριο της κανονικότητας ικανοποιείται (βλ. προηγούμενο Κεφάλαιο) μπορούμε να επιλέξουμε το πρώτο ως εξής:

Hide

```
library(stats)
pairwise.t.test(chickwts$weight, chickwts$feed, p.adjust.method = "fdr")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  chickwts$weight and chickwts$feed
##
##           casein  horsebean linseed meatmeal soybean
## horsebean 1.6e-08 -         -       -        -
## linseed   4.5e-05 0.0228    -       -        -
## meatmeal  0.0570  2.8e-05   0.0225  -        -
## soybean   0.0012  0.0007    0.2187  0.1991   -
## sunflower 0.8125  1.2e-08   2.8e-05 0.0360   0.0007
##
## P value adjustment method: fdr
```

Ας εξετάσουμε λίγο πιο αναλυτικά το output, που περιέχει το αποτέλεσμα όλων των ανά ζεύγος $t.test()$. Τα δεδομένα αναφέρονται στις αρχικές γραμμές και τα αποτελέσματα παρατίθενται με την μορφή πίνακα που περιέχει μόνο τα p-values. Ο πίνακας είναι φυσιολογικά τριγωνικός καθώς οι συγκρίσεις είναι συμμετρικές. Μια απλή επισκόπηση των τιμών μας βοηθάει να εντοπίσουμε τις κατηγορίες μεταξύ των οποίων υπάρχει σημαντική διαφορά. Από την άλλη δεν μπορούμε να πάρουμε διαστήματα εμπιστοσύνης και το κυριότερο δεν έχουμε πληροφορία για το ποια κατηγορία σε κάθε σύγκριση είχε την μεγαλύτερη μέση τιμή. Ακόμα κι έτσι όμως, σε ένα πρώτο επίπεδο, έχουμε επιτύχει τον αρχικό μας σκοπό που ήταν οι πολλαπλές συγκρίσεις. Βλέπουμε έτσι π.χ. ότι ο λιναρόσπορος διαφέρει σημαντικά από τα κουκιά σε ό,τι αφορά το βάρος όπως είχαμε δει με μια απλή σύγκριση στο προηγούμενο Κεφάλαιο:

Πολλαπλοί ζευγαρωτοί έλεγχοι (pairwise tests)

Hide

```
which(chickwts$feed=="linseed")->linseed
which(chickwts$feed=="horsebean")->horsebean
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
```

Τι συμβαίνει όμως εδώ; Γιατί το t-test εφαρμοσμένο στις δύο κατηγορίες δίνει p-value=0.006869 ενώ στο $pairwise.t.test()$ η αντίστοιχη τιμή είναι p-value=0.0228; Η διαφορά έγκειται στο γεγονός ότι στην περίπτωση του $pairwise.t.test()$ έχουμε εφαρμόσει μια διόρθωση στις τιμές p.value, η οποία δίνεται ως παράμετρος $p.adjust.method$. Για να καταλάβουμε τι ακριβώς κάνει αυτή η διόρθωση θα συζητήσουμε λίγο πιο αναλυτικά την έννοια των πολλαπλών υποθέσεων.

Έλεγχος Πολλαπλών Υποθέσεων

Φανταστείτε το εξής νοητικό πείραμα. Κάποιος σας δίνει ένα νόμισμα και σας ζητάει να το στρίψετε έναν μεγάλο αριθμό φορών (π.χ. N=100) με σκοπό να μπορέσει μετά να κάνει μια εκτίμηση για το αν το νόμισμα είναι κανονικό η κάλπικο. Έστω ότι εσείς φέρνετε σε 100 ρίψεις 60 φορές “κορώνα” κάτι που σας κάνει να υποψιαστείτε ότι κάτι δεν πάει καλά με το νόμισμα. Πράγματι, αν θεωρήσετε ότι η κορώνα είναι η επιτυχία και τα γράμματα η αποτυχία, μπορείτε να υπολογίσετε την πιθανότητα 60 επιτυχιών σε 100 δοκιμές για ένα νόμισμα που αναμένεται να έχει 50/50 πιθανότητες με την χρήση της διωνυμικής κατανομής. Η πιθανότητα υπολογίζεται στην R με την συνάρτηση $pbinom()$ και προκύπτει ότι για X=60, σε N=100 με πιθανότητα p=0.5 είναι ίση με:

Hide

```
pbinom(60, 100, prob=0.5, lower.tail = F)
```

```
## [1] 0.0176001
```

(Σημειώστε στο παραπάνω ότι η παράμετρος lower.tail=F λέει στην R να επιστρέψει την συμπληρωματική πιθανότητα από αυτήν που υπολογίζει ως default που είναι η αθροιστική πιθανότητα να φέρει κανείς **εως** 60 κορώνες σε 100 ρίψεις)

Η πιθανότητα αυτή είναι 0.0176, υπάρχει δηλαδή <2% πιθανότητα να φέρει κανείς 60 κορώνες σε 100 ρίψεις με ένα κανονικό νόμισμα. Ο ιδιοκτήτης του νομίσματος έχει κάθε λόγο να υποψιαστεί ότι πρόκειται για κάλπικο νόμισμα. Η αρχική, μηδενική του υπόθεση ότι το κέρμα είναι κανονικό μπορεί να απορριφθεί στη βάση του p=0.0176.

Ας υποθέσουμε τώρα ότι το νόμισμα έχει γι’ αυτόν πολύ μεγαλή αξία και δεν θέλει να το “ξεγράψει” τόσο εύκολα. Σκέφτεται τι θα γινόταν αν αντί για έναν είχε ζητήσει από περισσότερα άτομα να κάνουν το πείραμα των 100 ρίψεων. Επειδή δεν μπορεί να βρει αρκετούς πρόθυμους πειραματιστές αποφασίζει να προσομοιώσει τα δεδομένα με την χρήση της R ξεκινώντας με την μηδενική υπόθεση ότι το νόμισμα είναι κανονικό. Θα πραγματοποιήσει 1000 προσομοιώσεις εκατό ρίψεων και θα αποθηκεύσει σε ένα διάνυσμα το άθροισμα των επιτυχιών σε κάθεμία από αυτές. Κάτι τέτοιο μπορεί να γίνει με τον τρόπο που έχουμε ήδη δει μέσω ενός βρόχου επανάληψης. Στην συνέχεια σπεύδει να μετρήσει πόσες φορές παίρνει μια τιμή ίση η μεγαλύτερη του 60.

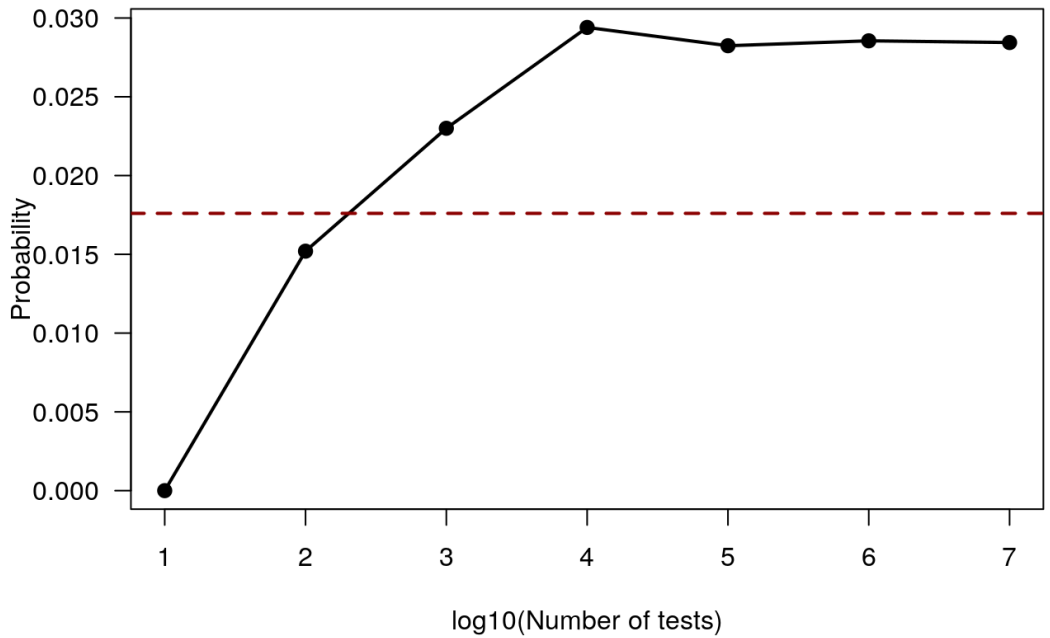
Hide

```
nheads<-vector("numeric", 1000)
for(i in 1:1000){
  nheads[i]=sum(rbinom(100, 1, 0.5))
}
length(which(nheads>=60))/1000
```

```
## [1] 0.03
```

Έλεγχος Πολλαπλών Υποθέσεων

εκτίμηση στη βάση ενός πολύ μεγαλύτερου αριθμού υποθέσεων. Στο παρακάτω διάγραμμα φαίνεται η πιθανότητα παρατήρησης >=60 επιτυχιών σε 100 ρίψεις για αυξανόμενο αριθμό πειραμάτων από 10 έως 10000000.



Με κόκκινη διακεκομμένη γραμμή φαίνεται η πιθανότητα που υπολογίζεται για ένα πείραμα. Από το διάγραμμα είναι προφανές ότι καθώς ο αριθμός των πειραμάτων (και κατά συνέπεια των ελέγχων) αυξάνεται, αυξάνεται και η πιθανότητα των ελέγχων να δώσουν ένα αποτέλεσμα που είναι φαινομενικά στατιστικά σημαντικό. Η βάση του ελέγχου πολλαπλών υποθέσεων βρίσκεται ακριβώς εδώ και υπαγορεύει μια διόρθωση των αναφερόμενων πιθανοτήτων (και p-values) προς τα πάνω όταν ο αριθμός των ελέγχων είναι μεγάλος.

Διόρθωση τιμής p-value

Η R διαθέτει μια χρήσιμη συνάρτηση για την διόρθωση (adjustment) των τιμών p-value, η οποία ενσωματώνει μια σειρά από διαφορετικές μεθόδους, η αναλυτική ερμηνεία των οποίων ξεφεύγει από τους σκοπούς του Κεφαλαίου. Η συνάρτηση ονομάζεται εύλογα $p.adjust()$ και δέχεται ως είσοδο ένα διάνυσμα τιμών p-value και το πλήθος των ελέγχων που έχουν πραγματοποιηθεί. Η μέθοδος διόρθωσης δίνεται μέσω της παραμέτρου $method$. Παρακάτω βλέπουμε ποια θα ήταν η διορθωμένη τιμή της αρχικής πιθανότητας των 60 επιτυχιών αν προερχόταν από ένα πείραμα που είχε επαναληφθεί 100 φορές, με διόρθωση τύπου Benjamini-Hochberg:

Hide

```
library(stats)
p.adjust(0.0176, method="BH", n=100)
```

```
## [1] 1
```

κάτι που σημαίνει ότι σε ένα δείγμα 100 πειραμάτων υπάρχει πρακτικά 100% πιθανότητα να πάρουμε τουλάχιστον 1 φορά μια τιμή 60/100 επιτυχιών ακόμα και με ένα απολύτως κανονικό νόμισμα.

Επανερχόμενοι στο αρχικό παράδειγμα των ζευγαρωτών ελέγχων των διατροφικών συμπληρωμάτων μπορούμε εύκολα να δούμε ότι όλες οι τιμές p-value είναι μεγαλύτερες από αυτές που θα προέκυπταν από μεμονωμένες συγκρίσεις ανα-δυο καθώς έχουμε επιβάλει στην $pairwise.t.test()$ να κάνει διόρθωση μέσω της μεθόδου FDR (False Discovery Rate). Ο γενικός κανόνας είναι ότι σε περιπτώσεις ενός αριθμού συγκρίσεων (N>10) πάντοτε κάνουμε διόρθωση των τιμών p-value με κάποια ενδεδειγμένη μεθοδολογία. Η διόρθωση αποκτά ιδιαίτερη σημασία στην ανάλυση διακύμανσης που θα εξετάσουμε στη συνέχεια καθώς εκεί ο αριθμός των ελέγχων είναι συχνά αρκετά μεγάλος.

Έλεγχος διακύμανσης

Η ανάλυση διακύμανσης (ή διασποράς) είναι ένα πολύ χρήσιμο μεθοδολογικό εργαλείο στη στατιστική ανάλυση, το οποίο βασίζεται στη σύγκριση των διασπορών μεταξύ πολλών δειγμάτων. Πριν όμως περάσουμε σε αυτήν θα συζητήσουμε το πιο απλό ερώτημα την σύγκρισης της διασποράς δύο δειγμάτων. Η σύγκριση διασποράς δύο δειγμάτων γίνεται με το F-test ή έλεγχο διασποράς. Στην R η αντίστοιχη συνάρτηση ονομάζεται $var.test()$ και η εφαρμογή της είναι απλή όπως φαίνεται παρακάτω:

Hide

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

```
##
##  F test to compare two variances
##
## data:  chickwts$weight[linseed] and chickwts$weight[horsebean]
## F = 1.8289, num df = 11, denom df = 9, p-value = 0.3739
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4674894 6.5617411
## sample estimates:
## ratio of variances
##           1.828854
```

Επισκόπηση του output δείχνει ότι είναι πολύ παρόμοιο αυτού του $t.test()$ με αναγραφή του στατιστικού F, των βαθμών ελευθερίας, της τιμής p-value και του διαστήματος εμπιστοσύνης. Η τιμή αναφοράς είναι εδώ ο λόγος των δύο διασπορών και η τιμή βάσης είναι η μονάδα. Στο συγκεκριμένο παράδειγμα παρότι οι δύο διασπορές διαφέρουν με αυτήν του λιναρόσπορου να είναι περίπου 2.8 φορές μεγαλύτερη, η διαφορά δεν θεωρείται στατιστικά σημαντική από τον έλεγχο καθώς η τιμή 1 βρίσκεται μέσα στο διάστημα εμπιστοσύνης 95%.

Από το παραπάνω παράδειγμα βλέπουμε ότι δύο δείγματα μπορεί να έχουν στατιστικά σημαντικές διαφορές σε ό,τι αφορά την μέση τιμή τους αλλά η διαφορά αυτή να μην αποτυπώνεται σε επίπεδο διασποράς, κάτι που είναι λογικό αν δεχτούμε ότι λίγο πολύ οι διασπορές μπορούν να είναι παρόμοιες. Για την ακρίβεια, οι παρόμοιες διασπορές εντός των δειγμάτων είναι βασική προϋπόθεση για την εφαρμογή της ανάλυσης διακύμανσης την οποία θα εξετάσουμε αμέσως μέτα.

Ανάλυση Διακύμανσης (ANOVA)

Με τον όρο **Ανάλυση Διακύμανσης** ή **Ανάλυση Διασποράς** (**An**alysis **o**f **Va**riance) αναφερόμαστε σε μια δέσμη από στατιστικές αναλύσεις μοντελοποίησης που σκοπό έχουν την εκτίμηση της διαφοράς των μέσων τιμών μεταξύ ομάδων σε ευρύτερα δείγματα. Όπως και πολλά άλλα μεθοδολογικά εργαλεία, η ANOVA αναπτύχθηκε από τον Ronald Fisher και στηρίζεται στην διαισθητικά απλή αλλά τεχνικά μάλλον πολύπλοκη βασική αρχή της ανάλυσης της παρατηρούμενης διακύμανσης μιας μεταβλητής σε διαφορετικές συνιστώσες που σχετίζονται με εγγενείς τάσεις και τυχαία σφάλματα. Η ANOVA είναι η μεθοδολογική προσέγγιση επιλογής για την σύγκριση περισσότερων από δύο δείγματα. Εννοιολογικά δε διαφέρει πολύ από μια προσέγγιση πολλαπλών ζευγαρωτών t-test, ωστόσο είναι γενικά πιο αυστηρή και έτσι οδηγεί σε πιο συντηρητικά αποτελέσματα με περιορισμένα σφάλματα τύπου I, οδηγεί δηλαδή σπανιότερα στην απόρριψη μηδενικών υποθέσεων που είναι αληθείς.

Η βάση της ANOVA είναι η εκτίμηση της διαφοράς στη μέση τιμή μιας παραμετρικής μεταβλητής σε σχέση με μια ή περισσότερες κατηγορικές μεταβλητές. Ανάλογα με το σχεδιασμό του πειράματος, των αριθμών των μεταβλητών και των ομάδων δειγμάτων υπάρχουν διάφορες παραλλαγές της ANOVA. Έτσι υπάρχει μονο- (one-way) και πολυ-παραγοντική ανάλυση (n-way ANOVA) ανάλογα με το αν η διακύμανση μελετάται σε σχέση με μία ή περισσότερες κατηγορικές μεταβλητές, η πολυμεταβλητή (mulitvariate ANOVA ή MANOVA) όταν η μετρούμενες μεταβλητές απόκρισεις είναι παραπάνω από μία, ενώ ανάλογα με το αν υπάρχουν επιπλέον παραμετρικές μεταβλητές που μπορούν να επηρεάσουν την μετρούμενη (target variable) χρησιμοποιείται ένας συνδυασμός ANOVA και παλινδρόμησης (regression) που ονομάζεται Analysis of Co-Variance (ANCOVA, βλ Κεφάλαιο 9).

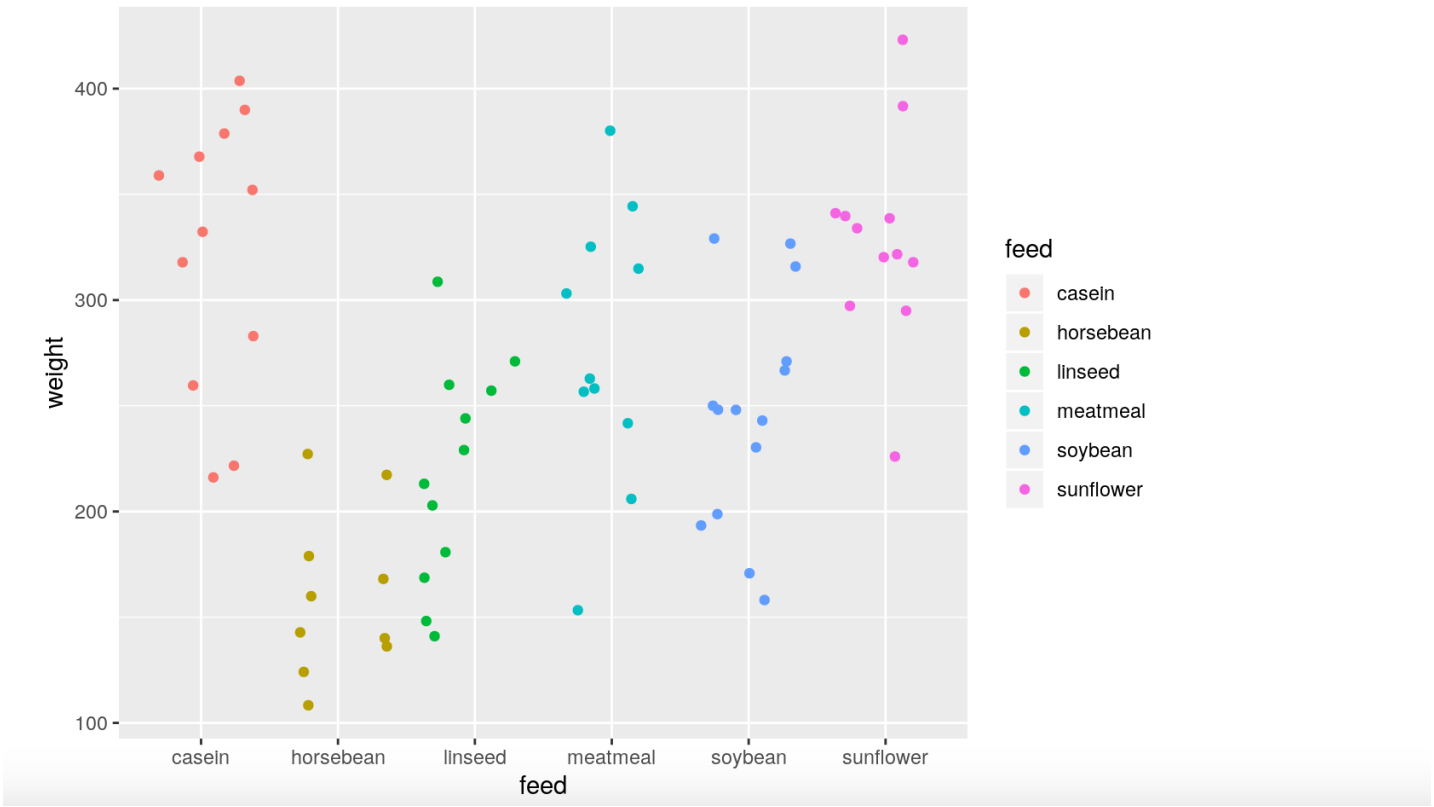
Σε γενικές γραμμές η ANOVA είναι μια σύνθετη εννοιολογικά μεθοδολογία που σε κάποιες περιπτώσεις πολύπλοκων πειραματικών σχεδιασμών γίνεται εξαιρετικά δύσκολη στη διδασκαλία και ερμηνεία. Για μια σε βάθος περιγραφή της παραπέμπουμε (όπως και για το σύνολο των στατιστικών εννοιών) τους αναγνώστες σε εγχειρίδια με επίκεντρο στη Στατιστική Ανάλυση. Για τις ανάγκες αυτού του Κεφαλαίου θα περιοριστούμε σε μια διαισθητική ανάλυση των βασικών εννοιών πίσω από την απλή μονο-παραγοντική ANOVA και στη συνέχεια σε μια περιγραφή των εντολών με τις οποίες μπορεί κανείς να διενεργήσει τους κυριότερους τύπους Ανάλυσης Διακύμανσης στην R.

Ανάλυση Διακύμανσης (ANOVA)

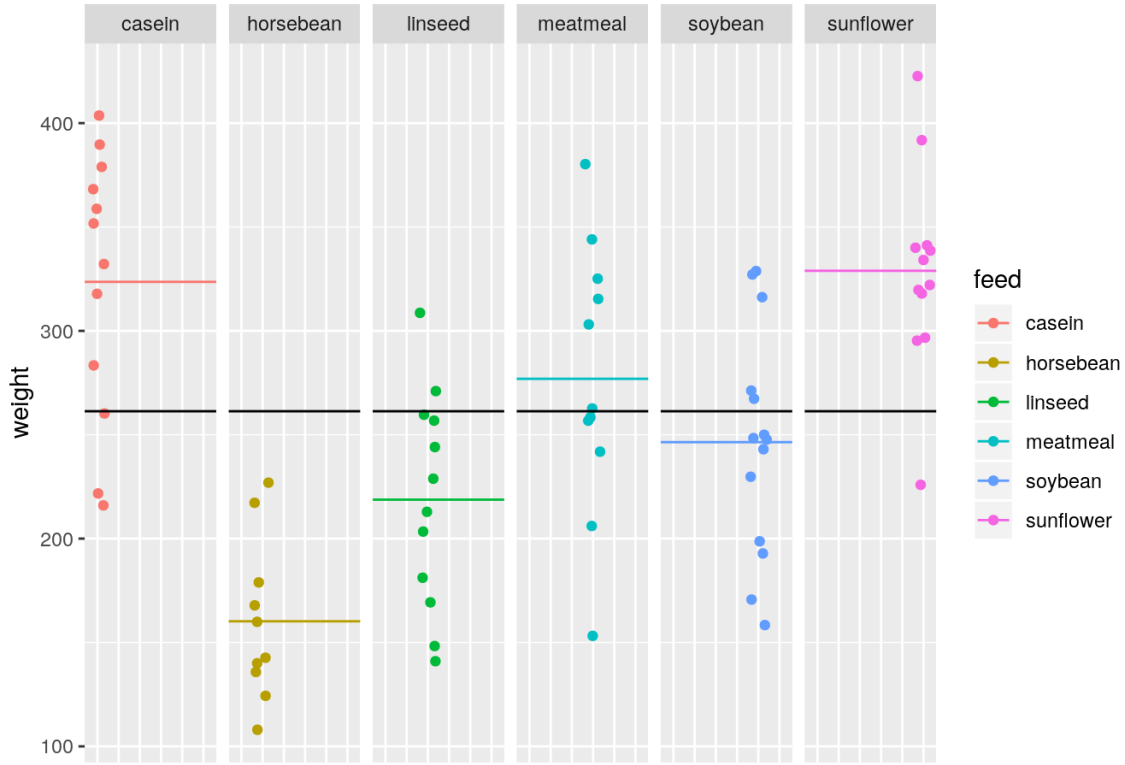
Ίσως το στοιχείο που προξενεί την μεγαλύτερη σύγχυση σχετικά με την ANOVA είναι το γεγονός ότι αποτελεί μια μέθοδο εκτίμησης διαφορών στη μέση τιμή ενώ υπολογίζει διασπορές (variances). Πράγματι, ο στατιστικός έλεγχος που διενεργείται στη βάση της ANOVA είναι ένα F-test δηλαδή ο έλεγχος διασποράς που είδαμε παραπάνω. Πώς φτάνουμε απο αυτον τον έλεγχο να βγάλουμε συμπεράσματα για τις μέσες τιμές; Ο καλύτερος τρόπος για να κατανοήσει κανείς πώς γίνεται κάτι τέτοιο είναι να δει ένα παράδειγμα βήμα-βήμα. Θα επιστρέψουμε για το σκοπό αυτό στο πλαίσιο δεδομένων *chickwts* που είδαμε στην περίπτωση των πολλαπλών t-test. Ας θυμηθούμε λίγο πώς μοιάζουν τα δεδομένα:

Hide

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



Ανάλυση Διακύμανσης (ANOVA)



Τι περισσότερο μαθαίνουμε από αυτό το διάγραμμα; Σε αυτό βλέπουμε ότι η συνολική διασπορά των τιμών κάθε κατηγορίας μπορεί να αναλυθεί σε δύο συνιστώσες. Η πρώτη είναι ο βαθμός στον οποίο η επιμέρους μέση τιμή τους διαφέρει από την γενική μέση τιμή και η δεύτερη είναι ο βαθμός στον οποίο διασπείρονται γύρω από την επιμέρους αυτή μέση τιμή της κατηγορίας τους. Μεταξύ των δύο μπορούμε να δούμε την πρώτη ως “σήμα” και την δεύτερη ως “θόρυβο”. Αυτό γιατί η μεν πρώτη είναι όντως δηλωτική της διαφοράς των κατηγοριών ως προς τις μέσες τιμές τους, ενώ η δεύτερη έχει να κάνει με την διασπορά των τιμών της κάθε κατηγορίας.

Σκοπός της ANOVA είναι ακριβώς να αξιολογήσει το βάρος της πρώτης σε σχέση με την δεύτερη. Αυτό το κάνει υπολογίζοντας έναν συνδυασμό των δύο σε έναν λόγο που ονομάζεται συνολική διασπορά (total variance) και ορίζεται ως το πηλίκο της διασποράς μεταξύ των κατηγοριών προς την διασπορά εντός των κατηγοριών. Ο λόγος αυτός ελέγχεται με ένα F-test όπως συμβαίνει στην ανάλυση διασποράς.

Ανάλυση Διακύμανσης (ANOVA)

Η εφαρμογή των συναρτήσεων που θα παρουσιάσουμε στη συνέχεια δεν διαφέρει μεταξύ των παραπάνω περιπτώσεων, είναι ωστόσο σημαντικό να έχουμε μια καλή εποπτεία των δεδομένων μας και των παραμέτρων που θέλουμε να εξετάσουμε.

Στην περίπτωση που θα δούμε παρακάτω έχουμε την απλούστερη περίπτωση μιας μεταβλητής απόκρισης (βάρος) και μιας κατηγορικής μεταβλητής (διατροφικό συμπλήρωμα). Με βάση τα παραπάνω η κλήση της συνάρτησης *aov*() θα είναι:

Hide

```
aov(weight~feed, data=chickwts)
```

```
## Call:
##    aov(formula = weight ~ feed, data = chickwts)
##
## Terms:
##                     feed Residuals
## Sum of Squares  231129.2  195556.0
## Deg. of Freedom        5        65
##
## Residual standard error: 54.85029
## Estimated effects may be unbalanced
```

Το output από μόνο του δε φαίνεται να μας λέει πολλά καθώς αναγράφει μόνο μέρος των στατιστικών που έχει υπολογίσει η ANOVA χωρς να γίνεται λόγος για στατιστική σημαντικότητα. Τις πληροφορίες αυτές παίρνουμε αν αποθηκεύσουμε το αποτέλεσμα της ανάλυσης σε μια μεταβλητή και εκτελέσουμε πάνω της την *summary*():

Ανάλυση Διακύμανσης (ANOVA)

Hide

```
aov(weight~feed, data=chickwts)->fit
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## feed         5 231129   46226   15.37 5.94e-10 ***
## Residuals   65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Το αποτέλεσμα σε αυτήν την περίπτωση είναι αρκετά πιο πλούσιο σε πληροφορία. Οι βαθμοί ελευθερίας υπήρχαν και πριν και είναι 5 για την κατηγορική μεταβλητή $feed$ (N-1, με N=6 διατροφικά συμπληρώματα) ενώ η αντίστοιχη τιμή df για τα σφάλματα είναι 65, όπως προκύπτει αν αφαιρέσουμε από το συνολικό πλήθος των τιμών (N-1, για N=71) τους βαθμούς ελευθερίας που έχουν ενσωματωθεί στην κατηγορική μεταβλητή (N=5).

Η τιμή που κυρίως μας ενδιαφέρει εδώ είναι η F και η πιθανότητα μια τέτοια η μεγαλύτερη τιμή να έχει προκύψει δεδομένου ότι οι κατανομές έχουν την ίδια διασπορά. Μπορούμε να υπολογίσουμε την πιθανότητα αυτή (που εδώ δίνεται κάτω από την τελευταία στήλη: Pr(>F)) με την χρήση της συνάρτησης $pf()$ που δίνει την πιθανότητα παρατήρησης μια τιμής F δεδομένων δύο βαθμών ελευθερίας.

Hide

```
pf(15.37, df1=5, df2=65, lower.tail = F)
```

```
## [1] 5.902455e-10
```

από την οποία βλέπουμε ότι (με μια μικρή απόκλιση λόγω στρογγυλοποίησης) παίρνουμε την ίδια τιμή που μας δίνει η $aov()$.

Τι μας λέει αυτή η πιθανότητα; Σε ένα πρώτο επίπεδο, ότι η τιμή της είναι πολύ μικρή, ώστε να δικαιολογεί την απόρριψη της μηδενικής υπόθεσης και άρα τα δεδομένα ανά κατηγορία να μην προέρχονται από την ίδια κατανομή. Ως αποτέλεσμα φαντάζει φτωχό καθώς δεν μας λέει τίποτα για το ποια διατροφικά συμπληρώματα συμπεριφέρονται καλύτερα. Μπορούμε να πάρουμε αυτήν την πληροφορια με την εφαρμογή ενός υπολογισμού διαστημάτων εμπιστοσύνης πάνω στις διαφορές των μέσων τιμών των επιμέρους κατηγοριών με την συνάρτηση $TukeyHSD()$ που δρα πάνω σε ένα αντικείμενο που έχει προκύψει από την ANOVA.

Ανάλυση Διακύμανσης (ANOVA)

Hide

```
aov(weight~feed, data=chickwts)->fit
TukeyHSD(fit, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = weight ~ feed, data = chickwts)
##
## $feed
##                             diff         lwr        upr     p adj
## horsebean-casein    -163.383333 -245.964363 -80.802304 0.0000000
## linseed-casein      -104.833333 -183.571256 -26.095411 0.0002100
## meatmeal-casein      -46.674242 -127.181777  33.833292 0.3324584
## soybean-casein       -77.154762 -153.028522  -1.281001 0.0083653
## sunflower-casein       5.333333  -73.404589  84.071256 0.9998902
## linseed-horsebean     58.550000  -24.031030 141.131030 0.1413329
## meatmeal-horsebean   116.709091   32.439113 200.979069 0.0001062
## soybean-horsebean     86.228571    6.373743 166.083400 0.0042167
## sunflower-horsebean  168.716667   86.135637 251.297696 0.0000000
## meatmeal-linseed      58.159091  -22.348444 138.666626 0.1276965
## soybean-linseed       27.678571  -48.195189 103.552332 0.7932853
## sunflower-linseed    110.166667   31.428744 188.904589 0.0000884
## soybean-meatmeal     -30.480519 -108.189144  47.228105 0.7391356
## sunflower-meatmeal    52.007576  -28.499959 132.515111 0.2206962
## sunflower-soybean     82.488095    6.614335 158.361856 0.0038845
```

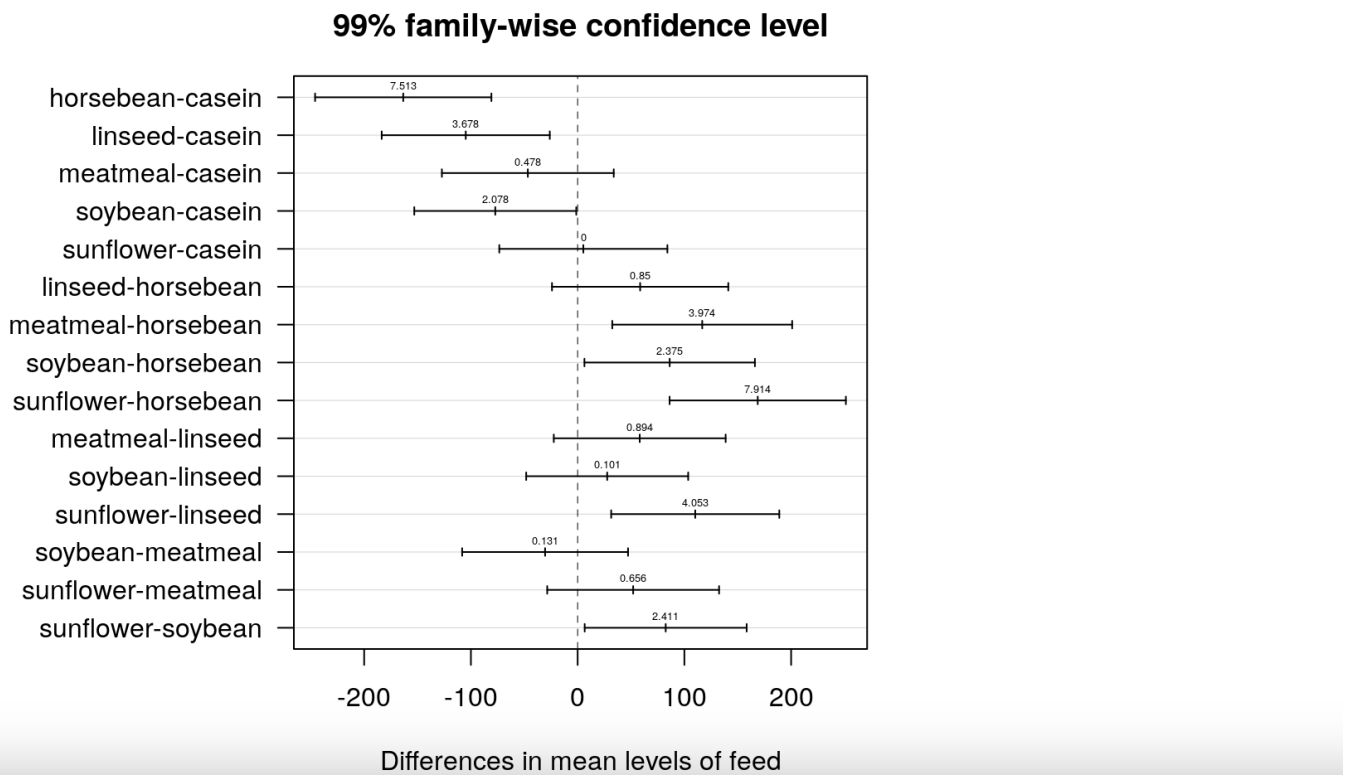
Η εφαρμογή μας δίνει ό,τι χρειαζόμαστε με την διαφορά μεταξύ των μέσων τιμών (diff), τα πάνω και κάτω όρια του διαστήματος εμπιστοσύνης που εμείς ορίσαμε (lwr, upr) και ένα διορθωμένο p-value στην τελευταία στήλη.

Ανάλυση Διακύμανσης (ANOVA)

Τα δεδομένα αυτά μπορούμε να αναπαραστήσουμε γραφικά με μια απλή κλήση *plot()*. Στο παρακάτω παράδειγμα, έχουμε προσθέσει μέσω ενός βρόχου επανάληψης την καταγραφή των τιμών -log10(p-adj) πάνω από τα αντίστοιχα διαστήματα εμπιστοσύνης με την συνάρτηση *text()*:

Hide

```
aov(weight~feed, data=chickwts)->fit
TukeyHSD(fit, conf.level = 0.99)->tukey
par(las=1)
par(mar=c(5,10,2,2))
plot(tukey)
for(i in 1:length(tukey$feed[,1])){
  text(x=tukey$feed[i,1],y=length(tukey$feed[,1])-i+1.3, labels=round(-log10(tukey$feed[i,4]), digit
s=3), cex=0.4)
}
```



Πολυ-παραγοντική ANOVA (multi-way ANOVA)

Όπως αναφέραμε παραπάνω, στόχος μας είναι να δείξουμε την βασική μεθοδολογία για την διενέργεια μιας ANOVA ανάλυσης χωρίς να κάνουμε αναλυτική περιγραφή όλων των διαφορετικών τύπων. Στο σημείο αυτό, ωστόσο, αξίζει να δούμε πώς η παραπάνω ανάλυση μπορεί να τροποποιηθεί στην περίπτωση που έχουμε περισσότερες από μία κατηγορικές μεταβλητές. Μπορούμε να δούμε πώς μετατρέπεται η ανάλυση στην περίπτωση δύο κατηγορικών μεταβλητών μελετώντας το σύνολο δεδομεων ToothGrowth από την βασική έκδοση της R. Το συγκεκριμένο σύνολο δεδομενων απαρτίζεται από τιμές μήκους οδοντοβλαστών (των αρχικών δομών που οδηγούν στην ανάπτυξη ενός ενήλικου δοντιού) σε ινδικά χοιρίδια κάτω από την επίδραση δύο διαφορετικών μορφών βιταμίνης C (χυμού πορτοκαλιού (OJ) και καθαρού ασκορβικού οξέος (VC)) σε τρεις διαφορετικές δοσολογίες 0.5, 1 και 2 mg ημερησίως. Μπορούμε να δούμε το dataframe παρακάτω:

Hide

```
TGdata<-ToothGrowth
str(TGdata)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Hide

```
TGdata$dose<-factor(TGdata$dose, levels = c(0.5, 1, 2))
str(TGdata)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

Πολυ-παραγοντική ANOVA (multi-way ANOVA)

Προσέξτε ότι ενώ αρχικά οι τιμές των δοσολογιών dose δεν αναγνωρίζονται ως κατηγορικές μεταβλητές από την R τις μετατρέπουμε σε τέτοιες μέσω της συνάρτησης $factor()$ δημιουργώντας ένα αντίγραφο του dataframe που ονομάζουμε TGdata. Αυτό, γιατί μας ενδιαφέρει να χρησιμοποιήσουμε την δοσολογία ως μια επιπλέον επεξηγηματική, κατηγορική μεταβλητή. Ας δούμε αρχικά αν τα δεδομένα είναι ισορροπημένα (αν δηλαδή έχουμε την ίδια εκπροσώπηση σε πλήθος μεταξύ των κατηγοριών), κάτι που όπως θα δούμε παρακάτω είναι μια σχετική προϋπόθεση για την καλύτερη διενέργεια του ελέγχου

Hide

```
table(TGdata[,2:3])
```

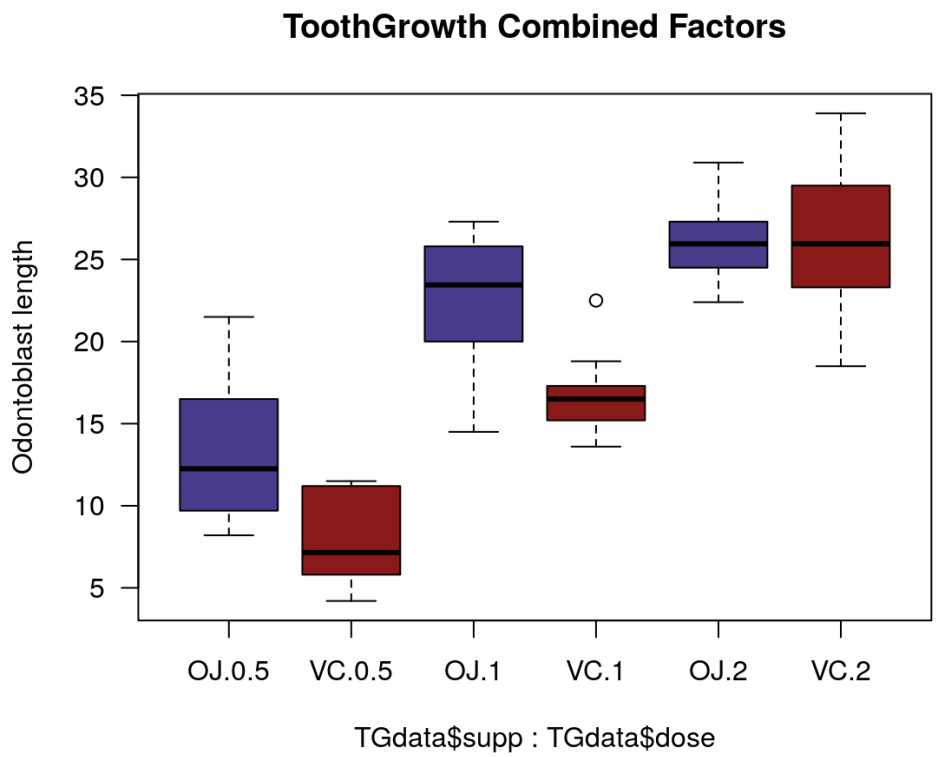
```
##     dose
## supp 0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

Εφόσον υπάρχει απόλυτη ισορροπία μεταξύ των κατηγοριών μπορούμε να συνεχίσουμε αρχικά στην εποπτική ανάλυση της επίδρασης των δύο παραγόντων στο μήκος των οδοντοβλαστών. Για δύο παραμέτρους μπορούμε αρχικά να δημιουργήσουμε ένα συνδυαστικό θηκόγραμμα με ομαδοποιημένες κατηγορίες. Μπορούμε να κάνουμε κάτι τέτοιο ζητώντας η μεταβλητή απόκρισης να αναπαρασταθεί ως συνδυασμός των δύο κατηγορικών μεταβλητών με τον τελεστή "*". Θυμηθείτε ότι συναντήσαμε τον συγκεκριμένο τελεστή πιο πάνω στην συζήτηση για συσχετιζόμενες μεταβλητές.

Πολυ-παραγοντική ANOVA (multi-way ANOVA)

Hide

```
boxplot(TGdata$len ~ TGdata$supp * TGdata$dose, col=c("slateblue4","firebrick4"), las=1, main="Tooth Growth Combined Factors", ylab="Odontoblast length")
```

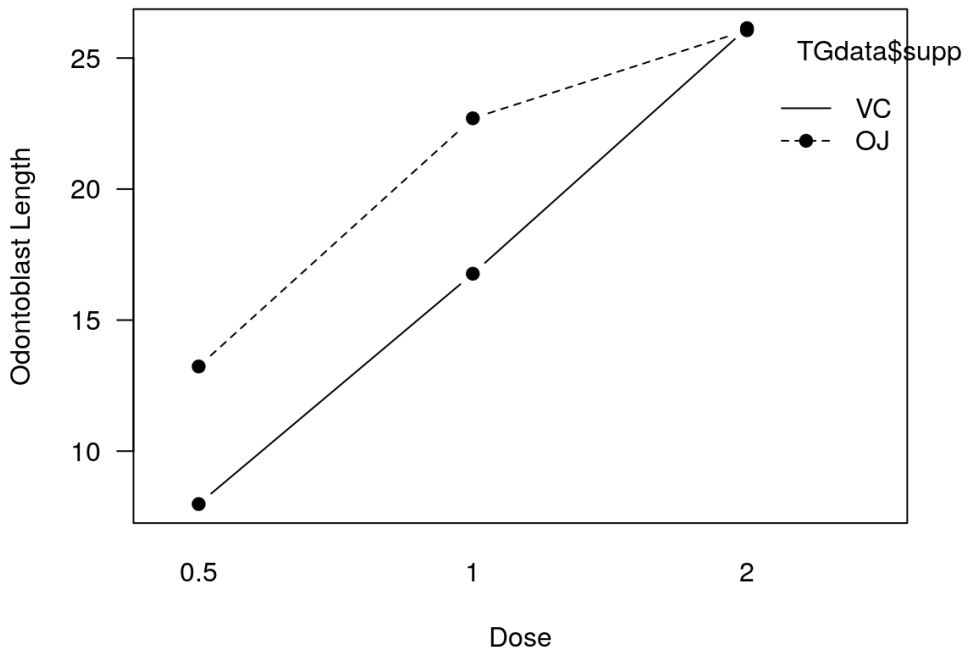


Από το θηκόγραμμα φαίνεται ότι ο χυμός πορτοκαλιού (OJ) έχει σημαντικά καλύτερα αποτελέσματα σε μικρότερες δόσεις, ωστόσο για 2 mg/ημέρα το πλεονέκτημα αυτό παύει να υπάρχει. Από το διάγραμμα αυτό λέμε ότι υπάρχει μια αλληλεπίδραση μεταξύ παραγόντων καθώς αυτοί δεν είναι τελείως ανεξάρτητοι μεταξύ τους. Η δόση δηλαδή επηρρεάζει την διαφορά στην απόκριση μεταξύ των δύο μορφών της βιταμίνης. Μπορούμε να δούμε καλύτερα αυτήν την επίδραση μέσω ενός ειδικού διαγράμματος που ονομάζεται *interaction. plot*()

Πολυ-παραγοντική ANOVA (multi-way ANOVA)

Hide

```
interaction.plot(TGdata$dose, TGdata$supp, response = TGdata$len,
                 fun = mean, type = "b",
                 xlab = "Dose", ylab="Odontoblast Length",
                 pch=c(19), las=1)
```



στο οποίο βλέπουμε πιο παραστατικά ότι τα δύο διατροφικά συμπληρώματα “συναντιούνται” για δόσεις =2mg/ημέρα.

Αυτή η παρατήρηση είναι σημαντική γιατί μας λέει ότι θα πρέπει να διενεργήσουμε μια ανάλυση η οποία θα λάβει υπ’ όψιν αυτήν την αλληλεπίδραση. Αν οι δύο καμπύλες του διαγράμματος αλληλεπίδρασης παρέμεναν παράλληλες θα μπορούσαμε να εκτελέσουμε μια ANOVA με “αθροιστικό” (additive) τρόπο ως εξής:

Πολυ-παραγοντική ANOVA (multi-way ANOVA)

Αυτή η παρατήρηση είναι σημαντική γιατί μας λέει ότι θα πρέπει να διενεργήσουμε μια ανάλυση η οποία θα λάβει υπ' όψιν αυτήν την αλληλεπίδραση. Αν οι δύο καμπύλες του διαγράμματος αλληλεπίδρασης παρέμεναν παράλληλες θα μπορούσαμε να εκτελέσουμε μια ANOVA με "αθροιστικό" (additive) τρόπο ως εξής:

Hide

```
aov(len ~ supp + dose, data=TGdata)->additive_fit
summary(additive_fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.4   205.4   14.02 0.000429 ***
## dose         2 2426.4  1213.2   82.81  < 2e-16 ***
## Residuals   56  820.4    14.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ο τύπος $len\ supp + dose$ επιβάλλει στην ANOVA να θεωρήσει ότι δεν υπάρχει συσχέτιση μεταξύ των δύο παραμέτρων. Ωστόσο όπως είδαμε παραπάνω αυτό δεν ισχύει. Ο σωστότερος τρόπος να τρέξουμε το μοντέλο ως "παραγοντικό" (factorial) με την χρήση του τελεστή "*" που είδαμε παραπάνω:

Hide

```
aov(len ~ supp * dose , data=TGdata)->factorial_fit
summary(factorial_fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *
## Residuals   54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Στο αποτέλεσμα του οποίου προσέξτε ότι εμφανίζεται ένα επιπλέον σημαντικός όρος supp:dose που αντιστοιχεί στην συνδυαστική επίδραση των δύο παραμέτρων στη μεταβλητή απόκρισης. Για να δούμε ποιοι συνδυασμοί είναι στατιστικά σημαντικά διαφορετικοί και κατά πόσο, θα καταφύγουμε στον έλεγχο πολλαπλών υποθέσεων Tukey, μόνο που αυτή την φορά θα πρέπει να υποδείξουμε στη συνάρτηση $TukeyHSD()$, μέσω της παραμέτρου $which$, αν θέλουμε να κάνει τους ελέγχους με βάση την μεταβλητή supp:

Πολυ-παραγοντική ANOVA (multi-way ANOVA)

Hide

```
TukeyHSD(factorial_fit, which="supp")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp * dose, data = TGdata)
##
## $supp
##       diff       lwr       upr     p adj
## VC-OJ -3.7 -5.579828 -1.820172 0.0002312
```

οπότε μας επιστρέφει ως αποτέλεσμα ότι συνολικά ο χυμός πορτοκαλιού έχει καλύτερα αποτελέσματα, ή αντίστοιχα με βάση την μεταβλητή της δοσολογίας dose:

Hide

```
TukeyHSD(factorial_fit, which="dose")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp * dose, data = TGdata)
##
## $dose
##         diff       lwr       upr   p adj
## 1-0.5  9.130  6.362488 11.897512 0.0e+00
## 2-0.5 15.495 12.727488 18.262512 0.0e+00
## 2-1    6.365  3.597488  9.132512 2.7e-06
```

από την οποία και πάλι βλέπουμε ότι η αύξηση της δόσης έχει πάντα καλύτερα αποτελέσματα.

Πολυ-παραγοντική ANOVA (multi-way ANOVA)

Αυτό που μας ενδιαφέρει περισσότερο, ωστόσο, είναι οι συνδυασμοί των δύο μεταβλητών τις οποίες μπορούμε να δούμε δίνοντας στην $TukeyHSD()$ τον συνδυασμό supp:dose:

Hide

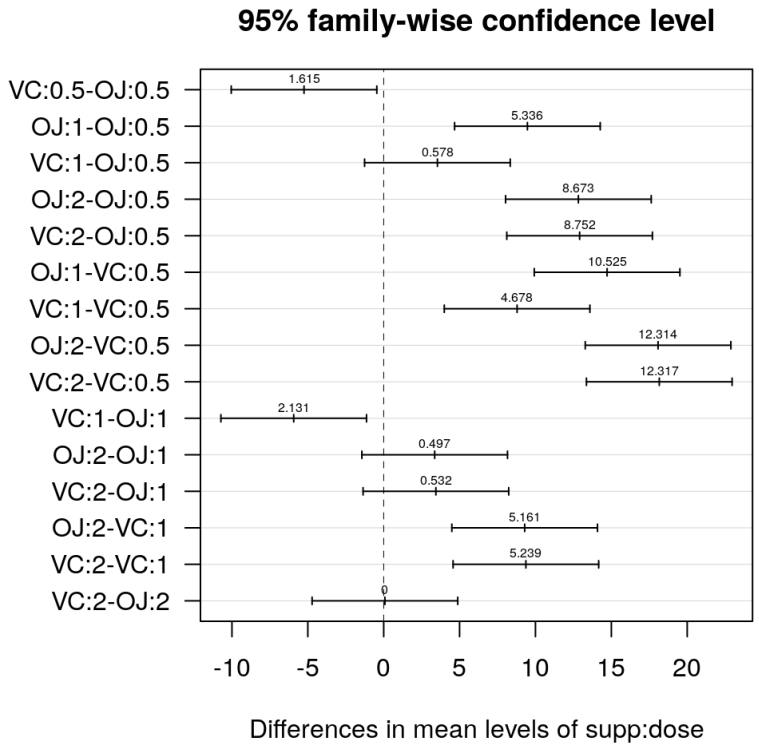
```
TukeyHSD(factorial_fit, which="supp:dose")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp * dose, data = TGdata)
##
## $`supp:dose`
##                 diff        lwr        upr     p adj
## VC:0.5-OJ:0.5 -5.25 -10.048124 -0.4518762 0.0242521
## OJ:1-OJ:0.5    9.47   4.671876 14.2681238 0.0000046
## VC:1-OJ:0.5    3.54  -1.258124  8.3381238 0.2640208
## OJ:2-OJ:0.5   12.83   8.031876 17.6281238 0.0000000
## VC:2-OJ:0.5   12.91   8.111876 17.7081238 0.0000000
## OJ:1-VC:0.5   14.72   9.921876 19.5181238 0.0000000
## VC:1-VC:0.5    8.79   3.991876 13.5881238 0.0000210
## OJ:2-VC:0.5   18.08  13.281876 22.8781238 0.0000000
## VC:2-VC:0.5   18.16  13.361876 22.9581238 0.0000000
## VC:1-OJ:1     -5.93 -10.728124 -1.1318762 0.0073930
## OJ:2-OJ:1      3.36  -1.438124  8.1581238 0.3187361
## VC:2-OJ:1      3.44  -1.358124  8.2381238 0.2936430
## OJ:2-VC:1      9.29   4.491876 14.0881238 0.0000069
## VC:2-VC:1      9.37   4.571876 14.1681238 0.0000058
## VC:2-OJ:2      0.08  -4.718124  4.8781238 1.0000000
```

Το αποτέλεσμα περιέχει διαφορές στη μέση τιμή, διαστήματα εμπιστοσύνης και διορθωμένα p-value για όλους τους συνδυασμούς. Από μια πρώτη επισκόπηση των τελευταίων βλέπουμε να επιβεβαιώνεται και ποσοτικά η παρατήρηση σύμφωνα με την οποία ο χυμός πορτοκαλιού είναι πιο αποδοτικός σε μικρές συγκεντρώσεις. Μπορούμε να δούμε γραφικά τα αποτελέσματα του παραπάνω πίνακα με τον τρόπο που είδαμε παραπάνω για την μονο-παραγοντική ANOVA.

Πολυ-παραγοντική ANOVA (multi-way ANOVA)

```
par(las=1)
par(mar=c(5,10,2,2))
TukeyHSD(factorial_fit, which="supp:dose")->tukey
plot(tukey)
for(i in 1:length(tukey$`supp:dose`[,1])){
  text(x=tukey$`supp:dose`[i,1],y=length(tukey$`supp:dose`[,1])-i+1.3, labels=round(-log10(tukey$`su
pp:dose`[i,4]), digits=3), cex=0.5)
}
```



Με απολύτως ανάλογο τρόπο μπορεί να εργαστεί κανείς για περισσότερες από δύο κατηγορικές μεταβλητές. Βασική προϋπόθεση είναι μια αρκετά καλή ανάλυση των επιδράσεων μεταξύ των εξηγηματικών -όπως τις λέμε- μεταβλητών ώστε ο τύπος με τον οποίο θα εκτελέσουμε την ANOVA να βρίσκεται όσο το δυνατόν πιο κοντά στην πραγματικότητα.

Προϋποθέσεις για την διενέργεια ANOVA

Όπως και στους ελέγχους που είδαμε στο προηγούμενο κεφάλαιο, έτσι και για την ANOVA υπάρχουν προϋποθέσεις σχετικά με την εφαρμογή της. Οι τρεις βασικοί περιορισμοί για την διενέργεια της ανάλυσης διακύμανσης είναι:

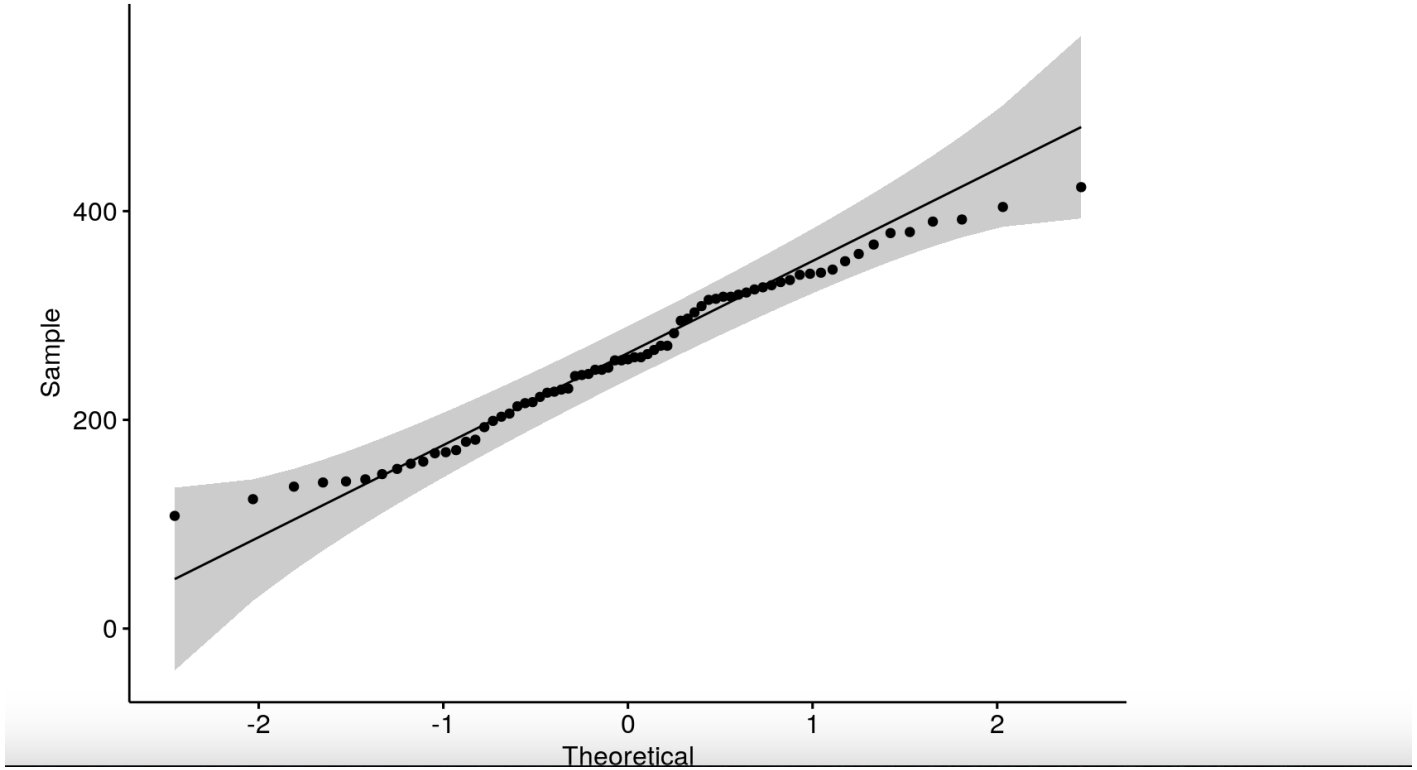
1. Η ανεξαρτησία των κατηγορικών μεταβλητών. Αυτή δεν πρέπει να συγχέεται με τις τυχόν αλληλεπιδράσεις μεταξύ τους. Με τον όρο ανεξαρτησία εννοούμε ότι οι κατηγορικές μεταβλητές δε θα πρέπει να σχετίζονται δομικά μεταξύ τους κατά τον σχεδιασμό του πειράματος.
2. Η κανονικότητα των κατανομών, δηλαδή τα σφάλματα των παραμετρικών μεταβλητών θα πρέπει να ακολουθούν την κανονική κατανομή
3. Η ομοσκεδαστικότητα, που σημαίνει ότι οι κατανομές θα πρέπει να έχουν την ίδια διασπόρα.

Προϋποθέσεις για την διενέργεια ANOVA

Σε ό,τι αφορά την κανονικότητα, έχουμε δει στο προηγούμενο κεφάλαιο ελέγχους για την εκτίμησή της. Καθώς η ανάλυση διακύμανσης είναι στην ουσία μέθοδος για την σύγκριση μέσων τιμών, προϋποθέτει κανονικά σφάλματα κι έτσι μπορούμε απλά να εφαρμόσουμε το Shapiro-Wilk, ή κάποιον ανάλογο έλεγχο κανονικότητας στα δεδομένα πριν ξεκινήσουμε την ανάλυση. Θυμηθείτε ότι μπορούμε επίσης να δείξουμε γραφικά την κανονικότητα μέσω *qqplots*:

Hide

```
library(ggpubr)
ggqqplot(chickwts$weight, pch=19)
```



Προϋποθέσεις για την διενέργεια ANOVA

Ο έλεγχος της ομοσκεδαστικότητας θα μπορούσε να διενεργηθεί τυπικά με μια σειρά από ζευγαρωτά F-test. Ένας έλεγχος που διενεργεί την ανάλυση σε όλες τις κατηγορίες λαμβάνοντας υπ' όψιν τις πολλαπλές συγκρίσεις είναι αυτός του Bartlett (Bartlett's test), που ελέγχει την μηδενική υπόθεση οι διασπορές της μεταβλητής απόκρισης ανά κατηγορία να είναι ίσες. Τον έλεγχο Bartlett μπορούμε να εκτελέσουμε στην R με την συνάρτηση $bartlett.test()$ από το πακέτο $stats$.

Hide

```
bartlett.test(chickwts$weight, chickwts$feed)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  chickwts$weight and chickwts$feed
## Bartlett's K-squared = 3.2597, df = 5, p-value = 0.66
```

Η σύνταξη είναι απλή και περιλαμβάνει την μεταβλητή απόκρισης ($weight$) ακολουθούμενη από την κατηγορική ($feed$) που ορίζει τις ομάδες. Το αποτέλεσμα περιλαμβάνει το στατιστικό (που στην περίπτωση αυτού του ελέγχου είναι η τιμή $K - squared$), τους βαθμούς ελευθερίας (που είναι όπως και παραπάνω ίσοι με τον αριθμό των κατηγοριών μείον 1, εδώ N=6-1). Και μια τιμή p-value που μας επιτρέπει να απορρίψουμε ή όχι την μηδενική υπόθεση. Στην συγκεκριμένη περίπτωση η τιμή p-value=0.66 δεν μας επιτρέπει να την απορρίψουμε, πράγμα που σημαίνει ότι μπορούμε να υποθέσουμε με σχετική ασφάλεια ότι το κριτήριο της ομοσκεδαστικότητας ισχύει.

Προϋποθέσεις για την διενέργεια ANOVA

Ένας επιπλέον ελεγχος, που μπορεί να χρησιμοποιηθεί και σε περιπτώσεις που δεν ισχύει το κριτήριο της κανονικότητας είναι ο έλεγχος του Levene μέσω της συνάρτησης *leveneTest*() από το πακέτο *car*.

Hide

```
library(car)
```

```
## Loading required package: carData
```

Hide

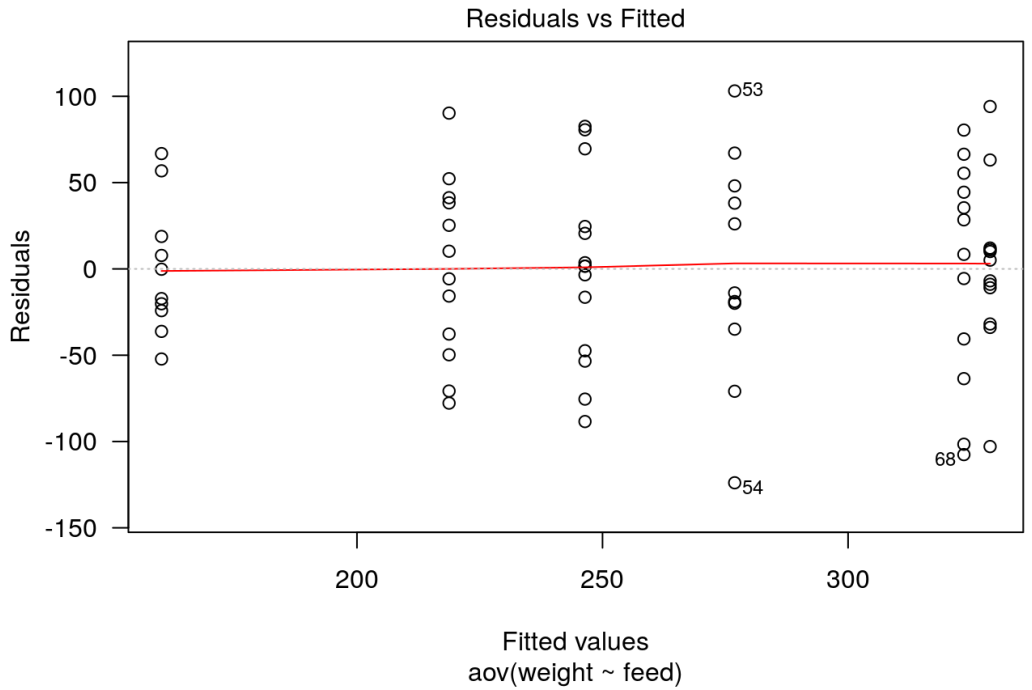
```
leveneTest(weight ~ feed, data = chickwts)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  0.7493 0.5896
##       65
```

Πέρα από τους ποσοτικούς ελέγχους, ένας γραφικός τρόπος για να ελεγξουμε την ομοσκεδαστικότητα είναι καλώντας το ίδιο το μοντέλο της ANOVA με την συνάρτηση *plot*()

Προϋποθέσεις για την διενέργεια ANOVA

```
fit<-aov(weight ~ feed, data=chickwts)
plot(fit,1, las=1)
```



Το συγκεκριμένο διάγραμμα αναπαριστά τις τιμές που έχει εκτιμήσει η ANOVA (fitted) σε σχέση με την απόκλισή τους από τις πραγματικές (residuals). Το αναμενόμενο εδώ είναι όλες οι τιμές (σημεία) να κατανέμονται ομοιόμορφα γύρω από μια τιμή κοντά στο 0 (που αντιστοιχεί σε μηδενική απόκλιση και αρά μέγιστη συμφωνία με τα πειραματικά δεδομένα). Όπως βλέπουμε εδώ αυτό ισχύει για την περίπτωσή μας.

Μια χρήσιμη εφαρμογή της παραπάνω γραφικής αναπαράστασης είναι ότι επισημαίνει τις τιμές (που φαίνονται στο διάγραμμα με τους αριθμητικούς δείκτες) που αντιστοιχούν σε ακραίες τιμές (outliers). Αφαίρεση των συγκεκριμένων τιμών από το δείγμα αναμένεται να αποδώσει καλύτερα αποτελέσματα, καθώς τα κριτήρια του ελέγχου θα ικανοποιούνται με πληρέστερο τρόπο.

Προϋποθέσεις για την διενέργεια ANOVA

Τέλος, μια τέταρτη προϋπόθεση που συνήθως παραλείπεται είναι η ισορροπία (balance) των δειγμάτων, που σημαίνει ότι οι κατηγορίες που ορίζονται θα πρέπει να έχουν ίσο ή παραπλήσιο αριθμό ατόμων. Σε περίπτωση που αυτό δεν ισχύει υπάρχουν σημαντικές αποκλίσεις στις διασπορές κι έτσι το κριτήριο της ομοσκεδαστικότητας δεν ικανοποιείται. Στην περίπτωση του σετ δεδομένων *chickwts* μπορούμε να δούμε ότι:

Hide

```
table(chickwts$feed)
```

```
##
##    casein horsebean   linseed  meatmeal   soybean sunflower
##        12         10        12        11        14        12
```

οι κατηγορίες έχουν όλες 10-14 άτομα συνεπώς σε μεγάλο βαθμό τα δείγματα είναι ισορροπημένα.

Τι συμβαίνει σε περιπτώσεις που τα δείγματα δεν είναι ισορροπημένα; Χωρίς να μπούμε σε πολλές τεχνικές λεπτομέρειες, αξίζει να αναφέρουμε ότι μια παραλλαγή της ANOVA επιτρέπει των υπολογισμό των αθροισμάτων τετραγώνων με διαφορετικό τρόπο ώστε να αποτιμηθεί αυτή η ανισορροπία. Η διόρθωση ενός μοντέλου ANOVA με την ομώνυμη συνάρτηση από το πακέτο *car*() επιτρέπει κάτι τέτοιο.

Προϋποθέσεις για την διενέργεια ANOVA

ANOVA σε μη ισορροπημένα δείγματα

Τι συμβαίνει σε περιπτώσεις που τα δείγματα δεν είναι ισορροπημένα; Μια παραλλαγή της ANOVA επιτρέπει των υπολογισμό των αθροισμάτων τετραγώνων με διαφορετικό τρόπο ώστε να αποτιμηθεί αυτή η ανισορροπία. Η διόρθωση ενός μοντέλου ANOVA με την ομώνυμη συνάρτηση από το πακέτο $car()$ επιτρέπει κάτι τέτοιο.

Αξίζει να δούμε ένα παράδειγμα για το πώς ένα μη-ισορροπημένο δείγμα μπορεί να επηρρεάσει την ανάλυση μας. Θυμηθείτε το σύνολο δεδομένων της Ρευματοδειδούς Αρθρίτιδας $Arthritis$ από το πακέτο vcd.

Μια παραμετρική μεταβλητή σε αυτό ήταν η ηλικία των ασθενών ενώ δύο από τις πιο ενδιαφέρουσες κατηγορικές μεταβλητές ήταν ο βαθμός βελτίωσης των συμπτωμάτων μετά την θεραπεία ($Improved$) και το φύλο τους (Sex).

Από την βιβλιογραφία γνωρίζουμε ότι τόσο η ασθένεια, όσο και η απόκριση στη θεραπεία συνδέεται με το φύλο, έτσι ένας έλεγχος ANOVA με τις δύο αυτές μεταβλητές ως εξηρτημένες δίνει το παρακάτω αποτέλεσμα:

Hide

```
library(vcd)
```

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

Προϋποθέσεις για την διενέργεια ANOVA

```
aov(Age ~ Improved * Sex, data=Arthritis)
```

```
## Call:
##    aov(formula = Age ~ Improved * Sex, data = Arthritis)
##
## Terms:
##                   Improved       Sex Improved:Sex Residuals
## Sum of Squares    1129.381    35.155       689.736 11679.014
## Deg. of Freedom          2         1             2        78
##
## Residual standard error: 12.23646
## Estimated effects may be unbalanced
```

Hide

```
summary(aov(Age ~ Improved * Sex, data=Arthritis))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Improved      2   1129   564.7   3.771 0.0273 *
## Sex           1     35    35.2   0.235 0.6294
## Improved:Sex  2    690   344.9   2.303 0.1067
## Residuals    78  11679   149.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Βλέπετε ότι η παράμετρος βελτίωσης (Improved) έχει μια σημαντική συνεισφορά ωστόσο η αρχική εκτέλεση της $aov()$ επιστρέφει την επισήμανση **"Estimated effects may be unbalanced"**.

Προϋποθέσεις για την διενέργεια ANOVA

Μπορούμε να αντιμετωπίσουμε αυτήν την ειδοποίηση με την διόρθωση που αναφέραμε παραπάνω. Η διαδικασία, μετά την εγκατάσταση του πακέτου *car* έχει ως εξής:

Hide

```
library(car)
```

```
## Loading required package: carData
```

Hide

```
afit<-aov(Age ~ Improved * Sex, data=Arthritis)
Anova(afit, type="III")
```

```
## Anova Table (Type III tests)
##
## Response: Age
##              Sum Sq Df   F value    Pr(>F)
## (Intercept)  159108  1 1062.6245 < 2.2e-16 ***
## Improved       1719  2    5.7406  0.004722 **
## Sex             121  1    0.8102  0.370837
## Improved:Sex    690  2    2.3033  0.106693
## Residuals     11679 78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Όπως βλέπετε η παράμετρος Improved είναι και πάλι σημαντική σε ό,τι αφορά την συνεισφορά στην διαφορά των μέσων τιμών, **η τιμή p-value όμως είναι μικρότερη**. Σε πιο έντονα μη-ισορροπημένα δείγματα η διαφορά αυτή μεταξύ των μεθοδολογιών θα μπορούσε να είναι σημαντικότερη και να οδηγήσει σε σφάλματα και των δύο τύπων.

Προϋποθέσεις για την διενέργεια ANOVA

ANOVA σε μη κανονικά κατανεμημένα δείγματα.

Μέχρι εδώ έχουμε εξετάσει τις δυνατότητες που μας προσφέρει η ANOVA, ωστόσο η εφαρμογή της μπορεί μόνο να γίνει σε περιπτώσεις που έχουμε κανονικά κατανεμημένα σφάλματα.

Όταν το κριτήριο της κανονικότητας δεν ικανοποιείται, το μη-παραμετρικό ανάλογο της ANOVA είναι ο έλεγχος Kruskall-Wallis που εκτελείται με την συνάρτηση $kruskal.test()$ ως εξής:

Hide

```
kruskal.test(weight ~ feed, data = chickwts)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight by feed
## Kruskal-Wallis chi-squared = 37.343, df = 5, p-value = 5.113e-07
```

Ο περιορισμός του συγκεκριμένου ελέγχου είναι ότι μπορεί μόνο να υποκαταστήσει την μονο-παραγοντική ANOVA.

Προϋποθέσεις για την διενέργεια ANOVA

Επίσης παρότι μας δίνει μια τιμή p-value για την αποδοχή ή απόρριψη της μηδενικής υπόθεσης δεν μας δίνει πληροφορία για τις επιμέρους διαφορές μεταξύ κατηγοριών. Για να έχουμε κάτι τέτοιο θα χρειαστεί να καταφύγουμε σε μια παραλλαγή των πολλαπλών ζευγαρωτών ελέγχων που είδαμε στην αρχή αυτού του κεφαλαίου, αντικαθιστώντας το $t.test()$ από το μη-παραμετρικό ανάλογό του $wilcox.test()$

Hide

```
pairwise.wilcox.test(chickwts$weight, chickwts$feed, p.adjust.method = "fdr")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum exact test
##
## data:  chickwts$weight and chickwts$feed
##
##           casein  horsebean linseed meatmeal soybean
## horsebean 0.00016 -         -       -        -
## linseed   0.00305 0.01191   -       -        -
## meatmeal  0.11355 0.00096   0.05451 -        -
## soybean   0.01110 0.00227   0.27306 0.28035  -
## sunflower 1.00000 9.3e-05   0.00025 0.09384  0.00334
##
## P value adjustment method: fdr
```

Ύλη

Ανάλυση δεδομένων με την R

Κεφάλαιο 12