|  |
| --- |
| Cane |

**Sugar-cane Disease Data**

**ΜΕΤΑΒΛΗΤΕΣ**

n

Ο συνολικός αριθμός των βλαστών σε κάθε οικόπεδο.

r

Ο αριθμός των άρρωστων βλαστών.

x

Ο αριθμός των τεμαχίων των μίσχων, από τα 50, που φυτεύτηκαν σε κάθε οικόπεδο.

var

Ένας παράγοντας που δείχνει την ποικιλία του ζαχαροκάλαμου σε κάθε παρατηρητήριο.

block

Ένας παράγοντας για τα μπλοκ.

|  |
| --- |
| Aids |

**Delay in AIDS Reporting in England and Wales**

**ΜΕΤΑΒΛΗΤΕΣ**

year

Το έτος της διάγνωσης.

quarter

Το τέταρτο του έτους κατά το οποίο έγινε η διάγνωση.

delay

Η χρονική καθυστέρηση (σε μήνες) μεταξύ διάγνωσης και αναφοράς. 0 σημαίνει ότι η υπόθεση αναφέρθηκε μέσα σε ένα μήνα. Οι μεγαλύτερες καθυστερήσεις ομαδοποιούνται σε διαστήματα 3 μηνών και η τιμή της καθυστέρησης είναι το μέσο του διαστήματος (επομένως η τιμή 2 δηλώνει ότι η αναφορά καθυστέρησε για διάστημα μεταξύ 1 και 3 μηνών).

dud

Ένας δείκτης λογοκρισίας. Πρόκειται για κατηγορίες για τις οποίες δεν υπάρχουν ακόμη πλήρεις πληροφορίες και ο αριθμός που καταγράφεται είναι μόνο χαμηλότερο όριο.

time

Το χρονικό διάστημα της διάγνωσης. Αυτός είναι ο αριθμός των τριμήνων από τον Ιούλιο του 1983 μέχρι το τέλος του τριμήνου κατά το οποίο διαγνωσθούν αυτές οι περιπτώσεις.

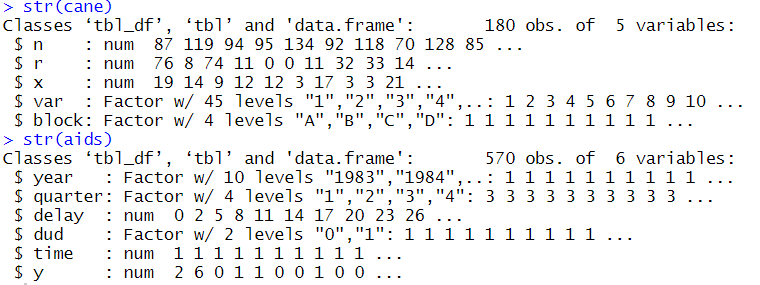
y

Ο αριθμός των περιπτώσεων AIDS που αναφέρθηκαν.

#1. Χαρακτηρίζουμε τις μεταβλητές ως προς το είδος του (για όσα σύνολα δεδομένων

# θα χρησιμοποιήσουμε).

Κανω αλλαγη τυπου μεταβλητης σε αυτες που χρειαζεται



#Για το CANE εχω τις ποσοτικες n,r,x (:numeric) και τις ποιοτικες var,block (:factor)

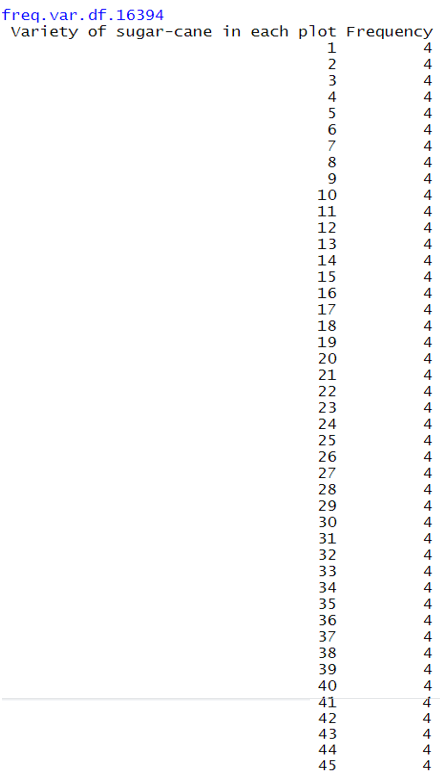
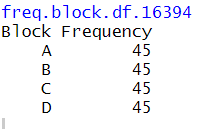
#Για το AIDS εχω τις ποσοτικες delay,time,y (:numeric) και τις ποιοτικες year,quarter,dud (:factor)

#2. Κατασκευάζουμε τον πίνακα συχνοτήτων και σχετικών συχνοτήτων για τις ποιοτικές

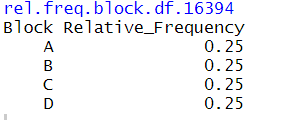
#μεταβλητές και ένα ραβδόγραμμα ή ένα κυκλικό διάγραμμα.

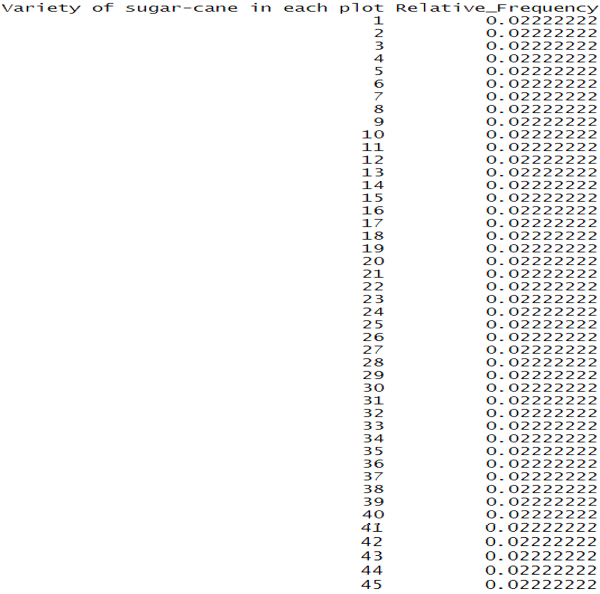
Επιλέγω να κάνω τη διαδικασία για τη μεταβλητή block και var του dataset cane

**# Πίνακας συχνοτήτων για τις ποσοτικες μεταβλητες block και var του dataset cane**

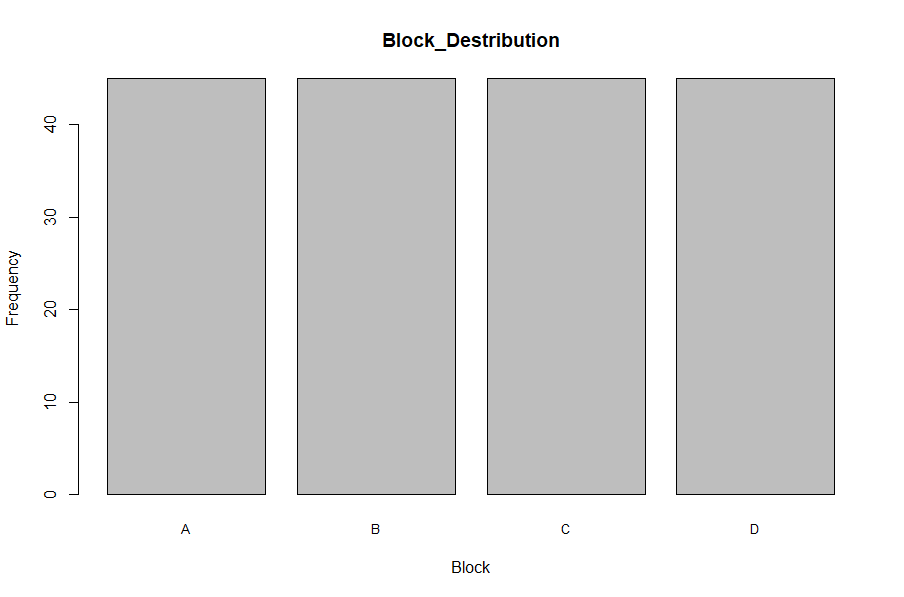
****

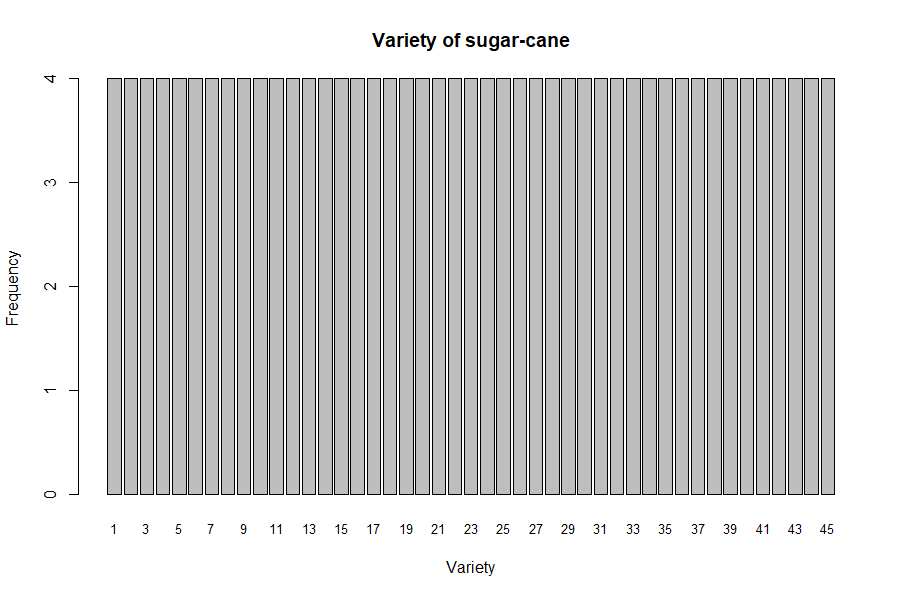
**# Πίνακας σχετικων συχνοτήτων για τις ποσοτικες μεταβλητες block και var του dataset cane**





**Ραβδόγραμμα** **για τις ποσοτικες μεταβλητες block και var του dataset cane**





#3. Κατασκευάζουμε Crosstabulation Matrix όπου στις στήλες θα έχει μια ποιοτική

#μεταβλητή και στις γραμμές μια άλλη ποιοτική μεταβλητή των δεδομένων σας.

#Επισης κατασκευάζουμε ένα συγκριτικό ραβδογράμμα. Στην συνέχεια με το chi-squared contingency

#table test εξετάζουμε αν οι δυο μεταβλητές είναι εξαρτημένες ή όχι.Οπτικοποιήουμε το

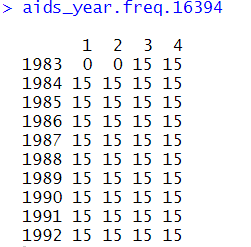
#αποτέλεσμα του ελέγχου με την gginference.

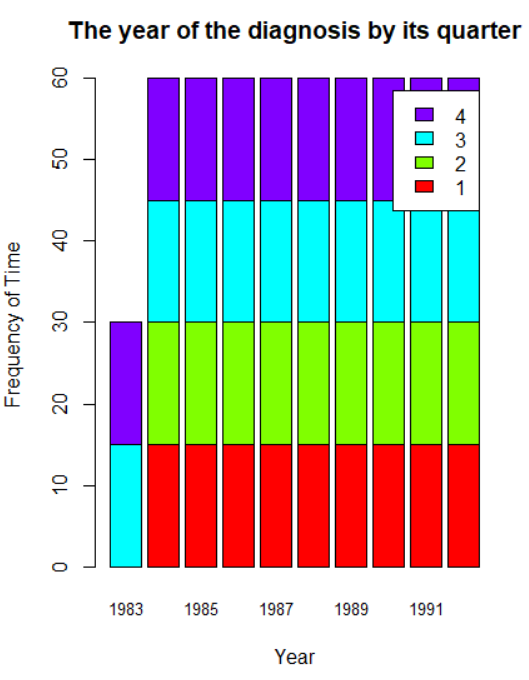
#Γράφουμε την μηδενική και την εναλλακτική υπόθεση καθώς και το συμπέρασμα του

#παραπάνω ελέγχου.

Επιλέγω να κάνω τη διαδικασία για τη ποιοτικη μεταβλητή year και quarter του dataset aids

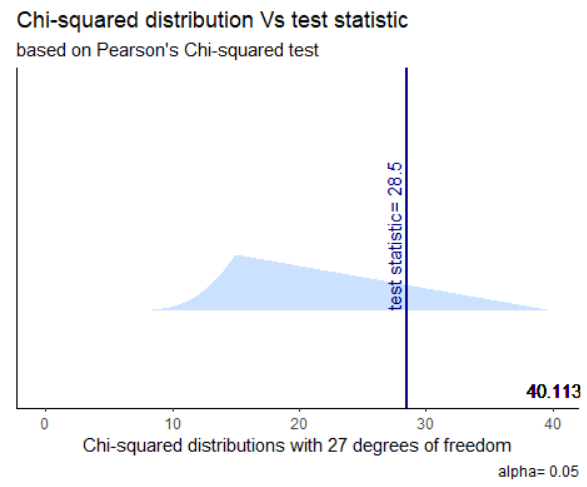
**Crosstabulation Matrix**



**#Συγκρητικο ραβδογραμμα**

To pvalue του χ^2 τεστ ανεξαρτησιας είναι μεγαλύτερο του 0,05 (p-value = 0.3856) άρα δέχομαι τη μηδενική υπόθεση και συνεπώς οι δύο μεταβλητές μου είναι ανεξάρτητες.

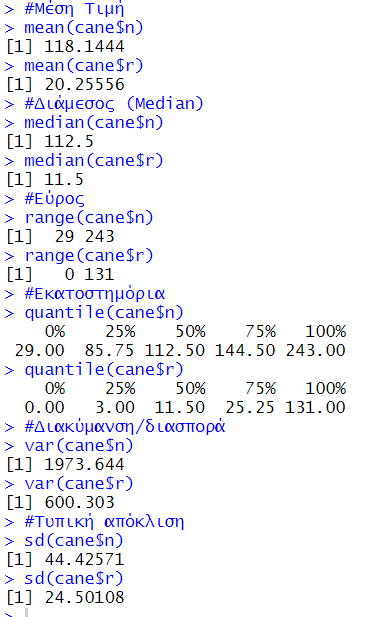
Ακολουθως θα προχωρησω σε ενα οπτικο ελεγχο του χ^2 τεστ ανεξαρτησιας

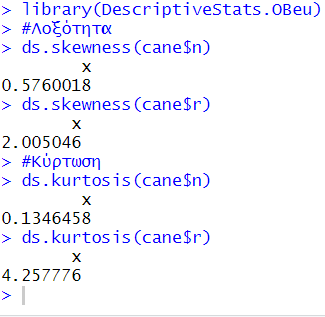


#4. Υπολογίζουμε για δυο ποσοτικές μεταβλητές τις εξής παραμέτρους: μέση τιμή, διάμεσο,

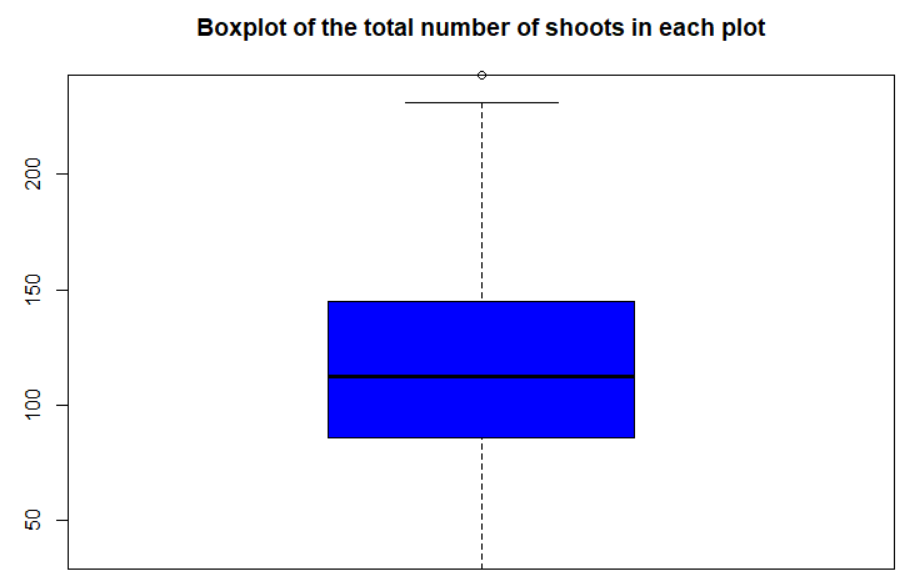
#τυπική απόκλιση, διασπορά, λοξότητα, κύρτωση, εύρος και ποσοστημόρια. Κάνουμε το

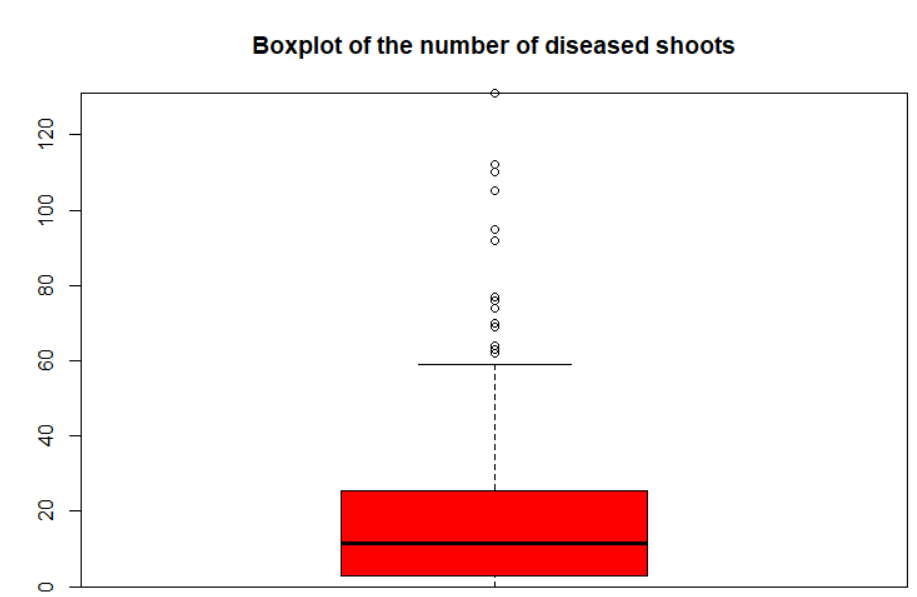
#αντίστοιχο θηκόγραμμα για κάθε μια ξεχωριστά καθώς και ένα σημειογραμμα (scatter plot).#Επιλέγω να κάνω τη διαδικασία για τη μεταβλητή n και r του dataset cane



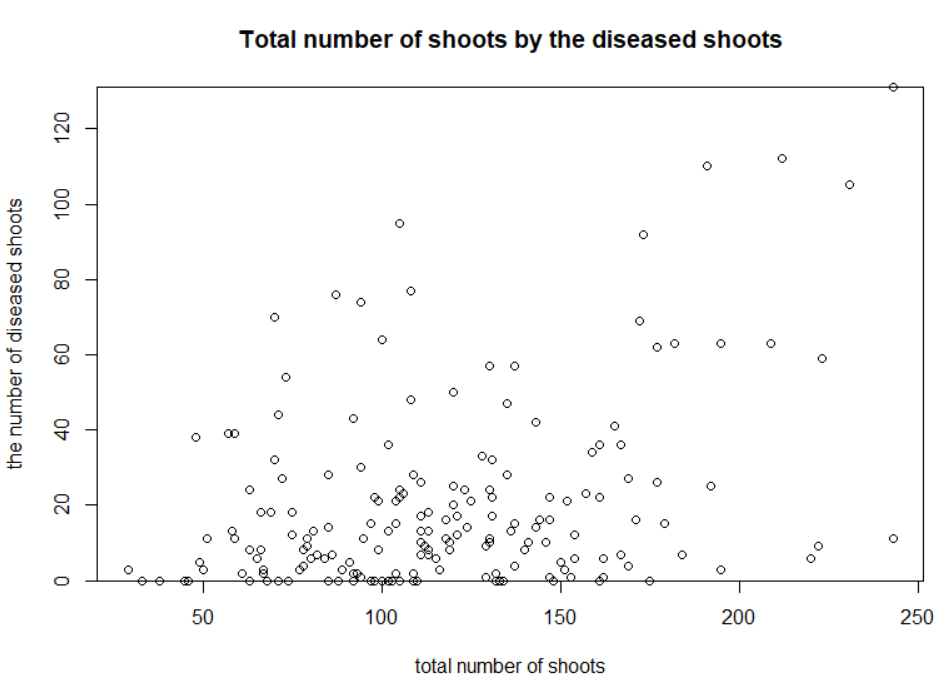


**#Θηκογραμμα για τη μεταβλητή n και r του dataset cane**





**#Συγκρητικο ραβδογραμμα**



**ΚΑΤΑΝΟΜΗ**

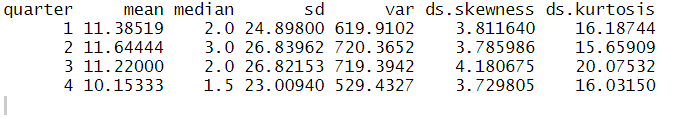
Παρατηρούμε ότι και στις δύο μεταβλητές ο δειγματικός μέσος είναι μεγαλύτερος από τη διάμεσο άρα έχουμε θετική ασυμμετρία (λοξότητα). Επίσης συντελεστής κύρτωσης στην μεταβλητή n είναι αρκετά μικρότερος από την τιμή 3 συνεπώς η κατανομή είναι πλατυκυρτη ενώ αντίθετα στην μεταβλητή r είναι μεγαλύτερος από 3 άρα η κατανομή είναι λεπτοκυρτη.

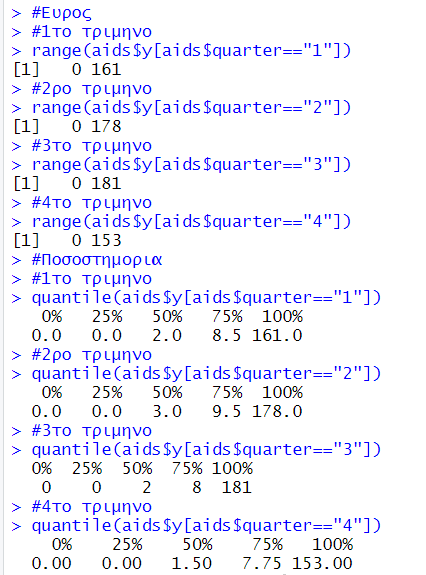
#5. Υπολογίζουμε για μια ποσοτική μεταβλητή σε κάθε δείγμα που προκύπτει από μια ποιοτική

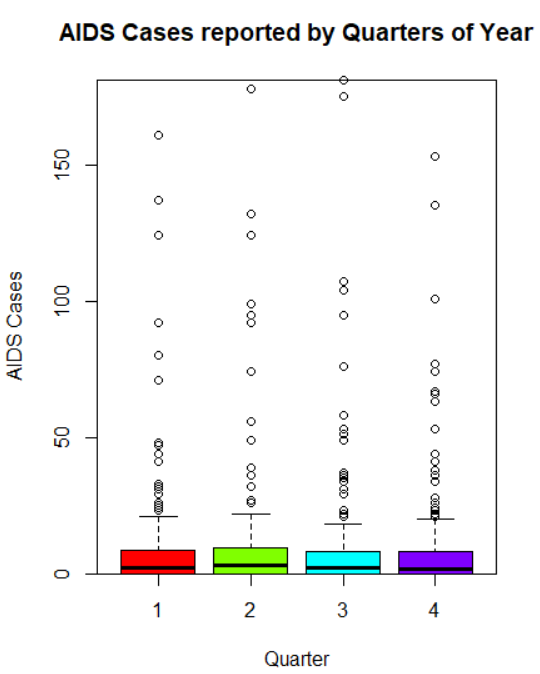
#(π.χ. age με case=0 και case=1) τις εξής παραμέτρους: μέση τιμή, διάμεσο, τυπική απόκλιση,

#διασπορά, λοξότητα, κύρτωση, εύρος και ποσοστημόρια. Κάνουμε επίσης και τα αντίστοιχο

#συγκριτικό θηκόγραμμα (1 γράφημα με δυο θηκογράμματα).ΤΡΙΜΗΝΟ ΜΕΣΗ ΤΙΜΗ ΔΙΑΜΕΣΟΣ ΤΥΠ.ΑΠΟΚΛΗΣΗ ΔΙΑΚΥΜΑΝΣΗ Σ.ΛΟΞΟΤΗΤΑΣ Σ.ΚΥΡΤΩΣΗΣ



Παρατηρώ ότι το εύρος και τα ποσοστημορια δεν μπορώ να το υπολογίσω με το compare st ats συνεπώς θα πρέπει να το βρω ξεχωριστά για κάθε περίπτωση

**Συγκριτικό θηκόγραμμα**

**ΚΑΤΑΝΟΜΗ**

Παρατηρούμε ότι και στις 4 περιπτωσεις ο δειγματικός μέσος είναι μεγαλύτερος από τη διάμεσο άρα έχουμε θετική ασυμμετρία (λοξότητα) σε ολες τις περιπτωσεις. Επίσης συντελεστής κύρτωσης και στις 4 περιπτωσεις είναι αρκετα μεγαλύτερος από 3 άρα η κατανομή είναι λεπτοκυρτη σε ολες τις περιπτωσεις.

ΑΣΚΗΣΗ 6

#6. Ελέγχουμε σε επίπεδο εμπιστοσύνης 95% με την βοήθεια του t-test τη μηδενική υπόθεση

#σύμφωνα με την οποία η μέση τιμή μιας ποσοτικής μεταβλητής δεν διαφέρει μεταξύ των δυο

#group μιας διχοτομικής ποιοτικής (δύο κατηγοριών). Οπτικοποιήουμε το αποτέλεσμα του

#ελέγχου με την gginference.

#Παρατηρούμε ότι δεν μπορούμε να χρησιμοποιήσουμε το data set cane γιατί δεν υπάρχει διχοτομική ποιοτική μεταβλητή.

#Άρα αναγκαστικά θα πρέπει να χρησιμοποιήσουμε το data set aids με τη διχοτομικη μεταβλητη dud

#Σύμφωνα με την άσκηση υποθέτουμε ότι η ποσοτική μεταβλητή delay ακολουθεί κανονική κατανομή,

#συνεπώς μένει να ελέγξουμε την ανεξαρτησία και την ισοτητας διασπορας των μεταβλητών για να προχωρήσουμε στο t test#Παρατηρούμε ότι δεν μπορούμε να χρησιμοποιήσουμε το data set cane γιατί δεν υπάρχει διχοτομική ποιοτική μεταβλητή.

#Άρα αναγκαστικά θα πρέπει να χρησιμοποιήσουμε το data set aids με τη διχοτομικη μεταβλητη dud

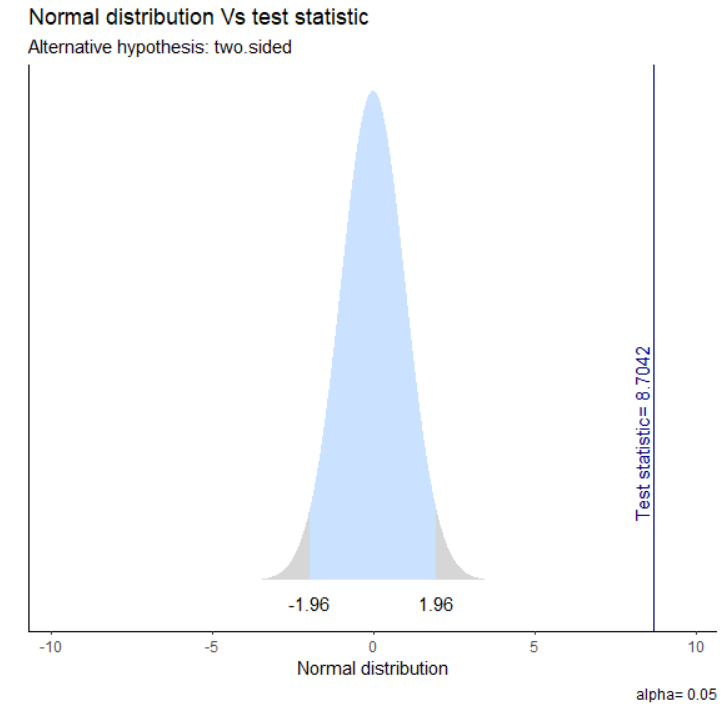
#Σύμφωνα με την άσκηση υποθέτουμε ότι η ποσοτική μεταβλητή delay ακολουθεί κανονική κατανομή,

#συνεπώς μένει να ελέγξουμε την ανεξαρτησία και την ισοτητας διασπορας των μεταβλητών για να προχωρήσουμε στο t test

#Το pvalue είναι αρκετα μικροτερο του 0.05 συνεπώς απορριπτω τη μηδενική υπόθεση ισοτητας διακυμανσεων άρα εχουν ανισες διακυμανσεις

#Το pvalue είναι μεγαλύτερο του 0.05 συνεπώς δέχομαι τη μηδενική υπόθεση του χ^2 τεστ ανεξαρτησιας άρα οι μεταβλητες ειναι ανεξαρτητες

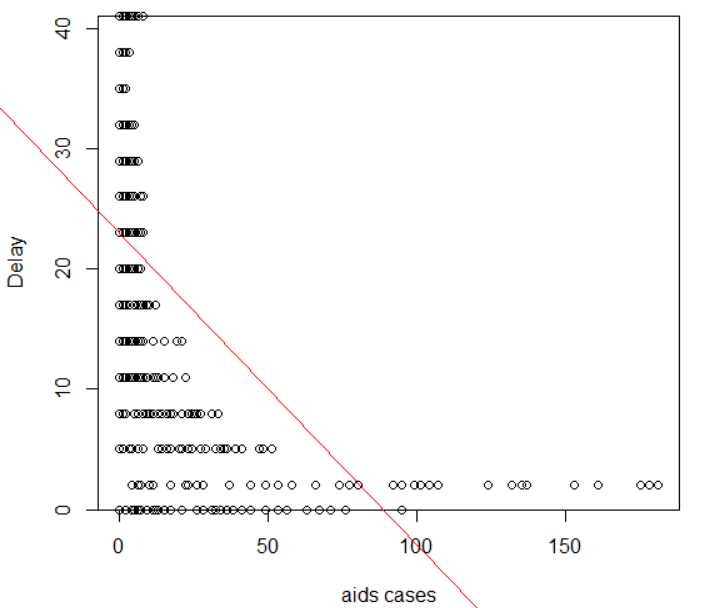
#Το pvalue είναι αρκετα μικροτερο του 0.05 συνεπώς απορριπτω τη μηδενική υπόθεση (Η0=μ1-μ2=0 ) άρα οι μεσες τιμες δεν ειναι ισες

Ακολουθως θα προχωρησω σε ενα οπτικο ελεγχο του t-test

#7. Κατασκευάζουμε το σημειόγραμμα (scatterplot) δυο μεταβλητών της επιλογής σας. Υπολογίζουμε τον συντελεστή συσχέτισης του Pearson.

**#διαγραμμα διασπορας**

**#σημειογραμα**



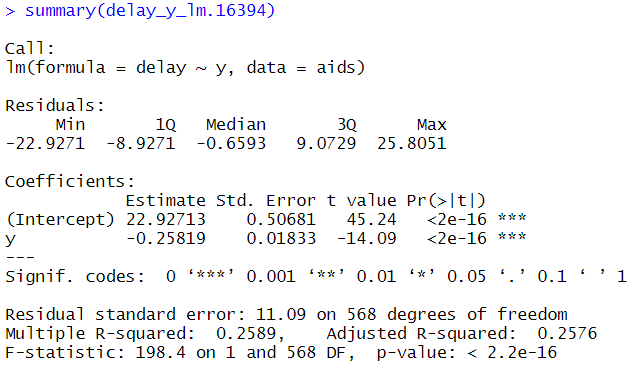
#Για να εξετάστει αν υπάρχει γραμμική σχέση μεταξύ των τιμών της μεταβλητής delay και y

#θα πρέπει να ελέγξουμε αν υπάρχει η ευθεία ψ = α + β Χ όπου Χ=delay και ψ = y ώστε

#να προβλέπουμε την τιμή της y (εξαρτημένη) γνωρίζοντας την τιμή της delay (ανεξάρτητη)

**#ευθεια γραμμικης παλινδρομισης**

#delay= 22.9271- 0.2582 y



#η μεση τιμη = -0.6593 δεν απεχει πολυ απτο 0

# τα σφάλματα που καθορίζουν την ακρίβεια των συντελεστών ειναι (Std. Error t value

#FOR delay 0.50681

#FOR y 0.01833

#Το p value είναι αρκετά μικρότερο του 0,05 και στις δύο περιπτώσεις συνεπώς απορρίπτω την μηδενική υπόθεση (Η0= ο συντελεστής είναι ίσος με μηδέν) αρα οι συντελεστες ειναι ανισοι του 0

# το τυπικό σφάλμα είναι σχετικα μεγαλο, συνεπως η ευθεία παλινδρόμησης δεν δίνει καλή περιγραφή της σχέσης μεταξύ των μεταβλητών

#Residual standard error: 11.09 on 568 degrees of freedom

#Η τιμή του συντελεστή προσδιορισμού είναι αρκετά μακριά στη μονάδα άρα το μοντέλο δεν έχει καλή προσαρμογή στα δεδομένα

#Multiple R-squared: 0.2589

#Ο συντελεστής συσχέτισης του PEARSON(περνει τιμες απο -1 μεχρι 1) είναι -0.5088282 συνεπώς βλέπουμε ότι υπάρχει αρνητική γραμμική συσχέτιση

#Από τα παραπάνω δεδομένα συμπεραίνουμε ότι εχουμε αρνητικη γραμμική συσχέτιση οχι τοσο σημαντικη ομως, άρα δεν θα ηταν τοσο αποτελεσματικο

#να κατασκευάσουμε μοντέλο γραμμικής παλινδρόμησης να προβλέπει μία μεταβλητή δίνοντας την άλλη.