Κεφάλαιο 9

Ανάλυση της Διακύμανσης

Η ανάλυση της διακύμανσης είναι μια από τις πλέον σημαντικές μεθόδους για ανάλυση δεδομένων. Η μέθοδος αυτή αναφέρετε στη διαμέριση του συνολικού αθροίσματος τετραγώνων σε αθροίσματα τετραγώνων λόγω των επιδράσεων των παραγόντων.

9.1 Ανάλυση Διακύμανσης κατά ένα Παράγοντα

Το πιο απλό είδος πειραμάτων είναι αυτά στα οποία μια απλή συνεχής εξαρτημένη μεταβλητή μετρείται έναν αριθμό φορών για κάθε ένα από τα διάφορα επίπεδα ενός πειραματικού παράγοντα. Για παράδειγμα, έστω τα δεδομένα στον Πίνακα 9.1, ο οποίος περιλαμβάνει τις τιμές του χρόνου πήξης του αίματος για τέσσερις διαφορετικές δίαιτες.

Ο χρόνος πήξης είναι η συνεχής εξαρτημένη μεταβλητή, ενώ η δίαιτα είναι ποιοτική μεταβλητή, ή παράγοντας, με τέσσερα επίπεδα: A, B, C, D. Ο κύριος στόχος είναι να εξεταστεί αν αν ο παράγοντας δίαιτα έχει οποιαδήποτε επίδραση στο μέσο χρόνο πήξης του αίματος. Για να γίνει η ανάλυση δεδομένων, πρέπει να γραφούν στην R με τέτοιο τρόπο έτσι ώστε να μπορούν να χρησιμοποιηθούν για ανάλυση της διακύμανσης. Αυτό επιτυγχάνεται με το σχεδιασμό ενός πλαισίου δεδομένων όπως πιο κάτω:

A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59

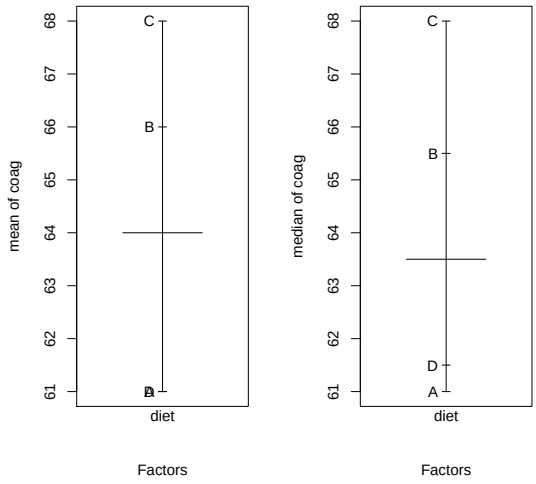
Πίνακας 9.1: Χρόνος πήξης του αίματος για τέσσερις δίαιτες.

```
> coag <- scan()
1: 62 60 63 59
5: 63 67 71 64 65 66
11: 68 66 71 67 68 68
17: 56 62 60 61 63 64 63 59
25:
> coag
 [1] 62 60 63 59 63 67 71 64 65 66 68 66 71 67 68 68 56 62 60 61
[21] 63 64 63 59
> diet <- factor(rep(LETTERS[1:4],c(4,6,6,8)))  #create a factor
> diet
 [1] A A A A B B B B B B C C C C C C D D D D D D D D
> coag.df <- data.frame(diet,coag)          #create a data frame
> coag.df
   diet coag
 1    A   62
 2    A   60
 3    A   63
 4    A   59
 5    B   63
 6    B   67
 7    B   71
 8    B   64
 9    B   65
```

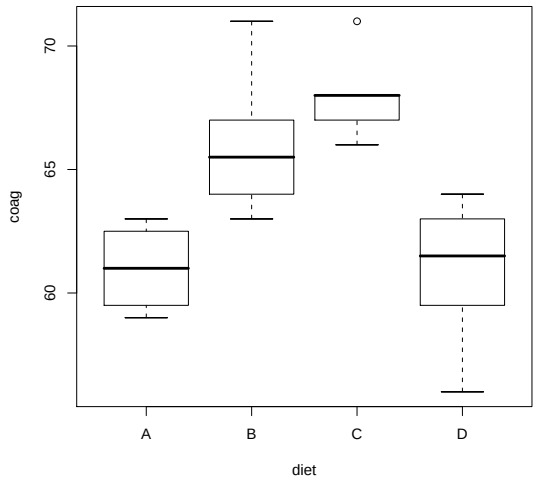
```
10    B   66
11    C   68
12    C   66
13    C   71
14    C   67
15    C   68
16    C   68
17    D   56
18    D   62
19    D   60
20    D   61
21    D   63
22    D   64
23    D   63
24    D   59
```

Το πρώτο βήμα στην ανάλυση δεδομένων είναι να ερευνηθεί γραφικά αν υπάρχουν ή όχι διαφορές ανάμεσα στα επίπεδα του παράγοντα. Τα Σχήματα 9.1 και 9.2 παρουσιάζουν τις μέσες τιμές και τις διαμέσους για κάθε επίπεδο του παράγοντα και το αντίστοιχο κυτιογράφημα. Η οριζόντια ευθεία στο αριστερό (δεξιό) γράφημα δίνει τη μέση τιμή (διάμεσο) όλων των δεδομένων. Είναι φανερό πως τα επίπεδα A και D σχηματίζουν μια κατηγορία, ενώ τα επίπεδα B και C μιαν άλλη κατηγορία.

```
> par(mfrow=c(1,2))
> plot.design(coag.df)
> plot.design(coag.df, fun= median)
> par(mfrow=c(1,1))
> plot(coag.df)
```



Σχήμα 9.1: Μέσες τιμές και διάμεσοι των επιπέδων του παράγοντα.



Σχήμα 9.2: Κυτιογράφημα των επιπέδων του παράγοντα.

Για να εφαρμοστεί η ανάλυση της διακύμανσης στην R χρησιμοποιείται η εντολή aov ως ακολούθως

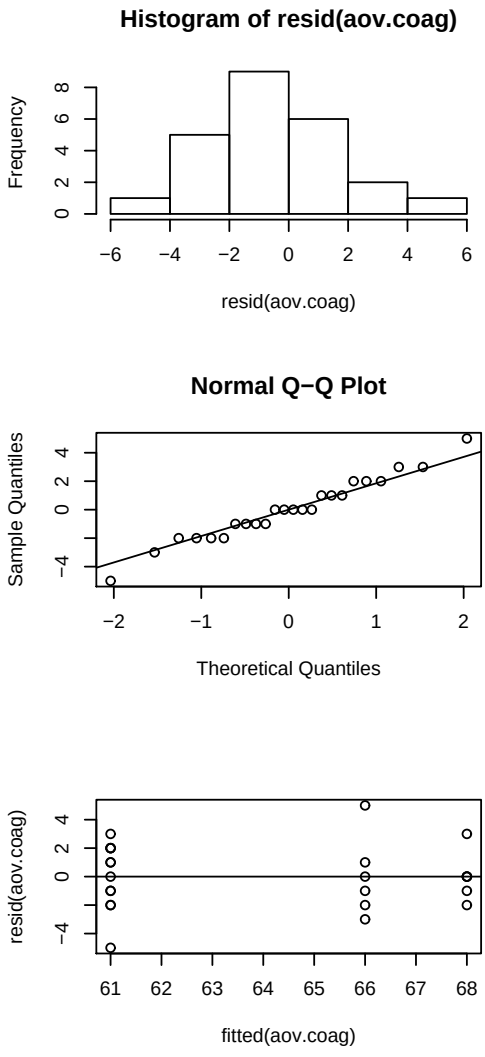
```
> aov.coag <- aov(coag ~ diet, coag.df)
> aov.coag
Call:
   aov(formula = coag ~ diet, data = coag.df)

Terms:
                diet Residuals
Sum of Squares   228       112
Deg. of Freedom    3        20

Residual standard error: 2.366432
Estimated effects may be unbalanced
> summary(aov.coag)
            Df Sum Sq Mean Sq F value    Pr(>F)
diet         3  228.0    76.0  13.571 4.658e-05 ***
Residuals   20  112.0     5.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Σημειώνεται ότι η εντολή aov χρησιμοποιείται με ανάλογο τρόπο όπως και η εντολή lm για τη γραμμική παλινδρόμηση. Ο τύπος coag ~ diet στο πρώτο όρισμα δίνει συμβολικά το μοντέλο της ανάλυσης της διακύμανσης κατά ένα παράγοντα, ενώ το δεύτερο όρισμα, coag.df, καθορίζει το πλαίσιο δεδομένων. Με την εντολή summary δίνεται ο πίνακας ΑΝΑΔΙΑ. Το αποτέλεσμα υποδεικνύει τη σημαντικότητα του παράγοντα οδηγώντας στο συμπέρασμα ότι υπάρχουν διαφορές ανάμεσα στις τέσσερις δίαιτες.

Όπως και στη γραμμική παλινδρόμηση, είναι χρήσιμη η γραφική ανάλυση των υπολοίπων για τον έλεγχο των υποθέσεων που απαιτούνται από την ανάλυση της διακύμανσης, δηλαδή να είναι ασυσχέτιστα, να έχουν σταθερή διακύμανση και να είναι κανονικά. Από τα γραφήματα (Σχήμα 9.3) φαίνεται ότι οι υποθέσεις αυτές ικανοποιούνται σε μεγάλο βαθμό.

```
> fitted.values(aov.coag)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
61 61 61 61 66 66 66 66 66 66 68 68 68 68 68 68 61 61 61 61 61 61
```

```
23 24
61 61
> par(mfrow=c(3.1))
> hist(resid(aov.coag))
> qqnorm(resid(aov.coag))
> qqline(resid(aov.coag))
> plot(fitted(aov.coag), resid(aov.coag))
> abline(h=0)
```



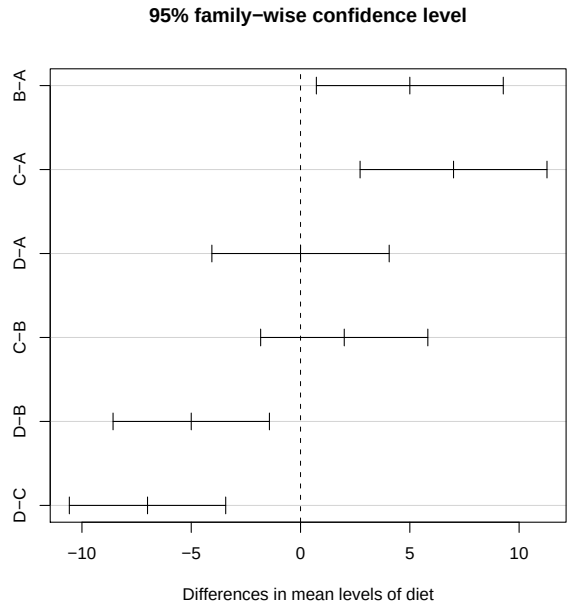
Σχήμα 9.3: Ανάλυση υπολοίπων.

9.2 Πολλαπλές Συγκρίσεις

Από την προηγούμενη ανάλυση επισημάνθηκε η διαφορά μεταξύ των επιπέδων του παράγοντα δίαιτα. Συνεπώς, είναι ενδιαφέρον να αναγνωριστούν αυτές οι διαφορές. Η κύρια μέθοδος πολλαπλών συγκρίσεων που χρησιμοποιείται στην R είναι η μέθοδος Tukey, η οποία εφαρμόζεται με την εντολή TukeyHSD. Η εντολή αυτή υπολογίζει τα 95% διαστήματα εμπιστοσύνης για όλα τα ζεύγη διαφορών ανάμεσα των μέσων τιμών των ειδών δίαιτας. Τα διαστήματα αυτά μπορούν να παρουσιαστούν και γραφικά για εποπτική σύγκριση, θέτοντας σαν όρισμα στην εντολή plot το αντικείμενο που παράγεται από την εντολή TukeyHSD. Όπως αναφέρθηκε και προηγουμένως, παρατηρείται ότι οι δίαιτες A και D σχηματίζουν μια κατηγορία, ενώ τα επίπεδα B και C μιαν άλλη κατηγορία, αφού το μηδέν περιέχεται στο διάστημα εμπιστοσύνης της διαφοράς τους.

```
> mca.coag <- TukeyHSD(aov.coag,"diet")
> mca.coag
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coag ~ diet, data = coag.df)

$diet
             diff         lwr       upr     p adj
B-A  5.000000e+00   0.7245544  9.275446 0.0183283
C-A  7.000000e+00   2.7245544 11.275446 0.0009577
D-A -1.421085e-14  -4.0560438  4.056044 1.0000000
C-B  2.000000e+00  -1.8240748  5.824075 0.4766005
D-B -5.000000e+00  -8.5770944 -1.422906 0.0044114
D-C -7.000000e+00 -10.5770944 -3.422906 0.0001268
> plot(mca.coag)
```

Γενικά, μπορούν να εφαρμοστούν και οι υπόλοιπες γνωστές μέθοδοι πολλαπλών συγκρίσεων (Dunnett, Sidak, Bonferroni και Scheffe) χρησιμοποιώντας τη βιβλιοθήκη της R multcomp.



Σχήμα 9.4: 95 % ταυτόχρονα διαστήματα εμπιστοσύνης των διαφορών των μέσων των επιπέδων του παράγοντα με τη μέθοδο Tukey.