ΟΝΟΜΑΤΕΠΩΝΥΜΟ : Κωνσταντίνος Σπυριδάκης

AM : csd3650

**HY-390.51 ΤΕΛΙΚΟ PROJECT 2020**

Θεωρητικές Ερωτήσεις

1. P-Value

Η τιμή σημαντικότητας(*p*-value) ορίζεται ως η πιθανότητα της απόκτησης ένος αποτελέσματος ίσου ή "πιο ακραίου" από ό,τι ήταν στην πραγματικότητα παρατηρήσιμο, όταν η μηδενική υπόθεση είναι αληθής.

1. α)Null-Hypothesis

Σε συμπεράσματα στατιστικών στοιχείων, η μηδενική υπόθεση (H0) είναι μια γενική δήλωση ή μια προεπιλεγμένη θέση οτι δεν υπάρχει σχέση μεταξύ δύο μετρημένων φαινομένων ή καμία συσχέτιση μεταξύ ομάδων.Δοκιμάζοντας την null hypothesis καταλήγουμε στο συμπέρασμα οτι υπάρχουν (ή όχι) αποδείξεις σχετικά με το αν υπάρχει σχέση μεταξύ δυο φαινομένων.Πιο συγκεκριμένα, το πεδίο της στατιστικής παρέχει ακριβή κριτήρια για την απόρριψη ή αποδοχή της Null hypothesis .

β)

* Στο πρώτο πλοτ βλέπουμε οτι ο μέσος όρος ύψους των γυναικών είναι μεγαλύτερος απο τον μέσο όρο ύψους των ανδρών.
* Στο δεύτερο πλοτ βλέπουμε οτι ο μέσος όρος ύψους των ανδρών είναι μεγαλύτερος απο των γυναικών.
* Στο τρίτο πλοτ παρατηρούμε διαφορές στους μέσους όρους καθώς και οι δυο αποκλίνουν απο το 0 .

1. Περίπτωση α

p1-value = 0.1884 , p2-value = 0.4245 , p3-value = 0.6056 , p4-value = 0.461 , p5-value = 0.9731

Περίπτωση β

p1-value = 2.2e-16, p2-value = 2.2e-16, p3-value = 2.2e-16, p4-value = 2.2e-16, p5-value = 2.2e-16

Το p-value είναι πάντα μικρότερο απο 0,05 στην 2η περίπτωση . Αυτό συμβαίνει διότι όσο αυξάνεται το πλήθος των δοκιμών με μέσο όρο 100 και 101 τόσο μικραίνει η πιθανότητα η rnorm() να παράξει μικρότερα/ανόμοια αποτελέσματα αρα το P-value είναι μικρότερο.

1. Α)Η συνάρτηση sapply εκτελείται 10.000 φορές και αποθηκεύει στην μεταβλητή (vector) pvalues τα p-values που εξάγει το t.test με greater(τα μεγαλύτερα) με ορίσματα a,b τα οποία είναι τυχαίοι αριθμοί (δείγματα) απο ίδιους πληθυσμούς με μέση τιμή 100 και τυπική απόκλιση 10.

Β)

a) Αφού παράγει random κάθε φορά

Ενα τεστ έβγαλε 480 pvlaues < 0.05 ( sum(pvalues < 0.05) ).

b) Εχουμε 10k pvalues \* 0.05 = 500

c)Οσες φορές και να τρέξει ο κώδικας θα εξάγει αποτελέσματα κοντά στο 500.

d)Οταν εμφανίζεται p value < 0.05 ουσιαστικά απορρίπτουμε την Null Hypothesis όμως και οτι η σχέση στο δείγμα αντικατοπτρίζει το λεγόμενο sampling error.Sampling error όταν παρατηρούμε μεταβλητότητα σε ένα στατιστικό στοιχείο απο δείγμα σε δείγμα.

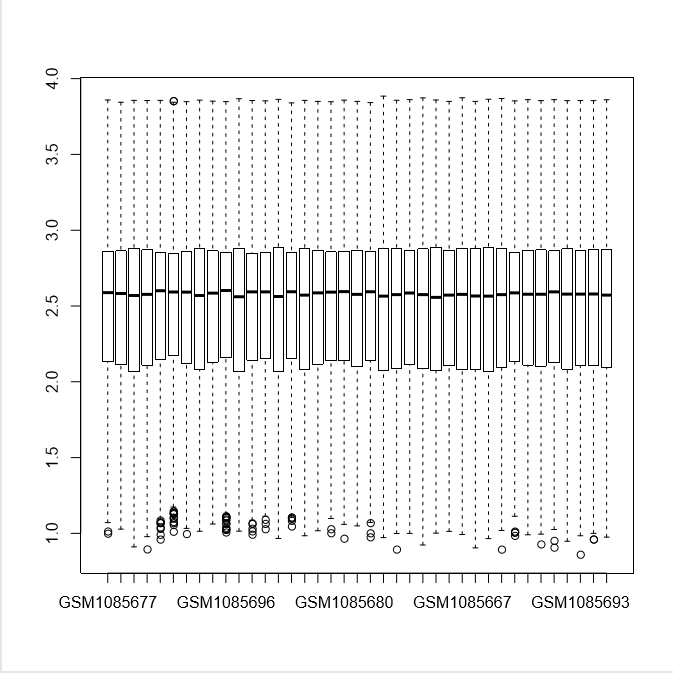
Κάθε στατιστική σχέση σε ένα δείγμα μεταφράζεται στα εξής

* Υπάρχει σχέση μεταξύ του πληθυσμού και η σχέση στο δείγμα το δείχνει αυτό.
* Δεν υπάρχει σχέση μεταξύ του πληθυσμού και η σχέση στο δείγμα αντικατοπτριζει μόνο το σφάλμα δειγματοληψίας (sampling error).

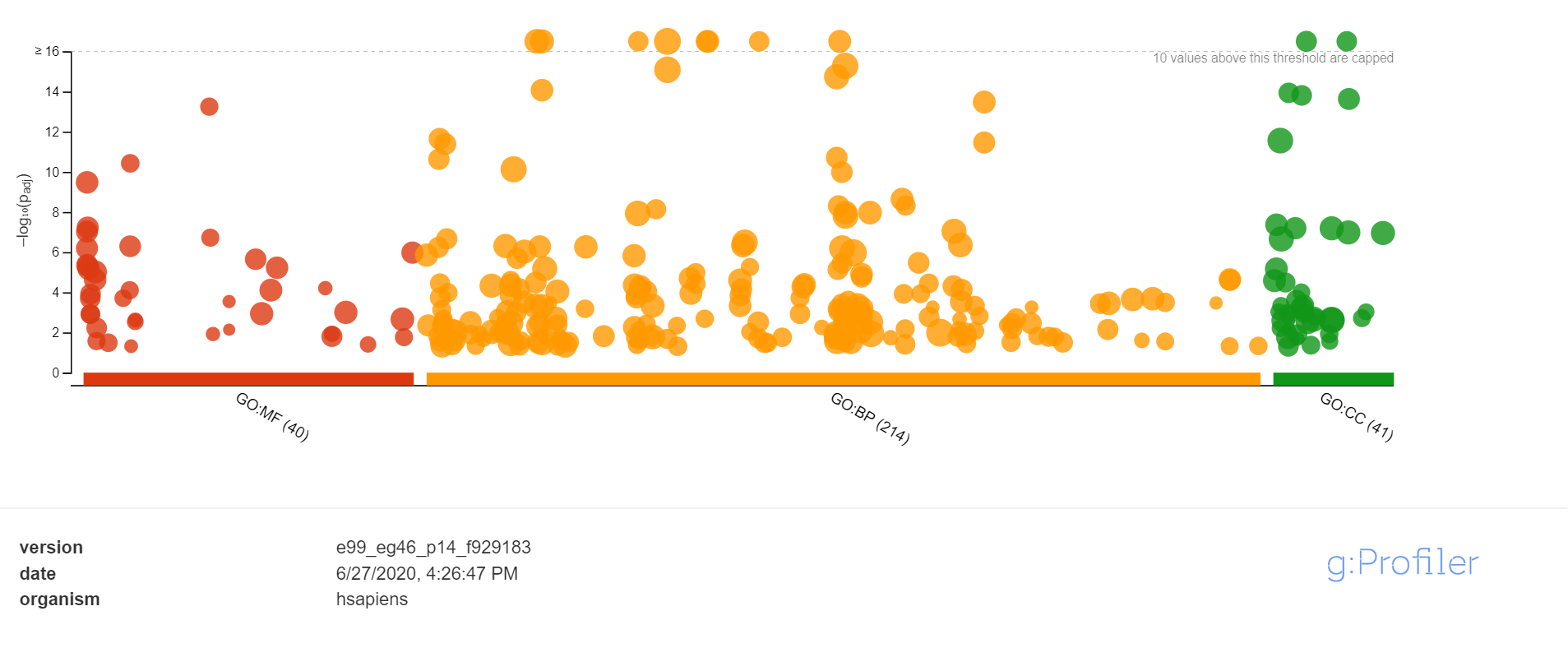
e)Τα δείγματα είναι απο ίδιους πληθυσμούς και είναι αλληλεξαρτόμενα.Ετσι οι διαφορές που βλέπουμε σε αποτελέσματα δείχνει οτι κάποια δεν είναι σωστά.Ετσι όταν κάνουμε multiple testing correction θα έχουμε αποτελέσματα που θα είναι λογικά.

Πρακτικό Μέρος.

2. Στο boxplot έχουμε μεγάλο εύρος τιμών απο πολύ μικρές έως πολύ μεγάλες.Ετσι δεν μπορούμε να έχουμε μια σαφή εικόνα των δεδομένων.Γι’αυτό μετατρέπουμε σε log2(x) τα δεδομένα μας και η εικόνα τους είναι πιο ομοιόμορφη.



5.Μετα την διόρθωση fdr βρίσκω 4 γονίδια ενώ με bonferroni 2 γονίδια με διορθωμένη τιμή και στις 2 περιπτώσεις.

6.Το request gprofiler(png\_fn = six.png) γυρνούσε fail λόγω μεγάλου όγκου δεδομένων , έτσι απλά έφτιαξα input για το <https://biit.cs.ut.ee/gprofiler/gost> με filters mp , cc,mf και export .

7.Ομαδοποίηση στα αποτελέσματα έχουμε περισσότερο στους μη αλκοολικούς και αλκοολικούς άντρες.Παρατηρούμε οτι πολλά είναι σχετικά κοντά (θέση μέση δεξιά).Για τα υπόλοιπα θεωρώ πως δεν μπορούμε να βγάλουμε σίγουρο συμπέρασμα καθώς είναι μικρό το δείγμα και διάσπαρτα τα αποτελέσματα.