**Final Project for HY390.51**

Tsainis Kyriakos

AM: 3170

Θεωρητικές ερωτήσεις:

1. Τιμή σημαντικότητας - pvalue: Στις στατιστικές μελέτες το p-value είναι η πιθανότητα της απόκτησης ενός αποτελέσματος ίσου ή πιο ακραίου από ότι ήταν στην πραγματικότητα παρατηρήσιμο, όταν η μηδενική υπόθεση είναι αληθής.

Tο pvalue είναι ευρέως χρησιμοποιούμενο στον στατιστικό έλεγχο υποθέσεων, ειδικά σε σημαντικές δοκιμές μηδενικής υπόθεσης. Σε αυτή τη μέθοδο, ως μέρος του πειραματικού σχεδιασμού, πριν από την εκτέλεση του πειράματος, πρώτα επιλέγει ένα μοντέλο ( null hypothesis) και μια οριακή τιμή για το σ, το οποίο ονομάζεται [επίπεδο σημαντικότητας](https://el.wikipedia.org/wiki/%CE%A3%CF%86%CE%AC%CE%BB%CE%BC%CE%B1%CF%84%CE%B1_%CF%84%CF%8D%CF%80%CE%BF%CF%85_%CE%99_%CE%BA%CE%B1%CE%B9_%CE%99%CE%99) του τεστ, παραδοσιακά, 5% ή 1% και συμβολίζεται ως α. **Εάν η τιμή σημαντικότητας είναι μικρότερη ή ίση με το επιλεγμένο επίπεδο σημαντικότητας α, η δοκιμή δείχνει ότι τα παρατηρούμενα δεδομένα δεν συνάδουν με τη μηδενική υπόθεση**, οπότε η μηδενική υπόθεση πρέπει να απορριφθεί.

Ωστόσο, αυτό δεν αποδεικνύει ότι η υπόθεση που εξετάζεται είναι η αλήθεια. Όταν η τιμή σημαντικότητας υπολογίζεται σωστά, αυτή η δοκιμή εγγυάται ότι το ποσοστό σφάλματος Τύπου I είναι το πολύ α.

Για τυπική ανάλυση, χρησιμοποιώντας το πρότυπο α = 0.05 διακοπών, η μηδενική υπόθεση απορρίπτεται όταν p < 0.05 και δεν απορρίπτεται όταν p > 0.05. Οι τιμές σημαντικότητας είναι μόνο ένα εργαλείο για να αποφασίσει εάν πρόκειται να απορρίψετε τη μηδενική υπόθεση.

1. α.

Η μηδενική υπόθεση είναι μία γενική δήλωση ότι δεν υπάρχει σχέση μεταξύ δύο μετρήσιμων αντικειμένων. Στο συγκεκριμένο παράδειγμα έχουμε φύλο και ύψος οπότε η μηδενική υπόθεση αναφέρει ότι τα αγόρια και τα κορίτσια έχουν το ίδιο ύψος.

Ακόμα, η μηδενική υπόθεση είναι μία δήλωση που κάνουμε γενικά για τους πληθυσμούς από τους οποίους λαμβάνεται το δείγμα και όχι για το δείγμα το ίδιο.

Η μηδενική υπόθεση είναι συνήθως το αντίθετο από αυτό που μας ενδιαφέρει.

β.

Γνωρίζουμε ότι έχουμε το MeanBoys(MB) , MeanGirls(MG),

t = (MG – MB)/ a\_function\_of\_VARIANCES   
και άρα πάμε να κάνουμε τις υποθέσεις μας.

Έστω ότι k = MG – MB

Έχοντας το διάγραμμα για την correct null hypothesis παίρνουμε τα παρακάτω.

Ας υποθέσουμε ότι το k = 1 και

* κάνουμε την υπόθεση ότι τα κορίτσια είναι ψηλότερα από τα αγόρια.

Τότε το pvalue μας θα είναι το μέρος του διαγράμματος για το οποίο ισχύει k>=1

* κάνουμε την υπόθεση ότι τα κορίτσια και τα αγόρια δεν έχουν το ίδιο ύψος. Τώρα το pvalue μας θα έιναι το μέρος του διαγράμματος για το οποίο ισχύει k<=1 ή k>=1.

Αυτό προκύπτει διότι αφού δεν έχουν το ίδιο ύψος θα πρέπει να πάρουμε και τα δύο cases, δηλαδή είτε τα κορίτσια να είναι ψηλότερα είτε τα αγόρια.

* Κάνουμε την υπόθεση ότι τα αγόρια είναι ψηλότερα από τα κορίτσια. Τώρα το pvalue μας θα είναι το μέρος του διαγράμματος για το οποίο ισχύει k<=1.

1. Τρέχοντας τις δύο περιπτώσεις επί 5 φορές πήραμε τα παρακάτω αποτελέσματα:

|  |  |
| --- | --- |
| **Α περίπτωση** | **Β περίπτωση** |
| 0,4558 | <2,2e-16 |
| 0,01779 | <2,2e-16 |
| 0,4136 | <2,2e-16 |
| 0,6469 | <2,2e-16 |
| 0,8631 | <2,2e-16 |

Με βάση τα αποτελέσματα που παίρνουμε για τα **p-values** βλέπουμε ότι στην **Β** **περίπτωση** το **p-value** θα είναι πάντα πολύ μικρότερο από **0.05**.

Παρόλο που και τα δύο δείγματα προκύπτουν από πληθυσμούς με μέση τιμή 100 και 101 στην πρώτη περίπτωση δεν μπορούμε να απορίψουμε την μηδενική υπόθεση.   
Απτην άλλη μερία στην **Β περίπτωση** βλέπουμε ότι μπορούμε να απορίψουμε την μηδενική υπόθεση αφού το **p-value** είναι **πάντα** μικρότερο του **0.05** και αυτό προκύπτει επειδή το μέγεθος του δείγματος **στην δευτερη περίπτωση είναι πολύ μεγαλύτερο από την Α** (10,100000) .

1. a) Στον κώδικα που μας δίνεται στο **pvalues** μπαίνει το αποτέλεσμα της **sapply**.

Η **sapply** φτιάνει ένα vector 10000 στοιχείων και κάνει apply σε κάθε ένα από τα στοιχεία την function(x).

Η function x είναι μία συνάρτηση που φτιάχνει δύο vector τα a,b τα οποία περιέχουν 100 δείγματα με μέση τιμή=100 και τυπική απόκλιση=10 και στην συνέχεια κάνει την t.test για να μας δώσει τα **pvalue** ελεγχοντας την υπόθεση ότι η διαφορά των μέσων όρων του **a** και του **b** είναι **μεγαλύτερη** από 0. Άρα στο τέλος η pvalues θα περιέχει όλα τα pvalue αφού έχουμε κάνει το t.test για καθε ένα από αυτά.

b)

a. Οι περιπτώσεις αυτές είναι περίπου 500.

b. Με βάση το δεδομένο που έχουμε ότι το επίπεδο σημαντικότητας είναι το 0.05 και

αφού εξετάζουμε 10.000 διαφορετικά στοιχεία τότε θα αναμένουμε περίπου   
0.05 \* 10.000 = 500 να έχουν pvalue < 0.05.

c. Η εκτίμηση μας από ότι φαίνεται ταυτίζεται με τον αριθμό των περιπτώσεων που βρήκαμε παραπάνω.

d. Την απάντηση σε αυτήν την ερώτηση την δίνει ο ορισμός του multiple testing problem όπως είχαμε δει κ σε προηγούμενες ασκήσεις. Το πρόβλημα αυτό προκύπτει όταν κάνουμε t-test σε πολλά δείγματα δημιουργώντας έτσι πολλά τυχαία pvalue < 0.05 απόρρίπτοντας πολλές φορές και true null hypothesis (type I,II errors).

e.

Με το multiple test correction προφανώς επιλύουμε το πρόβλημα που προέκυψε παραπάνω. Οσο μεγαλύτερο είναι το δείγμα που κάνουμε t-test τόσο μεγαλύτερη είναι η πιθανότητα για να προκύψει type I error και άρα rejection of a true null hypothesis. Για να το αποφύγουμε αυτό κάνουμε το multiple testing correction που μας βοηθάει να κάνουμε control τα false positives όταν κάνουμε πολλά t-test.

**Πρακτικό μέρος**

1)

data<- read.table("GDS4879.clean", header=TRUE) #read dataset

data <- data[,-c(1,2)] #delete first 2 columns

data<-log2(data) #do log2 to all data

dim(data)

2)

Οι τιμές έκφρασης πρέπει να μετατραπούν σε log2 κλίμακα διότι κάνει πολύ πιο ξεκάθαρη την κατανομή αφού τα δεδομένα μας έχουν πολύ μικρές διαφορές μεταξύ τους. Αν δεν το κάνουμε log2 παρατηρούμε ότι το σχήμα δεν είναι ευδιάκριτο καθώς όλες οι τιμές πέφτουν η μία πάνω στην άλλη.

raw.data<- read.table("GDS4879.clean",sep="\t", header=TRUE)

data <- raw.data[,-c(1,2)]

#boxplot without log2

pdf("boxplot.pdf")

boxplot(data)

dev.off()

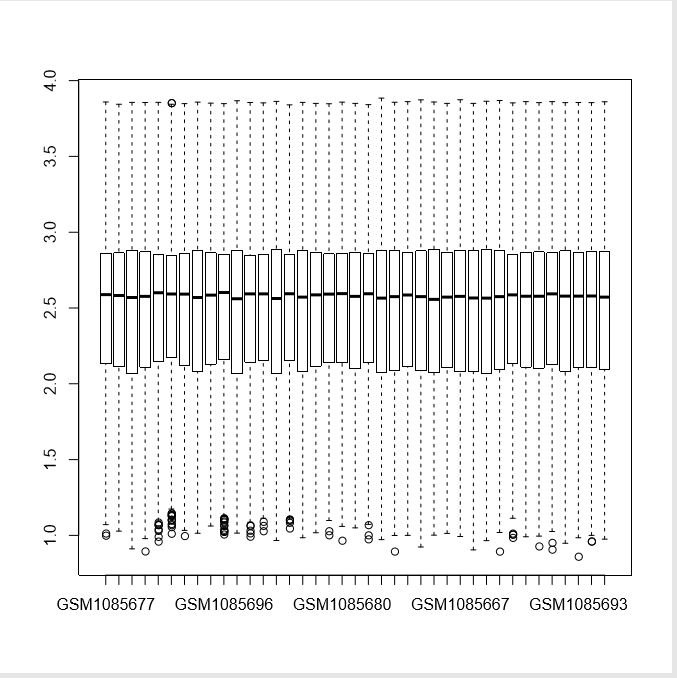
#correct boxplot with log2

data2 <- log2(data)

pdf("boxplot2.pdf")

boxplot(data2)

dev.off()



7)

