ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΕΛΕΓΧΟΣ ΠΟΛΛΑΠΛΩΝ ΥΠΟΘΕΣΕΩΝ ΣΕ ΜΙΚΡΟΣΥΣΤΟΙΧΙΕΣ DNA

Μαργαρίτα-Αργεντίνα Ν. Παπακωνσταντίνου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Ιούνιος 2007

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΛΕΓΧΟΣ ΠΟΛΛΑΠΛΩΝ ΥΠΟΘΕΣΕΩΝ ΣΕ ΜΙΚΡΟΣΥΣΤΟΙΧΙΕΣ DNA

Μαργαρίτα-Αργεντίνα Ν. Παπακωνσταντίνου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Ιούνιος 2007

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- **Ηλιόπουλος Γεώργιος** (Επιβλέπων)
- **Κατέρη Μαρία**
- **Καφφές Δημήτριος**

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

MULTIPLE HYPOTHESIS TESTING IN MICROARRAY EXPERIMENTS

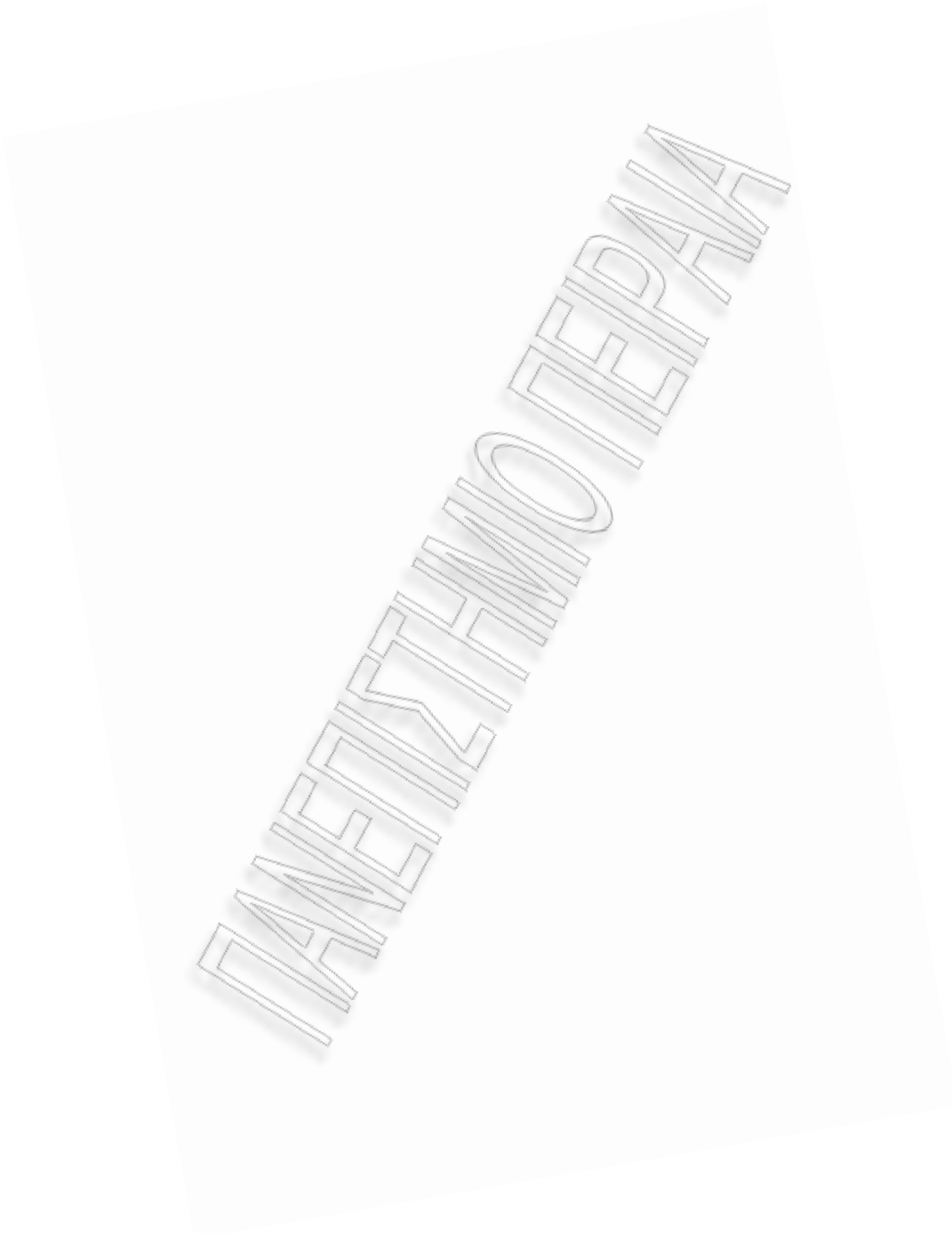
By

Margarita-Argentina Papakonstantinou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
June 2007


ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Στην οικογένειά μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κο Γεώργιο Ηλιόπουλο για την πολύτιμη βοήθειά του και την συνεργασία μας καθώς επίσης και τα μέλη της τριμελούς επιτροπής κα Μαρία Κατέρη και κο Δημήτριο Καφφέ. Επίσης, θα ήθελα να ευχαριστήσω την διδάκτορα Χαρά Δημοπούλου για τη βοήθειά της στη συγγραφή της ενότητας που αφορούσε στην επιστήμη της βιολογίας και της τεχνολογίας των μικροσυστοιχιών, καθώς επίσης και τον Κωνσταντίνο Παπακωνσταντίνου για τη βοήθειά του ως προς την παροχή των κατάλληλων υπολογιστικών προγραμμάτων.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, την Φραντζέσκα Παπά και την Αθηνά Παλιεράκη για την συμπαράστασή τους καθ' όλη την διάρκεια συγγραφής της διπλωματικής εργασίας.

Περίληψη

Οι μικροσυστοιχίες DNA αποτελούν τμήμα ενός καινούργιου και πολλά υποσχόμενου τομέα βιοτεχνολογίας ο οποίος μας επιτρέπει να παρατηρήσουμε ταυτοχρόνως χιλιάδες γονίδια και να καθορίσουμε ποια από αυτά είναι διαφορετικά εκφρασμένα σε ένα συγκεκριμένο τύπο κυττάρου. Το βιολογικό ερώτημα της διαφορετικής έκφρασης μπορεί να καθοριστεί εκ νέου σαν ένα πρόβλημα πολλαπλού ελέγχου υποθέσεων: του ταυτόχρονου ελέγχου, για κάθε γονίδιο, της μηδενικής υπόθεσης της μη ύπαρξης σχέσης μεταξύ των επιπέδων έκφρασης και ορισμένων αποκρίσεων ή συμμεταβλητών.

Σκοπός της παρούσας εργασίας είναι η διερεύνηση και η παρουσίαση διαφορετικών προσεγγίσεων του ελέγχου πολλαπλών υποθέσεων μέσα στο πλαίσιο των πειραμάτων μικροσυστοιχιών DNA και η παρουσίαση αλγορίθμων για την εφαρμογή της minP και maxT προσαρμογής με σκοπό τον έλεγχο της family wise πιθανότητας σφάλματος, όπως επίσης και ενός bootstrap αλγορίθμου για τον προσδιορισμό των ακατέργαστων τιμών-*p*.

Οι στατιστικές μέθοδοι εφαρμόστηκαν σε δύο πραγματικά σύνολα, από ασθενείς με λευχαιμία και από ασθενείς με καρκίνο του μαστού. Η ανάλυση πραγματοποιήθηκε χρησιμοποιώντας το Mathematica και το πρόγραμμα R.

Abstract

DNA microarrays are part of a new and promising class of biotechnologie that allows us to look at thousands of genes at once and determine which are expressed in a particular cell type. The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and some responses or covariates.

The purpose of the present study is to investigate and present different approaches to multiple hypothesis testing in the context of DNA microarray experiments and introduce algorithms for implementing the minP and maxT adjustment to control the family wise error rate as well as a bootstrap algorithm for the determination of the raw p-values.

These statistical methods were applied on two real data sets, from leukemia patients and patient with breast cancer. The analysis was carried out using the Mathematica and the R-package.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΕΙΣΑΓΩΓΗ

Η παρούσα διπλωματική εργασία με θέμα "Έλεγχος Πολλαπλών Υποθέσεων σε Μικροσυστοιχίες DNA" αποτελείται από τρία κεφάλαια. Στο πρώτο κεφάλαιο προσφέρονται στον αναγνώστη κάποιες βασικές γνώσεις από τη Βιολογία, όπως το τι είναι το DNA και το RNA, ποιος είναι ο τρόπος έκφρασης της γενετικής πληροφορίας που είναι αποθηκευμένη στο DNA, τι σημαίνει ο όρος έκφραση των γονιδίων. Επίσης, δίνεται μία περιγραφή της τεχνολογίας των μικροσυστοιχιών, τονίζεται η σημασίας της, και παρατίθενται τα κύρια είδη αυτής. Τέλος παρουσιάζονται τα δύο σύνολα δεδομένων που χρησιμοποιήθηκαν για την εφαρμογή των στατιστικών μεθόδων.

Στο δεύτερο κεφάλαιο παρουσιάζεται το πρόβλημα του ελέγχου πολλαπλών υποθέσεων. Ορίζονται οι πιθανότητες σφαλμάτων τύπου I, οι προσαρμοσμένες τιμές-p και δίνονται οι διαδικασίες για τη ρύθμιση της FWER και της FDR για πολλαπλά ανεξάρτητα αλλά και εξαρτημένα τεστ.

Στο τρίτο κεφάλαιο παρουσιάζονται διάφοροι αλγόριθμοι για την παραγωγή των προσαρμοσμένων και μη τιμών-p. Στους αλγορίθμους αυτούς η επαναδειγματοληψία γίνεται είτε με μεταθέσεις είτε με τη μέθοδο bootstrap. Τέλος πραγματοποιείται η ανάλυση των δεδομένων μέσω του Mathematica, της συνάρτησης MTP του προγράμματος R και του πακέτου multtest του Bioconductor και του πακέτου affylmGUI μέσω του προγράμματος R.

ΚΕΦΑΛΑΙΟ 1

1.1 DNA

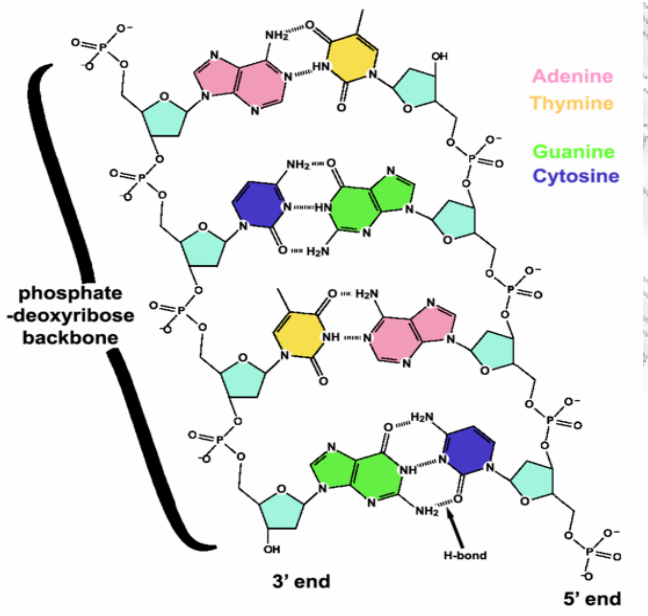
Το DNA (δεοξυριβονουκλεϊκό οξύ) είναι ένα μεγαλομόριο πολυδεοξυριβονουκλεοτιδίων, καθένα από τα οποία αποτελείται από ένα μεγάλο αριθμό μονομερών μονάδων, των δεοξυριβονουκλεοτιδίων. Τα δεοξυριβονουκλεοτίδια με τη σειρά τους αποτελούνται από μία αζωτούχο βάση, τύπου πουρίνης ή πυριμιδίνης, ένα σάκχαρο, τη β-D-2 δεοξυριβόζη και ένα φωσφορικό οξύ. Περιέχει τέσσερα είδη αζωτούχων βάσεων: τις αδενίνη (A) και γουανίνη (G) οι οποίες είναι τύπου πουρίνης και τις θυμίνη (T) και κυτοσύνη (C) οι οποίες είναι τύπου πυριμιδίνης. Τα δεοξυριβονουκλεοτίδια με τη σχετική αλληλουχία τους είναι οι φορείς της γενετικής πληροφορίας υπό τη μορφή γενετικού κώδικα.

Ο σκελετός του DNA, υπό μορφή αλυσίδας, αποτελείται από δεοξυριβόζες συνδεδεμένες με φωσφορικές ομάδες. Το ένα άκρο έχει μία ομάδα 5΄-OH και το άλλο άκρο του μία 3΄-OH ομάδα. Ειδικότερα η 3΄-OH (υδροξυλομάδα) του σακχάρου ενός δεοξυριβονουκλεοτιδίου ενώνεται με τη 5΄- OH (υδροξυλομάδα) του επόμενου σακχάρου με φωσφοδιεστερική γέφυρα. Κάθε δεοξυριβονουκλεοτίδιο του DNA έχει πολικότητα. Έχει επικρατήσει η σειρά των βάσεων να γράφεται με κατεύθυνση 5΄->3΄.

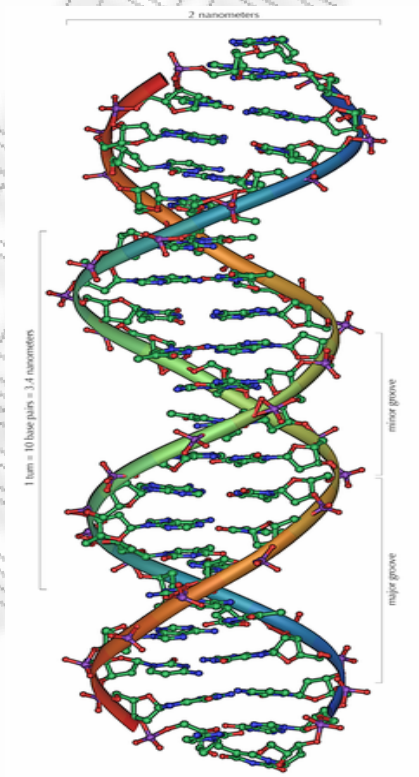
Δύο μονόκλωνες αλυσίδες DNA συγκρατούνται από δεσμούς υδρογόνου που αναπτύσσονται μεταξύ των βάσεων σύμφωνα με τον παρακάτω κανόνα: η γουανίνη ζευγαρώνει πάντα με την κυτοσίνη και η αδενίνη πάντα με την θυμίνη (σχήμα 1.1). Οι δύο αλυσίδες με συμπληρωματικές αλληλουχίες βάσεων και αντίθετη πολικότητα, περιελίσσονται η μία γύρω από την άλλη και δημιουργούν τη γνωστή έλικα του DNA (σχήμα 1.2).

Την τρισδιάστατη δομή του DNA που περιγράψαμε παραπάνω συμπέραναν οι James Watson και Francis Crick το 1953. Ανέλυσαν φωτογραφίες περίθλασης ακτίνων X από ίνες DNA που είχαν πάρει η Rosalind Franklin και ο Maurice Wilkins και πρότειναν ένα δομικό

μοντέλο που αποδείχτηκε αληθινό. Το επίτευγμα αυτό είναι από τα πιο σημαντικά στην ιστορία της βιολογίας διότι οδήγησε στην κατανόηση της λειτουργίας των γονιδίων σε μοριακό επίπεδο.



Σχήμα 1.1. Ειδικό ταίριασμα των βάσεων. Τα ζεύγη G-C είναι πιο σταθερά από τα A-T γιατί οι βάσεις συγκρατούνται με τρεις δεσμούς υδρογόνου και όχι με δύο.

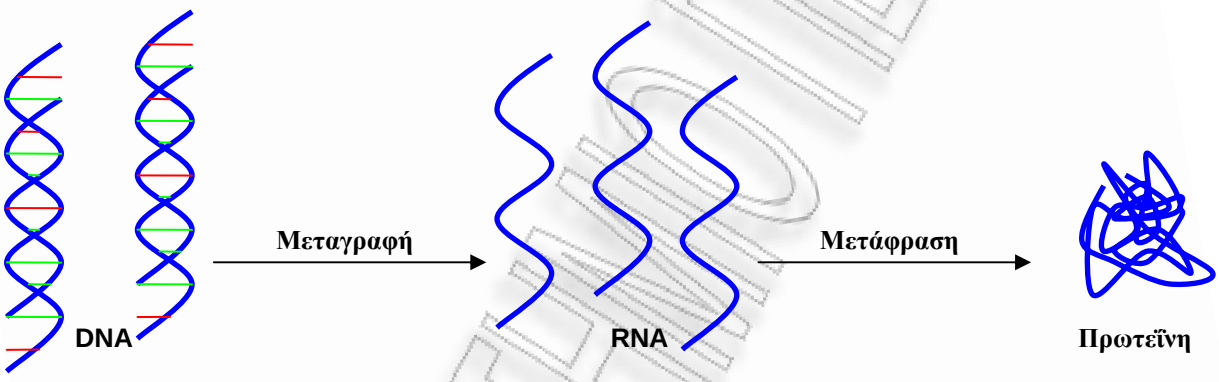


Σχήμα 1.2. Μοντέλο της διπλής έλικας του DNA.

Η έκφραση της γενετικής πληροφορίας που είναι αποθηκευμένη στο μόριο του DNA γίνεται σε δύο στάδια:

α) τη μεταγραφή (transcription), κατά την οποία τα εκμαγεία DNA δίνουν την πληροφορία για να συντεθούν τα αγγελιοφόρα RNA (mRNA) τα οποία είναι ενδιάμεσα μόρια που μεταφέρουν την πληροφορία για τη σύνθεση των πρωτεϊνών. Επίσης, άλλες μορφές κυτταρικού RNA που συμμετέχουν στο μηχανισμό αυτόν είναι το μεταφορικό RNA (tRNA) και το ριβοσωμικό RNA (rRNA).

β) την μετάφραση (translation), κατά την οποία οι πρωτεΐνες συντίθενται σύμφωνα με πληροφορίες που παρέχονται από τα εκμαγεία mRNA.



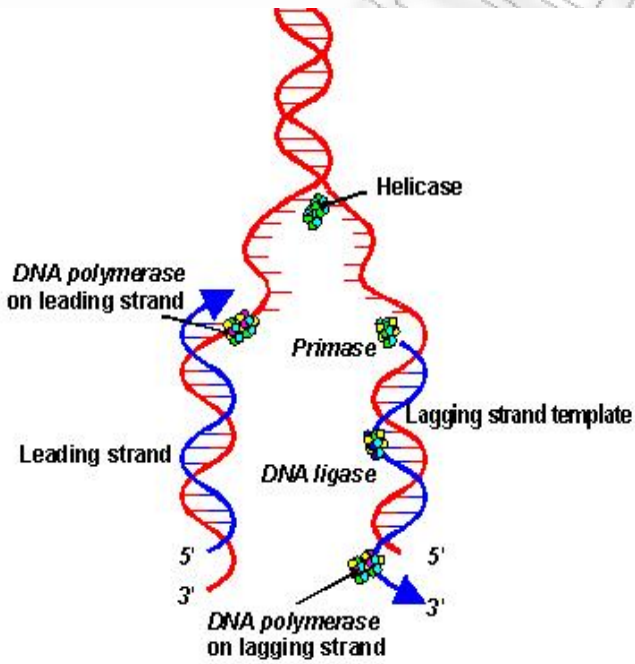
1.2 RNA

Το RNA (ριβονουκλεϊκό οξύ) είναι ένα πολυμερές χωρίς διακλαδώσεις που αποτελείται από νουκλεοτίδια ενωμένα με 3΄-5΄ φωσφοδιεστερικούς δεσμούς. Ξεχωρίζει βιοχημικά από το DNA σε δύο σημεία. Το σάκχαρο στο RNA, όπως φαίνεται και από το όνομά του, είναι η ριβόζη αντί της δεοξυριβόζης και η ουρακίλη αντικαθιστά τη θυμίνη ως συμπληρωματική της αδενίνης. Μία από τις κύριες λειτουργίες του RNA είναι να αντιγράφει τη γενετική πληροφορία και στη συνέχεια να τη μεταφράζει σε πρωτεΐνες.

Δομικά, το RNA δε διακρίνεται από το DNA, εκτός από την κρίσιμη παρουσία της υδροξυλικής ομάδας που είναι προσαρτημένη στο δακτύλιο της πεντόζης στη θέση 2΄ (το DNA έχει ένα άτομο υδρογόνου). Αυτή η υδροξυλική ομάδα κάνει το RNA λιγότερο σταθερό

από το DNA διότι καθιστά την υδρόλυση του φωσφορικού κορμού των σακχάρων πιο εύκολη.

Για τη σύνθεση του RNA συνήθως γίνεται κατάλυση από ένα ένζυμο, την RNA πολυμεράση, το οποίο χρησιμοποιεί το DNA σαν εκμαγείο. Το εκμαγείο DNA περιέχει περιοχές που λέγονται προαγωγείς και δεσμεύουν την RNA πολυμεράση, καθορίζοντας το σημείο που αρχίζει η μεταγραφή. Η διπλή έλικα του DNA ξετυλίγεται με τη βοήθεια της ελικάσης (σχήμα 1.3). Το ένζυμο προχωρεί κατά μήκος του προτύπου κλώνου με την κατεύθυνση 3΄->5΄, συνθέτοντας ένα συμπληρωματικό μόριο RNA. Η ακολουθία του DNA υπαγορεύει πότε θα λάβει χώρα η περάτωση της σύνθεσης του RNA.



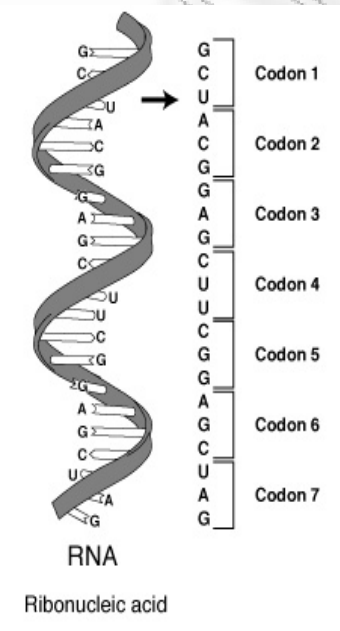
Σχήμα 1.3. Σύνθεση του RNA.

1.2.1 Αγγελιοφόρο RNA- mRNA

Το mRNA είναι το RNA που μεταφέρει την πληροφορία από το DNA στις ριβοσωμικές θέσεις της σύνθεσης της πρωτεΐνης σε ένα κύτταρο. Για κάθε γονίδιο ή ομάδα γονιδίων που χρειάζεται να εκφραστεί, παράγεται και ένα mRNA. Μόλις παράγεται το RNA από το DNA, εξάγεται από τον πυρήνα στο κυτταρόπλασμα, όπου δεσμεύεται από τα ριβοσώματα και μεταφράζεται σε πρωτεΐνη.

1.2.2 Μεταφορικό RNA- tRNA

Το tRNA περιέχει μία θέση δέσμευσης αμινοξέος και μία θέση αναγνώρισης του εκμαγείου. Κάθε μόριο tRNA μεταφέρει ένα είδος αμινοξέων στη θέση της πρωτεϊνοσύνθεσης, το ριβόσωμα, για το σχηματισμό πεπτιδικού δεσμού όπως καθορίζεται από το εκμαγείο mRNA. Η θέση αναγνώρισης του εκμαγείου στο tRNA είναι μία αλληλουχία τριών βάσεων που λέγεται αντικωδίκιο (anticodon) (ή αντικωδικόνιο). Το αντικωδίκιο στο tRNA αναγνωρίζει μία συμπληρωματική αλληλουχία τριών βάσεων στο mRNA που λέγεται κωδίκιο (codon) (σχήμα 1.4). Ο γενετικός κώδικας είναι η σχέση μεταξύ της αλληλουχίας των βάσεων και της αλληλουχίας των αμινοξέων στις πρωτεΐνες.



Σχήμα 1.4. Η έλικα του RNA και οι αλληλουχίες των βάσεων στο mRNA (κωδίκια).

1.2.3 Ριβοσωμικό RNA- rRNA

Το rRNA είναι το κύριο συστατικό των ριβοσωμάτων, αλλά ο ακριβής ρόλος του στη σύνθεση της πρωτεΐνης δεν είναι ακόμα γνωστός.

1.2.4 Συμπληρωματικό DNA- cDNA

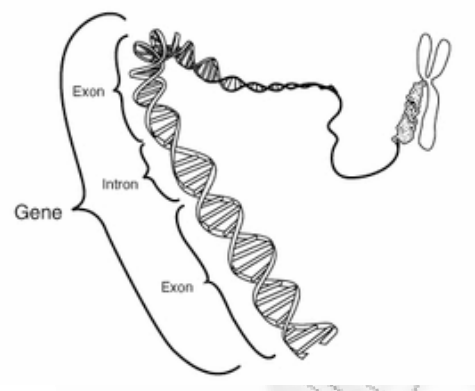
Το cDNA είναι το DNA το οποίο συντίθεται από ένα ώριμο πρότυπο mRNA και συνήθως παράγεται από το ένζυμο αντίστροφη μεταγραφάση. Το ένζυμο αυτό επιτρέπει σε ένα ώριμο mRNA να αποδοθεί σαν cDNA. Η συνύπαρξη του mRNA και του cDNA θεμελιώνει τη γενική αρχή ότι η πληροφορία στη μορφή του ενός τύπου ακολουθίας νουκλεϊκού οξέος

μπορεί να μετατραπεί στον άλλο τύπο. Στην τεχνολογία των μικροσυστοιχιών η διαδικασία της αντίστροφης μεταγραφής χρησιμοποιείται συχνά για την ενσωμάτωση φθοριζόντων χρωματισμών στο cDNA και συμπληρωματικά στα mRNA αντίγραφα.

$$\text{RNA} \xrightarrow[\text{αντίστροφη μεταγραφή}]{} \text{cDNA}$$

1.3 Έκφραση Γονιδίων

Τα γονίδια είναι οι μονάδες μιας αλληλουχίας DNA που ελέγχουν τα αναγνωρίσιμα κληρονομικά χαρακτηριστικά ενός οργανισμού. Κωδικοποιούνται στο γενετικό υλικό ενός οργανισμού (συνήθως στο DNA ή στο RNA) και ελέγχουν τη φυσική ανάπτυξη και συμπεριφορά του οργανισμού. Συγκεκριμένα, το γονιδίωμα είναι το σύνολο των γονιδίων που καθορίζουν τη γενετική δομή ενός οργανισμού ή ενός κυττάρου ή διαφορετικά το γονότυπό του.



Σχήμα 1.5. Ένα γονίδιο σε σχέση με τη διπλοελικομένη δομή του DNA και ενός χρωμοσώματος.
Τα ιντρόνια (introns) είναι παρεμβάλλουσες ακολουθίες που περιέχονται στα ευκαρυωτικά γονίδια.
Μόνο τα εξόνια (exons) κωδικοποιούν την πρωτεΐνη.

Η έκφραση των γονιδίων είναι η διαδικασία με την οποία το mRNA και τελικά η πρωτεΐνη συντίθενται από το εκμαγείο DNA κάθε γονιδίου. Στο πρώτο στάδιο αυτής της διαδικασίας γίνεται η μεταγραφή του DNA σε αγγελιοφόρο RNA. Στα ευκαρυωτικά mRNA συνοδεύεται από το μάτισμα του RNA, κατά το οποίο τα ιντρόνια (παρεμβάλλουσες ακολουθίες) αποκόβονται από το αρχικό αντίγραφο.

Το επόμενο στάδιο της έκφρασης ενός γονιδίου είναι η μετάφραση του mRNA σε πρωτεΐνη και συμβαίνει στο κυτταρόπλασμα.

Μεταβολές στη γονιδιακή έκφραση έχουν συσχετιστεί με διάφορες ασθένειες όπως είναι ο καρκίνος, η σκλήρυνση κατά πλάκας κ.α. Η γνώση για το ποια γονίδια είναι εκφρασμένα

κάτω από ορισμένες συνθήκες, εξηγεί για τις βιολογικές διαδικασίες σε ένα κύτταρο. Η ισχύς της τεχνολογίας των μικροσυστοιχιών υπόκειται στην ικανότητά της να μετρά την έκφραση χιλιάδων γονιδίων ταυτόχρονα.

1.4 Υβριδοποίηση νουκλεϊκού οξέος

Το ειδικό ταίριασμα των βάσεων των νουκλεϊκών οξέων αποτελεί θεμέλιο για την τεχνολογία των μικροσυστοιχιών. Το ειδικό ταίριασμα ενός τεχνητού ιχνηθέτη DNA με το βιολογικό του αντίστοιχο, επιτρέπει την ακριβή αναγνώριση της μοναδικής αλληλουχίας ή γονιδίου.

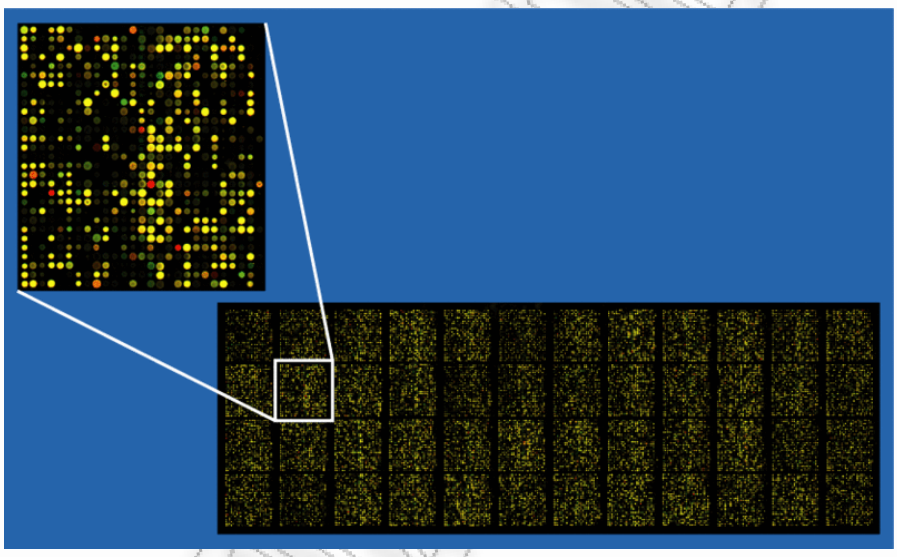
Εξαιτίας αυτού του ταιριάσματος, οι δύο κλώνοι του DNA μπορούν να χωριστούν και να ανασχηματιστούν πολύ γρήγορα κάτω από ορισμένες συνθήκες που διασπούν τους δεσμούς υδρογόνου μεταξύ των βάσεων αλλά και είναι αρκετά ήπιες για να απειλήσουν τους ομοιοπολικούς δεσμούς του σκελετού του DNA. Η διαδικασία του διαχωρισμού των κλώνων λέγεται αποδιάταξη ενώ του ανασχηματισμού τους επαναδιάταξη.

Η υβριδοποίηση είναι η βιοχημική μέθοδος πάνω στην οποία βασίζεται η τεχνολογία των μικροσυστοιχιών DNA. Οι αλληλουχίες των νουκλεϊκών οξέων μπορούν να συγκριθούν από άποψη συμπληρωματικότητας η οποία καθορίζεται από τους κανόνες ταιριάσματος των βάσεων και η συμπληρωματικότητα μπορεί να μετρηθεί επειδή η αποδιάταξη του DNA είναι αναστρέψιμη, κάτω από κατάλληλες συνθήκες. Η ανίχνευση και αναγνώριση ενός νουκλεϊκού οξέος - DNA ή mRNA - με ένα cDNA ιχνηθέτη το οποίο έχει κάποια σήμανση και το οποίο είναι συμπληρωματικό του, είναι μία εφαρμογή της υβριδοποίησης του νουκλεϊκού οξέος. Οι μικροσυστοιχίες DNA χρησιμοποιούν τις αντιδράσεις υβριδοποίησης μεταξύ των μονόκλωνων φθοριζόντων νουκλεϊκών οξέων που εξετάζονται και των μονόκλωνων αλληλουχιών που ακινητοποιούνται στην επιφάνεια της μικροδιάταξης.

1.5 Τεχνολογία Μικροσυστοιχιών

Μία μικροσυστοιχία DNA είναι ένα πλακίδιο κατασκευασμένο από ειδικό γυαλί πάνω στο οποίο παρατάσσονται μοριακοί ανιχνευτές σε συγκεκριμένες θέσεις, το πλήθος των οποίων μπορεί να κυμανθεί από μερικές εκατοντάδες έως μερικές χιλιάδες. Το προς εξέταση δείγμα

DNA χρωματίζεται με κατάλληλη χρωστική και μετά από ειδική επεξεργασία τοποθετείται πάνω στο πλακίδιο. Συγκεκριμένη αλληλουχία αυτού υβριδοποιείται με τους αντίστοιχους μοριακούς ανιχνευτές και στη θέση του κάθε ανιχνευτή ελευθερώνεται η χρωστική ουσία. Η ποσότητα της χρωστικής που ελευθερώνεται είναι ανάλογη της έκφρασης του αντίστοιχου γονιδίου. Στη συνέχεια, ένας οπτικός σαρωτής σαρώνει το πλακίδιο και στην έξοδό του παράγεται μία ψηφιακή εικόνα η οποία αποτελείται από ένα πλήθος κουκκίδων. Κάθε κουκκίδα αντιστοιχεί σε ένα διαφορετικό γονίδιο και η έντασή της αντιστοιχεί στο επίπεδο έκφρασής του. Η εικόνα αναλύεται με κατάλληλο λογισμικό με αποτέλεσμα την παραγωγή ενός διανύσματος, οι μεταβλητές του οποίου είναι οι μετρήσεις γονιδιακής έκφρασης των γονιδίων.



Με τη συμβολή της τεχνολογίας των μικροσυστοιχιών μπορούν να δημιουργηθούν μεγάλες συλλογές με δεδομένα εκφράσεων γονιδίων. Αυτοί οι κατάλογοι καλούνται “gene expression’’ ή “transcriptional profiling” και η διαδικασία συλλογής των δεδομένων “profiling”. Το transcriptional profiling βασίζεται είτε σε αλληλουχίες είτε σε υβριδοποίηση.

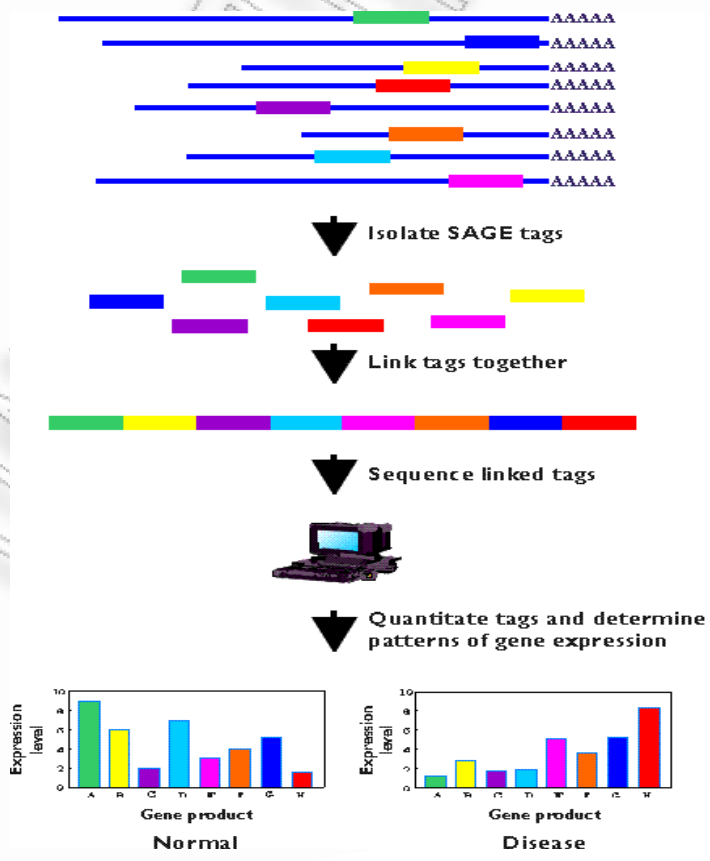
Οι προσεγγίσεις οι οποίες βασίζονται στις αλληλουχίες περιέχουν τη δημιουργία αλληλουχιών cDNA και την ακολουθιακή ανάλυση της έκφρασης του γονιδίου (SAGE). Το συμπληρωματικό DNA (cDNA) είναι το DNA το οποίο συντίθεται από το εκμαγείο mRNA με τη βοήθεια του ενζύμου αντίστροφη μεταγραφάση. Μόρια cDNA μπορούν να κλωνοποιηθούν σε φορείς οι οποίοι είναι γνωστοί ως φορείς έκφρασης. Για να έχουμε το μέγιστο της απόδοσης στην αντιγραφή, το cDNA εισάγεται στο φορέα, ο οποίος εξασφαλίζει τη μετάφραση κωδικοποιώντας θέσεις δέσμευσης στα ριβοσώματα κοντά στο mRNA. Οι

κλώνοι cDNA μπορούν να ανιχνευτούν βάσει της ικανότητάς τους να συνθέτουν ξένες πρωτεΐνες στα βακτήρια. Ένα ραδιενεργό αντίσωμα ειδικό για την πρωτεΐνη που μας ενδιαφέρει μπορεί να χρησιμοποιηθεί για να προσδιοριστούν οι αποικίες βακτηρίων που περιέχουν τον αντίστοιχο φορέα cDNA. Το αντίσωμα αυτό είναι σημασμένο και η αυτοραδιογραφία αποκαλύπτει τις θέσεις των αποικιών που μας ενδιαφέρουν.

Η βασική ιδέα της SAGE είναι να παραγάγει μικρές αλληλουχίες σημάνσεων cDNA από μία συγκέντρωση mRNA συνδυάζοντάς τις και δημιουργώντας πολλές αλληλουχίες σημάνσεων κάθε στιγμή. Οι τρεις βασικές αρχές της SAGE είναι:

1. Μία μικρή ακολουθία σημάνσεων 10 με 14 ζευγών βάσεων περιέχει αρκετή πληροφορία για τη μοναδική αναγνώριση ενός αντιγράφου, αρκεί η σήμανση να προέρχεται από μία μοναδική θέση μέσα από κάθε αντίγραφο.
2. Οι ακολουθίες σημάνσεων μπορούν να ενωθούν για να σχηματίσουν μεγάλα σειριακά μόρια, τα οποία μπορούν να κλωνοποιηθούν και να σχηματίσουν μια ακολουθία.
3. Η μέτρηση του αριθμού των φορών που μία συγκεκριμένη σήμανση παρατηρείται μας δίνει το επίπεδο έκφρασης του αντίστοιχου αντιγράφου.

Παρακάτω φαίνεται η σχηματική αναπαράσταση της μεθόδου SAGE.



Οι μέθοδοι υβριδοποίησης, όπως αυτή της Southern and Northern Αποτύπωσης, χρησιμοποιούνται για να αναγνωριστούν και να μετρηθούν τα νουκλεϊκά οξέα στα βιολογικά δείγματα. Η αποτύπωση κατά Southern αποτελεί μία τεχνική για την ανίχνευση μικρών διαφορών μεταξύ συγγενών μορίων DNA. Περιοριστικά ένζυμα αναγνωρίζουν ειδικές ακολουθίες βάσεων σε διπλή έλικα DNA και διασπούν και τις δύο έλικες σε συγκεκριμένες θέσεις. Τα περιοριστικά τμήματα DNA διαχωρίζονται και εμφανίζονται με ηλεκτροφόρηση σε πηκτή (το χαρακτηριστικό αυτών των πηκτών είναι η μεγάλη διαχωριστική τους ικανότητα). Οι ζώνες ή τα σημεία ραδιενεργού DNA στην πηκτή εμφανίζονται με αυτοραδιογραφία ή διαφορετικά μία πηκτή βάφεται με βρωμιούχο αιθίδιο, που δίνει έντονο πορτοκαλί χρώμα φθορισμού όταν είναι δεσμευμένο σε DNA. Ένα συγκεκριμένο τμήμα που περιέχει ειδικές αλληλουχίες βάσεων μπορεί να προσδιοριστεί υβριδοποιώντας το με ένα σημασμένο συμπληρωματικό DNA που είναι ο μοριακός ανιχνευτής DNA. Η αυτοραδιογραφία αποκαλύπτει τη θέση των θραυσμάτων περιορισμού που έχουν αλληλουχία συμπληρωματική με αυτή του μοριακού ανιχνευτή. Παρομοίως, μόρια RNA μπορούν να διαχωριστούν με ηλεκτροφόρηση σε πηκτή και ειδικές αλληλουχίες μπορούν να προσδιοριστούν με υβριδοποίηση. Αυτή η ανάλογη τεχνική για την ανάλυση του RNA έχει ονομαστεί αποτύπωση Northern. Αρχικά αυτές οι μέθοδοι προσπαθούσαν να αναγνωρίσουν και να μετρήσουν μόνο ένα γονίδιο κάθε φορά. Στη συνέχεια, αναπτύχθηκαν πολλές τεχνικές αναπτύχθηκαν έτσι ώστε να μπορούν να αναλυθούν χιλιάδες υβριδοποιήσεις παράλληλα. Σήμερα είναι δυνατόν να παραχθούν συστοιχίες μοριακών ανιχνευτών που παριστάνουν όλα τα γονίδια ενός γονιδιώματος πάνω σε μία μόνο διαφάνεια.

Υπάρχουν τρία κύρια είδη τεχνολογίας μικροσυστοιχιών: οι εναποτεθειμένες μικροσυστοιχίες cDNA (spotted cDNA arrays), οι εναποτεθειμένες μικροσυστοιχίες ολιγονουκλεοτιδίων (spotted oligonucleotide arrays) και οι in-situ μικροσυστοιχίες ολιγονουκλεοτιδίων (in-situ oligonucleotide arrays). Οι διαφορές τους οφείλονται στον τρόπο παραγωγής τους και στο είδος των ιχνηθετών που χρησιμοποιούν.

- Στα spotted cDNA arrays, ολόκληροι οι κλώνοι cDNA ή οι εκφρασμένες αλληλουχίες σημάνσεων είναι μηχανικά εναποτεθειμένες κουκίδες και ακινητοποιημένες στην επιφάνεια στήριξης. Οι συστοιχίες των εναποτεθειμένων κουκίδων cDNA έχουν το πλεονέκτημα, σε σχέση με άλλες συστοιχίες, ότι μπορούν να στιγματιστούν και άγνωστες αλληλουχίες. Έτσι, οι μικροσυστοιχίες των spotted

cDNA αποτελούν τη μόνη επιλογή για οργανισμούς για τους οποίους είναι διαθέσιμη μόνο περιορισμένη (ή και καμία) πληροφορία για την αλληλουχία του γονιδιώματος,.

- Τα spotted oligonucleotide arrays είναι αρκετά όμοια με τα spotted cDNA arrays, με τη διαφορά ότι ως μοριακοί ανιχνευτές, χρησιμοποιούνται τα συνθετικά ολιγονουκλεοτίδια (oligos) αντί των cDNA. Στην πραγματικότητα, χρησιμοποιείται η ίδια μηχανική εναπόθεση κουκίδας για την κατασκευή και των δύο μικροσυστοιχιών. Όταν η πληροφορία για την αλληλουχία είναι διαθέσιμη, οι ολιγονουκλεοτιδικοί μοριακοί ανιχνευτές των 20 ή 70 νουκλεοτιδίων μπορούν να σχεδιαστούν και να συντεθούν. Η χρήση των ολιγονουκλεοτιδικών μοριακών ανιχνευτών μάς δίνει καλύτερο έλεγχο για το ποιο μέρος του γονιδίου θα χρησιμοποιηθεί για την υβριδοποίηση.
- Τα in-situ oligonucleotide arrays χρησιμοποιούν ένα συνδυασμό φωτολιθογραφίας και ολιγονουκλεοτιδικής χημείας στερεάς φάσης για τη σύνθεση ολιγονουκλεοτιδικών μοριακών ανιχνευτών, περιορισμένων σε μήκος (short), κατευθείαν πάνω στη συμπαγή επιφάνεια στήριξης. Ο αριθμός των ολιγονουκλεοτιδίων (50.000 ιχνηθέτες σε $1.28\,cm^2$) σε ένα chip το οποίο είναι κατασκευασμένο με αυτόν τον τρόπο, ξεπερνά σε μεγάλο βαθμό το τι μπορεί να επιτευχθεί με τη λύση της μηχανικής εναπόθεσης κουκίδας. Η εταιρεία Affymetrix Inc. επέλεξε να χρησιμοποιήσει αυτό το πλεονέκτημα για να κατασκευάσει μία συστοιχία με διάφορους ολιγονουκλεοτιδικούς μοριακούς ανιχνευτές και πρότυπα δείγματα διασταυρούμενης υβριδοποίησης για κάθε γονίδιο στόχο.

Για τα in-situ oligonucleotide arrays τα υπό εξέταση δείγματα και τα πρότυπα δείγματα υβριδοποιούνται ξεχωριστά σε διαφορετικά chips. Αντιθέτως, για τα cDNA arrays ή τα spotted oligonucleotide arrays, το υπό εξέταση δείγμα και το πρότυπο δείγμα με σημάνσεις από δύο διαφορετικές χρωστικές υβριδοποιούνται συνήθως ταυτόχρονα πάνω στις ίδιες συστοιχίες.

1.6 Δεδομένα

Στην παρούσα εργασία θα χρησιμοποιήσουμε δύο σύνολα δεδομένων. Το πρώτο από αυτά αναφέρεται σε 38 δείγματα δύο διαφορετικών τύπων λευχαιμίας και μετρήσεις της έκφρασης 5000 γονιδίων. Προέρχονται από τη μελέτη που πραγματοποίησαν οι Golub et al. (1999) με

στόχο την παρουσίαση των δυνατοτήτων της τεχνολογίας των γονιδιωμάτων να παρέχει πιο ακριβείς τρόπους με τους οποίους οι νεοπλασίες μπορούν να χαρακτηριστούν και να ταξινομηθούν και έτσι να οδηγήσουν σε μία πιο αποτελεσματική διάγνωση και θεραπεία. Η σωστή διάγνωση της νεοπλασίας είναι απαραίτητη για τον προσδιορισμό της κατάλληλης θεραπείας και οι παραδοσιακοί τρόποι αναγνώρισης και ταξινόμησης των κακοηθειών βασίζονταν σε ιστολογική και ανοσοϊστολογική βαφή των παθολογικών δειγμάτων.

Τα δύο είδη λευχαιμίας είναι η οξεία λεμφωβλαστική λευχαιμία (*acute lemphoblastic leukemia*) (ALL, ομάδα 1) και η οξεία μυελώδης λευχαιμία (*acute myeloid leukemia*) (AML, ομάδα 2), οι οποίες ταξινομούνται εύκολα με τη βοήθεια των παραδοσιακών παθολογικών μεθόδων. Οι Golub et al. (1999) θέλησαν να αποδείξουν ότι αυτοί οι κακοήθεις όγκοι μπορούν να αναγνωριστούν και να διαχωριστούν με τη βοήθεια μετρήσεων της έκφρασης των γονιδίων σε μικροσυστοιχίες. Ένας από τους στόχους της στατιστικής ανάλυσης ήταν η αναγνώριση γονιδίων τα οποία διαφέρουν περισσότερο ανάμεσα στις δύο ομάδες. Το σύνολο μελέτης αποτέλεσαν 38 δείγματα μυελού των οστών από τα οποία 27 ήταν τύπου ALL και 11 τύπου AML και τα οποία προήλθαν από ασθενείς με οξεία λευχαιμία τη στιγμή της διάγνωσης. Το RNA το οποίο σχηματίστηκε από μονοπύρηνα κύτταρα μυελού των οστών υβριδοποιήθηκε σε πλακίδια ολιγονουκλεοτιδίων υψηλής πυκνότητας (της εταιρείας Affymetrix) τα οποία περιείχαν μοριακούς ανιχνευτές για 6.817 ανθρώπινα γονίδια και για κάθε γονίδιο λήφθηκε ένα επίπεδο έκφρασης. Τα δείγματα υπέστησαν έναν τυπικό ποιοτικό έλεγχο σχετικά με την ποσότητα του σημασμένου RNA και την ποιότητα που καταγράφηκε στην εικόνα της μικροσυστοιχίας.

Οι τιμές των εντάσεων φθορισμού παρουσιάζονται στον κανονικοποιημένο πίνακα που είναι διαθέσιμος στην ιστοσελίδα http://www.genome.wi.mit.edu/MPR. Η διαδικασία της κανονικοποίησης των δεδομένων αποτελεί προϋπόθεση για κάθε είδους ανάλυση μικροσυστοιχιών DNA και ορίζεται ως ο μετασχηματισμός των δεδομένων που απομακρύνει την τυχαία και συστηματική μεταβλητότητα. Η μεταβλητότητα αυτή μπορεί να προέρχεται από διαφορές πηγές όπως διαφορές στην κατασκευή των πλακιδίων, μη σταθερή εργαστηριακή προετοιμασία του δείγματος, πρωτόκολλα υβριδοποίησης, μη ακριβείς μετρήσεις του σήματος οφειλόμενες στον σαρωτή και διαφορές στην απόδοση της υβριδοποίησης από γονίδιο σε γονίδιο.

Σύμφωνα με τους Golub et al., σε αυτόν τον πίνακα εφαρμόζονται τρία προκαταρκτικά βήματα:

(i) Αντικατάσταση των τιμών που είναι μικρότερες του 100 με το 100 και των τιμών που είναι μεγαλύτερες του 16.000 με το 16.000.

(ii) Αποκλεισμός των γονιδίων για τα οποία ο λόγος της μεγαλύτερης προς την μικρότερη τιμή είναι μικρότερος ή ίσος του 5 ή η διαφορά της μεγαλύτερης και της μικρότερης τιμής είναι μικρότερη ή ίση του 500.

(iii) Λογαρίθμηση των τιμών.

Τέλος, σχεδιάζοντας τα θηκογράμματα των εκφράσεων των γονιδίων για κάθε ένα από τα 38 δείγματα διαπιστώνουμε την ανάγκη μετατροπής των τιμών των εκφράσεων των γονιδίων σε τιμές που ακολουθούν την τυπική κανονική κατανομή. Πραγματοποιώντας τα παραπάνω βήματα τα δεδομένα συνοψίζονται σε έναν πίνακα $X = (x_{ij}),\ i = 1,...,2.109,\ j = 1,...,38$ με τις γραμμές να αντιστοιχούν στα γονίδια και τις στήλες στα δείγματα και όπου x_{ij} η μέτρηση της έκφρασης του γονιδίου i στο δείγμα j.

Το δεύτερο σύνολο δεδομένων που θα χρησιμοποιήσουμε προέρχεται από το πείραμα που πραγματοποίησαν οι Scholtens et al. (2004) και είναι διαθέσιμο στο κοινό μέσα από την ιστοσελίδα του Bioconductor http://www.bioconductor.org/data/experimental.html. Στο πείραμα αυτό οι ερευνητές εστίασαν το ενδιαφέρον τους στην επίδραση του οιστρογόνου, κατά τη διάρκεια του χρόνου, σε γονίδια κυττάρων που είχαν προσβληθεί από καρκίνο του μαστού. Μελέτησαν τα επίπεδα έκφρασης 12625 γονιδίων σε Affymetrix πλακίδια για 8 δείγματα καρκίνου του μαστού. Εξέθεσαν τα 4 δείγματα σε οιστρογόνο και στη συνέχεια μέτρησαν την «αφθονία» του mRNA αντιγράφου μετά από 10 ώρες για τα δύο δείγματα και μετά από 48 ώρες για τα άλλα δύο. Τα υπόλοιπα 4 δείγματα τα άφησαν χωρίς καμία επεξεργασία και μέτρησαν την «αφθονία» του mRNA αντιγράφου μετά από 10 ώρες για τα δύο δείγματα και μετά από 48 ώρες για τα άλλα δύο. Το πείραμα αυτό έχει δύο παράγοντες, το οιστρογόνο και τον χρόνο και ο καθένας από αυτούς δύο επίπεδα. Παρουσία ή απουσία οιστρογόνου και 10 ή 48 ώρες, αποτελεί δηλαδή έναν 2×2 παραγοντικό σχεδιασμό γνωστό ως "Estrogen 2x2 Factorial Design". Σκοπός της μελέτης αυτής είναι η αναγνώριση των γονιδίων που αποκρίνονται στο οιστρογόνο και η ταξινόμηση αυτών σε σχέση με το χρόνο απόκρισης.

ΚΕΦΑΛΑΙΟ 2

2.1 Πολλαπλοί Έλεγχοι Υποθέσεων

Σε ένα πείραμα μικροσυστοιχιών, τα δεδομένα μας αποτελούνται από τις εκφράσεις *m* γονιδίων σε *n* δείγματα, κάθε ένα από τα οποία σχετίζεται με μία απόκριση ή συμμεταβλητή ενδιαφέροντος. Στην παρούσα εργασία, όπως αναφέρεται αναλυτικά στο Κεφάλαιο 1, το πρώτο σύνολο δεδομένων προέρχεται από τις μετρήσεις της έκφρασης γονιδίων σε βιοψία τμήματος όγκου από ασθενείς με λευχαιμία. Η απόκριση που μας ενδιαφέρει είναι ο τύπος του όγκου και στόχος μας είναι η ταυτοποίηση των γονιδίων που είναι διαφορετικά εκφρασμένα στους διαφορετικούς τύπους όγκου. Τα δεδομένα συγκεντρώνονται σε έναν $m \times n$ πίνακα $\mathbf{X} = \left(x_{ij}\right),\ i = 1,...,m,\ j = 1,...,n$, με τις γραμμές να αντιστοιχούν στα γονίδια και τις στήλες στα δείγματα: x_{ij} είναι η μέτρηση της έκφρασης του γονιδίου *i* στο δείγμα *j*.

Θεωρούμε ότι τα *n* τυχαία διανύσματα διάστασης-*m* $X_j,\ j = 1,...,n$, που αντιστοιχούν στις μετρήσεις έκφρασης των *m* γονιδίων στα *n* δείγματα είναι ανεξάρτητα και ισόνομα. Το τυχαίο διάνυσμα $X_j = \left(X_{1j},...,X_{mj}\right)$ ακολουθεί κάποια κατανομή *P* η οποία ανήκει σε μία οικογένεια κατανομών $\mathcal{M}$. Γενικότερα, σε μελέτες μικροσυστοιχιών που αφορούν σε έρευνες για τον καρκίνο, συμβολίζουμε με $\left(X_{1j},...,X_{gj}\right)$, $1 < g < m$, το διάνυσμα των μετρήσεων της έκφρασης των γονιδίων και με $\left(X_{(g+1)j},...,X_{mj}\right)$ το διάνυσμα των βιολογικών και κλινικών αποτελεσμάτων για τον ασθενή *j*, $j = 1,...,n$. Οι ποσότητες που μας ενδιαφέρουν είναι συναρτήσεις της άγνωστης κατανομής *P* και ενδέχεται να εκφράζουν μέσες τιμές, διαφορές μέσων τιμών ή συσχετίσεις. Θα τις συμβολίζουμε γενικά με $\mu_i,\ i = 1,...,m$. Πιο συγκεκριμένα, στην περίπτωση ανάλυσης των μικροσυστοιχιών μπορεί να είναι το μέσο επίπεδο έκφρασης $\mu_i = \mathrm{E}\left(X_{ij}\right),\ j = 1,...,n$, ή η διαφορά των μέσων επιπέδων έκφρασης $\mu_i = E\left(X_{ij_1}\right) - E\left(Y_{ij_2}\right),\ j_1 = 1,...,n_1,\ j_2 = 1,...,n_2$, του i γονιδίου, $i = 1,...,m$, σε δύο

πληθυσμούς (πλήθους n_1 και n_2). Μπορεί επίσης να είναι παράμετροι παλινδρόμησης για τη μέτρηση της σχέσης του επιπέδου έκφρασης X_{ij}, $i=1,...,g$, $j=1,...,n$, με αποτελέσματα ή αποκρίσεις X_{ij}, $i=g+1,...,m$, $j=1,...,n$.

Το βιολογικό ερώτημα της διαφορετικής έκφρασης μπορεί να διατυπωθεί ως ένα πρόβλημα ταυτόχρονου ελέγχου για τα *m* γονίδια, της μηδενικής υπόθεσης της μη ύπαρξης σχέσης μεταξύ των επιπέδων έκφρασης X_{ij} και της απόκρισης ή συμμεταβλητής, ή ισοδύναμα με τον πολλαπλό έλεγχο υποθέσεων:

H_{0i} : το γονίδιο *i* δεν είναι διαφορετικά εκφρασμένο

έναντι της εναλλακτικής υπόθεσης

H_{1i} : το γονίδιο *i* είναι διαφορετικά εκφρασμένο,

για $i=1,...,m$.

Η H_{0i} είναι αληθής $(H_{0i}=0)$ όταν $P\in M_i$ και ψευδής $(H_{0i}=1)$ διαφορετικά. Το $S_0=\{i: P\in M_i\}$ είναι το σύνολο των m_0 αληθών μηδενικών υποθέσεων, όπου $m_0=|S_0|$ το πλήθος των στοιχείων του συνόλου S_0 και το $S_o^c=\{i: P\notin M_i\}$ το σύνολο των $m_1=m-m_0$ ψευδών μηδενικών υποθέσεων.

Οι αποφάσεις για την απόρριψη ή όχι των μηδενικών υποθέσεων βασίζονται στο διάνυσμα διάστασης *m* των στατιστικών $\mathbf{T}_n=(T_{in},\ i=1,...,m)'$. Κάθε στατιστικό T_{in} αποτελεί συνάρτηση των $X_{i1}, X_{i2},..., X_{in}$ και μεγάλες ή μικρές τιμές του οδηγούν σε απόρριψη της μηδενικής υπόθεσης. Η από κοινού κατανομή των T_{in} θα συμβολίζεται με F_n.

Ο έλεγχος πολλαπλών υποθέσεων ουσιαστικά παράγει ένα σύνολο S_n το οποίο αποτελεί εκτίμηση του συνόλου S_0^c των ψευδών μηδενικών υποθέσεων. Το σύνολο

$$S_n=S(T_n,F_0,a)\equiv\{i:\mathrm{H_{0i}}\ \text{απορρίπτεται}\ \}\subseteq\{1,...,m\}$$

εξαρτάται από τα δεδομένα $(X_{1j},...,X_{mj})$, $j=1,...,n$, μέσω των στατιστικών συναρτήσεων, $\mathbf{T}_n$ από την από κοινού κατανομή F_n των $\mathbf{T}_n$ υπό την μηδενική υπόθεση και από το επίπεδο *α* του πολλαπλού ελέγχου.

2.1.1 Πιθανότητες Σφαλμάτων Τύπου Ι

Κατά τον στατιστικό έλεγχο των υποθέσεων μπορεί να εμφανιστούν δύο είδη σφαλμάτων: το λανθασμένα θετικό αποτέλεσμα ή σφάλμα τύπου Ι, το οποίο πραγματοποιείται όταν ένα

γονίδιο δηλωθεί ως διαφορετικά εκφρασμένο ενώ στην πραγματικότητα δεν είναι και το λανθασμένα αρνητικό αποτέλεσμα ή σφάλμα τύπου ΙΙ, το οποίο συμβαίνει όταν ο έλεγχος αποτυγχάνει να αναγνωρίσει τα διαφορετικά εκφρασμένα γονίδια. Στόχος μας είναι να έχουμε μεγάλη πιθανότητα να δηλώσουμε ορθά τη διαφορετική έκφραση ενός γονιδίου, κρατώντας την πιθανότητα να έχουμε μία λανθασμένη δήλωση διαφορετικής έκφρασης χαμηλή, δηλαδή να αυξήσουμε την ισχύ του ελέγχου.

Το γεγονός ότι σε κάθε πείραμα μικροσυστοιχιών εμφανίζονται ταυτόχρονα τα επίπεδα έκφρασης χιλιάδων γονιδίων έχει ως αποτέλεσμα την εμφάνιση έντονων προβλημάτων πολλαπλότητας. Συνήθως, κατά τον έλεγχο μιας υπόθεσης έχουμε μία προκαθορισμένη πιθανότητα σφάλματος τύπου Ι, η οποία όμως αυξάνεται με τον αριθμό των υποθέσεων. Συγκεκριμένα, μία τιμή-*p* ίση με 0.01 για ένα γονίδιο ανάμεσα σε αρκετές χιλιάδες δεν αποτελεί σημαντικό αποτέλεσμα, αφού μπορεί να έχει εμφανιστεί τυχαία κάτω από τη μηδενική υπόθεση θεωρώντας ένα τόσο μεγάλο σύνολο γονιδίων. Έτσι τα σφάλματα τύπου Ι και ΙΙ πρέπει να ρυθμιστούν για ολόκληρο το σύνολο των γονιδίων. Επομένως, αυτό που μας ενδιαφέρει σε ένα πρόβλημα ταυτόχρονου συμπεράσματος είναι *ο προσδιορισμός μιας κατάλληλης πιθανότητας σφαλμάτων τύπου Ι και η επινόηση ισχυρών διαδικασιών πολλαπλών υποθέσεων οι οποίες ελέγχουν την πιθανότητα σφάλματος τύπου Ι*.

Ένας συνηθισμένος τρόπος παρουσίασης του προβλήματος του ταυτόχρονου ελέγχου των *m* μηδενικών υποθέσεων είναι αυτός που φαίνεται στο πίνακα 2.1. (Benjamini and Hochberg, 1995).

	Αριθμός υποθέσεων που δεν έχουν απορριφθεί	Αριθμός υποθέσεων που έχουν απορριφθεί	
Ορθές μηδενικές υποθέσεις	U_n	V_n	m_0
Ψευδείς μηδενικές υποθέσεις	N_n	W_n	m_1
	$m - R_n$	R_n	$\boldsymbol{m}$

Πίνακας 2.1. *Το πρόβλημα του ταυτόχρονου ελέγχου των m μηδενικών υποθέσεων.*

Οι *m* υποθέσεις είναι γνωστές εκ των προτέρων, οι αριθμοί m_0 και $m_1 = m - m_0$ των ορθών και ψευδών μηδενικών υποθέσεων είναι άγνωστες παράμετροι, το R_n είναι μία

παρατηρήσιμη τυχαία μεταβλητή και τα U_n, V_n, N_n και W_n είναι μη παρατηρήσιμες τυχαίες μεταβλητές.

Στην περίπτωση του ελέγχου μίας υπόθεσης, η πιθανότητα σφάλματος τύπου I, δηλαδή της απόρριψης της μηδενικής υπόθεσης ενώ αυτή ισχύει, συνήθως *ρυθμίζεται* σε ένα προκαθορισμένο επίπεδο α. Στη συνέχεια, επιλέγεται μία κρίσιμη τιμή c_α τέτοια ώστε $\Pr\left(|T| \geq c_\alpha \middle| \mathrm{H}_0\right) \leq a$ για το στατιστικό T που έχουμε αποφασίσει να χρησιμοποιήσουμε και η H_0 απορρίπτεται όταν $|T| \geq c_\alpha$. Αντίστοιχα, *μία διαδικασία πολλαπλών υποθέσεων λέμε ότι ρυθμίζει μία συγκεκριμένη πιθανότητα σφάλματος τύπου I σε ένα επίπεδο α, αν αυτή η πιθανότητα είναι μικρότερη ή ίση με α όταν η δεδομένη διαδικασία εφαρμόζεται για να παραχθεί ένα σύνολο από R_n απορριφθείσες υποθέσεις*.

Στους πολλαπλούς ελέγχους υπάρχει μία ποικιλία γενικεύσεων για τον προσδιορισμό της πιθανότητας σφάλματος τύπου I. Εδώ θα μελετήσουμε τις πιθανότητες σφάλματος που ορίζονται ως συναρτήσεις της κατανομής ενός αριθμού σφαλμάτων τύπου I.

- Η **Per-Comparison Error Rate (PCER)** ορίζεται ως ο αναμενόμενος αριθμός των σφαλμάτων τύπου I διά το πλήθος των υποθέσεων.

$$PCER = \frac{E\left(V_n\right)}{m}.$$

- Η **Per-Family Error Rate (PFER)** ορίζεται ως ο αναμενόμενος αριθμός των σφαλμάτων τύπου I.

$$PFER = E\left(V_n\right).$$

- Η **Family-Wise Error Rate (FWER)** ορίζεται ως η πιθανότητα να εμφανιστεί τουλάχιστον ένα σφάλμα τύπου I.

$$FWER = \Pr\left(V_n \geq 1\right).$$

- Η **False Discovery Rate (FDR)** (Benjamini & Hochberg, 1995) ορίζεται ως το αναμενόμενο ποσοστό των σφαλμάτων τύπου I ανάμεσα στις υποθέσεις που έχουν απορριφθεί.

$$FDR = E\left(Q_n\right), \text{ όπου } Q_n = \frac{V_n}{R_n} \text{ αν } R_n > 0 \text{ και } Q_n = 0 \text{ αν } R_n = 0.$$

- Η **Positive False Discovery Rate (pFDR)** (Storey, 2002), ορίζεται ως το αναμενόμενο ποσοστό των σφαλμάτων τύπου I ανάμεσα στις υποθέσεις που έχουν απορριφθεί, δεδομένου ότι έχουμε μία τουλάχιστον απόρριψη.

$$pFDR = E\left(\frac{V_n}{R_n}\,\middle|\, R_n > 0\right)$$

2.1.2 Σύγκριση των πιθανοτήτων των σφαλμάτων τύπου I

Γενικά, για μία δεδομένη διαδικασία πολλαπλών ελέγχων, δηλαδή μία διαδικασία με τη ίδια περιοχή απόρριψης στον *m*-διάστατο χώρο, ισχύει ότι:

$$PCER \leq FDR \leq FWER \leq PFER.$$

Η σχέση αυτή δηλώνει ότι οι διαδικασίες που ρυθμίζουν την PFER είναι πιο συντηρητικές από αυτές που ρυθμίζουν την FWER και αντίστοιχα για τις υπόλοιπες, που σημαίνει ότι οδηγούν σε λιγότερες απορρίψεις. Επίσης, ισχύει $FDR \leq pFDR$. Τα παραπάνω αποδεικνύονται ως εξής:

- Αρχικά για την ανισότητα $PCER \leq FDR$ γνωρίζουμε ότι,

$$0 \leq V_n \leq R_n \leq m \text{ με } V_n = 0 \text{ όταν } R_n = 0.$$

Έτσι για $R_n > 0$ μπορούμε να γράψουμε ότι

$$\frac{V_n}{m} \leq \frac{V_n}{R_n}$$

Και αν πάρουμε τις αναμενόμενες τιμές καταλήγουμε στην ζητούμενη ανισότητα,

$$E\left(\frac{V_n}{m}\right) \leq E\left(\frac{V_n}{R_n}\,\middle|\, R_n > 0\right) \text{ δηλαδή } PCER \leq FDR.$$

- Για την ανισότητα $FDR \leq FWER$

(α) αν όλες οι μηδενικές υποθέσεις είναι αληθείς, δηλαδή $m = m_0$, τότε $FDR = FWER$.

Η παραπάνω σχέση ισχύει διότι:

έχουμε ότι $m = m_0$ οπότε συνεπάγεται ότι $W_n = 0 \Rightarrow V_n = R_n$.

αν $V_n = 0$ τότε $Q_n = 0 \Rightarrow \Pr(V_n \geq 1) = E(Q_n) = 0$

αν $V_n > 0$ τότε $Q_n = 1 \Rightarrow \Pr(V_n \geq 1) = E(Q_n) = 1$

(β) αν $m_0 < m$, τότε $FDR \leq FWER$.

Αν $V_n > 0 \Rightarrow \frac{V_n}{R_n} \leq 1 \Rightarrow Q_n \leq 1 = I(V_n \geq 1)$ όπου $I(V_n \geq 1)$ δείκτρια συνάρτηση. Από την τελευταία σχέση αν πάρουμε τις μέσες τιμές προκύπτει το ζητούμενο, δηλαδή ότι $E(Q_n) \leq \Pr(V_n \geq 1)$

- Τέλος, σύμφωνα με την ανισότητα του Markov $\Pr(V_n \geq 1) \leq E(V_n)$ άρα $FWER \leq RFER$.

2.1.3 Ισχυρή και Ασθενής Ρύθμιση

Οι αναμενόμενες τιμές και οι πιθανότητες που παρουσιάστηκαν παραπάνω ορίζονται κάτω από τη μηδενική και τυπικά άγνωστη κατανομή P των δεδομένων. Συγκεκριμένα, εξαρτώνται από το ποιο σύνολο $S_0 \subseteq \{1,...,m\}$ αντιστοιχεί στο σύνολο των αληθών μηδενικών υποθέσεων για αυτή την κατανομή. Για παράδειγμα, ρύθμιση της FWER σημαίνει ρύθμιση της πιθανότητας

$$\Pr(V_n \geq 1 | \bigcap_{i \in S_0} H_{0i}) = \Pr(\text{ απόρριψη μίας τουλάχιστον } H_{0i},\ i \in S_0 \mid \bigcap_{i \in S_0} H_{0i})$$

όπου $\bigcap_{i \in S_0} H_{0i}$ το σύνολο των αληθών μηδενικών υποθέσεων για την από κοινού κατανομή των δεδομένων. Η ρύθμιση αυτή ονομάζεται "ακριβής ρύθμιση".

Ένας βασικός διαχωρισμός είναι αυτός μεταξύ της ισχυρής και της ασθενούς ρύθμισης της πιθανότητας του σφάλματος τύπου Ι. Η ισχυρή ρύθμιση αναφέρεται στη ρύθμιση της πιθανότητας σφάλματος τύπου Ι κάτω από οποιοδήποτε συνδυασμό αληθών και ψευδών μηδενικών υποθέσεων, δηλαδή για κάθε υποσύνολο $S_0 \subseteq \{1,...,m\}$ αληθών μηδενικών υποθέσεων. Για την FWER, ισχυρή ρύθμιση σημαίνει ρύθμιση της $\max_{S_0} \Pr(V \geq 1 | H_{S_0})$. Αντίθετα η ασθενής ρύθμιση αναφέρεται στη ρύθμιση της πιθανότητας σφάλματος τύπου Ι κάτω από την πλήρη μηδενική υπόθεση $H_0^c = \bigcap_{i=1}^{m} H_{0i}$ με $m = m_0$, δηλαδή όταν όλες οι μηδενικές υποθέσεις είναι αληθείς. Για την FWER, ασθενής ρύθμιση σημαίνει ρύθμιση της $\Pr(V_n \geq 1 | H_0^c)$.

Στην πραγματικότητα, ορισμένες μηδενικές υποθέσεις μπορεί να είναι αληθείς και άλλες μπορεί να είναι ψευδείς, αλλά το υποσύνολο S_0 παραμένει άγνωστο. Σε ένα σύστημα μικροσυστοιχιών όπου δεν είναι πολύ πιθανό να μην υπάρχουν διαφορετικά εκφρασμένα γονίδια, η πλήρης μηδενική υπόθεση είναι σχεδόν απίθανο να ισχύει, κάτι που καθιστά την ασθενή ρύθμιση μη ικανοποιητική. Είναι σημαντικό λοιπόν να εξασφαλίζουμε την ισχυρή ρύθμιση της πιθανότητας σφάλματος τύπου Ι, διότι αυτή μας βεβαιώνει ότι ο έλεγχος γίνεται κάτω από την αληθή και άγνωστη κατανομή δεδομένων.

2.1.4 Προσαρμοσμένες και μη τιμές-*p*, τιμές-*q*

Θεωρούμε αρχικά τον έλεγχο μίας υπόθεσης H_0 σε ένα προκαθορισμένο επίπεδο α και περιοχή απόρριψης Γ_a τέτοια ώστε (α) $\Gamma_{a1} \subseteq \Gamma_{a2}$ για $0 \le a_1 \le a_2 \le 1$ και (β) $\Pr(T \in \Gamma_\alpha \mid H_0 = 0) \le \alpha$ για $0 \le \alpha \le 1$. Χρησιμοποιώντας το στατιστικό T για τον αμφίπλευρο έλεγχο, οι περιοχές απόρριψης $\Gamma_a = (-\infty, -c_\alpha] \cup [c_\alpha, \infty)$ είναι τέτοιες ώστε $\Pr(T \in \Gamma_a \mid H_0 = 0) = a$. Τότε η τιμή-*p* για την παρατηρούμενη τιμή $T = t$ δίνεται από τη σχέση:

$$p - value\,(t) = \min_{\{\Gamma_a : t \in \Gamma_a\}} \Pr(T \in \Gamma_a \mid H_0 = 0)$$

Δηλαδή, η τιμή-*p* ορίζεται ως η ελάχιστη τιμή του σφάλματος τύπου Ι για όλες τις περιοχές απόρριψης Γ_a που περιέχουν την παρατηρούμενη τιμή $T = t$. Για έναν αμφίπλευρο έλεγχο η τιμή-*p* δίνεται από την πιθανότητα $p = \Pr(|T| \ge |t| \,|\, H_0)$ και έτσι όταν αυτή είναι πολύ μικρή έχουμε ένδειξη για την απόρριψη της μηδενικής υπόθεσης. Σημειώνεται ότι η απόρριψη της H_0 όταν $p \le a$ μας παρέχει τη ρύθμιση του σφάλματος τύπου Ι σε επίπεδο α.

Αυτός ο ορισμός των τιμών-*p*, τις οποίες θα τις αναφέρουμε ως ακατέργαστες τιμές-*p* ή απλά τιμές-*p*, μπορεί να επεκταθεί και στα προβλήματα του πολλαπλού ελέγχου υποθέσεων. Αν $p_i = \Pr(|T_i| \ge |t_i| \,|\, H_{0i})$ είναι η τιμή-*p* για την υπόθεση $H_{0i}, i = 1,...,m$, η διαδικασία πολλαπλού ελέγχου μπορεί να καθοριστεί με δύο τρόπους. Ο πρώτος τρόπος είναι μέσω των τιμών-*p* των ατομικών υποθέσεων, δηλαδή την απόρριψη της $H_{0i}, i = 1,...,m$ όταν $p_i \le a_i$, όπου το a_i έχει επιλεγεί ώστε να ρυθμίζει την πιθανότητα σφάλματος τύπου Ι (FWER, PCER, RFER ή FDR) σε ένα επίπεδο α. Ο δεύτερος τρόπος είναι μέσω των προσαρμοσμένων τιμών-*p*, οι οποίες μπορούν να συγκριθούν κατευθείαν με το προκαθορισμένο επίπεδο α ή την πιθανότητα σφάλματος τύπου Ι. Οι προσαρμοσμένες τιμές-*p* ορίζονται ως εξής.

Ορισμός 2.1. Για κάθε διαδικασία πολλαπλού ελέγχου, η ***προσαρμοσμένη τιμή-p***, η οποία αντιστοιχεί στον έλεγχο μίας υπόθεσης $H_{0i}, i = 1,...,m$, μπορεί να οριστεί ως το επίπεδο σημαντικότητας ολόκληρης της διαδικασίας του ελέγχου στο οποίο η H_{0i} οριακά απορρίπτεται, δεδομένων των τιμών όλων των στατιστικών. □

Η προσαρμοσμένη τιμή-*p* ορίζεται σε σχέση με κάποια διαδικασία ελέγχου. Αν και η αναλυτική παρουσίαση των διαδικασιών ελέγχου γίνεται σε επόμενη παράγραφο, για την καλύτερη κατανόηση του ορισμού θα θεωρήσουμε τον έλεγχο της υπόθεσης $H_{0i}, i = 1,...,m$, χρησιμοποιώντας τη γνωστή διαδικασία Bonferroni.

Συγκεκριμένα, αν χρησιμοποιήσουμε την FWER=a, η H_{0i} απορρίπτεται αν

$$p_i \le a_i \text{ όπου } a_i = \frac{FWER}{m}$$

και από την παραπάνω ανισότητα μπορούμε να γράψουμε:

$$p_i \le \frac{FWER}{m} \Rightarrow mp_i \le FWER \qquad (2.1)$$

Έτσι αν θέλουμε να ρυθμίσουμε την FWER, η FWER προσαρμοσμένη τιμή-*p* για την υπόθεση H_{0i} είναι:

$$\tilde{p}_i = \inf\{ a \in [0,1] : H_{0i} \text{ απορρίπτεται για } FWER = a \}$$

δηλαδή, η ελάχιστη τιμή του επιπέδου σημαντικότητας όλης της διαδικασίας για την οποία απορρίπτεται η μηδενική υπόθεση. Από τη σχέση (2.1) προκύπτει ότι η προσαρμοσμένη τιμή-*p* είναι:

$$\tilde{p}_i = mp_i$$

και η υπόθεση απορρίπτεται, με $FWER = a$, αν $\tilde{p}_i \le a$.

Όμοια ορίζονται και οι προσαρμοσμένες τιμές-*p* για όλες τις πιθανότητες σφάλματος τύπου I.

Το πλεονέκτημα που μας δίνουν οι προσαρμοσμένες τιμές-*p* είναι ότι δεν απαιτούν την επιλογή μιας συγκεκριμένης πιθανότητας σφάλματος τύπου I ή τον προκαθορισμό του επιπέδου σημαντικότητας *α*, αλλά μας δίνουν την δυνατότητα να επιλέξουμε για κάθε πείραμα τον κατάλληλο συνδυασμό απορρίψεων και λανθασμένα θετικών αποτελεσμάτων.

Σε αναλογία με τις προσαρμοσμένες τιμές-*p* μπορούμε, μέσω μιας συγκεκριμένης πιθανότητας σφάλματος τύπου I, της pFDR, να ορίσουμε και τις τιμές-*q*,. Έχοντας υπόψη τον ορισμό που δώσαμε παραπάνω για τις τιμές-*p*, θεωρούμε τον ελάχιστο αριθμό των σφαλμάτων τύπου I για όλες τις δυνατές περιοχές απόρριψης Γ_a που περιέχουν την παρατηρούμενη τιμή $T = t$. Αν $pFDR(\Gamma_a)$ η pFDR στην περίπτωση που κάθε υπόθεση

απορρίπτεται από την ίδια περιοχή απόρριψης Γ_a, τότε η τιμή-*q* ορίζεται ανάλογα από τη σχέση:

$$q\text{-}\ value\ (t) = \inf_{\{\Gamma_a : t \in \Gamma_a\}} pFDR(\Gamma_a)$$

Για τις τιμές-*q* θα μιλήσουμε εκτενέστερα στην ενότητα 2.2.3.

2.2 Διαδικασίες Πολλαπλού Ελέγχου

Κάθε μία από τις πιθανότητες σφάλματος τύπου I που ορίσαμε παραπάνω ρυθμίζεται σε επίπεδο *α* από μία συγκεκριμένη διαδικασία πολλαπλών ελέγχων. Διακρίνουμε τρεις τύπους διαδικασιών πολλαπλού ελέγχου, τις single-step, τις step-down και τις step-up διαδικασίες. Στις single-step διαδικασίες κάθε μία υπόθεση εξετάζεται ανεξάρτητα από τα αποτελέσματα των υπολοίπων, ενώ στις stepwise διαδικασίες η απόρριψη μιας συγκεκριμένης υπόθεσης βασίζεται όχι μόνο στο πλήθος των υποθέσεων αλλά και στα αποτελέσματα που προκύπτουν από τον έλεγχο των προηγούμενων υποθέσεων. Συγκεκριμένα, στις step-down διαδικασίες ελέγχονται διαδοχικά οι υποθέσεις ξεκινώντας από αυτήν που αντιστοιχεί στο πιο σημαντικό στατιστικό. Στη συνέχεια, αν αυτή απορριφθεί πηγαίνουμε βηματικά στην υπόθεση που αντιστοιχεί στο επόμενο λιγότερο σημαντικό στατιστικό. Αν σε κάποιο βήμα ο έλεγχος αποτύχει να απορρίψει την μηδενική υπόθεση τότε κανένας από τους υπόλοιπους ελέγχους δεν θα την απορρίπτει. Αντίθετα, στις step-up διαδικασίες ξεκινάμε με τον έλεγχο της υπόθεσης που αντιστοιχεί στο λιγότερο σημαντικό στατιστικό και αν δεν απορριφθεί συνεχίζουμε βηματικά με το επόμενο πιο σημαντικό στατιστικό. Αν κάποια υπόθεση απορριφθεί, θα απορριφθούν και όλες οι υπόλοιπες. Για παράδειγμα, έστω $p_{(1)} \le p_{(2)} \le p_{(3)} \le p_{(4)} \le p_{(5)}$ οι διατεταγμένες τιμές-*p* και $H_{(1)}, H_{(2)}, H_{(3)}, H_{(4)}, H_{(5)}$ οι υποθέσεις που αντιστοιχούν σε αυτές. Μία step-down διαδικασία ξεκινά με τον έλεγχο της υπόθεσης $H_{(1)}$ που αντιστοιχεί στην μικρότερη τιμή-*p*, δηλαδή στην $p_{(1)}$. Αν αυτή απορριφθεί τότε ο έλεγχος συνεχίζεται για την υπόθεση $H_{(2)}$. Αν όχι, τότε σταματά και αποτυγχάνει να απορρίψει και τις τέσσερις υπόλοιπες υποθέσεις. Αντίθετα, μία step-up διαδικασία ξεκινά με τον έλεγχο της υπόθεσης $H_{(5)}$ που αντιστοιχεί στην μεγαλύτερη τιμή-*p*,

δηλαδή στην $p_{(5)}$. Αν αυτή απορριφθεί τότε ο έλεγχος σταματά και απορρίπτει και τις τέσσερις υπόλοιπες υποθέσεις . Αν όχι, τότε συνεχίζεται για την υπόθεση $H_{(4)}$.

2.2.1 Ρύθμιση της Family Wise Error Rate

2.2.1.α Μέθοδοι Single-step.

1. Η πιο παραδοσιακή μέθοδος στον έλεγχο πολλαπλών υποθέσεων είναι η μέθοδος *Bonferroni*. Για το σύνολο των υποθέσεων $H_{01},...,H_{0m}$ και για πιθανότητα σφάλματος τύπου Ι ίση με a, κάθε υπόθεση Η $_{0i}$ ελέγχεται σε επίπεδο σημαντικότητας a_i, τέτοιο ώστε $\sum a_i = a$. Εδώ θεωρούμε ότι $a_i = \frac{a}{m}$. Η ερμηνεία της διαδικασίας αυτής βασίζεται στην ανισότητα *Bonferroni*. Υποθέτουμε ότι οι μηδενικές υποθέσεις είναι αληθείς για όλα τα *m*, δηλαδή ότι $m = m_0$, και ότι το επίπεδο σημαντικότητας για κάθε υπόθεση χωριστά είναι κοινό και ίσο με a_0. Έτσι, η πιθανότητα μη απόρριψης των μηδενικών υποθέσεων για όλα τα *m* είναι 1- ma_0.

Αν p_i είναι η ακατέργαστη τιμή-*p* για την υπόθεση Η $_{0i}$, τότε η Η $_{0i}$ απορρίπτεται όταν $mp_i \leq a$. Έτσι, οι αντίστοιχες single-step *Bonferroni* προσαρμοσμένες τιμές-*p* δίνονται από τη σχέση:

$$\tilde{p}_i = \min(mp_i, 1) \tag{2.2}$$

2. Η διαδικασία *Sidak* είναι ανάλογη με τη μέθοδο *Bonferroni* και απορρίπτει την υπόθεση $H_{0i}, i = 1,...,m$, όταν η τιμή-*p* είναι μικρότερη ή ίση του $1-(1-a)^{1/m}$. Υποθέτοντας ότι έχουμε *m* ανεξάρτητα στατιστικά $T_1, T_2,...,T_m$ για τον έλεγχο των υποθέσεων $H_{01}, H_{02},...,H_{0m}$ σε επίπεδο σημαντικότητας a_0 για την κάθε μία, τότε η πιθανότητα να μην απορρίψουμε καμμία υπόθεση δεδομένου ότι είναι όλες αληθείς είναι:

$$\Pr\left[\bigcap_{i=1}^{m} \left(\text{η } H_{0i} \text{ δεν απορρίπτεται} \right) \mid \text{όλες οι } H_{0i} \text{ είναι αληθείς} \right] =$$

$$\prod_{i=1}^{m} \Pr\left(\text{η } H_{0i} \text{ δεν απορρίπτεται} \mid \text{η } H_{0i} \text{ είναι αληθής} \right) =$$

$$\prod_{i=1}^{m}(1-a_0)=(1-a_0)^m$$

Έτσι, το επίπεδο σημαντικότητας για τον πολλαπλό έλεγχο σημαντικότητας θα είναι:

$$a=1-(1-a_0)^m \Rightarrow (1-a_0)^m=1-a \Rightarrow a_0=1-(1-a)^{1/m}.$$

Οι αντίστοιχες single-step *Sidak* προσαρμοσμένες τιμές-*p* δίνονται από τη σχέση:

$$\tilde{p}_i=1-(1-p_i)^m,\ i=1,...,m \qquad (2.3)$$

3. Μία άλλη μέθοδος η οποία λαμβάνει υπόψη την εξάρτηση μεταξύ των ελέγχων, είναι αυτή των Westfall and Young (1993). Σύμφωνα με αυτή, οι single-step *minP* προσαρμοσμένες τιμές-*p* δίνονται από τη σχέση:

$$\tilde{p}_i=\Pr\left(\min_{1\le k\le m}P_k\le p_i \mid H_0^c\right)$$

όπου H_0^c είναι η πλήρης μηδενική υπόθεση ($H_0^c=\bigcap_{i=1}^{m}H_{0i}$) και P_k η τυχαία μεταβλητή για τις ακατέργαστες τιμές-*p* της υπόθεσης *k*.

4. Ανάλογα ορίζονται και οι single-step *maxT* προσαρμοσμένες τιμές-*p*, με βάση τα στατιστικά T_i, από τη σχέση:

$$\tilde{p}_i=\Pr\left(\max_{1\le k\le m}|T_k|\ge|t_i|\,|H_0^c\right)$$

2.2.1.β Μέθοδοι Step-down.

Στις step-down μεθόδους θεωρούμε τις διατεταγμένες ακατέργαστες τιμές-*p* $p_{(1)}\le p_{(2)}\le\ldots\le p_{(m)}$ και τις αντίστοιχες υποθέσεις $H_{0(1)},H_{0(2)},...,H_{0(m)}$.

1. Σύμφωνα με τη διαδικασία *Holm* (1979) πραγματοποιείται ένας ακολουθιακός έλεγχος Bonferroni στο επίπεδο a ως εξής:

Ξεκινάμε συγκρίνοντας τη μικρότερη τιμή-*p* με την τιμή a/m. Αν $p_{(1)}\le a/m$ απορρίπτουμε την υπόθεση $H_{0(1)}$ και συνεχίζουμε τον έλεγχο με $m-1$ υποθέσεις, διαφορετικά δεν απορρίπτουμε καμμία υπόθεση. Συνεχίζοντας με τον ίδιο τρόπο, αν $p_{(2)}\le a/(m-1)$ απορρίπτουμε την υπόθεση $H_{0(2)}$, αν όχι δεν απορρίπτουμε καμία από τις υπόλοιπες υποθέσεις. Η διαδικασία αυτή συνεχίζεται για $m-2,...,1$ υποθέσεις μέχρι το

σημείο όπου καμμία απόρριψη δεν θα μπορεί να γίνει, δηλαδή είτε όταν αποδεχτούμε όλες τις υποθέσεις που απομένουν είτε όταν απορρίψουμε και την τελευταία υπόθεση $H_{0(m)}$. Στον ακολουθιακό έλεγχο *Bonferroni*, οι τιμές-*p* συγκρίνονται με τους αριθμούς $\frac{a}{m}, \frac{a}{m-1}, \ldots, \frac{a}{1}$ ενώ στην κλασσική μέθοδο *Bonferroni*, με $\frac{a}{m}$. Το γεγονός αυτό δηλώνει ότι η πιθανότητα απόρριψης για ένα σύνολο μηδενικών υποθέσεων χρησιμοποιώντας την κλασική μέθοδο *Bonferroni* είναι μικρότερη ή ίση από την αντίστοιχη πιθανότητα απόρριψης αν χρησιμοποιήσουμε τον ακολουθιακό έλεγχο *Bonferroni*.

Η χρήση του συγκεκριμένου ελέγχου βασίζεται στο παρακάτω θεώρημα.

Θεώρημα 2.1. *Εάν είναι δυνατόν να εμφανιστούν όλα τα υποσύνολα μηδενικών υποθέσεων τότε ο ακολουθιακός έλεγχος απόρριψης Bonferroni έχει πολλαπλό επίπεδο σημαντικότητας α.*
Απόδειξη. **(**Υπενθυμίζουμε ότι το S_0 είναι το σύνολο των m_0 αληθών μηδενικών υποθέσεων.)

Η πιθανότητα απόρριψης τουλάχιστον μίας από τις επιμέρους μηδενικές υποθέσεις είναι

$$\Pr\left(\bigcup_{i\in S_0}\left\{P_i \le \frac{a}{m_0}\right\}\right) \le \sum_{i\in S_0}\Pr\left(P_i \le \frac{a}{m_0}\right) = \sum_{i\in S_0}\frac{a}{m_0} = m_0\frac{a}{m_0} = a\,.$$

Αν το ενδεχόμενο $\left\{P_i > \frac{a}{m}\right\}$ συμβεί για όλα τα $i \in S_0$, δηλαδή αν αποδεχθούμε όλες τις αληθείς μηδενικές υποθέσεις, τότε $P_{(m-m_0+1)} > \frac{a}{m_0}$ και ο ακολουθιακός έλεγχος σταματά στο βήμα $m - m_0 + 1$ ή νωρίτερα, έχοντας δημιουργήσει ένα σύνολο από μη απορριφθείσες υποθέσεις. Αυτό περιέχει το σύνολο των αληθών μηδενικών υποθέσεων. □

Σημείωση. Με τη φράση για "όλα τα υποσύνολα μηδενικών υποθέσεων" θέλουμε να δηλώσουμε ότι οποιαδήποτε από τις *m* μηδενικές υποθέσεις ενδέχεται να είναι αληθής. Η έννοια αυτή χαρακτηρίζεται από τη φράση "*για ανεξάρτητους συνδυασμούς*".

Σύμφωνα με τα παραπάνω, αν ορίσουμε ως $J = \min\left\{i : p_{(i)} > a/(m-i+1)\right\}$ τότε απορρίπτουμε την υπόθεση $H_{0(i)}$, για τα $i = 1,\ldots,J-1$. Αν δε βρεθεί ένα τέτοιο J τότε

απορρίπτουμε όλες τις υποθέσεις. Οι αντίστοιχες step-down *Holm* προσαρμοσμένες τιμές-*p* δίνονται από τη σχέση:

$$\tilde{p}_{(i)} = \max_{k=1,\ldots,i} \left\{ \min\left[(m-k+1)\, p_{(k)}, 1 \right] \right\} \tag{2.4}$$

2. Μία άλλη διαδικασία ακολουθιακού ελέγχου προκύπτει αν αντικαταστήσουμε τις σταθερές $\frac{a}{m}, \frac{a}{m-1}, \ldots, \frac{a}{1}$ που χρησιμοποιήσαμε στον ακολουθιακό έλεγχο *Bonferroni* με τις σταθερές $1-(1-a)^{1/m}, 1-(1-a)^{1/(m-1)}, \ldots, 1-(1-a)^{1}$ που εμφανίζονται στη μέθοδο *Sidak*. Να σημειωθεί ότι οι τελευταίες είναι πάντοτε μεγαλύτερες από τις αντίστοιχες του ακολουθιακού ελέγχου *Bonferroni*.

Οι step-down *Sidak* προσαρμοσμένες τιμές-*p* δίνονται από τη σχέση:

$$\tilde{p}_{(i)} = \max_{k=1,\ldots,i} \left\{ 1-(1-p_{(k)})^{(m-k+1)} \right\} \tag{2.5}$$

3. Αντίστοιχα ορίζονται και οι Westfall and Young (1993) step-down *minP* προσαρμοσμένες τιμές-*p*:

$$\tilde{p}_{(i)} = \max_{k=1,\ldots,i} \left\{ P\left(\min_{l \in \{k,\ldots,m\}} P_l \le p_{(k)} \mid H_0^c \right) \right\}, \tag{2.6}$$

4. Τέλος οι step-down *maxT* προσαρμοσμένες τιμές-*p* ορίζονται ως

$$\tilde{p}_{(i)} = \max_{k=1,\ldots,i} P\left(\max_{l \in \{k,\ldots,m\}} |T_l| \ge |t_{(i)}| \mid H_0^c \right), \tag{2.7}$$

όπου $t_{(1)} \ge t_{(2)} \ge \ldots \ge t_{(m)}$ τα διατεταγμένα παρατηρούμενα στατιστικά.

2.2.1.γ Μέθοδοι Step-up.

Στις step-up μεθόδους θεωρούμε πάλι τις διατεταγμένες ακατέργαστες τιμές-*p* $p_{(1)} \le p_{(2)} \le \ldots \le p_{(m)}$ και τις αντίστοιχες υποθέσεις $H_{0(1)}, H_{0(2)}, \ldots, H_{0(m)}$, αλλά ξεκινάμε τον έλεγχο από τη μεγαλύτερη τιμή-*p*, δηλαδή την $p_{(m)}$.

1. Οι step-up μέθοδοι στηρίζονται κυρίως στη διαδικασία που παρουσίασε ο Simes (1986) και η οποία αποτελεί βελτίωση της διαδικασίας Bonferroni. Σύμφωνα με αυτή, συμβολίζουμε

με $p_{(1)} \le p_{(2)} \le ... \le p_{(m)}$ τις διατεταγμένες τιμές-p και με H_0 την τομή των υποθέσεων $H_{0(1)}, H_{0(2)}, ..., H_{0(m)}$, δηλαδή το ενδεχόμενο να απορρίψουμε ή να αποδεχτούμε ταυτόχρονα και τις m υποθέσεις. Απορρίπτουμε την H_0 αν για οποιοδήποτε $i = 1,...,m$ ισχύει $p_{(i)} \le ia/m$. Αυτή η διαδικασία έχει σφάλμα τύπου Ι ίσο με a για ανεξάρτητους ελέγχους, όπως αποδεικνύεται από το παρακάτω θεώρημα.

Θεώρημα 2.2. *Έστω $P_1, P_2, ..., P_m$ ανεξάρτητες $U(0,1)$ τυχαίες μεταβλητές και $P_{(1)}, P_{(2)}, ..., P_{(m)}$ οι αντίστοιχες διατεταγμένες τυχαίες μεταβλητές. Τότε ισχύει*

$$A_m(a) = \Pr\left(P_i > i\frac{a}{m} \; ; \; i = 1,...,m \right) = 1 - a \, .$$

Απόδειξη. Το αποτέλεσμα θα αποδειχθεί με επαγωγή. Για $m = 1$ είναι προφανές. Έστω ότι $A_{m-1}(a) = 1 - a, \forall a \in (0,1)$. Η $P_{(m)}$ έχει πυκνότητα mp^{m-1}, $0 < p < 1$, και είναι ανεξάρτητη από τις $P_{(1)}/P_{(m)}, ..., P_{(m-1)}/P_{(m)}$, η από κοινού κατανομή των οποίων είναι η κατανομή $m-1$ διατεταγμένων ανεξάρτητων $U(0,1)$ τυχαίων μεταβλητών.

Επομένως,

$$\Pr\left(\frac{P_{(1)}}{P_{(m)}} > \frac{a}{mp}, ..., \frac{P_{(m-1)}}{P_{(m)}} > \frac{(m-1)a}{mp} \,\middle|\, P_{(m)} = p \right) = \Pr\left(\frac{P_{(1)}}{P_{(m)}} > \frac{a}{mp}, ..., \frac{P_{(m-1)}}{P_{(m)}} > \frac{(m-1)a}{mp} \right)$$

$$= A_{m-1}\left(\frac{(m-1)a}{mp} \right)$$

και

$$A_m(a) = \Pr\left(P_{(1)} > \frac{a}{m}, ..., P_{(m-1)} > \frac{(m-1)a}{m}, P_{(m)} > a \right)$$

$$= \int_a^1 \Pr\left(\frac{P_{(1)}}{P_{(m)}} > \frac{a}{mp}, ..., \frac{P_{(m-1)}}{P_{(m)}} > \frac{(m-1)a}{mp} \mid P_{(m)} = p \right) mp^{m-1} dp$$

$$= \int_a^1 A_{m-1}\left(\frac{a(m-1)}{p \cdot m} \right) m \, p^{m-1} \, dp = \int_a^1 \left[1 - \frac{(m-1)a}{mp} \right] mp^{m-1} dp$$

$$= 1 - a^m - a\int_a^1 (m-1)\, p^{m-2} dp = 1 - a^m - a + a^{m-1} = 1 - a \, . \qquad \square$$

2. Αν και ο Simes (1986) απέδειξε ότι η διαδικασία αυτή ελέγχει την FWER υπό την μηδενική υπόθεση της τομής H_0 σε επίπεδο a, δεν έκανε λόγο για το πρόβλημα ελέγχου της ατομικής υπόθεσης H_{0i}. Για το σκοπό αυτό ο Hommel (1988) πρότεινε μία άλλη διαδικασία, επέκταση της θεωρίας του Simes, η οποία βασίζεται στην αρχή της «κλειστής διαδικασίας ελέγχου» των Marcus, Peritz και Gabriel (1976) που περιγράφουμε στη συνέχεια.

Αρχή Κλειστής Διαδικασίας Ελέγχου. Έστω $H = \{H_{0(1)}, H_{0(2)}, ..., H_{0(m)}\}$ ένα σύνολο m υποθέσεων. Ορίζουμε όλους τους δυνατούς συνδυασμούς υποθέσεων με το σύνολο $H_I = \{H_i : i \in I\}$ για όλα τα $I \in K$, όπου K το σύνολο των μη κενών υποσυνόλων του $\{1,...,m\}$. Για κάθε H_I θεωρούμε ότι γίνεται ένας έλεγχος που βασίζεται στο στατιστικό T_I. Τότε για ένα ορισμένο επίπεδο α, η H_I απορρίπτεται αν,

κάθε H_J απορρίπτεται σε επίπεδο α από το αντίστοιχο στατιστικό T_J, όπου

$$J \in K \text{ και } J \supseteq I.$$

Η πιθανότητα λανθασμένης απόρριψης μίας ή περισσοτέρων υποθέσεων κατά τον έλεγχο της H_I είναι μικρότερη ή ίση του q.

Κάνοντας χρήση της διαδικασίας αυτής, ξεκινάμε με τον έλεγχο της $H_I = \{H_i : i \in I\}$. Αν αυτός οδηγήσει σε απόρριψη της H_I σε επίπεδο α, συνεχίζουμε με τον έλεγχο κάθε υποσυνόλου των $m-1$ υποθέσεων. Εφόσον οι υποθέσεις συνεχίζουν να απορρίπτονται σε επίπεδο α συνεχίζουμε τον έλεγχο μέχρι να φτάσουμε σε υποσύνολα μίας υπόθεσης. Η διαδικασία αυτή μπορεί να επαναδιατυπωθεί με σκοπό την απόδοση των προσαρμοσμένων τιμών-p για κάθε H_I με τον παρακάτω τρόπο.

Έστω p_I η τιμή-p για τον έλεγχο της υπόθεσης H_I. Τότε η H_I απορρίπτεται αν $p_J \le a$ για όλες τις H_J με $J \supseteq I$. Έτσι, η προσαρμοσμένη τιμή-p για την H_I θα είναι η μεγαλύτερη από τις τιμές p_J.

Το μειονέκτημα της διαδικασίας είναι ότι στη γενική περίπτωση που κανείς θέλει να λάβει μία τιμή-p για κάθε ατομική υπόθεση H_{0i}, θα πρέπει να πραγματοποιήσει τον έλεγχο για κάθε πιθανό υποσύνολο υποθέσεων, δηλαδή θα πρέπει να πραγματοποιήσει $\sum_{i=1}^{m} \binom{m}{i} = 2^m - 1$

ελέγχους. Παρόλα αυτά, για ορισμένες περιπτώσεις, υπάρχουν πιο σύντομες προσεγγίσεις. Για τον έλεγχο της ατομικής υπόθεσης H_{0i}, ο Hommel (1988) πρότεινε μία διαδικασία που βασίζεται στην κλειστή διαδικασία ελέγχου και ως κριτήριο απόρριψης χρησιμοποιεί τον έλεγχο του Simes. Η διαδικασία αυτή περιγράφεται ως εξής:

Έστω j ο αριθμός των υποθέσεων στο μεγαλύτερο υποσύνολο υποθέσεων που δεν απορρίπτεται από τη διαδικασία του Simes. Θα είναι δηλαδή,

$$j = \max\left\{ i' \in \{1,...,m\} : p_{(m-i'+k)} > ka/i', k = 1,...,i' \right\}$$

και θα απορρίπτουμε την H_{0i} όταν $p_i \le \frac{a}{j}$. Διαφορετικά, αν δεν υπάρχουν υποσύνολα που να απορρίπτονται από τη διαδικασία του Simes τότε όλες οι H_{0i} απορρίπτονται.

3. Ο Hochberg (1988) παρουσίασε έναν παρόμοιο τρόπο επέκτασης της διαδικασίας του Simes για την ισχυρή ρύθμιση της FWER προτείνοντας την παρακάτω διαδικασία.

Υποθέτουμε ότι το σύνολο των μηδενικών υποθέσεων $H = \{H_{0(1)}, H_{0(2)},..., H_{0(m)}\}$, ικανοποιεί τη συνθήκη των ανεξάρτητων συνδυασμών [Holm (1979)]. Για κάθε σύνολο $H' \subseteq H$ με $m' \le m$ υποθέσεις, διατάσσουμε τις αντίστοιχες τιμές-p, $p_{(i_1)} \le ... \le p_{(i_{m'})}$. Η επέκταση της διαδικασίας του Simes, απορρίπτει κάθε υποσύνολο της τομής των υποθέσεων H_0', όταν για όλα τα $H'' \supseteq H'$

$$p_{(i_k)} \le k\frac{a}{m''}, \quad \text{για κάθε } H_{0(i_k)} \in H'', \quad k = 1,...,m',$$

όπου m'' ο αριθμός των υποθέσεων στο H''.

Έτσι ενώ η κλασσική διαδικασία του Simes έχει πιθανότητα σφάλματος FWER ίση με α υπό τη μηδενική υπόθεση, η επέκταση της διαδικασίας του Simes ρυθμίζει την FWER με την ισχυρή έννοια, δηλαδή για κάθε υποσύνολο H_0'.

Παρακάτω παρουσιάζουμε τη διατύπωση της διαδικασίας του Hochberg (1988) με την οποία μπορούμε να βγάλουμε συμπεράσματα για συγκεκριμένες υποθέσεις.

Η H_{0i} απορρίπτεται αν υπάρχει κάποιο i', με $1 \le i' \le m$, τέτοιο ώστε $p_{(i')} \le a/(m - i' + 1)$ και $p_i \le p_{(i')}$.

Οι δύο παραπάνω διαδικασίες αποτελούν βελτίωση της διαδικασίας Bonferroni. Σύμφωνα με τον Hommel (1989), η διαδικασία του Hommel (1988) είναι το ίδιο ισχυρή όσο και η

διαδικασία του Hochberg, ενώ σε ορισμένες περιπτώσεις οδηγεί σε ακόμα περισσότερες απορρίψεις.

2.2.2 Ρύθμιση της False Discovery Rate

Οι Benjamini και Hochberg (1995) πρότειναν μία διαφορετική προσέγγιση του πολλαπλού ελέγχου υποθέσεων μέσω της ρύθμισης της FDR. Όπως ορίστηκε παραπάνω, η FDR είναι το αναμενόμενο ποσοστό των σφαλμάτων τύπου I μεταξύ των υποθέσεων που έχουν απορριφθεί. Αν όλες οι υποθέσεις που εξετάζουμε είναι αληθείς, τότε η ρύθμισή της συνεπάγεται και τη ρύθμιση της FWER (όταν $m = m_0$ έχουμε δείξει ότι $FDR = FWER$). Όταν όμως ο αριθμός των απορρίψεων είναι μεγάλος και επομένως πολλές από τις υποθέσεις μπορεί να είναι ψευδείς, είναι προτιμότερο να ρυθμίσουμε το ποσοστό των σφαλμάτων, αντί του σφάλματος από μία λανθασμένη απόρριψη.

Η διαδικασία που πρότειναν οι Benjamini και Hochberg (1995) για τη ρύθμιση της FDR στο επίπεδο $\frac{m_0}{m} q^* \le q^*$ είναι η εξής.

Αν $p_{(1)} \le p_{(2)} \le \ldots \le p_{(m)}$ οι διατεταγμένες παρατηρούμενες τιμές-p και $J = \max\left\{ i : p_{(i)} \le \frac{i}{m} q^* \right\}$, τότε απορρίπτουμε τις μηδενικές υποθέσεις $H_{0(i)}, i = 1,\ldots,J$. Αν δε βρεθεί τέτοιο i, δεν απορρίπτουμε καμμία υπόθεση.

Στην περίπτωση που όλες οι υποθέσεις είναι αληθείς, δηλαδή $m_0 = m$, η διαδικασία περιορίζεται στην θεωρία του Simes για την τομή των υποθέσεων.

Οι αντίστοιχες step-up προσαρμοσμένες τιμές-p δίνονται από τη σχέση:

$$\tilde{p}_{(i)} = \min_{k=i,\ldots,m} \left\{ \min\left(\frac{m}{k} p_{(k)}, 1 \right) \right\}.$$

Θεώρημα 2.3. *Για ανεξάρτητα στατιστικά και για κάθε περιορισμό από ψευδείς μηδενικές υποθέσεις, η παραπάνω διαδικασία ρυθμίζει την FDR στο* q^*.

Η απόδειξη του θεωρήματος βασίζεται στο παρακάτω λήμμα.

Λήμμα 2.1. *Για κάθε m_0 ανεξάρτητες τιμές-p, $0 \leq m_0 \leq m$, οι οποίες αντιστοιχούν στις ορθές μηδενικές υποθέσεις και για οποιεσδήποτε τιμές των $m_1 = m - m_0$ τιμών-p που αντιστοιχούν στις ψευδείς μηδενικές υποθέσεις, η διαδικασία πολλαπλού ελέγχου των Benjamini και Hochberg ικανοποιεί την ανισότητα:*

$$E\left(Q \mid P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) \leq \frac{m_0}{m} q^*$$

όπου Q το ποσοστό των σφαλμάτων τύπου I μεταξύ των υποθέσεων που έχουν απορριφθεί. (*Απόδειξη του λήμματος στο Παράρτημα*)

Ολοκληρώνοντας την παραπάνω ανισότητα έχουμε:

$$E(Q) \leq \frac{m_0}{m} q^* \leq q^*$$

και η FDR ρυθμίζεται στο επίπεδο q^*.

□

Συγκρίνοντας τη διαδικασία του Hochberg (1988) για τη ρύθμιση της FWER που περιγράψαμε στην ενότητα 2.2.2γ και τη διαδικασία των Benjamini και Hochberg (1995) για τη ρύθμιση της FDR, παρατηρούμε ότι υπάρχει μία σχέση μεταξύ αυτών όταν το q^* επιλεγεί να είναι ίσο με *α*. Βλέπουμε ότι και οι δύο είναι step-up διαδικασίες οι οποίες αρχίζουν με τη σύγκριση της $p_{(m)}$ με το α. Αν $p_{(m)} \leq a$ απορρίπτουν όλες τις υποθέσεις, διαφορετικά συνεχίζουν τον έλεγχο με μικρότερες τιμές-*p*. Επίσης, και οι δύο διαδικασίες καταλήγουν (αν δεν έχουν σταματήσει νωρίτερα) στη σύγκριση του $p_{(1)}$ με το a/m. Η διαφορά μεταξύ αυτών των μεθόδων φαίνεται στα ενδιάμεσα βήματα αυτών, όπου στη μεν πρώτη οι τιμές $p_{(i)}$ συγκρίνονται με τις τιμές $a/(m-i+1)$, ενώ στη δεύτερη με τις τιμές ia/m. Οι σταθερές στη διαδικασία των Benjamini και Hochberg ελαττώνεται γραμμικά και είναι πάντα μεγαλύτερες από τις αντίστοιχες σταθερές της διαδικασίας του Hochberg οι οποίες ελαττώνονται με πιο γρήγορο ρυθμό. Από το γεγονός αυτό συμπεραίνουμε ότι η διαδικασία ελέγχου της FDR μας δίνει τουλάχιστον ίσες απορρίψεις με αυτή για τον έλεγχο της FWER και είναι ακόμα πιο ισχυρή από άλλες διαδικασίες ελέγχου της FWER, όπως για παράδειγμα αυτή του Holm (1979).

Οι Benjamini και Liu (1999) πρότειναν και μία step-down διαδικασία για τη ρύθμιση της FDR όταν τα στατιστικά είναι ανεξάρτητα. Συγκρίνοντάς τη με τη step-up διαδικασία των Benjamini και Hochberg (1995), απέδειξαν ότι αυτή η step-down διαδικασία δεν υπερτερεί έναντι της step-up, παρά μόνο στην περίπτωση που ο αριθμός των υποθέσεων είναι μικρός και το ποσοστό των ψευδών μηδενικών υποθέσεων ως προς τις αληθείς είναι μεγάλο.

Η step-down διαδικασία των Benjamini και Liu έχει ως εξής. Θεωρούμε τα m κρίσιμα σημεία

$$\delta_i = 1 - \max\left\{0, 1 - \frac{m}{m-i+1}q\right\}^{1/(m-i+1)}, 1 \le i \le m.$$

Παρατηρούμε ότι $0 \le \delta_1 \le \ldots \le \delta_m \le 1$.

Ορίζουμε ως J το $\min\{i : p_{(i)} > \delta_i\}$ και απορρίπτουμε τις υποθέσεις $H_{0(1)},\ldots,H_{0(J-1)}$. Αν δε βρεθεί τέτοιο J τότε απορρίπτονται όλες οι υποθέσεις.

Η διαδικασία αυτή λειτουργεί ως εξής: ξεκινώντας από τη μικρότερη τιμή-p ελέγχουμε αν $P_{(1)} > \delta_1$. Αν ισχύει σταματάμε και δεν απορρίπτουμε καμία υπόθεση διαφορετικά απορρίπτουμε την $H_{(1)}$ και συνεχίζουμε με τον έλεγχο των υπολοίπων.

Οι αντίστοιχες step-down προσαρμοσμένες τιμές-p δίνονται από τη σχέση:

$$\tilde{p}_{(i)} = \max_{k=1,\ldots,i}\left\{\min\left(\frac{P_{(k)}}{\delta_k}, 1\right)\right\}$$

Η διαδικασία αυτή ρυθμίζει την FDR σε επίπεδο q όπως αποδεικνύεται από το παρακάτω θεώρημα.

Θεώρημα 2.4. *Αν οι τιμές-p $P_1,\ldots,P_m$ είναι ανεξάρτητες, τότε η step-down διαδικασία ρυθμίζει την FDR στο επίπεδο q.*

Απόδειξη. Αν $m_0 = 0$ τότε $V = 0, Q = 0$ και $E(Q) = 0$. Επίσης, αν $m_0 = m$ τότε $V = R$ και

$$E(Q) = E\left(I(V > 0)\right) = P(V > 0) = \Pr\left(P_{(1)} < \delta_1\right) = 1 - \left[\Pr\left(P_{(1)} > \delta_1\right)\right]^m = 1 - (1 - \delta_1)^m.$$

Θεωρούμε τώρα την περίπτωση όπου $1 \le m_0 \le m - 1$. Έστω $m_1 = m - m_0 > 0$, $P_1', \ldots, P_{m_1}'$ οι τιμές-p που αντιστοιχούν στις m_1 ψευδείς υποθέσεις και $P_1^*,\ldots,P_{m_0}^*$ οι τιμές-p που αντιστοιχούν στις m_0 αληθείς υποθέσεις. Θα δείξουμε ότι $E\left(Q \mid P_1', \ldots, P_{m_1}'\right) \le q$.

Έστω $P'_{(1)} \le ... \le P'_{(m_1)}$ οι διατεταγμένες τιμές-*p*. Ορίζουμε με $S, 0 \le S \le m_1$ το μεγαλύτερο ακέραιο *j* που ικανοποιεί τις συνθήκες $P'_{(1)} \le \delta_1, ..., P'_{(j)} \le \delta_j$. Αν $P'_{(1)} > \delta_1$ τότε $S = 0$. Έτσι έχουμε:

$$
\begin{aligned}
E\left(Q \,|\, P'_1, ..., P'_{m_1} \right) &= E\left(\frac{V}{R} I\left(V > 0\right) | \, P'_1, ..., P'_{m_1} \right) \\
&\le E\left(\frac{V}{S+V} I\left(V > 0\right) | \, P'_1, ..., P'_{m_1} \right) \\
&\le \frac{m_0}{S+m_0} \cdot E\left(I\left(V > 0\right) | \, P'_1, ..., P'_{m_1} \right) \\
&\le \frac{m_0}{S+m_0} \cdot \Pr\left(\min\left(P_1^*, ..., P_{m_0}^* \right) \le \delta_{S+1} \right) \\
&\le \frac{m_0}{S+m_0} \cdot \left[1 - \left(1 - \delta_{S+1}\right)^{m-S} \right] \\
&= \frac{m_0}{S+m_0} \cdot \left[1 - \left(1 - \delta_{S+1}\right)^{m} \right] \\
&\le \frac{m_0 \cdot m}{\left(S+m_0\right)\left(m-S\right)} q \le q \\
&= \frac{m_0}{S+m_0} \cdot \min\left(1, \frac{m}{m-S} q \right)
\end{aligned}
$$

□

Οι Benjamini και Liu (1999) πραγματοποίησαν τη σύγκριση μεταξύ της καινούργιας step-down μεθόδου και της step-up μεθόδου των Benjamini και Hochberg (1995) για τη ρύθμιση της FDR, στηριζόμενοι στις κρίσιμες τιμές αυτών. Συγκεκριμένα, η step-up διαδικασία για τη ρύθμιση της FDR στο *q* χρησιμοποιεί τις

$$c_i = \left(\frac{i}{m} \right) q, \; 1 \le i \le m \,,$$

ενώ η step-down διαδικασία χρησιμοποιεί τις κρίσιμες τιμές

$$\delta_i = 1 - \max\left\{ 0 \,, 1 - \frac{m}{m-i+1} q \right\}^{1/(m-i+1)}, \; 1 \le i \le m \,.$$

Παρόλο που οι της step-down διαδικασίας είναι μεγαλύτερες από τις αντίστοιχες κρίσιμες τιμές της step-up διαδικασίας, η πιο ισχυρή κατεύθυνση βημάτων της step-up δεν εγγυάται

ότι η step-down διαδικασία είναι πιο ισχυρή. Σύμφωνα με τους Benjamini και Liu η σύγκριση των δύο διαδικασιών οδηγεί στις παρακάτω παρατηρήσεις. Προφανώς, $\delta_i > c_i$ για $i = 1,...,m$, με τον περιορισμό όμως ότι το m είναι μικρό. Για $m \geq 4$, $\delta_i < c_i$ για όλα τα i μέχρι κάποιο $i_0(m)$ και τότε το δ_i γίνεται μεγαλύτερο. Συγκρίνοντας τα $(1-c_i)^{m+i-1}$ και $(1-\delta_i)^{m+i-1}$ διαπιστώνεται ότι είναι περίπου ίσα όταν το $i_0(m)$ ικανοποιεί τη σχέση $i_0(m)(m+1-i_0(m))^2 \approx m^2$. Στην περίπτωση όπου το m είναι μικρό θα ισχύει $i_0(m) \approx (m+1)-\sqrt{m+1}$, ενώ αν είναι μεγάλο $i_0(m) = m(1-q)$. Συμπεραίνουμε λοιπόν ότι, εκτός από την περίπτωση που το m είναι μικρό, όταν ο αριθμός των ψευδών υποθέσεων είναι το πολύ $i_0(m)$, η step-up διαδικασία είναι πιο ισχυρή από τη step-down.

2.2.3 Ρύθμιση της Positive False Discovery Rate

Όπως αναφέραμε και σε προηγούμενη παράγραφο, η positive false discovery rate ορίζεται από τη σχέση:

$$pFDR = E\left(\frac{V}{R} \mid R > 0\right).$$

Με άλλα λόγια, εκφράζει το αναμενόμενο ποσοστό των σφαλμάτων τύπου I ανάμεσα στις υποθέσεις που έχουν απορριφθεί, δεδομένου ότι έχουμε μία τουλάχιστον απόρριψη.

Γενικά δεν υπάρχει κάποια διαδικασία που να μπορεί να μας δώσει τον ισχυρό ή τον ασθενή έλεγχο της pFDR, διότι για $m_0 = m$ η pFDR ισούται με τη μονάδα. Παρόλα αυτά, στην περίπτωση των μικροσυστοιχιών είναι απίθανο να μην υπάρχει κάποιο γονίδιο που να είναι διαφορετικά εκφρασμένο, οπότε $m_0 \neq m$ και η pFDR εκτιμάται κάτω από την άγνωστη μηδενική υπόθεση H_{S_0} ($H_{S_0} = \bigcap_{i \in S_0} H_{0i}$, $S_0 \subseteq \{1,...,m\}$). Η εκτίμηση αυτή μπορεί να γίνει με τη χρήση των τιμών-q, οι οποίες κατά μία έννοια, όπως θα δούμε παρακάτω, μας παρέχουν ένα τρόπο προσαρμογής των τιμών-p κάτω από την H_{S_0} ο οποίος οδηγεί στη ρύθμιση της pFDR.

Υποθέτουμε ότι θέλουμε να πραγματοποιήσουμε m παρόμοια τεστ για μία μηδενική υπόθεση έναντι μιας εναλλακτικής βασιζόμενοι στα στατιστικά $T_1, T_2, ..., T_m$. Για μία περιοχή απόρριψης Γ, ορίζουμε την pFDR από τη σχέση:

$$pFDR\left(\Gamma\right)=E\left(\frac{V\left(\Gamma\right)}{R\left(\Gamma\right)}\mid R\left(\Gamma\right)>0\right)$$

όπου

$$V\left(\Gamma\right)=\#\left\{\text{ αληθών μηδενικών υποθέσεων }:T_i\in\Gamma\right\}$$

και

$$R\left(\Gamma\right)=\#\left\{\text{ μηδενικών υποθέσεων }:T_i\in\Gamma\right\}.$$

Με $H_i=0$ δηλώνουμε ότι η υπόθεση για το *i* γονίδιο είναι αληθής και με $H_i=1$ ότι είναι ψευδής για $i=1,...,m$. Θεωρούμε επίσης ότι η εκ των προτέρων πιθανότητα μία υπόθεση να είναι αληθής είναι π_0. Επομένως, αν οι H_i είναι ανεξάρτητες και ισόνομες (*i.i.d*) τυχαίες μεταβλητές, τότε αποτελούν τυχαίο δείγμα από την κατανομή Bernoulli με $\Pr\left(H_i=0\right)=\pi_0$ και $\Pr\left(H_i=1\right)=1-\pi_0=\pi_1$.

Θα δείξουμε ότι η pFDR ισούται με την πιθανότητα

$$\Pr\left(H=0\mid T\in\Gamma\right), \tag{2.8}$$

η οποία δεν εξαρτάται από το πλήθος των υποθέσεων *m*. Γράφουμε $\Pr\left(H=0\mid T\in\Gamma\right)$ παραλείποντας το δείκτη, διότι η πιθανότητα $\Pr\left(H_i=0\mid T_i\in\Gamma\right)$ είναι ίδια για κάθε $i=1,...,m$.

Ας δούμε αρχικά την περίπτωση $m=1$. Κάτω από τις παραπάνω προϋποθέσεις, η πιθανότητα ενός λανθασμένα θετικού αποτελέσματος είναι ίση με $\Pr\left(H=0\mid T\in\Gamma\right)$. Δεδομένου ότι $T\in\Gamma$, το πηλίκο $\frac{V\left(\Gamma\right)}{R\left(\Gamma\right)}$ είναι ίσο με 0 ή 1 ανάλογα με το αν έχουμε ορθά θετικό ή λανθασμένα θετικό αποτέλεσμα, αντίστοιχα. Έτσι, μπορούμε εύκολα να διαπιστώσουμε ότι $pFDR=\Pr\left(H=0\mid T\in\Gamma\right)$. Για $m>1$ παρουσιάζουμε το παρακάτω θεώρημα του Storey (2001).

Θεώρημα 2.5. *Έστω ότι για τους m παρόμοιους ελέγχους υποθέσεων με τα στατιστικά* $T_1,T_2,...,T_m$ *και περιοχή απόρριψης Γ ισχύουν τα εξής:*

(α) Οι $\left(T_i,H_i\right)$ *είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές.*

(β) $T_i\mid H_i\sim\left(1-H_i\right)\cdot F_0+H_i\cdot F_1$.

(γ) $H_i \sim Bernoulli(\pi_1)$ *για* $i = 1,...,m$, *όπου* $\pi_0 = 1 - \pi_1$ *η εκ των προτέρων πιθανότητα. Τότε:*

$$pFDR = \Pr(H = 0 \mid T \in \Gamma)$$

Απόδειξη. Η pFDR μπορεί να γραφεί στην εξής μορφή:

$$\begin{aligned} pFDR(\Gamma) &= E\left(\frac{V(\Gamma)}{R(\Gamma)} \mid R(\Gamma) > 0 \right) \\ &= E\left[E\left(\frac{V(\Gamma)}{R(\Gamma)} \middle| R(\Gamma) \right) \middle| R(\Gamma) > 0 \right] \\ &= \sum_{k=1}^{m} E\left(\frac{V(\Gamma)}{R(\Gamma)} \mid R(\Gamma) = k \right) \cdot \Pr\left(R(\Gamma) = k \mid R(\Gamma) > 0 \right) \\ &= \sum_{k=1}^{m} E\left(\frac{V(\Gamma)}{k} \mid R(\Gamma) = k \right) \cdot \Pr\left(R(\Gamma) = k \mid R(\Gamma) > 0 \right) \end{aligned}$$

Εφόσον τα στατιστικά είναι ανεξάρτητα, η $V(\Gamma) \mid R(\Gamma) = k$ είναι μία διωνυμική τυχαία μεταβλητή με πιθανότητα επιτυχίας $\Pr(H = 0 \mid T \in \Gamma)$. Επομένως,

$$E\left(V(\Gamma) \mid R(\Gamma) = k \right) = k \cdot \Pr(H = 0 \mid T \in \Gamma).$$

Έτσι προκύπτει ότι

$$\begin{aligned} pFDR(\Gamma) &= \sum_{k=1}^{m} \frac{k \cdot \Pr(H = 0 \mid T \in \Gamma)}{k} \cdot \Pr\left(R(\Gamma) = k \mid R(\Gamma) > 0 \right) \\ &= \Pr(H = 0 \mid T \in \Gamma) \end{aligned}$$

□

Εφόσον η pFDR εκφράζεται από την πιθανότητα $\Pr(H = 0 \mid T \in \Gamma)$ είναι λογικό να συσχετιστεί με το σφάλμα τύπου Ι. Θα μπορούσε κανείς να την ονομάσει ως "Μπεϋζιανό εκ των υστέρων σφάλμα τύπου Ι" (Storey, 2001).

Η $pFDR = \Pr(H = 0 \mid T \in \Gamma)$ μας δίνει ένα γενικό μέτρο αλλά δεν μας παρέχει συγκεκριμένες πληροφορίες για τις τιμές του κάθε στατιστικού. Ένα μέτρο σημαντικότητας για την τιμή του κάθε στατιστικού μπορούμε να πάρουμε από την τιμή-*q*. Στην παράγραφο 2.1.3 ορίσαμε τις τιμές-*q* από τη σχέση:

$$q - value(t) = \inf_{\{\Gamma:\, t \in \Gamma\}} pFDR(\Gamma). \tag{2.9}$$

Η τιμή-*q* είναι ένα μέτρο της ισχύος ενός παρατηρούμενου στατιστικού σε σχέση με την pFDR. Είναι η μικρότερη τιμή που μπορεί να πάρει η pFDR όταν απορρίπτουμε ένα στατιστικό με τιμή t, για ένα σύνολο περιοχών απόρριψης. Κάτω από τις υποθέσεις του Θεωρήματος 2.5, μπορούμε να δούμε ότι η τιμή-*q* έχει μία ακόμα πιο ερμηνεύσιμη σχέση με την τιμή-*p*. Στην ενότητα 2.1.3 ορίσαμε την τιμή-*p* από την σχέση

$$p-value\,(t)=\min_{\{\Gamma_a:t\in\Gamma_a\}}\Pr(T\in\Gamma_a\,|\,H_0=0).$$

Συγκρίνοντας τη σχέση αυτή με την (2.9), φαίνεται ότι η τιμή-*q* είναι το Μπεϋζιανό ανάλογο της τιμής-*p* και καλείται "εκ των υστέρων Μπεϋζιανή τιμή-*p*" (posterior Bayesian *p*-value). Είναι η μικρότερη εκ των υστέρων πιθανότητα να απορριφθεί η υπόθεση H ($H=0$) για όλες τις περιοχές απόρριψης που περιέχουν το στατιστικό. Καλείται τιμή-*q* διότι είναι παρόμοια με την τιμή-*p* με τα γεγονότα $\{T\in\Gamma\}$ και $\{H=0\}$ αντεστραμμένα.

Σημείωση: Από τον ορισμό της προσαρμοσμένης τιμής-*p* προκύπτει ότι η τιμή-*q* δεν είναι η "pFDR προσαρμοσμένη τιμή-*p*". Ο Shaffer (1995) αναφέρει:

> "Δεδομένης μιας διαδικασίας ελέγχου, η προσαρμοσμένη τιμή-*p* που αντιστοιχεί στον έλεγχο μιας υπόθεσης H_i μπορεί να οριστεί ως το επίπεδο ολόκληρης της διαδικασίας στο οποίο η H_i οριακά απορρίπτεται, δεδομένων των τιμών όλων των στατιστικών".

Έτσι, εφόσον η pFDR δεν μπορεί να ρυθμιστεί από κάποια διαδικασία, δεν μπορεί και να χρησιμοποιηθεί για να ορίσουμε την προσαρμοσμένη τιμή-*p*.

Ο Storey (2001) για να εισαγάγει την έννοια της τιμή-*q* παρουσίασε το παρακάτω παράδειγμα.

Παράδειγμα 2.1. Θεωρούμε ότι οι *m* υποθέσεις είναι της μορφής $\theta=0$ έναντι του $\theta=2$ και ότι οι τυχαίες μεταβλητές $T_1,\ldots,T_m$ ακολουθούν την κανονική κατανομή $N(\theta,1)$. Συγκεκριμένα, οι (T_i,H_i) είναι ανεξάρτητες και ισόμονες τυχαίες μεταβλητές με

$$T_i\,|\,H_i\sim(1-H_i)\cdot N(0,1)+H_i\cdot N(2,1)$$

Η τιμή-*p* για την τιμή $T_i=t_i$ μπορεί να υπολογιστεί από τη σχέση

$$p-value(t_i)=\Pr(T\geq t_i\,|\,H=0)=1-\Phi(t_i),$$

όπου Φ η σ.κ. της τυπικής κανονικής κατανομής.

Δηλαδή, η p - $value(t_i)$ μας δίνει το σφάλμα τύπου Ι αν απορρίψουμε κάποια υπόθεση με στατιστικό μεγαλύτερο ή ίσο του t_i.

Από το Θεώρημα 2.5 η pFDR μπορεί να γραφεί ως εξής:

$$pFDR\left(\{T \geq t_i\}\right) = \Pr\left(H = 0 \mid T \geq t_i\right).$$

Άρα, η pFDR βρίσκεται σε αναλογία με την τιμή $p - value(t_i) = \Pr(T \geq t_i \mid H = 0)$. Για το λόγο αυτό λέμε ότι η $pFDR\left(\{T \geq t_i\}\right)$ είναι το Μπεϋζιανό ανάλογο της τιμής $p - value(t_i)$.

Η $pFDR\left(\{T \geq t_i\}\right)$ είναι η τιμή-q, $q - value(t_i)$, που ορίσαμε προηγουμένως. Σε πολλές περιπτώσεις, η τιμή-q είναι η pFDR που προκύπτει από την απόρριψη μιας υπόθεσης η οποία αντιστοιχεί σε ένα στατιστικό μεγαλύτερο ή ίσο του t_i, μεταξύ όλων των m υποθέσεων, αλλά μπορεί να οριστεί και με έναν πιο γενικό τρόπο, ανάλογο με αυτόν των τιμών-p. □

Ο Storey (2002) πρότεινε μία διαφορετική διαδικασία όσον αφορά στη ρύθμιση της πιθανότητας pFDR αλλά και της FDR. Οι διαδικασίες που έχουμε αναφέρει μέχρι στιγμής όπως για παράδειγμα, η ρύθμιση της FDR η οποία συνεπάγεται την μέθοδο της ακολουθιακής απόρριψης των τιμών-p βασισμένες σε παρατηρούμενα δεδομένα, συνήθως προκαθορίζουν το ποσοστό σφάλματος και έπειτα εκτιμούν την αντίστοιχη περιοχή απόρριψης. Η καινούργια προσέγγιση που θα περιγράψουμε παρακάτω καθορίζει αρχικά την περιοχή απόρριψης και στη συνέχεια εκτιμά το αντίστοιχο ποσοστό σφάλματος.

Όταν όλες οι μηδενικές υποθέσεις είναι αληθείς, η συνηθισμένη ακολουθιακή μέθοδος των τιμών-p δεν μπορεί να εφαρμοστεί διότι $pFDR = 1$. Έτσι για να ρυθμίσουμε την pFDR πρέπει να την εκτιμήσουμε για μία συγκεκριμένη περιοχή απόρριψης.

Μία ακολουθιακή μέθοδος που χρησιμοποιεί τις τιμές-p μάς επιτρέπει να καθορίσουμε το ποσοστό σφάλματος και να εκτιμήσουμε την περιοχή απόρριψης. Στην περίπτωση ρύθμισης της FWER αυτή η μέθοδος έχει έννοια. Επειδή η FWER μετράει την πιθανότητα να κάνουμε ένα η περισσότερα σφάλματα τύπου Ι, μπορούμε να καθορίσουμε εκ των προτέρων την πιθανότητα να συμβεί αυτό. Αντίθετα, δεν μπορούμε να πούμε το ίδιο για την FDR. Για παράδειγμα, ας θεωρήσουμε ότι ελέγχουμε 1000 υποθέσεις και αποφασίζουμε να ρυθμίσουμε την FDR σε επίπεδο 5%. Το αν αυτό είναι μια καλή επιλογή εξαρτάται από τον αριθμό των υποθέσεων που απορρίπτονται. Αν απορριφθούν 100 υποθέσεις είναι μια καλή επιλογή, αν απορριφθούν μόνο 2 υποθέσεις τότε δεν είναι.

Έτσι, ο εξ αρχής καθορισμός της περιοχής απόρριψης μπορεί να φανεί πιο χρήσιμος όταν χρησιμοποιούμε την pFDR ή την FDR. Για παράδειγμα, όταν πραγματοποιούμε πολλούς ελέγχους υποθέσεων, έχει νόημα να απορρίψουμε όλες τις υποθέσεις με τιμές-*p* μικρότερες από 0.05 ή 0.01.

Θεωρούμε πάλι τη σχέση (2.8) που αποδείξαμε στο Θεώρημα 2.5 για την pFDR,

$$pFDR = \Pr\left(H = 0 \mid T \in \Gamma\right).$$

Επειδή οι απορρίψεις βασίζονται στις τιμές-*p*, συνεπάγεται ότι όλες οι περιοχές απόρριψης είναι της μορφής $[0,\gamma]$ για κάποιο $\gamma > 0$. Με άλλα λόγια, μία υπόθεση απορρίπτεται αν η αντίστοιχη τιμή-*p* ανήκει στο διάστημα $[0,\gamma]$. Έτσι, συμβολίζουμε την περιοχή απόρριψης με γ, το οποίο αναφέρεται στο διάστημα $[0,\gamma]$ και γράφουμε την pFDR ως εξής:

$$pFDR(\gamma) = \frac{\pi_0 \Pr\left(P \le \gamma \mid H = 0\right)}{\Pr\left(P \le \gamma\right)} = \frac{\pi_0 \gamma}{\Pr\left(P \le \gamma\right)}$$

όπου *P* η τυχαία μεταβλητή που αντιστοιχεί στην τιμή-*p* που προκύπτει από κάθε έλεγχο. Στην περίπτωση που τα στατιστικά είναι ανεξάρτητα, οι τιμές-*p* είναι ανταλλάξιμες (exchangeable) (δες ορισμό 3.1, σελίδα 49) δηλαδή κάθε μία προέρχεται από την κατανομή υπό τη μηδενική υπόθεση με πιθανότητα π_0 και υπό την εναλλακτική με πιθανότητα π_1.

Εφόσον οι $m_0 p$ από τις τιμές-*p* αναμένεται να είναι μηδενικές, τότε είναι πιο πιθανό οι μεγαλύτερες τιμές-*p* να αντιστοιχούν στις υποθέσεις που είναι αληθείς και το π_0 μπορεί να εκτιμηθεί από την:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m} = \frac{W(\lambda)}{(1-\lambda)m} p$$

όπου το λ είναι γνωστή σταθερά, $p_1,...,p_m$ οι παρατηρούμενες τιμές-*p* και $W(\lambda) = \#\{p_i > \lambda\}$. Επίσης, μία εκτίμηση για την $\Pr(P \le \gamma)$ είναι η

$$\widehat{\Pr}\left(P \le \gamma\right) = \frac{\#\{p_i \le \gamma\}}{m} = \frac{R(\gamma)}{m}$$

όπου $R(\gamma) = \#\{p_i \le \gamma\}$ και μία καλή εκτίμηση για την pFDR είναι η

$$\widehat{pFDR}(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\Pr\left(P \le \gamma\right)} = \frac{W(\lambda)\gamma}{(1-\lambda)R(\gamma)}$$

Για την εκτίμηση της pFDR θα κάνουμε δύο μικρές προσαρμογές. Αντικαθιστούμε το $R(\gamma)$ με το $R(\gamma)\vee 1 \ = \max\{R(\gamma),1\}$, διότι όταν το $R(\gamma)=0$ η παραπάνω σχέση δεν ορίζεται. Επίσης, επειδή η ποσότητα $1-(1-\gamma)^m$ είναι ένα κάτω φράγμα για την $P\{R(\gamma)>0\}$ και η pFDR δεσμεύεται στο $R(\gamma)>0$ μπορούμε να διαιρέσουμε με $1-(1-\gamma)^m$. Με άλλα λόγια, το πηλίκο $\dfrac{\gamma}{1-(1-\gamma)^m}$ είναι μια συντηρητική εκτίμηση του σφάλματος τύπου I δοθέντος ότι $R(\gamma)>0$. Έτσι, ο τύπος για την εκτίμηση της pFDR γίνεται

$$\widehat{pFDR}(\gamma)=\frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{P}(\mathrm{P}\le\gamma)\left\{1-(1-\gamma)^m\right\}}=\frac{W(\lambda)\gamma}{(1-\lambda)\{R(\gamma)\vee 1\}\left\{1-(1-\gamma)^m\right\}}. \qquad (2.10)$$

Επειδή η FDR δεν δεσμεύεται από το ενδεχόμενο να έχουμε τουλάχιστο μία απόρριψη, μπορούμε να θέσουμε

$$\widehat{FDR}(\gamma)=\frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{\Pr}(\mathrm{P}\le\gamma)}=\frac{W(\lambda)\gamma}{(1-\lambda)\{R(\gamma)\vee 1\}}\ . \qquad (2.11)$$

Για μεγάλες τιμές *m*, αυτές οι δύο εκτιμήσεις είναι ισοδύναμες.

2.3 Ρύθμιση της FDR για πολλαπλά εξαρτημένα τεστ

Στην πράξη, και ειδικότερα στην περίπτωση δεδομένων που προέρχονται από μικροσυστοιχίες DNA, τα στατιστικά που αφορούν σε μετρήσεις έκφρασης γονιδίων σε διαφορετικά δείγματα είναι εξαρτημένα. Οι Benjamini και Yekutieli (2001) απέδειξαν ότι η διαδικασία των Benjamini και Hochberg (1995) που παρουσιάσαμε παραπάνω ρυθμίζει την FDR σε μία πιο γενική κατάσταση όπου τα στατιστικά παρουσιάζουν μία μορφή θετικής εξάρτησης, όπως για παράδειγμα στην περίπτωση της πολυμεταβλητής κανονικής κατανομής με θετικές συσχετίσεις. Συγκεκριμένα, ρυθμίζει την FDR σε οικογένειες με θετικά εξαρτημένα στατιστικά ενώ σε άλλες περιπτώσεις εξάρτησης αποδεικνύεται ότι η διαδικασία μπορεί να τροποποιηθεί έτσι ώστε να ρυθμίσει την FDR με έναν πιο συντηρητικό τρόπο.

Ορισμός 2.2. Ένα σύνολο $C \subseteq R^n$ καλείται *αύξον* (*φθίνον*) αν και μόνον αν το $x = (x_1,...,x_n) \in C$ συνεπάγεται ότι $x' = (x_1',...,x_n') \in C$ για κάθε $x_i \le x_i'$ ($x_i \ge x_i'$), $i = 1,...,n$. □

Ορισμός 2.3. Έστω $I_0 \subseteq \{1,...,n\}$. Ένα τυχαίο διάνυσμα $X = (X_1,...,X_n)$ (ή η αντίστοιχη πολυμεταβλητή κατανομή του) λέγεται ότι είναι *θετικά εξαρτημένο λόγω παλινδρόμησης* σε ένα υποσύνολο $\{X_i, i \in I_0\}$ ή απλά στο I_0 αν η πιθανότητα $P(X \in C \mid X_i)$ είναι μη φθίνουσα (μη αύξουσα) στο X_i για κάθε $i \in I_0$ και για κάθε αύξον (φθίνον) σύνολο. □

Ιδιότητα PRDS. Η ιδιότητα η οποία εκφράζει την έννοια της θετικής εξάρτησης ονομάζεται *θετική εξάρτηση λόγω παλινδρόμησης σε κάθε ένα από τα υποσύνολα* I_0 (*positive regression dependency on each one from the subset* I_0), ή PRDS στο I_0.

Πρώτος ο Lehmann (1966) πρότεινε έναν σχεδιασμό για την θετική εξάρτηση μιας διδιάστατης τυχαίας μεταβλητής (X,Y) παρουσιάζοντας τρεις ορισμούς οι οποίοι είναι διαδοχικά ο ένας ισχυρότερος από τον άλλον, με σκοπό την διερεύνηση των συνεπειών και της ισχύος τους.

Κατά τον πρώτο ορισμό, θέλησε να συγκρίνει την πιθανότητα κάθε τεταρτημόριου $X \le x$, $Y \le y$ κάτω από την κατανομή F των (X,Y) με την αντίστοιχη πιθανότητα στην περίπτωση της ανεξαρτησίας. Λέμε ότι το ζεύγος των μεταβλητών (X,Y), (ή η κατανομή τους F), είναι *θετικά quadrant εξαρτημένες* αν

$$\Pr(X \le x, Y \le y) \ge \Pr(X \le x)\Pr(Y \le y)$$

για όλα τα x, y, με την απόλυτη εξάρτηση να εμφανίζεται στην περίπτωση που ισχύει η ισότητα για τουλάχιστον ένα ζεύγος (x,y). Όμοια, το ζεύγος των μεταβλητών (X, Y) (ή η κατανομή τους F) είναι *αρνητικά quadrant εξαρτημένες* αν η ανισότητα ισχύει με την αντίθετη φορά.

Αναπτύσσοντας τον παραπάνω τύπο καταλήγουμε στον πρώτο ορισμό της θετικής εξάρτησης. Η σχέση

$$\Pr(Y \le y \mid X \le x) \ge \Pr(Y \le y) \qquad (2.12)$$

εκφράζει το γεγονός ότι η πιθανότητα η τυχαία μεταβλητή Y να πάρει μικρές τιμές αυξάνει με την πληροφορία ότι η τυχαία μεταβλητή X παίρνει μικρές τιμές.

Ο δεύτερος τρόπος με τον οποίο μπορούμε να ορίσουμε τη θετική εξάρτηση είναι μέσω της παρακάτω, πιο ισχυρής, συνθήκης:

$$\Pr\left(Y \le \mathrm{y} \,\middle|\, X \le x \right) \ge \Pr\left(Y \le \mathrm{y} \,\middle|\, X \le x' \right) \tag{2.13}$$

για όλα τα $x \le x'$ και για όλα τα y.

Τέλος ο τρίτος, ακόμα πιο ισχυρός ορισμός δίνεται από τη συνθήκη:

$$\text{η πιθανότητα } \Pr\left(Y \le y \,\middle|\, X = x \right) \text{ είναι φθίνουσα στο } x. \tag{2.14}$$

Αν ισχύει η σχέση (2.14) τότε λέμε ότι η Y είναι θετικά εξαρτημένη λόγω παλινδρόμησης στη X.

Λήμμα 2.2. *Ο ορισμός (2.14) συνεπάγεται τον (2.13) και αυτός συνεπάγεται τον (2.12).*
(*Απόδειξη του λήμματος στο Παράρτημα*)

Η γενίκευση αυτού του σχεδιασμού από τις διμεταβλητές κατανομές στις πολυμεταβλητές έγινε από τον Sarkar (1969).

Ορισμός 2.4. Μία πολυμεταβλητή κατανομή λέμε ότι έχει θετική εξάρτηση λόγω παλινδρόμησης αν για κάθε αύξον σύνολο C και για κάθε $i \in \{1,...,n\}$

$$\text{η πιθανότητα } \Pr\left(X \in C \,\middle|\, X_1 = x_1,..., X_i = x_i \right) \text{ είναι μη φθίνουσα στο } \left(x_1,..., x_i \right). \qquad \square$$

Η ιδιότητα της θετικής εξάρτησης λόγω παλινδρόμησης αποτελεί την αυστηρότερη διατύπωση της ιδιότητας PRDS στο I_0. Η διαφορά μεταξύ των δύο, είναι ότι στην PRDS η δέσμευση θεωρείται για μία μεταβλητή κάθε φορά και ότι απαιτείται να ισχύει για ένα υποσύνολο μεταβλητών.

Τα βασικά συμπεράσματα που αφορούν στην ιδιότητα PRDS και τα εξαρτημένα τεστ συνοψίζονται στα θεωρήματα που ακολουθούν.

Θεώρημα 2.6. *Αν η από κοινού κατανομή των στατιστικών είναι PRDS πάνω στο υποσύνολο των στατιστικών που αντιστοιχούν στις αληθείς μηδενικές υποθέσεις, η διαδικασία των Benjamini και Hochberg ρυθμίζει την FDR σε επίπεδο μικρότερο ή ίσο του* $\frac{m_0}{m} q^*$.

Απόδειξη. Αρχικά θέτουμε για ευκολία τις σταθερές που εμφανίζονται στη διαδικασία των Benjamini και Hochberg ίσες με $q_i = \frac{i}{m} q, i = 1,...,m$.

Έστω $A_{u,s}$ το ενδεχόμενο η διαδικασία των Benjamini και Hochberg να απορρίψει ακριβώς *u* αληθείς και *s* ψευδείς μηδενικές υποθέσεις. Τότε η FDR θα είναι

$$E(Q) = E\left(\frac{V}{R}\right) = \sum_{s=0}^{m_1} \sum_{u=1}^{m_0} \frac{u}{u+s} \Pr\left(A_{u,s}\right).$$

Στο λήμμα που ακολουθεί η πιθανότητα $P\left(A_{u,s}\right)$ εκφράζεται ως ένας μέσος όρος.

Λήμμα 2.3. *Ισχύει ότι* $\Pr\left(A_{u,s}\right) = \frac{1}{u} \sum_{i=1}^{m_0} \Pr\left(\left\{P_i \le q_{u+s}\right\} \cap A_{u,s}\right)$.

Απόδειξη. Για ορισμένο αριθμό αληθών και ψευδών μηδενικών υποθέσεων *u* και *s*, έστω *ω* ένα υποσύνολο του $\{1,...,m_0\}$ με *u* στοιχεία και $A_{u,s}^{\omega}$ το ενδεχόμενο το *ω* να αποτελείται από τις *u* μηδενικές υποθέσεις του συνόλου $A_{u,s}$ που έχουν απορριφθεί. Σημειώνεται ότι η

$$P\left(\left\{P_i \le q_{u+s}\right\} \cap A_{u,s}^{\omega}\right) = \begin{cases} P\left(A_{u,s}^{\omega}\right) & , i \in \omega \\ 0 & , \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

Επομένως,

$$\begin{aligned} \sum_{i=1}^{m_0} P\left(\left\{P_i \le q_{u+s}\right\} \cap A_{u,s}\right) &= \sum_{i=1}^{m_0} \sum_{\omega} P\left(\left\{P_i \le q_{u+s}\right\} \cap A_{u,s}^{\omega}\right) \\ &= \sum_{\omega} \sum_{i=1}^{m_0} P\left(\left\{P_i \le q_{u+s}\right\} \cap A_{u,s}^{\omega}\right) \\ &= \sum_{\omega} \sum_{i=1}^{m_0} I\left(i \in \omega\right) \cdot P\left\{A_{u,s}^{\omega}\right\} \\ &= \sum_{\omega} u \cdot P\left(A_{u,s}^{\omega}\right) \\ &= u \cdot P\left\{A_{u,s}\right\} \end{aligned}$$

Έτσι, αντικαθιστώντας στον τύπο της FDR έχουμε:

$$\begin{aligned} E(Q) &= \sum_{s=0}^{m_1} \sum_{u=1}^{m_0} \frac{u}{u+s} \left\{\frac{1}{u} \sum_{i=1}^{m_0} P\left(\left\{P_i \le q_{u+s}\right\} \cap A_{u,s}\right)\right\} \\ &= \sum_{i=1}^{m_0} \left\{\sum_{s=0}^{m_1} \sum_{u=1}^{m_0} \frac{1}{u+s} P\left(\left\{P_i \le q_{u+s}\right\} \cap A_{u,s}\right)\right\} \end{aligned} \qquad (2.14)$$

Με αυτόν τον τρόπο, η αναμενόμενη τιμή του *u* εξαρτάται μόνο από το $A_{u,s}$. Αυτό που θα κάνουμε στη συνέχεια είναι να εκφράσουμε το $A_{u,s}$ σε ενδεχόμενα που εξαρτώνται από το *i* και το $k = u + s$ έτσι ώστε και η FDR να μπορεί να εκφραστεί με παρόμοιο τρόπο.

Για $i = 1,...,m_0$ έστω $\mathbf{P}^{(i)}$ οι $m - 1$ τιμές-*p* που απομένουν μετά την αφαίρεση του P_i. Έστω επίσης $C_{u,s}^{(i)}$ το ενδεχόμενο κατά το οποίο αν το P_i απορρίπτεται τότε απορρίπτονται παράλληλα και οι $u - 1$ αληθείς και *s* ψευδείς μηδενικές υποθέσεις, δηλαδή η τομή $\{P_i \le q_{u+s}\} \cap A_{u,s}$. Έτσι

$$\{P_i \le q_{u+s}\} \cap A_{u,s} = \{P_i \le q_{u+s}\} \cap C_{u,s}^{(i)}$$

Ορίζουμε με $C_k^{(i)}$ την ένωση των $\{C_{u,s}^{(i)} : u + s = k\}$. Για κάθε *i* τα $C_k^{(i)}$ είναι ξένα μεταξύ τους, και η FDR μπορεί να πάρει τη μορφή

$$E(Q) = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} P\left(\{P_i \le q_k\} \cap C_k^{(i)}\right). \qquad (2.16)$$

Στην τελευταία έκφραση η FDR δεν εξαρτάται πλέον από τα *u*, *s*.

Στο τελευταίο μέρος της απόδειξης, κατασκευάζουμε μία σειρά από αύξοντα σύνολα στα οποία χρησιμοποιούμε την ιδιότητα PRDS για να φράξουμε το εσωτερικό άθροισμα στη σχέση (2.15) από το $\frac{q}{m}$. Για το σκοπό αυτό ορίζουμε το σύνολο

$$D_k^{(i)} = \cup\{C_j^{(i)} : j \le k\} \text{ για } k = 1,...,m$$

Το $D_k^{(i)}$ μπορεί επίσης να περιγραφεί χρησιμοποιώντας το διατεταγμένο σύνολο των τιμών-*p* στην σειρά των $\mathbf{P}^{(i)}$, $\{p_{(1)}^{(i)} \le ... \le p_{(m)}^{(i)}\}$ με τον παρακάτω τρόπο:

$$D_k = \{\mathbf{p} : q_{k+1} \le p_{(k)}^{(i)},\ q_{k+2} \le p_{(k+1)}^{(i)},...,\ q_m \le p_{(m-1)}^{(i)}\}$$

για $k = 1,...,m - 1$. Παρατηρούμε ότι $D_m^{(i)}$ είναι πλήρης χώρος. Με το να εκφράσουμε το $D_k^{(i)}$ με αυτόν τον τρόπο γίνεται φανερό ότι πρόκειται για ένα μη φθίνον σύνολο.

Θα κάνουμε τώρα χρήση της ιδιότητας PRDS, σύμφωνα με την οποία για $p \le p'$ έχουμε

$$\Pr\left(D \mid P_i = p\right) \le \Pr\left(D \mid P_i = p'\right).$$

Σύμφωνα με τον Lehmann (1966) είναι εύκολο να διαπιστώσουμε ότι για $j \le l$ (και άρα $q_j \le q_l$) ισχύει

$$\Pr\left(D \mid P_i \le q_j\right) \le \Pr\left(D \mid P_i \le q_l\right)$$

για ένα μη φθίνον σύνολο D, ή ισοδύναμα,

$$\frac{\Pr\left(\{P_i \le q_k\} \cap D_k^{(i)}\right)}{\Pr\left(P_i \le q_k\right)} \le \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap D_k^{(i)}\right)}{\Pr\left(P_i \le q_{k+1}\right)}.$$

Από τη σχέση αυτή, αν λάβουμε υπόψη ότι $D_{j+1}^{(i)} = D_j^{(i)} \cup C_{j+1}^{(i)}$, προκύπτει ότι για όλα τα $k = 1, \ldots, m-1$:

$$\frac{\Pr\left(\{P_i \le q_k\} \cap D_k^{(i)}\right)}{\Pr\left(P_i \le q_k\right)} + \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap C_{k+1}^{(i)}\right)}{\Pr\left(P_i \le q_{k+1}\right)} \le \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap D_k^{(i)}\right)}{\Pr\left(P_i \le q_{k+1}\right)} + \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap C_{k+1}^{(i)}\right)}{\Pr\left(P_i \le q_{k+1}\right)}$$

$$= \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap D_{k+1}^{(i)}\right)}{\Pr\left(P_i \le q_{k+1}\right)}$$

Ξεκινώντας με $C_1 = D_1$, εφαρμόζουμε την παραπάνω σχέση για $i = 1, \ldots, m-1$ και παίρνουμε

$$\sum_{k=1}^{m} \frac{\Pr\left(\{P_i \le q_k\} \cap C_k^{(i)}\right)}{\Pr\left(P_i \le q_k\right)} \le \frac{\Pr\left(\{P_i \le q_m\} \cap D_m^{(i)}\right)}{\Pr\left(P_i \le q_m\right)} = 1$$

όπου η τελευταία ισότητα ισχύει διότι ο χώρος $D_m^{(i)}$ είναι πλήρης.

Επομένως, η σχέση (2.16) γίνεται:

$$E(Q) = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \Pr\left(\{P_i \le q_k\} \cap C_k^{(i)}\right) \le \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{q}{m} \frac{\Pr\left(\{P_i \le q_k\} \cap C_k^{(i)}\right)}{\Pr\left(P_i \le q_k\right)}$$

επειδή $\Pr\left(P_i \le q_k\right) \le q_k = \frac{k}{m} q$ κάτω από τη μηδενική υπόθεση, καταλήγοντας έτσι στην επιθυμητή ανισότητα

$$\frac{q}{m} \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{\Pr\left(\{P_i \le q_k\} \cap C_k^{(i)}\right)}{\Pr\left(P_i \le q_k\right)} \le \frac{m_0}{m} q. \qquad \square$$

Οι Benjamini και Yekutieli (2001) πρότειναν και μία απλή παραλλαγή της κλασσικής διαδικασίας των Benjamini και Hochberg (1995) η οποία ρυθμίζει την FDR κάτω από κάποιες αυθαίρετες περιπτώσεις εξάρτησης. Οι προσαρμοσμένες τιμές-p για αυτήν τη τροποποιημένη διαδικασία είναι:

$$\tilde{p}_{(i)} = \min_{k=i,\ldots,m} \left\{ \min \left(\frac{m \sum_{i=1}^{k} 1/i}{k} p_{(k)}, 1 \right) \right\}.$$

Ουσιαστικά, πρόκειται για μία step-up διαδικασία όμοια με αυτή των Benjamini και Hochberg (1995), με μόνη διαφορά την αντικατάσταση του m/k με το $\frac{m \sum_{i=1}^{k} i^{-1}}{k}$.

Θεώρημα 2.7. *Όταν η διαδικασία Benjamini- Hochberg εκτελείται με αντικατάσταση του* q *με το* $q / \sum_{j=1}^{m} \left(j^{-1} \right)$, *ρυθμίζει πάντα την FDR σε επίπεδο μικρότερο ή ίσο του* $\frac{m_0}{m} q^*$.

Απόδειξη. Συμβολίζουμε

$$p_{ikj} = \Pr\left(\left\{ P_i \in \left[\frac{(j-1)}{m} q, \frac{j}{m} q \right] \right\} \cap C_k^{(i)} \right)$$

και παρατηρούμε ότι

$$\sum_{k=1}^{m} p_{ikj} = \Pr\left(\left\{ P_i \in \left[\frac{(j-1)}{m} q, \frac{j}{m} q \right] \right\} \cap \left(\bigcup_{k=1}^{m} C_k^{(i)} \right) \right) = \frac{q}{m}.$$

Από την (2.16) η FDR μπορεί να εκφραστεί ως εξής:

$$E(Q) = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \sum_{j=1}^{k} p_{ijk} = \sum_{i=1}^{m_0} \sum_{j=1}^{m} \sum_{k=j}^{m} \frac{1}{k} p_{ijk} \le \sum_{i=1}^{m_0} \sum_{j=1}^{m} \sum_{k=j}^{m} \frac{1}{j} p_{ijk} \le \sum_{i=1}^{m_0} \sum_{j=1}^{m} \frac{1}{j} \sum_{k=1}^{km} p_{ijk} = m_0 \sum_{j=1}^{m} \frac{1}{j} \frac{q}{m}$$

□

Μία πιο αυστηρή συνθήκη η οποία συνεπάγεται την θετική εξάρτηση λόγω παλινδρόμησης είναι η πολυμεταβλητή ολική θετικότητα τάξης 2 (*multivariate total positivity of order 2*, MTP_2).

Ορισμός 2.5. Μία μη αρνητική συνάρτηση $f(x,y)$ ορισμένη στο $\mathsf{X} \times \mathsf{Y} \subseteq \mathsf{R}^2$, καλείται *ολικά θετική τάξης 2* (*total positive of order 2*, TP_2) αν για όλα τα $x_1 < x_2$, $y_1 < y_2$, με $x_i, y_j \in \mathsf{X} \times \mathsf{Y}$, οι 2×2 ορίζουσες είναι μη αρνητικές, δηλαδή αν

$$f(x_1, x_2) f(y_1, y_2) \ge f(x_1, y_2) f(y_1, x_2)$$ □

Ορισμός 2.6. Έστω $\mathbf{x}$ ένα τυχαίο m-διάστατο διάνυσμα με πυκνότητα $f(\mathbf{x})$. Η $f(\mathbf{x})$ καλείται *πολυμεταβλητά ολικά θετική τάξης 2* (*multivariate totally positive of order 2*, MTP_2) αν

$$f(\mathbf{x}\vee\mathbf{y})f(\mathbf{x}\wedge\mathbf{y})\geq f(\mathbf{x})f(\mathbf{y})$$

για όλα τα $\mathbf{x},\mathbf{y}\in\Re^n$, όπου

$$\mathbf{x}\vee\mathbf{y}=\left(\max(x_1,y_1),...,\max(x_n,y_n)\right) \text{ και } \mathbf{x}\wedge\mathbf{y}=\left(\min(x_1,y_1),...,\min(x_n,y_n)\right). \qquad \square$$

Ο Sarkar το 2002 συνέχισε τη δουλειά των Benjamini και Yekutieli (2001) δείχνοντας ότι οι κρίσιμες τιμές αυτής της διαδικασίας μπορούν να χρησιμοποιηθούν σε μία πιο γενική stepwise διαδικασία κάτω από την ίδια θετική συσχέτιση. Επίσης απέδειξε ότι η step-down διαδικασία ρύθμισης της FDR των Benjamini και Liu μπορεί να χρησιμοποιηθεί και στην περίπτωση των θετικά εξαρτημένων στατιστικών.

Θα περιγράψουμε αρχικά τη γενικευμένη step-down διαδικασία που πρότειναν οι Tamhane, Liu και Dunnett το 1998.

Έστω ότι έχουμε m μηδενικές υποθέσεις $H_{0,1},...,H_{0,m}$ οι οποίες εξετάζονται ταυτόχρονα χρησιμοποιώντας τις αντίστοιχες τιμές-p, $p_1,...,p_m$. Διατάσσουμε τις τιμές-p έτσι ώστε $p_{(1)}\leq...\leq p_{(m)}$ με τις αντίστοιχες μηδενικές υποθέσεις να είναι $H_{(1)},...,H_{(m)}$. Δεδομένων των κρίσιμων τιμών $0\leq a_{(1)}\leq...\leq a_{(m)}\leq 1$ η γενικευμένη step-down-step-up διαδικασία τάξης r πραγματοποιείται ως εξής:

Αν $p_{(m-r+1)}>a_{(m-r+1)}$ ο έλεγχος αποδέχεται τις $H_{(m-r+1)},...,H_{(m)}$ και πηγαίνει στο γενικό βήμα (α), διαφορετικά πηγαίνει στο γενικό βήμα (β).

Γενικό βήμα (α): Για $i=m-r+1$ και αν $p_{(i-1)}\leq a_{(i-1)}$, ο έλεγχος σταματά απορρίπτοντας τις $H_{(1)},...,H_{(i-1)}$, διαφορετικά αποδέχεται την $H_{(i-1)}$, θέτει $i=i-1$ και επιστρέφει στο αρχικό βήμα. Για $i=1$ ο έλεγχος σταματά.

Γενικό βήμα (β): Για $i = m - r + 1$ και αν $p_{(i+1)} \geq a_{(i+1)}$ ο έλεγχος σταματά και αποδέχεται τις $H_{(i+1)},...,H_{(m)}$, διαφορετικά απορρίπτει την $H_{(i+1)}$, θέτει $i = i + 1$ και επιστρέφει στο αρχικό βήμα. Για $i = m$ ο έλεγχος σταματά.

Για $r = 1$ ή $r = m$ η γενικευμένη step-down-step-up διαδικασία σειράς *r* ταυτίζεται με την step-up ή την step-down διαδικασία αντίστοιχα.

Αν $H_{(1)},...,H_{(m)}$ οι υποθέσεις που αντιστοιχούν στα διατεταγμένα στατιστικά $T_{(1)},...,T_{(m)}$ τότε η step-down-step-up διαδικασία σειράς *r* με $\Pr\left(T_{(i)} \geq c_{(i)}\right) = a_{(m-i+1)}, i = 1,...,m$, ξεκινά με το στατιστικό $T_{(r)}$. Αν $T_{(r)} < c_{(r)}$ τότε δεχόμαστε τις $H_{(1)},...,H_{(r)}$ και συνεχίζουμε με τον έλεγχο των $H_{(r+1)},...,H_{(m)}$ με τον τρόπο που διενεργούνται οι step-up διαδικασίες. Διαφορετικά, απορρίπτουμε τις υποθέσεις $H_{(r)},...,H_{(m)}$ και συνεχίζουμε με τον έλεγχο των $H_{(r+1)},...,H_{(m)}$ με τον τρόπο που διενεργούνται οι step-down διαδικασίες.

Τα δύο βασικά συμπεράσματα του Sarkar (2002) παρουσιάζονται και αποδεικνύονται στα παρακάτω θεωρήματα.

Θεώρημα 2.8. *Έστω ότι έχουμε μία γενικευμένη step-down-step-up διαδικασία τάξης r για δεξιά μονόπλευρους ελέγχους που βασίζονται σε συνεχή στατιστικά* $(T_1,...,T_m)$ *τα οποία είναι ισόνομα κατανεμημένα με συνάρτηση κατανομής* $F(t)$ *κάτω από τις αληθείς μηδενικές υποθέσεις. Αν η από κοινού κατανομή των* $(T_1,...,T_m)$ *είναι PRDS στο υποσύνολο των στατιστικών που αντιστοιχούν στις αληθείς μηδενικές υποθέσεις δεδομένου ότι* $\{T_i > a, i = 1,...,m\}$ *για κάποιο ορισμένο α και οι κρίσιμες τιμές* $c_{(1)} \leq ... \leq c_{(m)}$ *ικανοποιούν την* $F\left(c_{(j)}\right) = 1 - (m - j + 1)\frac{a}{m}$ *για* $j = 1,...,m$*, τότε η FDR της διαδικασίας αυτής είναι μικρότερη ή ίση του* $m_0 \frac{a}{m}$ *όπου* $0 < a < F(a)$.

Απόδειξη. Η απόδειξη του θεωρήματος βασίζεται στο παρακάτω λήμμα.

Λήμμα 2.4. *Η FDR μιας γενικευμένης step-down-step-up διαδικασίας σειράς r για τον δεξίπλευρο έλεγχο των m μηδενικών υποθέσεων που βασίζεται στα στατιστικά* $(T_1,...,T_m)$ *και τις κρίσιμες τιμές* $c_{(1)} \leq ... \leq c_{(m)}$ *είναι:*

$$FDR = \frac{1}{m-r+1}\sum_{i=1}^{m_0}\Pr\left\{T_i \geq c_{(r)}\right\}$$

$$+\sum_{i=1}^{m_0}\sum_{j=1}^{r-1}E\left[\Pr\left\{T_{(j)} \geq c_{(j)},...,T_{(r)} \geq c_{(r)} \,\middle|\, T_i\right\}\times\left\{\frac{I\left(T_i \geq c_{(j)}\right)}{m-j+1}-\frac{I\left(T_i \geq c_{(j+1)}\right)}{m-j}\right\}\right]$$

$$+\sum_{i=1}^{m_0}\sum_{j=r}^{m-1}E\left[\Pr\left\{T_{(r)} < c_{(r)},...,T_{(j)} < c_{(j)} \,\middle|\, T_i\right\}\times\left\{\frac{I\left(T_i \geq c_{(j+1)}\right)}{m-j}-\frac{I\left(T_i \geq c_{(j)}\right)}{m-j+1}\right\}\right]$$

όπου οι πιθανότητες ορίζονται υπό τις αληθείς μηδενικές υποθέσεις $H_{0,1},...,H_{0,m_0}$.

Αρχικά σημειώνεται ότι η διαφορά $\frac{I\left(T_i \geq c_{(j)}\right)}{m-j+1}-\frac{I\left(T_i \geq c_{(j+1)}\right)}{m-j}$ είναι $\leq$ ή ≥ 0 ανάλογα με το αν η T_i είναι $\leq$ ή $\geq c_{(j+1)}$. Επίσης, επειδή $c_{(j)} > a$ για όλα τα $j = 1,...,m$ και το $\left\{T_{(j)} \geq c_{(j)},...,T_{(r)} \geq c_{(r)}\right\}$ είναι ένα αύξον σύνολο, η δεσμευμένη πιθανότητα $\Pr\left(T_{(j)} \geq c_{(j)},...,T_{(r)} \geq c_{(r)} \,\middle|\, T_i\right)$ είναι φθίνουσα στο $T_i > a$ για όλα τα $i = 1,...,m_0$, $j = 1,...,r$ - 1 σύμφωνα με την ιδιότητα PRDS. Έτσι έχουμε ότι

$$E\left[\Pr\left\{T_{(j)} \geq c_{(j)},...,T_{(r)} \geq c_{(r)} \,\middle|\, T_i\right\}\times\left\{\frac{I\left(T_i \geq c_{(j)}\right)}{m-j+1}-\frac{I\left(T_i \geq c_{(j+1)}\right)}{m-j}\right\}\right]$$

$$\leq \Pr\left\{T_{(j)} \geq c_{(j)},...,T_{(r)} \geq c_{(r)} \,\middle|\, T_i = c_{(j+1)}\right\}\times\left\{\frac{\Pr\left(T_i \geq c_{(j)}\right)}{m-j+1}-\frac{\Pr\left(T_i \geq c_{(j+1)}\right)}{m-j}\right\}$$

για όλα τα $i = 1,...,m_0$, $j = 1,...,r$ - 1.

Όμοια σκεφτόμαστε και για τη δεύτερη μέση τιμή και έτσι προκύπτει ότι

$$E\left[\Pr\left\{T_{(r)} \geq c_{(r)},...,T_{(j)} < c_{(j)} \,\middle|\, T_i\right\}\times\left\{\frac{I\left(T_i \geq c_{(j+1)}\right)}{m-j}-\frac{I\left(T_i \geq c_{(j)}\right)}{m-j+1}\right\}\right]$$

$$\leq \Pr\left\{T_{(r)} \geq c_{(r)},...,T_{(j)} < c_{(j)} \,\middle|\, T_i = c_{(j+1)}\right\} \times \left\{\frac{\Pr\left(T_i \geq c_{(j+1)}\right)}{m-j} - \frac{\Pr\left(T_i \geq c_{(j)}\right)}{m-j+1}\right\}$$

για όλα τα $i = 1,...,m_0$, $j = r,...,n-1$.

Αν αντικαταστήσουμε τις παραπάνω σχέσεις στο Λήμμα 2.4 καταλήγουμε στο ζητούμενο αποτέλεσμα. □

Πριν περάσουμε στο δεύτερο αποτέλεσμα του Sarkar (2002), θα παρουσιάσουμε ένα Λήμμα που μας παρέχει ένα άνω φράγμα για την FDR μιας step-down διαδικασίας. Έστω $T_{m_0+1},...,T_m$ τα στατιστικά που αντιστοιχούν στις ψευδείς μηδενικές υποθέσεις. Τότε, με $I_1 = \{m_0+1,...,m\}$, ορίζουμε το ενδεχόμενο

$$B_{(j)} = \left\{T_{(j)} \geq c_{(m_0+j)}, T_{(j+1)} \geq c_{(m_0+j+1)},...,T_{(m_1)} \geq c_{(m)}\right\}.$$

Αυτό είναι το ενδεχόμενο η step-down διαδικασία βασισμένη στα στατιστικά $T_{m_0+1},...,T_m$ και στα κρίσιμα σημεία $c_{(m_0+1)},...,c_{(m)}$ να δεχτεί j και να απορρίψει $m_1 - j$ από τις ψευδείς μηδενικές υποθέσεις για $j = 0,...,m_1$.

Λήμμα 2.5. *Για μία step-down διαδικασία για τον δεξιά μονόπλευρο έλεγχο των m μηδενικών υποθέσεων $H_{0,1},...,H_{0,m}$ που βασίζεται στα στατιστικά $(T_1,...,T_m)$ και τις κρίσιμες τιμές $c_{1:m} \leq ... \leq c_{m:m}$ έχουμε:*

$$FDR \leq \sum_{j=0}^{m_1} \frac{m_0 + j}{m} P\left\{T_{m_0:m} \geq c_{(m_0+j)}, B_{j,I_1}\right\}. \qquad (2.17)$$

□

Θεώρημα 2.9. *Η step-down διαδικασία των Benjamini και Liu ρυθμίζει την FDR αν τα στατιστικά είναι MTP_2 κάτω από οποιαδήποτε εναλλακτική και ανταλλάξιμα όταν οι υποθέσεις είναι αληθείς.*

Απόδειξη. Η απόδειξη βασίζεται στη σχέση (2.17) του Λήμματος 2.5.

Το $B_{(j)}$ για $j = 0,1,...,m_1$ μπορεί να γραφεί ως διαφορά των συνόλων $C_{(j+1)}$ και $C_{(j)}$, δηλαδή $B_{(j)} = C_{(j+1)} - C_{(j)}$ όπου $C_{(j)} = \left\{T_{(j)} \geq c_{(m_0+j)},...,T_{(m)} \geq c_{(m)}\right\}$ με $C_{(0)}$ και $C_{(m_1+1)}$ το

αδύνατο και το βέβαιο ενδεχόμενο αντίστοιχα. Έτσι, το άθροισμα στην παραπάνω σχέση του Λήμματος 2.5 μπορεί να γραφεί ως εξής:

$$\frac{1}{m}\sum_{j=0}^{m_0}(m_0+j)\cdot \Pr\left\{T_{(m_0)} \ge c_{(m_0+j)}, C_{(j+1)} - C_{(j)}\right\}$$

$$= \Pr\left\{T_{(m_0)} \ge c_{(m)}\right\}$$

$$+\frac{1}{m}\sum_{j=1}^{m_1}\left[(m_0+j-1)\cdot \Pr\left\{C_{(j)}, T_{(m_0)} \ge c_{(m_0+j-1)}\right\} - (m_0+j)\cdot \Pr\left\{C_{(j)}, T_{(m_0)} \ge c_{(m_0+j)}\right\}\right]$$

$$= \Pr\left\{T_{(m_0)} \ge c_{(m)}\right\}$$

$$+\frac{1}{m}\sum_{j=1}^{m_1} E\left[\Pr\left\{C_{(j)} \,\middle|\, T_{(m_0)}\right\}\cdot\left[(m_0+j-1)\cdot I\left\{T_{(m_0)} \ge c_{(m_0+j-1)}\right\} - (m_0+j)\cdot I\left\{T_{(m_0)} \ge c_{m_0+j}\right\}\right]\right] \tag{2.18}$$

Θεωρούμε τώρα ότι ισχύει η συνθήκη

$$\text{η } E\left\{\phi\left(T_{m_0+1},...,T_m\right) \,\middle|\, T_{(m_0)} = t\right\} \text{ είναι μη φθίνουσα συναρτήση του } t \tag{2.19}$$

για κάθε κατά συντεταγμένη μη φθίνουσα συνάρτηση ϕ των $T_{n_0+1},...,T_n$.

Επειδή το σύνολο $C_{(j)}$ είναι μη φθίνον στο $\left(T_{n_0+1},...,T_n\right)$ μπορούμε να γράψουμε

$$E\left[\Pr\left\{C_{(j)} \,\middle|\, T_{(m_0)}\right\}\left\{(m_0+j-1)I\left(T_{(m_0)} \ge c_{(m_0+j-1)}\right) - (m_0+j)I\left(T_{(m_0)} \ge c_{(m_0+j)}\right)\right\}\right]$$

$$\le \Pr\left\{C_{(j)} \,\middle|\, T_{(m_0)} = c_{(m_0+j)}\right\}\left\{(m_0+j-1)\Pr\left(T_{(m_0)} \ge c_{(m_0+j-1)}\right) - (m_0+j)I\left(T_{(m_0)} \ge c_{m_0+j}\right)\right\}.$$

Χρησιμοποιώντας την παραπάνω ανισότητα στην τελευταία γραμμή της σχέσης (2.18) η (2.17) γίνεται:

$$FDR \le \frac{1}{m}\sum_{j=0}^{m_0}(m_0+j)\Pr\left\{T_{(m_0)} \ge c_{(m_0+j)}\right\}$$

$$\times\left[\Pr\left\{C_{(j+1)} \,\middle|\, T_{(m_0)} = c_{(m_0+j-1)}\right\} - \Pr\left\{C_{(j)} \,\middle|\, T_{(m_0)} = c_{(m_0+j)}\right\}\right]$$

$$\le \frac{1}{m}\max_{0\le j\le m_1}\left[(m_0+j)\Pr\left\{T_{(m_0)} \ge c_{(m_0+j)}\right\}\right] \tag{2.20}$$

$$\times\sum_{j=0}^{m_1}\left[\Pr\left\{C_{(j+1)} \,\middle|\, T_{(m_0)} = c_{(m_0+j-1)}\right\} - \Pr\left\{C_{(j)} \,\middle|\, T_{(m_0)} = c_{(m_0+j)}\right\}\right]$$

$$\leq \frac{1}{m}\max_{0\leq j\leq m_1}\left[\left(m_0+j\right)\cdot \Pr\left\{T_{(m_0)}\geq c_{(m_0+j)}\right\}\right].$$

Η δεύτερη ανισότητα προκύπτει από το γεγονός ότι

$$\Pr\left\{C_{(j+1)}\,\middle|\, T_{(m_0)}=c_{(m_0+j-1)}\right\}-\Pr\left\{C_{(j)}\,\middle|\, T_{(m_0)}=c_{(m_0+j)}\right\}\geq \Pr\left\{B_{(j)}\,\middle|\, T_{(m_0)}=c_{(m_0+j)}\right\}$$

λόγω της συνθήκης (2.19) και επομένως είναι μη αρνητική για όλα τα $j=0,1,...,m_1$. Επειδή $T_{(m_0)}<T_{(|J|)}$ για κάθε $J\supseteq J_0$, όπου $|J|$ το πλήθος των στοιχείων του συνόλου *J*, η FDR στη σχέση (2.20) είναι μικρότερη ή ίση με *α* για κάθε m_0 αν οι κρίσιμες τιμές $c_{(j)}$ ικανοποιούν την

$$\frac{j}{m}\Pr\left\{\max\left(T_1,...,T_j\right)\geq c_{(m)}\right\}\leq a \quad \text{για όλα τα } j=1,...,m\,, \tag{2.20}$$

υποθέτοντας ότι όλες οι υποθέσεις είναι αληθείς.

Όταν τα T_j είναι ανεξάρτητα, οι κρίσιμες τιμές $c_{(j)}$ που ικανοποιούν την συνθήκη (2.21) είναι αυτές που δίνονται από την σχέση

$$F\left(c_{(j)}\right)=\max\left\{0,1-\frac{m}{j}a\right\}^{1/j},\ j=1,...,m\,,$$

των Benjamini και Liu (1999). Οι ίδιες κρίσιμες τιμές θα ικανοποιούν τη συνθήκη (2.21) στην περίπτωση των εξαρτημένων X_j, αν ισχύει η ακόλουθη ιδιότητα υπό τις μηδενικές υποθέσεις.

$$\Pr\left\{T_j\leq c_{(j)}, j=1,...,m\right\}\geq\prod_{j=1}^{m}\Pr\left\{T_j\leq c_{(j)}\right\} \text{ για όλα τα } \left(c_{(1)},...,c_{(m)}\right). \tag{2.22}$$

Για να διαπιστώσουμε ότι οι συνθήκες (2.19) και (2.22) ισχύουν για τις πολυμεταβλητές κατανομές που θεωρήσαμε στο θεώρημα, σημειώνεται ότι:

$$E\left\{\phi\left(T_{m_0+1},...,T_m\right)\,\middle|\, T_{(m_0)}=t\right\}=\sum_{k=1}^{m_0}E\left\{\phi\left(T_{m_0+1},...,T_m\right)\,\middle|\, T_{(m_0)}=t, T_k=T_{(m_0)}\right\}\times\Pr\left\{T_k=T_{(m_0)}, T_{(m_0)}=t\right\}.$$

Αυτό είναι μικρότερο ή ίσο από

$$\frac{1}{m_0}\sum_{k=1}^{m_0}E\left\{\phi\left(T_{m_0+1},...,T_m\right)\,\middle|\, T_{(m_0)}=t, T_m=T_{(m_0)}\right\}$$

λόγω της ιδιότητας της ανταλλαξιμότητας μεταξύ των X_j κάτω από την μηδενική υπόθεση.

Επειδή τα X_j είναι επίσης MTP_2 κάτω από οποιαδήποτε εναλλακτική, κάθε μία από τις δεσμευμένες αναμενόμενες τιμές στην τελευταία σχέση όπως και στην συνθήκη (2.18), με

την ϕ μη φθίνουσα, είναι μη φθίνουσα στο x. Η συνθήκη (2.21) είναι η συνθήκη της θετικής quadrant εξάρτησης και αποτελεί συνέπεια της ιδιότητας MTP_2 (Karlin & Rinott (1980)).

□

ΚΕΦΑΛΑΙΟ 3

Στις περισσότερες περιπτώσεις η από κοινού κατανομή F_n των στατιστικών $T_1, T_2, ..., T_m$ είναι άγνωστη. Έτσι, για την εκτίμηση των τιμών-*p* και των προσαρμοσμένων τιμών-*p* χρησιμοποιούμε μεθόδους επαναδειγματοληψίας, όπως για παράδειγμα τη μέθοδο των μεταθέσεων και την bootstrap. Στο κεφάλαιο αυτό θα παρουσιάσουμε αλγόριθμους μεταθέσεων για τον προσδιορισμό των ακατέργαστων και των προσαρμοσμένων τιμών-*p*, και αλγόριθμους που χρησιμοποιούν τη μέθοδο επαναδειγματοληψίας bootstrap για τον προσδιορισμό των ακατέργαστων τιμών-*p*. Παράλληλα, θα υλοποιήσουμε κάποιους από αυτούς τους αλγορίθμους στο Mathematica. Επίσης, θα πραγματοποιήσουμε την ανάλυση των δεδομένων μας χρησιμοποιώντας συναρτήσεις και πακέτα του προγράμματος R, ειδικά σχεδιασμένα για τον πολλαπλό έλεγχο υποθέσεων και την ανάλυση των γονιδιακών δεδομένων.

3.1 Επαναδειγματοληψία με μεταθέσεις για τον έλεγχο της FWER

Θεωρούμε την περίπτωση που οι *n* πειραματικές μονάδες είναι χωρισμένες τυχαία σε δύο ομάδες πλήθους n_1 και n_2 αντίστοιχα, με $n_1 + n_2 = n$. Έστω ότι η πρώτη είναι η ομάδα θεραπείας και η δεύτερη η ομάδα ελέγχου. Μία κατάλληλη μέτρηση καταγράφεται για την μονάδα $j = 1,2,...,n_k$ της ομάδας $k = 1, 2$.

Σύμφωνα με τη μηδενική υπόθεση H_{0i} για κάθε $i = 1,...,m$, δεν υπάρχει διαφορά μεταξύ των μετρήσεων που αντιστοιχούν στην ομάδα θεραπείας και στην ομάδα ελέγχου. Αν όλες οι μηδενικές υποθέσεις είναι αληθείς, τότε η τυχαία ανάθεση στις ομάδες θεραπείας και ελέγχου σημαίνει ότι όλοι οι δυνατοί συνδυασμοί των *n* πειραματικών μονάδων στις ομάδες 1 και 2 είναι ισοπίθανοι. Το πλήθος των μεταθέσεων των *n* μονάδων σε δύο ομάδες πλήθους n_1 και n_2 είναι ίσο με

$$\mathrm{A} = \frac{(n_1 + n_2)!}{n_1!\, n_2!}. \tag{3.1}$$

Συμβολίζουμε με π την κάθε μετάθεση.

Θέλουμε τώρα να ελέγξουμε αν οι μετρήσεις για τις καταστάσεις θεραπείας και ελέγχου έχουν ίδια κάποια χαρακτηριστικά, όπως για παράδειγμα ίδια μέση τιμή ή ίδια διασπορά. Έστω ότι επιλέγουμε να εξετάσουμε την παρατηρούμενη διαφορά των μέσων των δύο ομάδων

$$d^* = \frac{\sum_{j=1}^{n_2} x_{i2}}{n_2} - \frac{\sum_{j=1}^{n_1} x_{i1}}{n_1} = \bar{x}_1 - \bar{x}_2$$

όπου $\bar{x}_1$ και $\bar{x}_2$ οι μέσες αποκρίσεις για τις δύο ομάδες. Η διαφορά αυτή d^* είναι η παρατηρούμενη διαφορά των μέσων των ομάδων για τα πραγματικά δεδομένα που έχουν συγκεντρωθεί για τη μελέτη.

Αν η H_0 είναι αληθής, τότε κάθε μετάθεση π των n μετρήσεων μπορεί να θεωρηθεί σαν μία πραγματοποίηση του πειράματος. Το πείραμα επαναλαμβάνεται Α φορές και έστω d_π η διαφορά που αντιστοιχεί σε κάθε μετάθεση π. Μία από αυτές τις Α μεταθέσεις αντιπροσωπεύει τα πραγματικά αρχικά δεδομένα. Αν συμβολίσουμε με π^* αυτή τη μετάθεση, τότε η $d_{\pi^*} = d^*$ θα είναι η διαφορά των μέσων που αντιστοιχεί σε αυτή και η οποία συμπίπτει με την παρατηρούμενη διαφορά.

Η τιμή-p σε ένα τεστ μεταθέσεων, όπως θα αποδειχθεί και στη συνέχεια, δίνεται από το πηλίκο του αριθμού των διαφορών d_π που είναι μεγαλύτερες ή ίσες κατά απόλυτη τιμή από την παρατηρούμενη διαφορά d^* προς το Α, δηλαδή από τη σχέση:

$$p - value = \frac{\#\left\{ |d_\pi| \geq |d^*| \right\}}{\mathrm{A}} \qquad (3.2)$$

Στο πρώτο σύνολο δεδομένων της εργασίας δεν εξετάζουμε τη διαφορετική έκφραση των γονιδίων μεταξύ ομάδων θεραπείας και ελέγχου αλλά μεταξύ δύο ομάδων θεραπείας. Τα δεδομένα αυτά προέρχονται από 38 δείγματα μυελού των οστών από τα οποία 27 ήταν τύπου Acute Lemphoblastic Leukemia (ALL) και 11 τύπου Acute Myeloid Leukemia (AML). Επομένως από τη σχέση (3.1) προκύπτει ότι το πλήθος των δυνατών μεταθέσεων είναι $\mathrm{A} = \frac{38!}{27!11!} = 1203322288$.

Για την κατανόηση των παραπάνω και ειδικότερα των τεστ των μεταθέσεων που εφαρμόζονται στα σύνθετα πειραματικά δεδομένα όπως αυτά που συναντάμε σε μελέτες μικροσυστοιχιών, είναι χρήσιμο να ορίσουμε την έννοια της ανταλλαξιμότητας των τυχαίων μεταβλητών.

Ορισμός 3.1. *Έστω* $\mathbf{x} = (x_1, ..., x_n)$ *ένα διάνυσμα n τυχαίων μεταβλητών με από κοινού πυκνότητα πιθανότητας* $f(\mathbf{x})$ *και* $\mathbf{x}_\pi = \pi(\mathbf{x})$ *μία μετάθεση των συνιστωσών του* $\mathbf{x}$. *Οι n τυχαίες μεταβλητές καλούνται ανταλλάξιμες (exchangeable) αν* $f(\mathbf{x}) = f(\mathbf{x}_\pi)$ *για όλα τα διανύσματα* $\mathbf{x}$ *και όλες τις μεταθέσεις* π. □

Η ιδιότητα της ανταλλαξιμότητας σχετίζεται με τα τεστ των μεταθέσεων με τον παρακάτω τρόπο. Η αλήθεια της μηδενικής υπόθεσης συνεπάγεται ότι κάποιες παρατηρήσεις που δε θα ήταν ανταλλάξιμες αν η μηδενική υπόθεση δεν ίσχυε, θα είναι. Στα τεστ των μεταθέσεων θέλει κανείς να αναγνωρίσει αυτά τα *πρόσθετα* σύνολα ανταλλάξιμων παρατηρήσεων υπό τη μηδενική υπόθεση και να υπολογίσει τα στατιστικά για κάθε μία από τις *πρόσθετες* μεταθέσεις που προσφέρει η ιδιότητα της ανταλλαξιμότητας. Ορισμένες παρατηρήσεις βέβαια, μπορεί να είναι ανταλλάξιμες είτε η μηδενική υπόθεση ισχύει είτε όχι.

Συγκεκριμένα, για τη περίπτωση που μελετάμε, ας συμβολίσουμε με A_0 τις αρχικές δυνατές μεταθέσεις, $A_0 = 27!11!$. Κάτω από τη μηδενική υπόθεση οι δύο θεραπείες έχουν ίδιες αποκρίσεις και οι 38 παρατηρήσεις γίνονται ανταλλάξιμες δίνοντας έτσι $A_1 = 38!$ δυνατές μεταθέσεις. Έτσι οι πρόσθετες μεταθέσεις που προέρχονται από την υπόθεση αλήθειας της μηδενικής υπόθεσης είναι $A = A_1 / A_0 = 38!/27!11!$ όπως προκύπτει και από τη σχέση (3.1).

Ερμηνεία της σχέσης (3.2). Η σχέση (3.2) μπορεί, μέσα από ένα πιο γενικευμένο συμβολισμό, να αιτιολογηθεί ως εξής. Θεωρούμε το σύνολο των παρατηρούμενων δεδομένων $\mathbf{x} = \{x_1, ..., x_n\}$ το οποίο προέρχεται από ένα συμβολικό πείραμα που εκτελείται *n* φορές, ως πραγματοποιήσεις αντίστοιχων τυχαίων μεταβλητών που παίρνουν τιμές σε κάποιο δειγματικό χώρο X. Για το σκοπό της ανάλυσης, το σύνολο των δεδομένων είναι χωρισμένο σε ομάδες (ή δείγματα), σύμφωνα με τα επίπεδα θεραπείας του πειράματος. Για κάθε γενικό πρόβλημα ελέγχου η μηδενική υπόθεση H_0 υποθέτει ότι τα δεδομένα προέρχονται από μία

άγνωστη κατανομή P. Όλο το σύνολο των παρατηρούμενων δεδομένων $\mathbf{x}$ θεωρείται ως ένα δείγμα το οποίο παίρνει τιμές από ένα δειγματικό χώρο X^n, όπου $\mathbf{x}$ είναι μία παρατήρηση της n-διάστατης μεταβλητής $\mathbf{X}^{(n)}$. Αυτό το δείγμα δεν αποτελείται απαραίτητα από ανεξάρτητες και ισόνομες συνιστώσες. Θεωρούμε λοιπόν ότι η H_0 είναι αληθής και ότι όλα τα μέλη μιας μη παραμετρικής οικογένειας κατανομών $\mathcal{M}$ κυριαρχούνται από ένα μέτρο ξ (π.χ. αν η οικογένεια αποτελείται από απόλυτες συνεχείς κατανομές, τότε το ξ είναι το μέτρο Lebesgue) . Επίσης συμβολίζουμε με f_P την πυκνότητα πιθανότητας της P και με $f_P^{(n)}(\mathbf{x})$ την πυκνότητα πιθανότητας του τυχαίου διανύσματος $\mathbf{X}^{(n)}$.

Σύμφωνα με την αρχή της πιθανοφάνειας, αν για κάποια $\mathbf{x},\mathbf{x}^* \in \mathsf{X}^n$ ο λόγος πιθανοφανειών $\dfrac{f_P^{(n)}(\mathbf{x})}{f_P^{(n)}(\mathbf{x}^*)} = \rho(\mathbf{x},\mathbf{x}^*)$ δεν εξαρτάται από την f_P για οποιαδήποτε κατανομή $P \in M$, τότε τα $\mathbf{x}$ και $\mathbf{x}^*$ περιέχουν την ίδια πληροφορία, με την έννοια ότι είναι ισοδύναμα ως προς τα συμπεράσματα που θα προκύψουν από αυτά. Το σύνολο των σημείων που είναι ισοδύναμα με το $\mathbf{x}$ καλείται *τροχιά* (τροχιά) που σχετίζεται με το $\mathbf{x}$ και το συμβολίζουμε με $\mathsf{X}_{/\mathbf{x}}^n$, τέτοιο ώστε

$$\mathsf{X}_{/\mathbf{x}}^n = \{\mathbf{x}^* : \rho(\mathbf{x},\mathbf{x}^*) \text{ ανεξάρτητη της } f_P\}.$$

Σημειώνουμε ότι, όταν τα δεδομένα προέρχονται από τυχαία δειγματοληψία με ανεξάρτητες και ισόνομες παρατηρήσεις, τέτοια ώστε $f_P^{(n)}(\mathbf{x}) = \prod_{1\le i\le n} f_P(x_i)$ τότε η τροχιά $\mathsf{X}_{/\mathbf{x}}^n$ που σχετίζεται με το $\mathbf{x}$ περιέχει όλες τις μεταθέσεις του $\mathbf{x}$ και ο λόγος πιθανοφανειών ικανοποιεί την ισότητα $\rho(\mathbf{x},\mathbf{x}^*)=1$. Το ίδιο συμπέρασμα λαμβάνουμε και αν υποθέσουμε ότι η $f_P^{(n)}(\mathbf{x})$ θεωρείται αμετάβλητη ως προς τις μεταθέσεις των στοιχείων του $\mathbf{x}$. Αυτό συμβαίνει όταν η υπόθεση της ανεξαρτησίας των δεδομένων αντικαθίσταται από αυτήν της ανταλλαξιμότητας. Τα τεστ των μεταθέσεων είναι υπό συνθήκη στατιστικές διαδικασίες, όπου η συνθήκη αφορά στην τροχιά $\mathsf{X}_{/\mathbf{x}}^n$. Έτσι το $\mathsf{X}_{/\mathbf{x}}^n$ παίζει το ρόλο του συνόλου αναφοράς για το υπό συνθήκη συμπέρασμα. Κάτω από την μηδενική υπόθεση και θεωρώντας ότι ισχύει η ιδιότητα της ανταλλαξιμότητας, η υπό συνθήκη κατανομή πιθανότητας του σημείου $\mathbf{x}' \in \mathsf{X}^n$ για οποιαδήποτε $P \in M$ είναι

$$\Pr\left(\mathbf{x}^* = \mathbf{x}' \mid \mathsf{X}_{/\mathbf{x}}^n\right) = \frac{\sum_{\mathbf{x}^* = \mathbf{x}'} f_P^{(n)}(\mathbf{x}) \cdot d\xi^n}{\sum_{\mathbf{x}^* \in \mathsf{X}_{/\mathbf{x}}^n} f_P^{(n)}(\mathbf{x}) \cdot d\xi^n} = \frac{\#\left\{\mathbf{x}^* = \mathbf{x}', \mathbf{x}^* \in \mathsf{X}_{/\mathbf{x}}^n\right\}}{\#\left\{\mathbf{x}^* \in \mathsf{X}_{/\mathbf{x}}^n\right\}}.$$

Τα παραπάνω λοιπόν, μας επιτρέπουν να θεωρήσουμε ότι τα συμπεράσματα που προκύπτουν από τις μεταθέσεις είναι αμετάβλητα σε σχέση με την κατανομή *P*.

Επιστρέφοντας στον αρχικό συμβολισμό μας, η τιμή-*p* αποτελεί επίσης μία υπό συνθήκη κατανομή πιθανότητας για οποιαδήποτε $P \in M$

$$p - value = \Pr\left(|d_\pi| \geq |d^*| \mid H_0\right).$$

Έτσι, στη σχέση (3.2) έχουμε στον αριθμητή τον $\#\left\{|d_\pi| \geq |d^*|\right\}$ και στον παρονομαστή το πλήθος των πρόσθετων μεταθέσεων που προέρχονται από την υπόθεση αλήθειας της μηδενικής υπόθεσης, δηλαδή το πλήθος *A*.

Στη συνέχεια της εργασίας θα παρουσιάσουμε διάφορους αλγορίθμους μεταθέσεων για την παραγωγή των ακατέργαστων και των προσαρμοσμένων τιμών-*p* και την αναγνώριση της διαφορετικής έκφρασης των γονιδίων. Ως στατιστικό θα χρησιμοποιήσουμε το στατιστικό του Welch για δύο δείγματα που δίνεται από τη σχέση:

$$t_i = \frac{\bar{x}_{2i} - \bar{x}_{1i}}{\sqrt{\frac{s_{2i}^2}{n_2} + \frac{s_{1i}^2}{n_1}}}, \quad i = 1, \ldots, m,$$

όπου $\bar{x}_{1i}$ και $\bar{x}_{2i}$ τα μέσα επίπεδα έκφρασης του γονιδίου *i* για τις δύο ομάδες πλήθους n_1 και n_2.

3.1.1 Ακατέργαστες τιμές-*p*

Ο Αλγόριθμος 1 μας δίνει τις ακατέργαστες τιμές-*p* μέσω των μεταθέσεων. Οι αντίστοιχες προσαρμοσμένες τιμές-*p* σύμφωνα με τις διαδικασίες Bonferroni, Sidak και Holm προκύπτουν με αντικατάσταση των p_i στις σχέσεις (2.2)-(2.5) που παρουσιάσαμε στο Κεφάλαιο 2.

Αλγόριθμος 1. Αλγόριθμος μεταθέσεων για τον υπολογισμό των ακατέργαστων τιμών-*p*

Για την τυχαία μετάθεση $b = 1, \ldots, B$:

ΒΗΜΑ 1: Μεταθέτουμε τις n στήλες του πίνακα X .

ΒΗΜΑ 2: Υπολογίζουμε τα στατιστικά $t_{1,b},...,t_{m,b}$ για κάθε υπόθεση.

Μετά από B μεταθέσεις , η τιμή-p για την υπόθεση H_{0i} είναι

$$p_i = \frac{\#\left\{b : |t_{i,b}| \geq |t_i|\right\}}{B}, \quad \text{για} \quad i = 1,...,m .$$

3.1.2 Step-down maxT προσαρμοσμένες τιμές-*p*

Ο Αλγόριθμος 2 είναι ο αλγόριθμος παραγωγής των step-down maxT προσαρμοσμένων τιμών-*p* των Westfall & Young. Σε αυτόν απαιτείται η εκτίμηση της κατανομής των διαδοχικών μέγιστων $\max_{k=1,...,m} |T_{(k)}|$.

3.1.3 Step-down minP προσαρμοσμένες τιμές-*p*

Οι single-step και step-down minP προσαρμοσμένες τιμές-*p* των Westfall και Young είναι πολύ πιο δύσκολο να υπολογιστούν διότι απαιτούν την από κοινού κατανομή των $P_1,...,P_m$ υπό την μηδενική υπόθεση.

Ο Αλγόριθμος 3 είναι ο αλγόριθμος παραγωγής των step-down minP προσαρμοσμένων τιμών-*p* και καλείται διπλός αλγόριθμος διότι η μέθοδος της επαναδειγματοληψίας πραγματοποιείται δύο φορές. Το πλήθος των επαναλήψεων είναι πολύ μεγάλο και για το λόγο αυτό θεωρείται υπολογιστικά σχεδόν αδύνατος. Συγκεκριμένα, αν θεωρήσουμε τους $m \times B$ πίνακες $T = \left[t_{i,b}\right]$ των στατιστικών, $P = \left[p_{i,b}\right]$ των ακατέργαστων τιμών-*p* και $Q = \left[q_{i,b}\right]$ των ελάχιστων τιμών των ακατέργαστων τιμών-*p*, όπου $q_{i,b} = \min_{l=1,...,m} p_{l,b}$, ο αλγόριθμος αυτός υπολογίζει κάθε φορά μία από τις στήλες των πινάκων *T*, *P* και *Q*. Οι τιμές-*p* στη στήλη *b* του πίνακα *P* προκύπτουν θεωρώντας τις *B* μεταθέσεις των στηλών του πίνακα X_b και υπολογίζοντας τον πίνακα *T* ξανά από την αρχή.

Αλγόριθμος 2. Αλγόριθμος μεταθέσεων για τον υπολογισμό των step-down maxT προσαρμοσμένων τιμών-*p*.

Αρχικά, για τα πραγματικά δεδομένα μας διατάσσουμε τα στατιστικά έτσι ώστε $|t_{(1)}| \geq |t_{(2)}| \geq ... \geq |t_{(m)}|$

Για την τυχαία μετάθεση $b = 1,...,B$:

ΒΗΜΑ 1: Μεταθέτουμε τις n στήλες του πίνακα X.

ΒΗΜΑ 2: Υπολογίζουμε τα στατιστικά $t_{1,b},...,t_{m,b}$ για κάθε υπόθεση.

ΒΗΜΑ 3: Υπολογίζουμε τα $u_{i,b} = \max_{k=1,...,m} |t_{(k),b}|$, δηλαδή τα διαδοχικά μέγιστα των στατιστικών χρησιμοποιώντας τις σχέσεις

$$u_{m,b} = |t_{(k),b}|$$

$$u_{i,b} = \max\left(u_{i+1,b}, |t_{(i),b}|\right) \quad \text{για } i = m - 1,...,1.$$

Τα παραπάνω βήματα επαναλαμβάνονται B φορές και οι προσαρμοσμένες τιμές-p εκτιμώνται από τη σχέση:

$$\tilde{p}_{(i)} = \frac{\#\left\{ b: u_{i,b} \geq |t_{(i)}| \right\}}{\mathrm{B}} \quad \text{για } i = 1,...,m.$$

Αλγόριθμος 3. Κλασικός διπλός αλγόριθμος μεταθέσεων για τον υπολογισμό των step-down minP προσαρμοσμένων τιμών-*p*.

Για την τυχαία μετάθεση $b = 1,...,B$:

ΒΗΜΑ1: Μεταθέτουμε τις n στήλες του πίνακα X.

ΒΗΜΑ 2: Υπολογίζουμε τις ακατέργαστες τιμές-p, $p_{1,b},...,p_{m,b}$ για κάθε μία υπόθεση σύμφωνα με τον Αλγόριθμο 1.

ΒΗΜΑ 3: Διατάσσουμε τις ακατέργαστες τιμές-p έτσι ώστε $p_{(1),b} \leq p_{(2),b} \leq ... \leq p_{(m),b}$.

ΒΗΜΑ 4: Υπολογίζουμε τα $q_{i,b} = \min_{l=1,...,m} p_{(l),b}$ δηλαδή τα διαδοχικά ελάχιστα των ακατέργαστων τιμών-p.

$$q_{m,b} = p_{(l),b}$$

$$q_{i,b} = \min\left(q_{i+1,b}, p_{(i),b}\right), \text{ για } i = m - 1,...,1.$$

Τα παραπάνω βήματα επαναλαμβάνονται B φορές και οι προσαρμοσμένες τιμές-p εκτιμώνται από τη σχέση:

$$\tilde{p}_{(i)} = \frac{\#\left\{b: q_{i,b} \geq p_{(i)}\right\}}{B} \quad \text{για } i = 1,...,m .$$

Ο Αλγόριθμος 4 αποτελεί έναν καινούργιο αλγόριθμο μεταθέσεων για τον υπολογισμό των step-down minP προσαρμοσμένων τιμών-*p*. Υπολογίζει τον πίνακα *T* μόνο μία φορά και αντιμετωπίζει τις γραμμές των πινάκων *T*, *P* και *Q* διαδοχικά, ξεκινώντας από την τελευταία. Η βασική ιδέα είναι να προχωράει με μία υπόθεση αντί μίας μετάθεσης κάθε φορά και να υπολογίζει τις *B* ακατέργαστες τιμές-*p* για κάθε υπόθεση ταξινομώντας τα *B* στατιστικά χρησιμοποιώντας ένα μικρό και σύντομο αλγόριθμο. Σημειώνουμε ότι αυτός ο αλγόριθμος παράγει τις ίδιες τιμές-*p* με αυτές που παράγονται από τον Αλγόριθμο 3.

Αλγόριθμος 4. Νέος αλγόριθμος μεταθέσεων για τον υπολογισμό των step-down minP προσαρμοσμένων τιμών-*p*.

ΒΗΜΑ 1: Υπολογίζουμε τις ακατέργαστες τιμές-*p* για κάθε υπόθεση. Υποθέτουμε ότι $p_1 \leq p_2 \leq ... \leq p_m$, διαφορετικά διατάσσουμε τις γραμμές του πίνακα *X* σύμφωνα με τα διατεταγμένα p_i^*.

Χρησιμοποιούμε τις σχέσεις $q_{m+1,b} = 1$ για $b = 1,...,B$ και $i = m$.

ΒΗΜΑ 2: Για την υπόθεση H_i (γραμμή *i*), υπολογίζουμε τα στατιστικά $t_{i,1},...,t_{i,B}$ από τις *B* μεταθέσεις και χρησιμοποιούμε ένα γρήγορο αλγόριθμο για τις ακατέργαστες τιμές-*p* $p_{i,1},...,p_{i,B}$.

ΒΗΜΑ 3: Παίρνουμε τα διαδοχικά ελάχιστα $q_{i,b}$.

$$q_{i,b} \leftarrow \min\left(q_{i+1,b}, p_{i,b}\right) \text{ για } b = 1,...,B .$$

ΒΗΜΑ 4: Υπολογίζουμε τις προσαρμοσμένες τιμές-*p* για την υπόθεση H_i από τη σχέση

$$\tilde{p}_i = \frac{\#\left\{b: q_{i,b} \geq p_i\right\}}{B}$$

ΒΗΜΑ 5: Διαγράφουμε τη γραμμή *i* του πίνακα *P* και τη γραμμή $i+1$ του πίνακα *Q*.

ΒΗΜΑ 6: Κινούμαστε μία γραμμή επάνω, $i \leftarrow i - 1$

Αν $i = 0$ πηγαίνουμε στο βήμα 6, διαφορετικά πηγαίνουμε στο βήμα 1.

3.1.4 Τιμές-*q*

Οι Storey και Tibshirani (2001) (ST) επέκτειναν τη θεωρία του Storey (2002) για την εκτίμηση της pFDR έτσι ώστε να μπορεί να εφαρμοστεί κάτω από ορισμένες πιο γενικές υποθέσεις εξάρτησης.

Γράφοντας τη σχέση (2.10) για μία γενική οικογένεια περιοχών απόρριψης $\{\Gamma_a\}$ έχουμε:

$$pFDR_{\Gamma_{a_0}}(\Gamma_a) = \frac{\pi_0(\Gamma_{a_0})a}{\Pr(T \in \Gamma_a)\Pr(R>0)} = \frac{W(\Gamma_{a_0})a}{(1-a_0)\{R(\Gamma_a) \vee 1\}\Pr(R>0)}$$

όπου $R(\Gamma) = \#\{i : T_i \in \Gamma\}$, $W(\Gamma) = \#\{i : T_i \notin \Gamma\} = m - W(\Gamma)$ και Γ_a η περιοχή απόρριψης επιπέδου a.

Στη συνέχεια, θεωρούμε μία γενική περιοχή απόρριψης Γ, όπως για παράδειγμα η $[-\infty, -c] \cup [c, +\infty]$ και εκτιμούμε την pFDR μέσω της παραπάνω σχέσης με επαναδειγματοληψία. Επιλέγουμε μία περιοχή Γ_0 που πιστεύεται ότι περιέχει τις περισσότερες μηδενικές υποθέσεις και συμβολίζουμε με B το πλήθος των μεταθέσεων που προκύπτουν και με $t_{i,b}, i = 1,...,m, b = 1,...,B$ τα αντίστοιχα στατιστικά. Τότε οι εκτιμήσεις των a, a_0 και $\Pr(R>0)$ δίνονται από τις σχέσεις:

$$\hat{a} = \frac{1}{Bm}\sum_{b=1}^{B} R_b(\Gamma) = \frac{\overline{R}(\Gamma)}{m}$$

$$\hat{a}_0 = \frac{1}{Bm}\sum_{b=1}^{B} R_b(\Gamma_0) = \frac{\overline{R}(\Gamma_0)}{m}$$

$$\widehat{\Pr}(R>0) = \frac{\#\{b : R_b(\Gamma) > 0\}}{B} = \overline{I}_{\{R(\Gamma)>0\}}$$

όπου $R_b(\Gamma) = \#\{i : t_{i,b} \in \Gamma\}$, $\overline{R}(\Gamma) = \frac{1}{B}\sum_{b=1}^{B} R_b(\Gamma)$ και όμοια για τα $W_b(\Gamma)$ και $\overline{W}(\Gamma)$.

Αντικαθιστώντας τους παραπάνω τύπους στην αρχική μας σχέση έχουμε:

$$\widehat{pFDR}_{\Gamma_0}(\Gamma) = \frac{W(\Gamma_0)\bar{R}(\Gamma)}{\left(m-\bar{R}(\Gamma_0)\right)\left(R(\Gamma)\vee 1\right)\widehat{\Pr}(R>0)}$$

$$= \frac{W(\Gamma_0)\bar{R}(\Gamma)}{\overline{W}(\Gamma_0)\left(R(\Gamma)\vee 1\right)\bar{I}_{\{R(\Gamma)>0\}}} \tag{3.3}$$

Τέλος, αγνοώντας την εκτίμηση της $\Pr(R>0)$, έχουμε μια συντηρητική εκτίμηση της $FDR(\Gamma)$,

$$\widehat{FDR}_{\Gamma_0}(\Gamma) = \frac{W(\Gamma_0)\bar{R}(\Gamma)}{\left(m-\bar{R}(\Gamma_0)\right)\left(R(\Gamma)\vee 1\right)}$$

$$= \frac{W(\Gamma_0)\bar{R}(\Gamma)}{\overline{W}(\Gamma_0)\left(R(\Gamma)\vee 1\right)} \tag{3.4}$$

Η σχέση (2.10) μπορεί να χρησιμοποιηθεί για τον υπολογισμό των τιμών-q με σκοπό τη ρύθμιση της pFDR και η μέθοδος αυτή ονομάζεται Storey-q διαδικασία. Όμοια η διαδικασία ST (Storey-Tibshirani) χρησιμοποιεί τη σχέση (3.3) για τη ρύθμιση της FDR κάτω από μία γενική εξάρτηση, ενώ η διαδικασία ST-q χρησιμοποιεί τη σχέση (3.4) για τη ρύθμιση της pFDR.

Οι διαδικασίες αυτές περιγράφονται στον Αλγόριθμο 5.

Αλγόριθμος 5. Αλγόριθμος μεταθέσεων για τις διαδικασίες ST και ST-q βασισμένες στους Storey και Tibshirani (2001)

Αρχικά επιλέγουμε μία τιμή τ_0 τέτοια ώστε η περιοχή $[-\infty,-\tau_0]\cup[\tau_0,+\infty]$ να περιέχει τις περισσότερες μηδενικές υποθέσεις, για παράδειγμα $\tau_0 = 0.2$. Από τα αρχικά δεδομένα υπολογίζουμε το two-sample στατιστικό-t ώστε $\tau_i = |t_i|$ και διατάσσουμε έτσι ώστε $\tau_1 \geq \ldots \geq \tau_m$.

Υπολογίζουμε τα $R_i = \#\{k : |t_k| \geq \tau_i\}$ και $W_0 = \#\{k : |t_k| \geq \tau_0\}$.

Για τη μετάθεση $b = 1,\ldots,B$:

ΒΗΜΑ 1. Μεταθέτουμε τις n στήλες του πίνακα X.

ΒΗΜΑ 2. Υπολογίζουμε τα στατιστικά $t_{1,b},...,t_{m,b}$ για κάθε υπόθεση.

ΒΗΜΑ 3. Υπολογίζουμε τα

$$R_{i,b} = \#\left\{\lambda : \left|t_{\lambda,b}\right| \geq \tau_i\right\}$$

και

$$W_{0,b} = \#\left\{k : \left|t_{i,b}\right| \geq \tau_0\right\} \text{ για } i = 1,...,m\,.$$

Τα παραπάνω βήματα επαναλαμβάνονται *B* φορές και έπειτα για $i = 1,...,m$ υπολογίζουμε τα

$$\overline{R}_i = \frac{1}{B}\sum_{b=1}^{B} R_{i,b}$$

$$\overline{I}_i = \frac{1}{B}\sum_{b=1}^{B} I\left(R_{i,b} > 0\right)$$

$$\overline{W}_0 = \frac{1}{B}\sum_{b=1}^{B} W_{0,b}$$

Έτσι η pFDR για το τ_i είναι

$$pFDR_i = \frac{\overline{W}_0 \overline{R}_i}{\overline{W}_0 \left(R_i \vee 1\right)\overline{I}_i}, \quad i = 1,...,m\,,$$

και η FDR είναι

$$FDR_i = \frac{\overline{W}_0 \overline{R}_i}{\overline{W}_0 \left(R_i \vee 1\right)}, \quad i = 1,...,m\,.$$

Οι τιμές-*q* (για την ST-*q* διαδικασία) και οι τιμές-*p* (για την ST-διαδικασία) εκτιμώνται από τις σχέσεις:

$$q_m = pFDR_m, \qquad q_i = \min\left(q_{i+1}, pFDR_i\right), \quad i = m-1,...,1\,,$$

$$\hat{p}_m = FDR_m, \qquad \hat{p}_i = \min\left(\hat{p}_{i+1}, FDR_i\right), \quad i = m-1,...,1\,.$$

3.2 Επαναδειγματοληψία με τη μέθοδο Bootstrap

Έστω n_1 πειραματικές μονάδες μεγέθους *m* που αποτελούν το δείγμα 1 και n_2 πειραματικές μονάδες μεγέθους *m* που αποτελούν το δείγμα 2. Συμβολίζουμε με $X_{i1},...,X_{in_1}$ και $Y_{i1},...,Y_{in_2}$ τις τυχαίες μεταβλητές που αντιστοιχούν στις μετρήσεις για τα δύο δείγματα

για $i = 1,...,m$ και θέλουμε να ελέγξουμε αν υπάρχει διαφορά μεταξύ των δύο ομάδων, δηλαδή αν $F_X \neq F_Y$.

Δεδομένου ότι θέλουμε να χρησιμοποιήσουμε τη μέθοδο επαναδειγματοληψίας Bootstrap, μία πρώτη προσέγγιση στο πρόβλημα της εκτίμησης της από κοινού κατανομής Q_n των στατιστικών $T_{1n}, T_{2n},..., T_{mn}$, είναι η εξής: επιλέγουμε τυχαία n_1 τιμές $X^*_{i1},..., X^*_{in_1}$ από τους $X_{i1},..., X_{in_1}$ και n_2 τιμές $Y^*_{i1},..., Y^*_{in_2}$ από τους $Y_{i1},..., Y_{in_2}$ με επανάθεση. Από αυτές μπορούμε να εκτιμήσουμε την κατανομή τής

$$T_i = T\left(X_{i1},..., X_{in_1}, Y_{i1},..., Y_{in_2}\right) = \overline{X}_i - \overline{Y}_i \text{ όπου } X_i \sim F_X, Y_i \sim F_Y$$

από την κατανομή της

$$T^*_i = T\left(X^*_{i1},..., X^*_{in_1}, Y^*_{i1},..., Y^*_{in_2}\right) = \overline{X}^*_i - \overline{Y}^*_i \text{ όπου } X^*_i \sim \hat{F}_X, Y^*_i \sim \hat{F}_Y.$$

Οι $\hat{F}_X$, $\hat{F}_Y$ είναι οι εμπειρικές συναρτήσεις κατανομής που προκύπτουν από τα δείγματα $X_{i1},..., X_{in_1}$, $Y_{i1},..., Y_{in_2}$ και αποτελούν εκτιμήσεις των συναρτήσεων κατανομής F_X, F_Y αντίστοιχα.

Η Bootstrap εκτίμηση της τιμής-*p* δίνεται από τη σχέση

$$p - value_i = P\left(\left|T^*_i\right| > \left|T_i\right| \mid H_{0i}\right) = E\left[I\left(\left|T^*_i\right| > \left|T_i\right| \mid H_{0i}\right)\right].$$

Η (Monte Carlo) εκτίμηση της παραπάνω μέσης τιμής μπορεί να υπολογιστεί μέσω προσομοίωσης επαναλαμβάνοντας *B* φορές τη μέθοδο της επαναδειγματοληψίας και υπολογίζοντας διαδοχικά τα $T^*_{i1},..., T^*_{ik}$. Έτσι, η εκτίμηση της τιμής-*p* για το *i* γονίδιο δίνεται από τη σχέση:

$$p - value_i = \frac{1}{k}\sum_{r=1}^{k} I\left(\left|T^*_i\right| > \left|T_i\right|\right) = \frac{\#\left\{\left|T^*_i\right| > \left|T_i\right|\right\}}{k}$$

Παρακάτω παρουσιάζουμε τον αλγόριθμο για την παραγωγή των ακατέργαστων τιμών-*p*.

Αλγόριθμος 6. Αλγόριθμος για την παραγωγή των ακατέργαστων τιμών-*p* με τη μέθοδο Bootstrap.

Για $i = 1,2,..., m$:

ΒΗΜΑ 1: θέτουμε $s = 0$, $m = 0$.

ΒΗΜΑ 2: Για $j = 1,2,...,n_1$ παράγουμε τυχαίους αριθμούς $U_{1,j} \sim U(0,1)$ και θέτουμε $Z_{1,j} = \left[n_1 U_{1,j} \right] + 1$ και $X^*_{i,j} = X_{i,Z_{1,j}}$.

ΒΗΜΑ 3: Για $j = 1,2,...,n_2$ παράγουμε τυχαίους αριθμούς $U_{2,j} \sim U(0,1)$ και θέτουμε $Z_{2,j} = \left[n_2 U_{2,j} \right] + 1$ και $Y^*_{i,j} = Y_{2,Z_{2,j}}$.

ΒΗΜΑ 4: Υπολογίζουμε τη διαφορά $T^*_i = \left| \overline{X}^*_i - \overline{Y}^*_i \right| = \frac{1}{n_1} \sum_{j=1}^{n_1} X^*_{i,j} - \frac{1}{n_2} \sum_{j=1}^{n_2} Y^*_{i,j}$.

ΒΗΜΑ 5: Αν $T^*_i > T_i$, θέτουμε $s = s + 1$, $m = m + 1$ και αν $m < k$ επιστρέφουμε στο βήμα 1 αλλιώς πάμε στο βήμα 6.

ΒΗΜΑ 6: Επιστρέφουμε την εκτίμηση s / k για την τιμή-p που αντιστοιχεί στο i γονίδιο.

Ο αλγόριθμος αυτός υλοποιήθηκε στο Mathematica για το σύνολο των δεδομένων της λευχαιμίας και σαν αποτέλεσμα μας έδωσε μία λίστα με τις τιμές-p για τα 2109 γονίδια. Οι τιμές αυτές φαίνεται να είναι αρκετά μεγάλες. Σύμφωνα με τον Boos (2003) η συγκεκριμένη μέθοδος επαναδειγματοληψίας δεν είναι η κατάλληλη καθώς δεν θέτει κανένα περιορισμό στα δεδομένα. Το καθοριστικό σημείο για το σωστό υπολογισμό της τιμής-p είναι *η επαναδειγματοληψία να πραγματοποιηθεί κάτω από την κατάλληλη μηδενική υπόθεση* και ένας τρόπος για να γίνει αυτό είναι να πάρουμε και τα δύο δείγματα (με επανάθεση) από το ίδιο σύνολο $\{X_{i1},...,X_{in_1},Y_{i1},...,Y_{in_2}\}$. Με αυτόν τον τρόπο δημιουργούμε την μηδενική υπόθεση

$$H_{0i} : P\left(X^*_i \le t\right) = P\left(Y^*_i \le t\right) = H_n(t)$$

όπου $P\left(X^*_i \le t\right)$ η συνάρτηση κατανομής του X^*_i, $P\left(Y^*_i \le t\right)$ η συνάρτηση κατανομής του Y_i και $H_n(t)$ η εμπειρική συνάρτηση κατανομής του "συνδυασμένου" συνόλου των $n = n_1 + n_2$ μετρήσεων για το γονίδιο $i = 1,...,m$.

Έτσι ένας πιο αξιόπιστος αλγόριθμος για την παραγωγή των ακατέργαστων τιμών-p είναι ο αλγόριθμος 7.

Αλγόριθμος 7. Αλγόριθμος για την παραγωγή των ακατέργαστων τιμών-*p* με τη μέθοδο Bootstrap υπό τη μηδενική υπόθεση.

Για $i = 1,2,...,m$:

ΒΗΜΑ 1: θέτουμε $s = 0$, $m = 0$.

ΒΗΜΑ 2: Για $j = 1,2,...,n$ παράγουμε τυχαίους αριθμούς $U_{1,j} \sim U(0,1)$ και θέτουμε $Z_{1,j} = \left[nU_{1,j} \right] + 1$ και $X^*_{i,j} = X_{i,Z_{1,j}}$.

ΒΗΜΑ 3: Για $j = 1,2,...,n$ παράγουμε έναν τυχαίο αριθμό $U_{2,j} \sim U(0,1)$ και θέτουμε $Z_{2,j} = \left[nU_{2,j} \right] + 1$ και $Y^*_{i,j} = Y_{2,Z_{2,j}}$.

ΒΗΜΑ 4: Υπολογίζουμε τη διαφορά $T^*_i = \left| \overline{X}^*_i - \overline{Y}^*_i \right| = \frac{1}{n_1} \sum_{j=1}^{n_1} X^*_{i,j} - \frac{1}{n_2} \sum_{j=1}^{n_2} Y^*_{i,j}$.

ΒΗΜΑ 5: Αν $T^*_i > T_i$, θέτουμε $s = s + 1$, επαναλαμβάνουμε *m* φορές.

ΒΗΜΑ 6: Επιστρέφουμε την εκτίμηση s / k για την τιμή-*p* που αντιστοιχεί στο *i* γονίδιο.

Αντί της διαφοράς των μέσων μπορούμε επίσης να χρησιμοποιήσουμε το στατιστικό

$$t_p = \frac{\left(\overline{X} - \overline{Y} \right)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

όπου $s_p^2 = \frac{(n_1 - 1) s_X^2 + (n_2 - 1) s_Y^2}{n_1 + n_2 - 2}$ και s_X^2, s_Y^2 οι δειγματικές διασπορές των $X_{i1},...,X_{in_1}$ και $Y_{i1},...,Y_{in_2}$ αντίστοιχα.

Στο σημείο αυτό θα μπορούσαμε να κάνουμε μία σύγκριση των μεθόδων μεταθέσεων και Bootstrap. Οι έλεγχοι των μεταθέσεων περιορίζονται σε ένα μικρό αριθμό περιπτώσεων ελέγχου όπου οι μεταθέσεις υπό τη μηδενική υπόθεση έχουν την ίδια κατανομή. Βέβαια, για αυτές τις περιπτώσεις δίνουν ακριβή αποτελέσματα. Αντίθετα, το εύρος εφαρμογών των ελέγχων bootstrap είναι μεγάλο, αλλά τα αποτελέσματα είναι προσεγγιστικά. Τα bootstrap studentized στατιστικά, όπως το t_p, είναι προτιμότερα από τα bootstrap στατιστικά όπως η διαφορά των μέσων $\overline{X} - \overline{Y}$, λόγω της πιο γρήγορης σύγκλισης των κατανομών.

Στην περίπτωση που θέλουμε να κάνουμε τον έλεγχο της υπόθεσης

$$H_{0i} : \mu_X - \mu_Y = 0$$

χωρίς κανέναν άλλο περιορισμό για τις κατανομές των στατιστικών (επιτρέποντας π.χ. οι κατανομές να έχουν διαφορετικές διακυμάνσεις), μπορούμε να υπολογίσουμε το στατιστικό του Welch,

$$t_W = \left(\overline{X} - \overline{Y}\right) / \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}$$

και ένας τρόπος να χρησιμοποιήσουμε τη μέθοδο Bootstrap με μία κατάλληλη μηδενική υπόθεση είναι να επιλέξουμε τυχαία με επανάθεση n_1 τιμές από τις $X_{i1} - \overline{X}, ..., X_{in_1} - \overline{X}$ και n_2 τιμές από τις $Y_{i1} - \overline{Y}, ..., Y_{in_2} - \overline{Y}$. Με αυτόν τον τρόπο ενισχύουμε την παραδοχή υπό την μηδενική υπόθεση ότι οι μέσες τιμές των *X* και *Y* έχουν μηδενική διαφορά.

3.3 Ανάλυση Δεδομένων

Στην παρούσα ενότητα θα παρουσιάσουμε τρεις διαδικασίες για την αναγνώριση των διαφορετικά εκφρασμένων γονιδίων. Αυτές είναι:

- Η χρήση του Mathematica για την υλοποίηση ορισμένων από τους αλγορίθμους που περιγράψαμε παραπάνω.
- Η χρήση της συνάρτησης MTP μέσω του προγράμματος R και του πακέτου multtest του Bioconductor.
- Η χρήση του πακέτου affylmGUI μέσω του προγράμματος R.

Με τους δύο πρώτους θα αναλύσουμε τα δεδομένα που προέρχονται από τα 38 δείγματα μυελού των οστών τύπου ALL και 11 τύπου AML, ενώ με τον τρίτο τρόπο θα αναλύσουμε τα δεδομένα που προέρχονται από τα 8 δείγματα καρκίνου του μαστού.

3.3.1 Mathematica

Χρησιμοποιώντας τη μέθοδο bootstrap, ο πιο αξιόπιστος αλγόριθμος για την παραγωγή των ακατέργαστων τιμών-*p* είναι ο Αλγόριθμος 7. Η υλοποίηση του αλγορίθμου στο Mathematica, με τη χρήση του bootstrap studendized στατιστικού t_p, φαίνεται παρακάτω.

Με την εκτέλεση του αλγορίθμου, λαμβάνουμε τη λίστα P με τις τιμές των ακατέργαστων τιμών-*p* για τα 2109 γονίδια. Διατάσσουμε τις τιμές αυτές και εφαρμόζουμε τις διαδικασίες Holm (1979), Sidak, Hochberg (1988), Benjamini-Hochberg (1995), Benjamini- Liu (1999), Benjamini- Yekutieli (2001) για τον έλεγχο των υποθέσεων

H_{0i} : το γονίδιο *i* δεν είναι διαφορετικά εκφρασμένο.

Υλοποίηση του Αλγορίθμου 7 στο Mathematica:

```
Import["LeukemiaData.txt"];
data = ReadList["LeukemiaData.txt", Number];
A = Partition[data, 38];
n1 = 27; n2 = 11;
A1 = Table[0, {i, 1, 2109}, {j, 1, n1}];
A2 = Table[0, {i, 1, 2109}, {j, 1, n2}];
Do[A1[[i, j]] = A[[i, j]], {i, 1, 2109}, {j, 1, n1}];
Do[A2[[i, j]] = A[[i, j + 27]], {i, 1, 2109}, {j, 1, n2}];

<< Statistics`DescriptiveStatistics`
k = 10000;
X = Table[0, {i, 1, 2109}, {j, 1, n1}];
Y = Table[0, {i, 1, 2109}, {j, 1, n2}];
i = 1; P = {};
Do[s1 = 0;
  sp = ((n1 - 1) Variance[A1[[i]]] + (n2 - 1) Variance[A2[[i]]]) / (n1 + n2 - 2);
  t = (Mean[A1[[i]]] - Mean[A2[[i]]]) / Sqrt[sp (1 / n1 + 1 / n2)];
  Do[
   Do[Z1 = Floor[38 * Random[]] + 1; X[[i, j]] = A[[i, Z1]], {j, 1, n1}];
   Do[Z2 = Floor[38 * Random[]] + 1; Y[[i, j]] = A[[i, Z2]], {j, 1, n2}];
   SP = ((n1 - 1) Variance[X[[i]]] + (n2 - 1) Variance[Y[[i]]]) / (n1 + n2 - 2);
   T = (Mean[X[[i]]] - Mean[Y[[i]]]) / Sqrt[SP (1 / n1 + 1 / n2)];
   If[T >= t, s1 = s1 + 1];
   , {k}];
  AppendTo [P, N[s1 / k]];
  , {i, 1 , 2109}];
Print[P]
```

Στον πίνακα που ακολουθεί βλέπουμε για κάθε διαδικασία ποια πιθανότητα σφάλματος τύπου I ρυθμίζει, αν είναι step-down ή step-up κα τέλος τον αριθμό των υποθέσεων που απορρίπτει. Έτσι, για παράδειγμα, η διαδικασία του Holm, η οποία είναι μία step-down

διαδικασία που ρυθμίζει την FWER, απορρίπτει 87 υποθέσεις, Οι υποθέσεις αυτές αντιστοιχούν στα γονίδια που εμφάνισαν τις 87 μικρότερες ακατέργαστες τιμές-*p*.

Διαδικασία	Πιθανότητα Σφάλματος Τύπου Ι	Μέθοδος	Αριθμός Διαφορετικά Εκφρασμένων Γονιδίων
Holm (1979)	FWER	Step-down	87
Sidak	FWER	Step-down	87
Hochberg (1988)	FWER	Step-up	87
Benjamini- Hochberg (1995)	FDR	Step-up	343
Benjamini- Liu (1999)	FDR	Step-down	87
Benjamini- Yekutieli (2001)	FDR	Step-up	158

Από τα αποτελέσματα του πίνακα μπορούμε να διαπιστώσουμε ότι οι step-down διαδικασίες οδηγούν σε λιγότερες απορρίψεις από ότι οι step-up. Επίσης, οι διαδικασίες που ρυθμίζουν την FWER φαίνεται ότι είναι πιο συντηρητικές από αυτές που ρυθμίζουν την FDR. Αυτό είναι φυσικό αφού στο Κεφάλαιο 2 είχαμε αποδείξει ότι $FDR \leq FWER$.

3.3.2 Πρόγραμμα R - Πακέτο multtest

Το πρόγραμμα R αποτελεί μία γλώσσα προγραμματισμού και ένα περιβάλλον εργασίας για στατιστικούς υπολογισμούς και γραφικά, τα οποία είναι παρόμοια με τη γλώσσα S και το περιβάλλον του προγράμματος S-Plus. Παρέχει μια μεγάλη ποικιλία στατιστικών τεχνικών (γραμμική και μη γραμμική μοντελοποίηση, στατιστικούς ελέγχους, time-series ανάλυση, ταξινόμηση, ομαδοποίηση κ.α.) και γραφικών τεχνικών, αλλά συγχρόνως περιέχει κάποια ειδικά πακέτα για την ανάλυση δεδομένων μικροσυστοιχιών DNA, με κύρια πηγή τον Bioconductor. Ο Bioconductor είναι μια ανοικτή πηγή και ένα ανοικτό εξελισσόμενο λογισμικό για την ανάλυση και την κατανόηση των γονιδιακών δεδομένων.

Ορισμένες από τις διαδικασίες πολλαπλού ελέγχου που περιγράψαμε μπορούν να εφαρμοστούν μέσω του πακέτου multtest του προγράμματος R, το οποίο βρίσκεται στον Bioconductor. Το εύρος εφαρμογής του είναι αρκετά μεγάλο και για αυτόν τον λόγο θα εστιάσουμε σε διαδικασίες πολλαπλών υποθέσεων που χρησιμοποιούν την bootstrap

εκτίμηση η οποία είναι διαθέσιμη μέσω της βασικής συνάρτησης του πακέτου, **MTP**. Διαδικασίες πολλαπλών ελέγχων βασισμένες στις μεταθέσεις είναι επίσης διαθέσιμες. Συγκεκριμένα, οι step-down maxT και step-up minP διαδικασίες για τη ρύθμιση της FWER μπορούν να εφαρμοστούν μέσω των συναρτήσεων **mt.maxT** και **mt.minP** αντίστοιχα, αλλά και μέσω της συνάρτησης **MTP**.

Χρησιμοποιώντας τη συνάρτηση **MTP** πρέπει κανείς να καθορίσει αρχικά τα δεδομένα $X_1,...,X_n$, ένα κατάλληλο στατιστικό T_n, για τις μηδενικές υποθέσεις που εξετάζονται, την πιθανότητα σφάλματος τύπου I και την κατάλληλη από κοινού κατανομή υπό τη μηδενική υπόθεση F_0 (ή την εκτίμηση αυτής, F_{0n}). Δεδομένων των παραπάνω, η διαδικασία πολλαπλού ελέγχου $S_n = S(T_n, F_{0n}, \alpha)$, ρυθμίζει την πιθανότητα σφάλματος σε επίπεδο a. Σε αναλογία με τα παραπάνω, το πακέτο multtest έχει υιοθετήσει μία εκτεταμένη προσέγγιση για την ερμηνεία των πολλαπλών ελέγχων υποθέσεων, με τις παρακάτω τέσσερις κύριες συναρτήσεις.

- Συναρτήσεις για τον υπολογισμό των στατιστικών, T_n. Υπάρχουν *εσωτερικές συναρτήσεις*, που δεν καλούνται κατευθείαν από το χρήστη, αλλά ορίζονται μέσα από τη συνάρτηση **MTP.** Υπάρχει επίσης η δυνατότητα ο χρήστης να προσθέσει μόνος του κάποιες συναρτήσεις μέσα στη βιβλιοθήκη (library) αυτών των εσωτερικών συναρτήσεων.
- Συναρτήσεις που παρέχουν την από κοινού κατανομή F_0 (ή την εκτίμηση αυτής F_{0n}).
- Συναρτήσεις για την εφαρμογή της διαδικασίας πολλαπλού ελέγχου $S_n = S(T_n, F_{0n}, \alpha)$ με στόχο τον καθορισμό περιοχών απόρριψης, διαστημάτων εμπιστοσύνης και προσαρμοσμένων τιμών-*p*. Η κύρια συνάρτηση είναι η **MTP**, η οποία εφαρμόζει τις single-step και step-down maxT και minP διαδικασίες για τη ρύθμιση της FWER. Μέσω των προσαρμοσμένων τιμών που προκύπτουν από τη ρύθμιση της FWER προκύπτει και η ρύθμιση της FDR χρησιμοποιώντας το κατάλληλο όρισμα στη συνάρτηση **MTP**.
- Συναρτήσεις για αριθμητικές και γραφικές απεικονίσεις της διαδικασίας πολλαπλού ελέγχου.

Για να καλέσουμε τη συνάρτηση **MTP** πρέπει αρχικά να εισαχθούμε στον Bioconductor, να ανοίξουμε το πακέτο multtest εκτελώντας τις παρακάτω εντολές.

```
> library(Biobase)
Loading required package: tools
Welcome to Bioconductor
    Vignettes contain introductory material.
    To view, simply type 'openVignette()' or start with 'help(Biobase)'.
    For details on reading vignettes, see the openVignette help page.

> library(multtest)

> args(MTP)
function (X, W = NULL, Y = NULL, Z = NULL, Z.incl = NULL, Z.test = NULL,
    na.rm = TRUE, test = "t.twosamp.unequalvar", robust = FALSE,
    standardize = TRUE, alternative = "two.sided", psi0 = 0,
    typeone = "fwer", k = 0, q = 0.1, fdr.method = "conservative",
    alpha = 0.05, smooth.null = FALSE, nulldist = "boot", B = 1000,
    method = "ss.maxT", get.cr = FALSE, get.cutoff = FALSE, get.adjp = TRUE,
    keep.nulldist = TRUE, seed = NULL)
NULL
```

Όπως φαίνεται και από το output η συνάρτηση **MTP** παίρνει τα παρακάτω ορίσματα.

1. *Δεδομένα.*

Τα δεδομένα X αποτελούνται από ένα *m*-διάστατο τυχαίο διάνυσμα, το οποίο παίρνει τιμές για τις *n* πειραματικές μονάδες και αποθηκεύονται σε έναν $m \times n$ πίνακα. Αν υπάρχει κάποιο διάνυσμα βαρών διάστασης *m* που σχετίζεται με κάθε παρατήρηση, αποθηκεύεται σε έναν $m \times n$ πίνακα W. Το όρισμα Y είναι ένα διάνυσμα ή παράγοντας που περιέχει ένα αποτέλεσμα ενδιαφέροντος. Αυτό μπορεί να είναι ο προσδιορισμός της ομάδας, μία συνεχής ή πολυδιάστατη εξαρτημένη μεταβλητή ή δεδομένα επιβίωσης. Σε μερικές μελέτες, *L* επιπλέον μεταβλητές που χρησιμοποιούνται σε μοντέλα παλινδρόμησης, μπορεί να μετρηθούν για κάθε πειραματική μονάδα και να αποθηκευτούν σε ένα $n \times L$ πίνακα Z. Τέλος τα Z.incl και Z.test είναι οι δείκτες ή τα ονόματα των στηλών του Z που περιλαμβάνονται στο μοντέλο. Από αυτά τα πέντε πρώτα ορίσματα μόνο ο πίνακας X των δεδομένων είναι απαραίτητο να ορίζεται πάντα, τα υπόλοιπα είναι εξ ορισμού (by default) μηδενικά (NULL). Τέλος το όρισμα na.rm δηλώνει αν θα αφαιρεθούν οι παρατηρήσεις που δεν είναι διαθέσιμες (NA- not available). Εξ ορισμού είναι NULL.

2. *Στατιστικά.*

Το στατιστικό επιλέγεται από το όρισμα test. Η default τιμή του είναι το **t.twosamp.unequalvar,** δηλαδή το two-sample στατιστικό του Welch. Υπάρχει η δυνατότητα εισαγωγής και άλλων συναρτήσεων όπως το one-sample στατιστικό για τον έλεγχο των μέσων τιμών (**t.onesamp**), το two-sample στατιστικό για τον έλεγχο της διαφοράς των μέσων και ίσες διακυμάνσεις (**t.onesamp. unequalvar**), το στατιστικό F για τον έλεγχο της ισότητας των πληθυσμιακών μέσων σε *n* δείγματα (**f**) και άλλα.

3. *Πιθανότητες σφαλμάτων τύπου Ι.*

Η συνάρτηση **MTP** ρυθμίζει εξ ορισμού την FWER (typeone = "fwer"). Παρέχεται επίσης η ρύθμιση της FDR και άλλων πιθανοτήτων σφαλμάτων μέσω του ορίσματος typeone. Τέλος το επίπεδο του ελέγχου καθορίζεται από το όρισμα alpha το οποίο εξ ορισμού είναι ίσο με 0.05.

4. *Κατανομή υπό τη μηδενική Υπόθεση*.

Από το όρισμα nulldist καθορίζεται η μέθοδος εκτίμησης της κατανομής υπό τη μηδενική υπόθεση. Η μέθοδος bootstrap είναι η default επιλογή (nulldist = "boot"), αλλά υπάρχει η δυνατότητα χρήσης της μεθόδου των μεταθέσεων (nulldist = "perm"). Το πλήθος των μεταθέσεων καθορίζεται από το B του οποίου η default τιμή του είναι ίση με 1000.

5. *Διαδικασία πολλαπλών ελέγχων*.

Για τη ρύθμιση της FWER είναι δυνατόν να εφαρμοστούν οι single-step και step-down maxT και minP διαδικασίες πολλαπλού ελέγχου, μέσω του ορίσματος method. Η default τιμή είναι single-step maxT (method = "ss.maxT"). Για τη ρύθμιση της FDR, υπάρχει η δυνατότητα χρήσης της συνάρτησης **fwer2fdr** η οποία παίρνει τις FWER προσαρμοσμένες τιμές-*p* και επιστρέφει τις αντίστοιχες προσαρμοσμένες τιμές-*p* για τη ρύθμιση της FDR.

6. *Output*.

Τα υπόλοιπα ορίσματα της συνάρτησης αφορούν στα αποτελέσματα που θέλουμε να πάρουμε κάθε φορά. Για τις προσαρμοσμένες τιμές-*p* κρατάμε την εξ ορισμού επιλογή, get.adjp = TRUE.

Σημειώνεται ότι το πακέτο multtest παρέχει και άλλες διαδικασίες πολλαπλών υποθέσεων για τη ρύθμιση της FWER, όπως η Bonferroni, η Holm (1979), η Hochberg (1988), Sidak (1967), αλλά και για τη ρύθμιση της FDR, όπως η μέθοδος των Benjamini και Hochberg

(1995) και των Benjamini και Yekutieli (2001). Οι διαδικασίες αυτές πραγματοποιούνται μέσω της συνάρτησης **mt.rawp2adjp** η οποία δέχεται ένα διάνυσμα μη προσαρμοσμένων τιμών-*p* και επιστρέφει τις αντίστοιχες προσαρμοσμένες τιμές-*p*.

Για την ανάλυση των δεδομένων της λευχαιμίας θα τρέξουμε τη συνάρτηση MTP ορίζοντας τις διαδικασίες step-down minP και step-down maxT. Παρακάτω φαίνονται οι εντολές που χρησιμοποιούμε στο R, τα χαρακτηριστικά και τα αποτελέσματα που προκύπτουν για κάθε μία διαδικασία.

Α) Για την step-down minP διαδικασία:

```
> filepath<-system.file("data","LeukemiaData.txt",package="datasets")
> filepath
[1] "C:/PROGRA~1/R/R-23~1.0/library/datasets/data/LeukemiaData.txt"
> data<-read.table(filepath)
> group<-c(rep(1,27),rep(0,11))
> m1<-MTP(X=data,Y=group,standardize=FALSE,B=1000,method="sd.minP")
> print(m1)
      Multiple Testing Procedure

Object of class:  MTP
sample size = 38
number of hypotheses = 2109

test statistics = t.twosamp.unequalvar
type I error rate = fwer
nominal level alpha = 0.05
multiple testing procedure = sd.minP

Call: MTP(X = data, Y = group, standardize = FALSE, B = 1000, method = "sd.minP")

> summary(m1)
MTP:  sd.minP
Type I error rate:  fwer

  Level   Rejections
1  0.05          375

             Min.  1st Qu.  Median     Mean      3rd Qu.  Max.
adjp        0.000  0.6060  1.00000  7.566e-01  1.0000   1.000
rawp        0.000  0.0030  0.08600  2.423e-01  0.4450   1.000
statistic  -2.948 -0.2925  0.03332  3.932e-06  0.3384   2.198
estimate   -2.948 -0.2925  0.03332  3.932e-06  0.3384   2.198

 > slotNames("MTP")
 [1] "statistic" "estimate"  "sampsize"   "rawp"      "adjp"      "conf.reg"  "cutoff"    "reject"
```

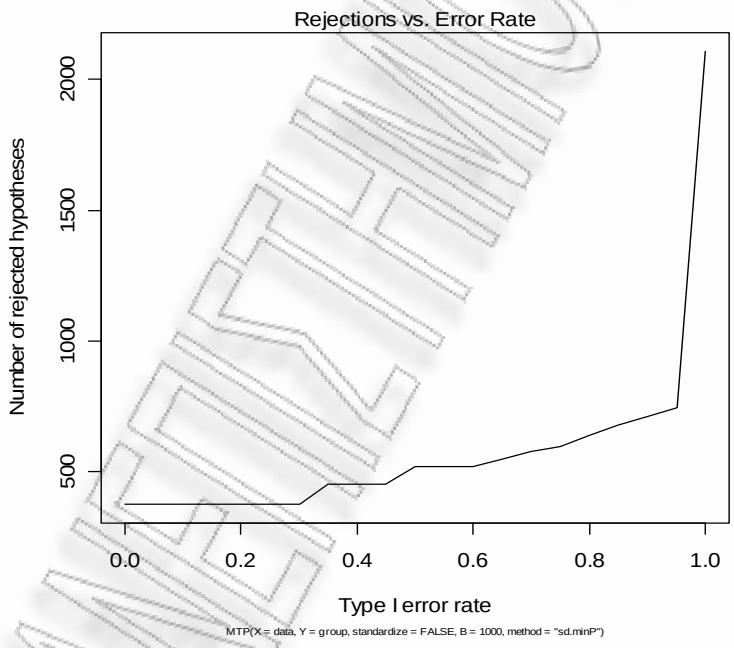
```
 [9] "nulldist"  "call"      "seed"
> slot(m1,"reject")
> plot(m1)
```

Βλέπουμε ότι η step-down minP μέθοδος για τη ρύθμιση της FWER απορρίπτει 375 υποθέσεις.

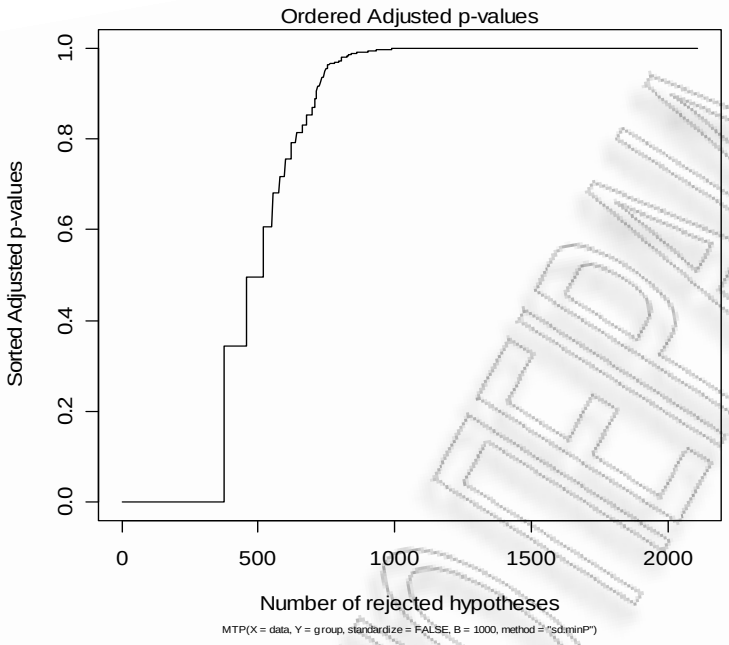
gene2	gene197	gene363	gene580	gene756	gene990	gene1175	gene1400	gene1623	gene1914
gene7	gene205	gene369	gene590	gene767	gene995	gene1194	gene1408	gene1629	gene1922
gene14	gene210	gene370	gene598	gene770	gene996	gene1201	gene1409	gene1635	gene1929
gene22	gene216	gene376	gene604	gene775	gene1002	gene1204	gene1415	gene1639	gene1935
gene23	gene218	gene377	gene605	gene776	gene1003	gene1211	gene1432	gene1640	gene1944
gene24	gene222	gene381	gene608	gene783	gene1004	gene1222	gene1452	gene1655	gene1951
gene25	gene225	gene390	gene614	gene797	gene1014	gene1224	gene1453	gene1664	gene1957
gene30	gene226	gene394	gene615	gene798	gene1038	gene1226	gene1454	gene1667	gene1958
gene31	gene228	gene398	gene617	gene810	gene1039	gene1227	gene1461	gene1669	gene1966
gene34	gene230	gene407	gene622	gene811	gene1045	gene1230	gene1464	gene1675	gene1969
gene48	gene231	gene420	gene624	gene835	gene1048	gene1232	gene1468	gene1687	gene1976
gene52	gene232	gene421	gene626	gene836	gene1051	gene1233	gene1470	gene1688	gene1981
gene56	gene245	gene422	gene633	gene843	gene1054	gene1236	gene1473	gene1693	gene1985
gene64	gene248	gene433	gene636	gene864	gene1056	gene1247	gene1482	gene1708	gene1986
gene73	gene253	gene440	gene637	gene867	gene1059	gene1256	gene1490	gene1709	gene2001
gene78	gene263	gene441	gene644	gene869	gene1076	gene1275	gene1517	gene1712	gene2002
gene79	gene264	gene451	gene645	gene870	gene1077	gene1286	gene1518	gene1719	gene2007
gene86	gene266	gene465	gene651	gene880	gene1080	gene1297	gene1521	gene1733	gene2011
gene91	gene276	gene474	gene655	gene882	gene1084	gene1298	gene1523	gene1746	gene2016
gene94	gene278	gene475	gene661	gene887	gene1085	gene1305	gene1529	gene1754	gene2026
gene96	gene280	gene488	gene665	gene891	gene1092	gene1306	gene1535	gene1757	gene2027
gene102	gene282	gene494	gene667	gene904	gene1096	gene1308	gene1543	gene1767	gene2031
gene108	gene288	gene497	gene669	gene911	gene1110	gene1310	gene1544	gene1769	gene2034
gene114	gene300	gene510	gene674	gene921	gene1114	gene1314	gene1546	gene1800	gene2038
gene127	gene313	gene520	gene677	gene926	gene1115	gene1315	gene1553	gene1805	gene2045
gene130	gene314	gene522	gene683	gene930	gene1118	gene1316	gene1555	gene1817	gene2054
gene136	gene316	gene523	gene686	gene935	gene1119	gene1325	gene1557	gene1820	gene2062
gene141	gene327	gene524	gene690	gene950	gene1123	gene1328	gene1568	gene1827	gene2063
gene151	gene328	gene529	gene691	gene955	gene1129	gene1331	gene1574	gene1833	gene2066
gene153	gene331	gene530	gene693	gene959	gene1130	gene1343	gene1575	gene1854	gene2080
gene158	gene334	gene533	gene704	gene960	gene1143	gene1349	gene1579	gene1855	gene2089
gene160	gene336	gene534	gene711	gene962	gene1145	gene1351	gene1581	gene1858	gene2101
gene170	gene340	gene537	gene713	gene963	gene1148	gene1353	gene1583	gene1861	gene2107

gene174	gene345	gene543	gene722	gene971	gene1154	gene1354	gene1587	gene1865	
gene178	gene347	gene548	gene723	gene974	gene1159	gene1361	gene1596	gene1866	
gene184	gene349	gene558	gene727	gene979	gene1167	gene1375	gene1599	gene1878	
gene187	gene351	gene561	gene741	gene982	gene1172	gene1382	gene1614	gene1894	
gene190	gene355	gene574	gene746	gene985	gene1174	gene1396	gene1621	gene1907	

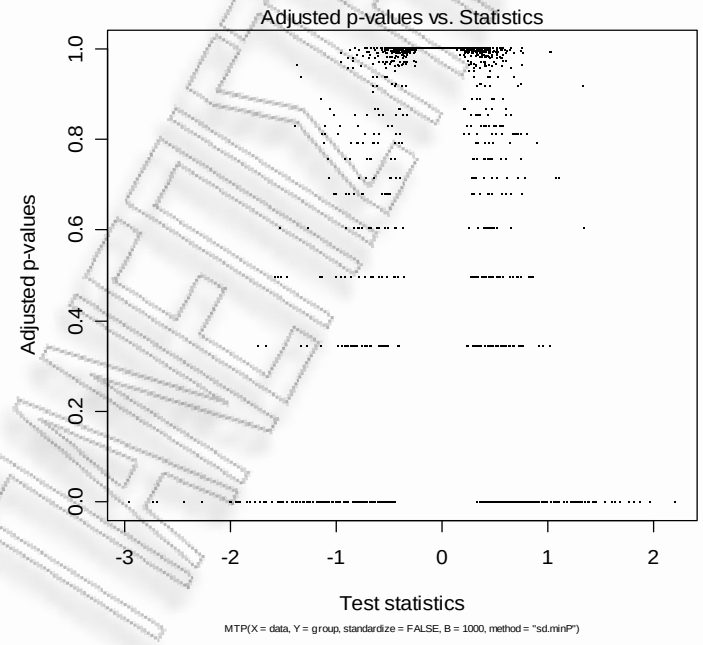
Η συνάρτηση MTP μας παρέχει και τα παρακάτω γραφήματα. Στο σχήμα 3.1 φαίνεται ο αριθμός των υποθέσεων που απορρίπτονται για διάφορες τιμές του σφάλματος τύπου I. Στο σχήμα 3.2 ο αριθμός των υποθέσεων που απορρίπτονται ως προς τις διατεταγμένες προσαρμοσμένες τιμές-*p*. Στο σχήμα 3.3 οι προσαρμοσμένες τιμές-*p* ως προς τις τιμές του στατιστικού (two-sample στατιστικό του Welch) και στο σχήμα 3.4 οι τιμές των προσαρμοσμένων τιμών-*p* για τα 2109 γονίδια.



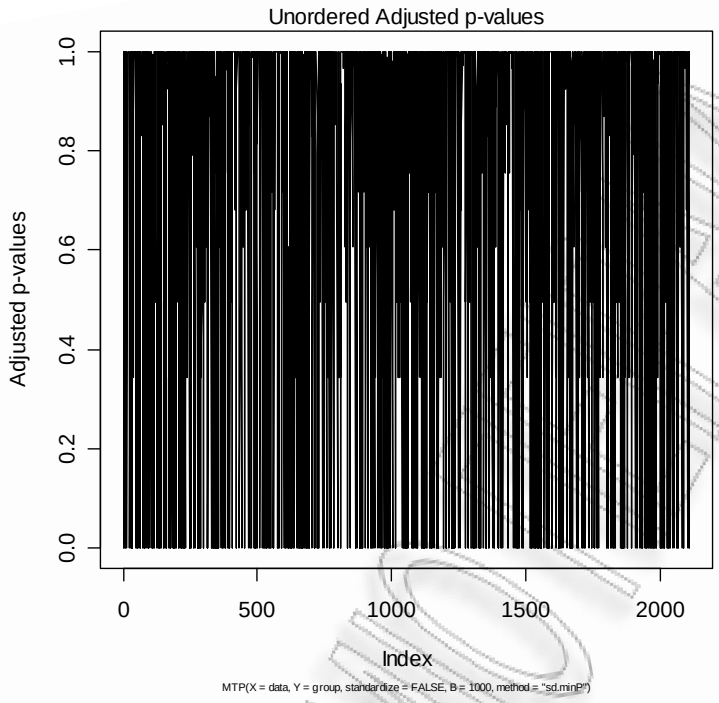
Σχήμα 3.1. Αριθμός των υποθέσεων που απορρίπτονται για διάφορες τιμές του σφάλματος τύπου I (μέθοδος step-down minP).



Σχήμα 3.2. Διατεταγμένες προσαρμοσμένες τιμές-*p* ως προς τον αριθμό των υποθέσεων που απορρίπτονται (μέθοδος step-down minP).



Σχήμα 3.3. Προσαρμοσμένες τιμές-*p* ως προς τις τιμές του στατιστικού (μέθοδος step-down minP).



Σχήμα 3.4. Τιμές των προσαρμοσμένων τιμών-*p* για τα 2109 γονίδια (μέθοδος step-down minP).

Στη συνέχεια εφαρμόζουμε τη συνάρτηση MTP αλλάζοντας τη μέθοδο step-down minP σε step-down maxT. Τα υπόλοιπα ορίσματα παραμένουν ίδια.

```
m2<-MTP(X=data,Y=group,standardize=FALSE,B=1000,method="sd.maxT")
print(m2)

      Multiple Testing Procedure

Object of class:  MTP
sample size = 38
number of hypotheses = 2109

test statistics = t.twosamp.unequalvar
type I error rate = fwer
nominal level alpha = 0.05
multiple testing procedure = sd.maxT

Call: MTP(X = data, Y = group, standardize = FALSE, B = 1000, method = "sd.maxT")

> summary(m2)
MTP:  sd.maxT
Type I error rate:  fwer

  Level Rejections
1  0.05         32
```

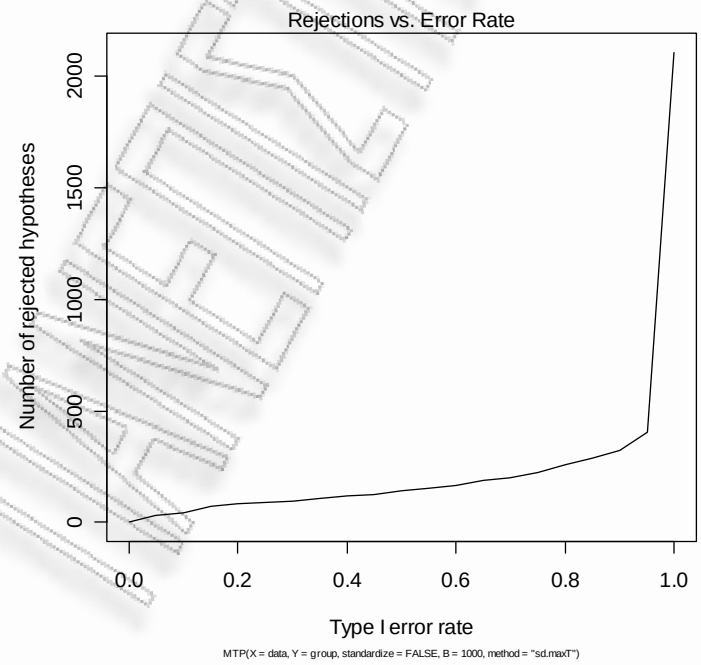
```
            Min. 1st Qu.  Median      Mean 3rd Qu.  Max.
adjp       0.000  0.9900 1.00000 9.198e-01  1.0000 1.000
rawp       0.000  0.0030 0.08600 2.423e-01  0.4450 1.000
statistic -2.948 -0.2925 0.03332 3.932e-06  0.3384 2.198
estimate  -2.948 -0.2925 0.03332 3.932e-06  0.3384 2.198

> slot(m2,"reject")
> plot(m2)
```

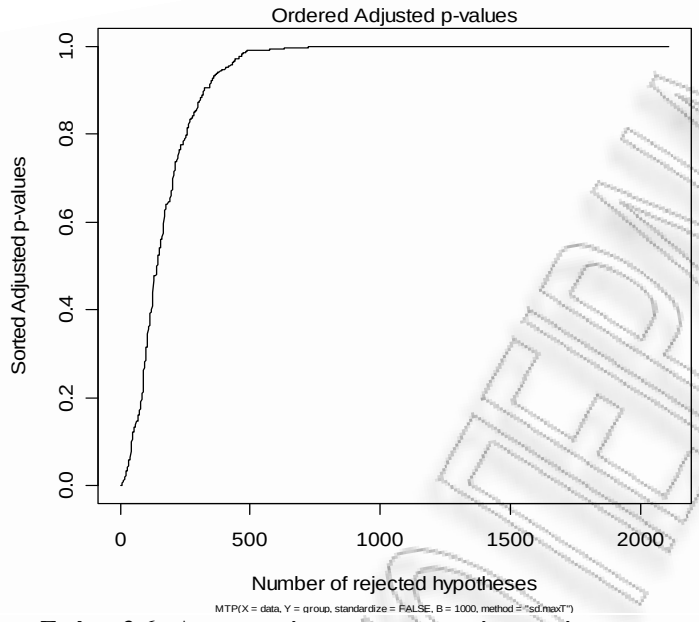
Η step-down maxT μέθοδος για τη ρύθμιση της FWER βλέπουμε ότι απορρίπτει 32 υποθέσεις. Είναι πολύ πιο συντηρητική από τη step-down minP. (μέθοδος step-down maxT)

gene64	gene153	gene278	gene451	gene655	gene1123	gene1474	gene1746
gene91	gene197	gene376	gene534	gene880	gene1286	gene1517	gene2038
gene96	gene225	gene390	gene580	gene935	gene1298	gene1655	gene2045
gene136	gene276	gene434	gene626	gene1004	gene1375	gene1668	gene2062

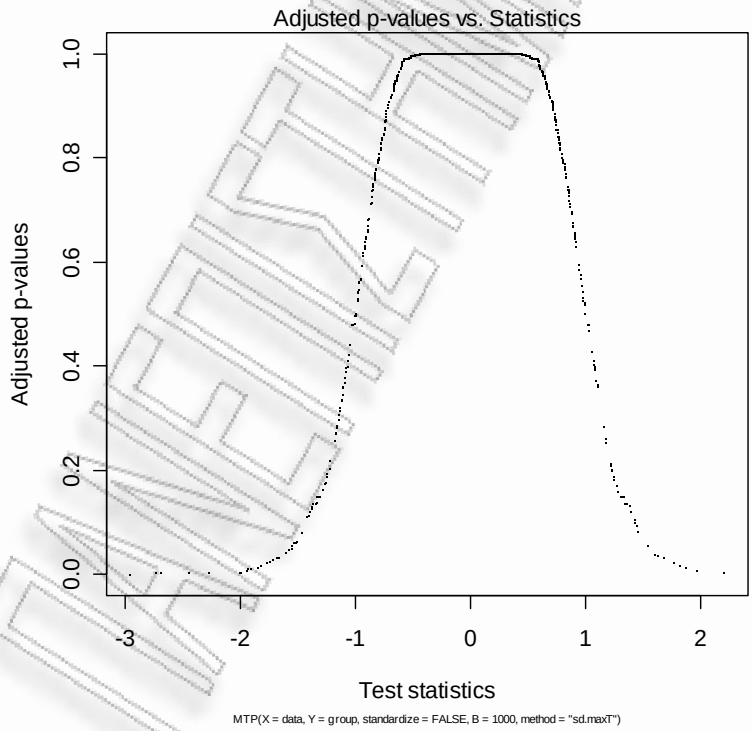
Από αυτά τα 32 γονίδια τα 29 εμφανίζονται ως διαφορετικά εκφρασμένα και με τη step-down minP διαδικασία. Ακολουθούν τα αντίστοιχα διαγράμματα.



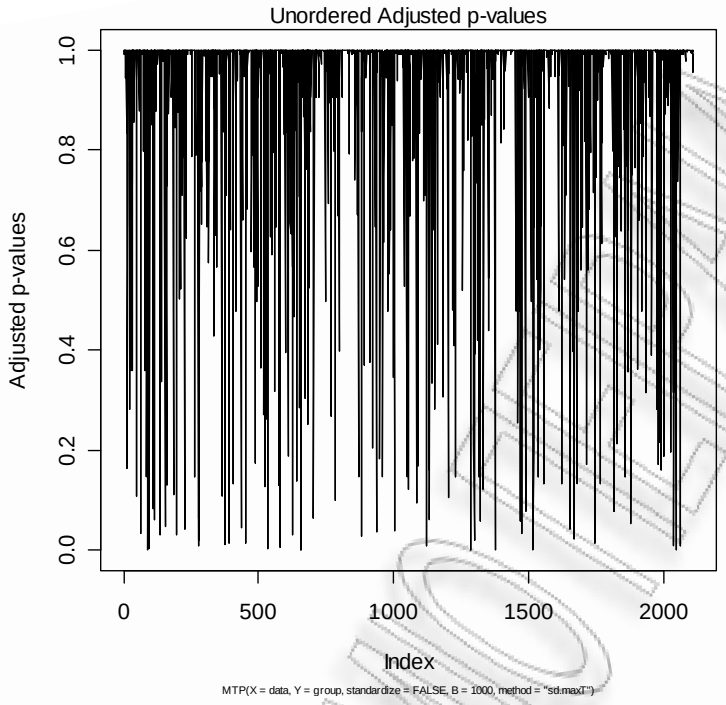
Σχήμα 3.5. Αριθμός των υποθέσεων που απορρίπτονται για διάφορες τιμές του σφάλματος τύπου I(μέθοδος step-down maxT).



Σχήμα 3.6. Διατεταγμένες προσαρμοσμένες τιμές-*p* ως προς τον αριθμό των υποθέσεων που απορρίπτονται (μέθοδος step-down maxT).



Σχήμα 3.7. Προσαρμοσμένες τιμές-*p* ως προς τις τιμές του στατιστικού (μέθοδος step-down maxT).

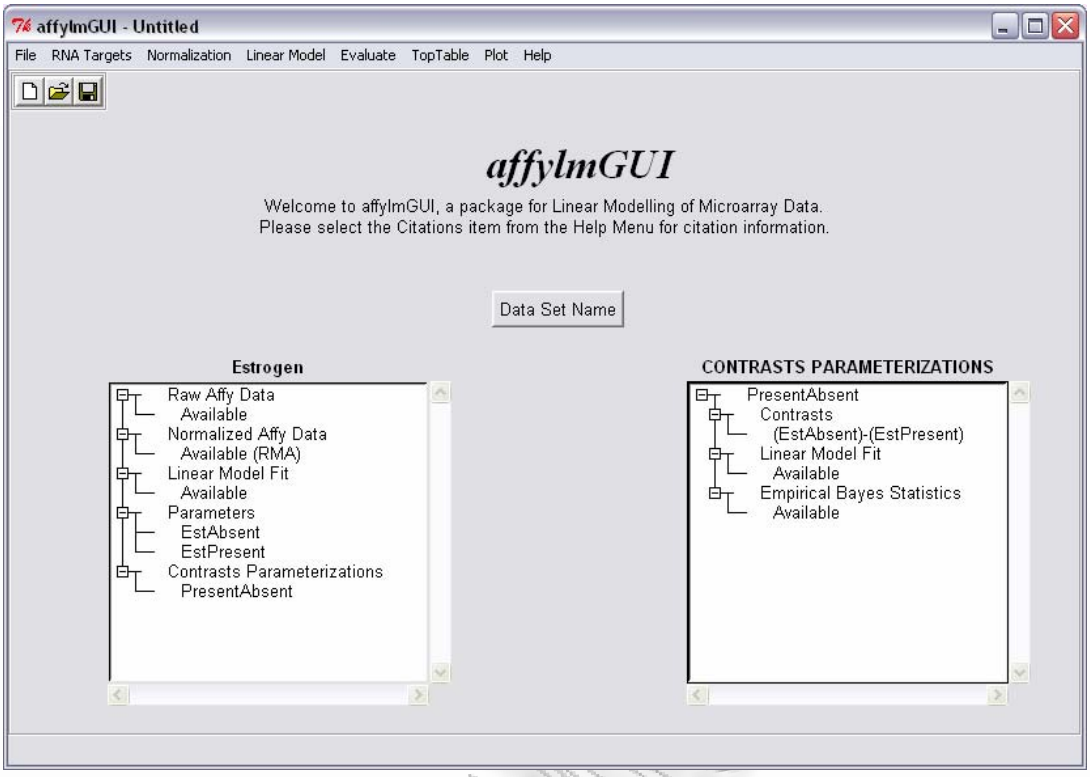


Σχήμα 3.8. Τιμές των προσαρμοσμένων τιμών-*p* για τα 2109 γονίδια (μέθοδος step-down maxT).

3.3.2 Πρόγραμμα R - Πακέτο affylmGUI

Το affylmGUI (Affymetrix Linear Modeling Graphical User Interface) είναι ένα πακέτο του προγράμματος R ειδικά σχεδιασμένο για την ανάλυση των Affymetrix δεδομένων μικροσυστοιχιών. Για την εκτέλεση των αναλύσεων χρησιμοποιεί το πακέτο limma (Linear Models for Microarray Data) το οποίο περιέχει τεχνικές για τη γραμμική μοντελοποίηση των δεδομένων και την αναγνώριση των διαφορετικά εκφρασμένων γονιδίων. Οι δυνατότητες που παρέχει στο χρήστη είναι εκτελέσιμες μέσω του κεντρικού παραθύρου εντολών του affylmGUI πακέτου το οποίο φαίνεται παρακάτω.

Για την παρουσίαση της λειτουργίας του affylmGUI θα χρησιμοποιήσουμε τα δεδομένα για τον καρκίνο του μαστού. Όπως περιγράψαμε και στο Κεφάλαιο 1, τα δεδομένα αυτά αντιστοιχούν σε ένα 2×2 παραγοντικό σχεδιασμό για κύτταρα που έχουν προσβληθεί από καρκίνο του μαστού. Οι παράγοντες του πειράματος είναι το οιστρογόνο (παρουσία ή απουσία) και ο χρόνος έκθεσης του κάθε δείγματος (10 ή 48 ώρες).



Το affylmGUI μας παρέχει τη δυνατότητα να πραγματοποιήσουμε την ανάλυση των δεδομένων για όλες τις δυνατές αντιθέσεις. Παρόλα αυτά, για την απλούστευση της ανάλυσης και δεδομένου ότι ενδιαφερόμαστε για τη διερεύνηση των διαδικασιών ελέγχου πολλαπλών υποθέσεων, θα αγνοήσουμε την επίδραση του χρόνου και θα εστιάσουμε στη μέτρηση των διαφορών στην έκφραση των γονιδίων εξαιτίας της παρουσίας (ή απουσίας) του οιστρογόνου. Θα θεωρήσουμε δηλαδή δύο ομάδες δεδομένων αποτελούμενες από 4 δείγματα η καθεμία.

Ξεκινώντας την ανάλυση στο affylmGUI, δημιουργούμε ένα ευρετήριο (working directory) για τα δεδομένα μας και αποθηκεύουμε εκεί τα κυτταρικά αρχεία, low10-1.cel, low10-2.cel, high10-1.cel, high10-2.cel, low48-1.cel, low48-2.cel, high48-1.cel και high48-2.cel. Τα αρχεία αυτά περιέχουν τις μετρήσεις έκφρασης των 12625 γονιδίων στα 8 δείγματα. Επίσης, αποθηκεύουμε στον ίδιο χώρο το αρχείο “RNA Targets”, το οποίο δημιουργείται είτε σε ένα λογιστικό φύλλο (όπως το Excel) είτε σε έναν επεξεργαστή κειμένου και έχει πάντα τη μορφή που βλέπουμε στον παρακάτω πίνακα.

Οι επικεφαλίδες των στηλών πρέπει να εμφανίζονται ακριβώς όπως φαίνονται στον πίνακα. Στη στήλη Name σε κάθε πλακίδιο αντιστοιχεί ένα και μόνο όνομα. Το Affymetrix κυτταρικό

αρχείο για κάθε πλακίδιο φαίνεται στη στήλη FileName, ενώ στη στήλη Target καθορίζει τα πλακίδια που έχουν αντιγραφεί.

Name	FileName	Target
Abs10.1	**low10-1.cel**	**EstAbsent**
Abs10.2	**low10-2.cel**	**EstAbsent**
Pres10.1	**high10-1.cel**	**EstPresent**
Pres10.2	**high10-2.cel**	**EstPresent**
Abs48.1	**low48-1.cel**	**EstAbsent**
Abs48.2	**low48-2.cel**	**EstAbsent**
Pres48.1	**high48-1.cel**	**EstPresent**
Pres48.2	**high48-2.cel**	**EstPresent**

Χρησιμοποιώντας λοιπόν το πακέτο affylmGUI για τα δεδομένα του καρκίνου του μαστού, ας δούμε αρχικά μερικά από τα γραφικά αποτελέσματα που μπορούμε να πάρουμε.

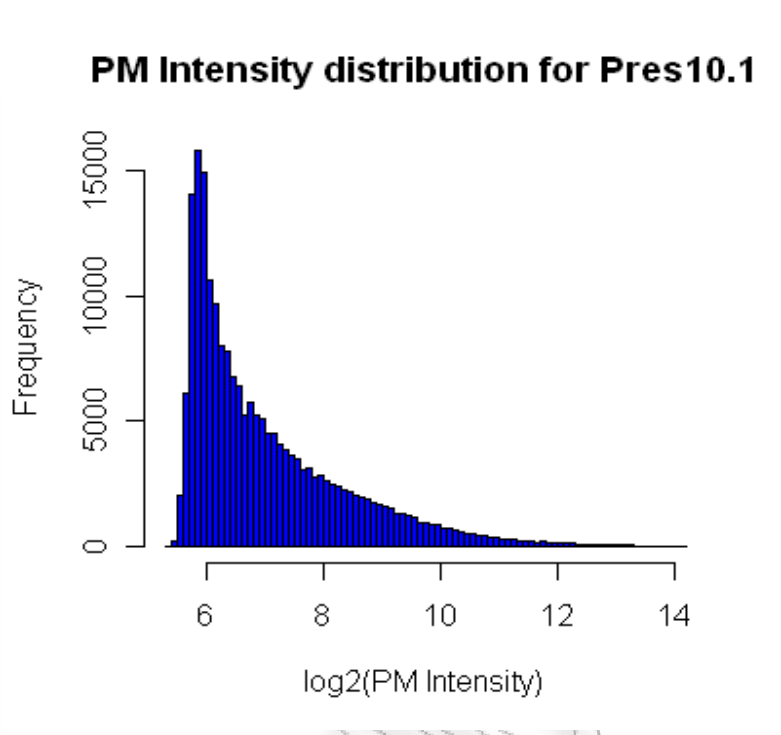
- Σχήμα 3.9. Είναι το διάγραμμα συχνοτήτων για τις μετρήσεις έκφρασης των γονιδίων στο δείγμα Pres10.1 το οποίο μας δίνει μία επισκόπηση της κατανομής των δεομένων.
- Σχήμα 3.10. Είναι το γράφημα στο οποίο φαίνονται τα θηκογράμματα για τα ακατέργαστα δεδομένα των 8 δειγμάτων τα οποία δείχνουν την κεντρική τάση και μεταβλητότητα των δεδομένων.
- Σχήμα 3.11. Είναι το γράφημα στο οποίο φαίνονται τα αντίστοιχα θηκογράμματα μετά την κανονικοποίηση των δεδομένων.
- Σχήμα 3.12. Ονομάζεται MA plot και δείχνει την εξαρτημένη αναλογία της έντασης των δεδομένων των μικροσυστοιχιών. Είναι ουσιαστικά ένα διάγραμμα διασποράς με μετασχηματισμένους άξονες. Ο άξονας x παριστάνει τη μέση ένταση των δύο χρωστικών (R, G),

$$\mathrm{A} = \frac{1}{2}\left(\log_2 R + \log_2 G\right)$$

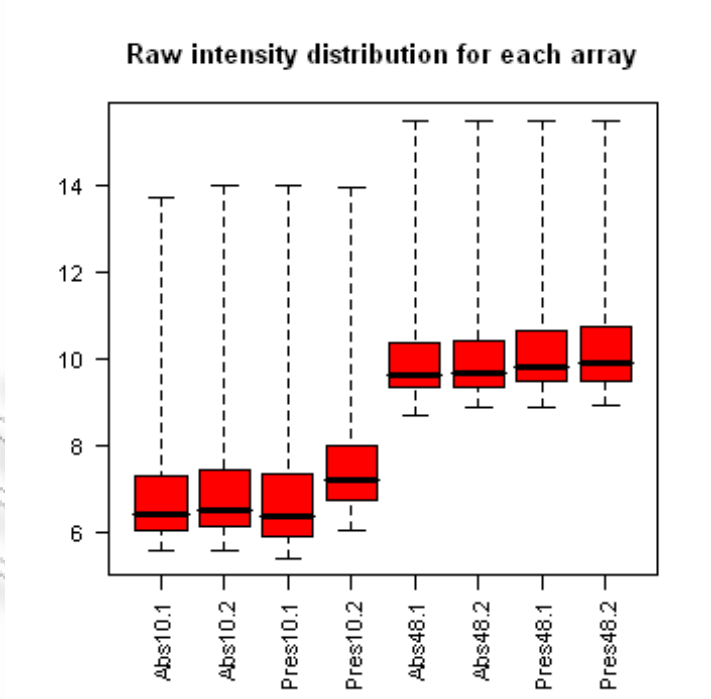
ενώ ο άξονας y την αναλογία μεταξύ των χρωστικών,

$$\mathrm{M} = \log_2 \frac{R}{G} = \log_2 R - \log_2 G\,.$$

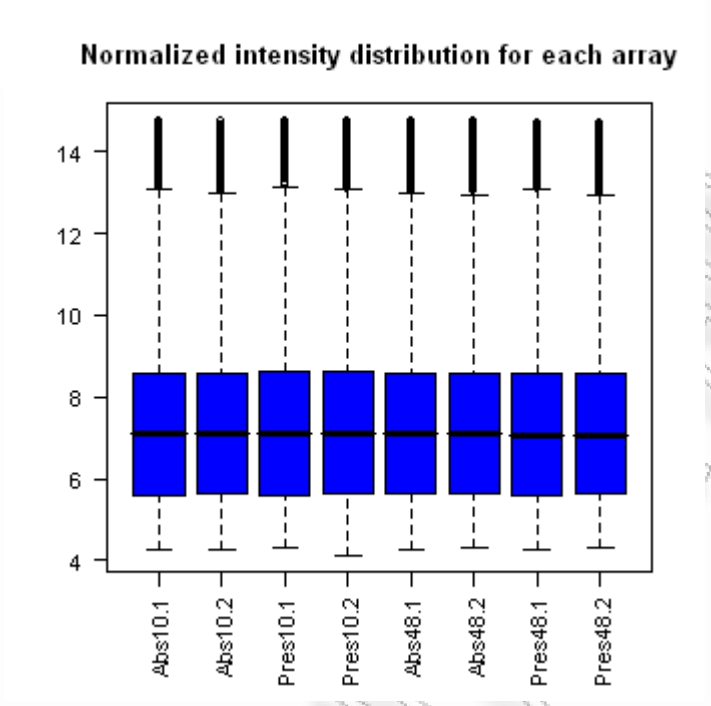
- Σχήμα 3.13. Είναι το QQ plot το οποίο συγκρίνει την κατανομή των δεδομένων με την κατανομή t. Το γράφημα αυτό δηλώνει έντονα τη διαφορετική έκφραση των γονιδίων, αφού υπάρχουν πολλά σημεία που αποκλίνουν από τη γραμμή με κλίση 1.



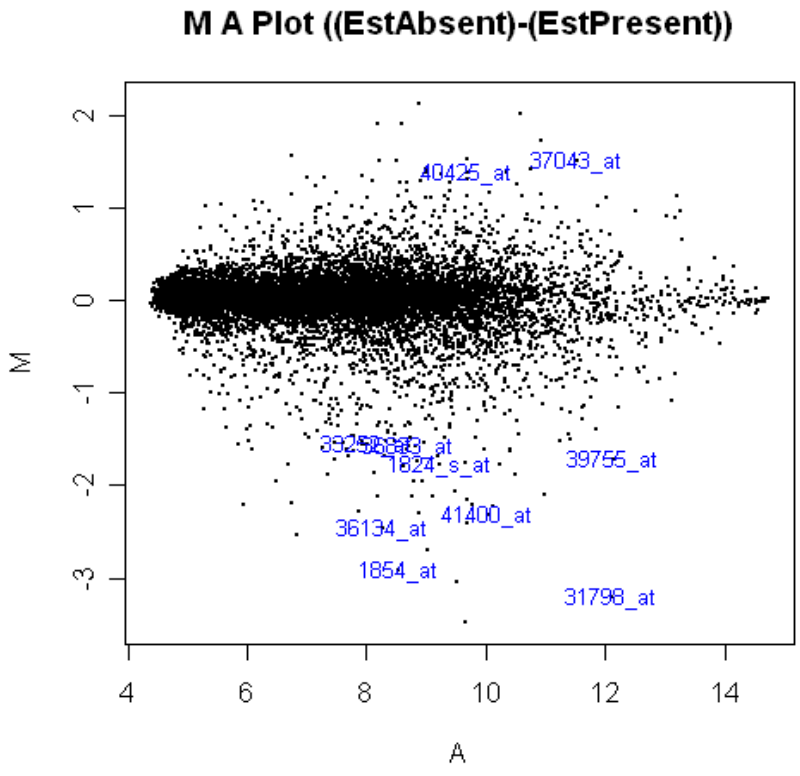
Σχήμα 3.9. Διάγραμμα συχνοτήτων.



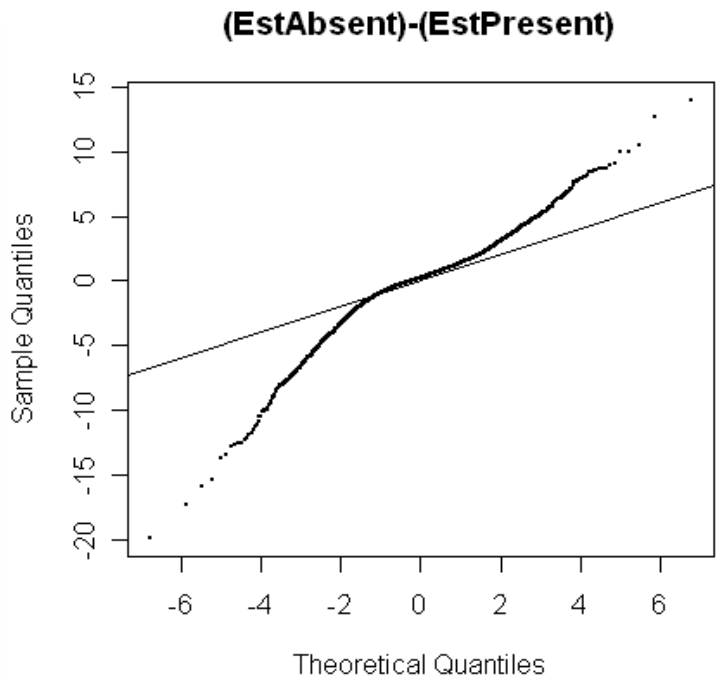
Σχήμα 3.10. Θηκογράμματα που αντιστοιχούν στα 8 δείγματα καρκίνου του μαστού.



Σχήμα 3.11. Θηκογράμματα μετά την κανονικοποίηση των δεδομένων.

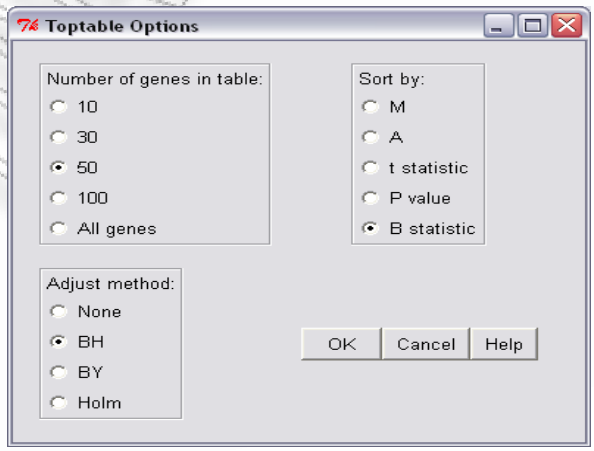


Σχήμα 3.12. Διάγραμμα διασποράς.



Σχήμα 3.13. Q-Q Plot. Δηλώνει έντονα τη διαφορετική έκφραση των γονιδίων

Στη συνέχεια παρουσιάζουμε τους πίνακες με τα 50 πρώτα υποψήφια γονίδια που είναι πιθανό να είναι διαφορετικά εκφρασμένα. Ουσιαστικά το affylmGUI δεν κάνει τον έλεγχο των υποθέσεων για κάποιο επίπεδο a, αλλά μας παρέχει τα 10, 30, 50, 100 ή όλα τα υποψήφια γονίδια τα οποία είναι δυνατόν να διαταχθούν με 5 τρόπους: ως προς τις τιμές M, τις τιμές A, τις προσαρμοσμένες τιμές-p, τις τιμές του στατιστικού t ή τις τιμές του στατιστικού B (log odds της διαφορετικής έκφρασης). Έχουμε επίσης τη δυνατότητα να χρησιμοποιήσουμε τρεις μεθόδους προσαρμογής των τιμών-p, όπως φαίνεται και στο αντίστοιχο παράθυρο του affylmGUI.



Αυτές είναι η διαδικασία του Holm (1979), των Benjamini και Hochberg (1995) και των Benjamini και Yekutieli (2001) και τα αποτελέσματα για κάθε μία διαδικασία φαίνονται στους πίνακες 3.1, 3.2, 3.3 αντίστοιχα.

ID	M	A	t	P.Value	B
41400_at	-2.31225	10.04155	-19.9521	0.000116	9.823686
1824_s_at	-1.77799	9.238868	-17.3131	0.000202	8.959766
39755_at	-1.7223	12.13182	-15.9005	0.00028	8.40337
36134_at	-2.4666	8.275738	-15.4105	0.00028	8.19219
37043_at	1.496959	11.52942	13.98329	0.000494	7.51622
36833_at	-1.58154	8.709889	-13.7742	0.000494	7.408872
1854_at	-2.92426	8.532097	-13.4443	0.000521	7.234845
33252_at	-1.55986	8.000345	-12.8708	0.000557	6.91802
31798_at	-3.19872	12.11577	-12.7541	0.000557	6.851221
40425_at	1.376222	9.69109	12.72879	0.000557	6.836617
1505_at	-2.12252	8.764746	-12.6261	0.000557	6.777006
480_at	-1.0624	10.13993	-12.5302	0.000557	6.72075
1515_at	-1.95606	8.946283	-12.3313	0.00058	6.602286
2042_s_at	-2.11266	8.178136	-12.2486	0.00058	6.55234
947_at	-2.1023	10.97252	-11.9121	0.000686	6.344429
33255_at	-1.40414	8.367932	-11.7722	0.000711	6.255751
1884_s_at	-2.69304	9.034773	-11.5777	0.000771	6.13026
35312_at	-2.22932	10.12458	-11.3175	0.000883	5.958293
673_at	-1.47642	9.828868	-11.2073	0.000909	5.883954
1803_at	-1.33586	8.72651	-10.9644	0.001039	5.717035
1462_s_at	-0.9343	9.341027	-10.5255	0.001333	5.403896
39542_at	1.164624	10.0553	10.52498	0.001333	5.403522
41583_at	-2.21467	9.771343	-10.4626	0.00134	5.357783
40117_at	-2.41576	9.676526	-10.1615	0.001584	5.132482
1474_s_at	-1.15505	6.033744	-10.1541	0.001584	5.126826
1476_s_at	-0.99127	8.471458	-9.99964	0.001588	5.008176
38422_s_at	-1.54015	7.474208	-9.98855	0.001588	4.999582
37294_at	1.900328	8.194959	9.955109	0.001588	4.973588
37576_at	-1.28122	6.918036	-9.94924	0.001588	4.969015
1687_s_at	1.376881	8.996093	9.930995	0.001588	4.954785
910_at	-3.48441	9.660231	-9.82262	0.001684	4.869651
1516_g_at	-2.14961	9.685739	-9.59752	0.001975	4.689453
36499_at	-1.09687	9.216045	-9.56253	0.001975	4.661029
1376_at	-1.29333	8.990026	-9.47572	0.002067	4.590018
39059_at	-0.94874	11.28948	-9.4047	0.002137	4.531399
38551_at	2.013988	10.59207	9.12614	0.002661	4.296839
36813_at	-1.78936	8.606042	-9.05064	0.002771	4.231964
982_at	-1.6087	7.677224	-9.00745	0.002799	4.194598

967_g_at	-1.00433	8.180096	-8.97876	0.002799	4.169665
31831_at	0.769261	9.499375	8.954107	0.002799	4.148185
685_f_at	-0.90668	8.484465	-8.85896	0.002979	4.064688
1005_at	0.960637	8.623236	8.724912	0.003237	3.94549
36317_at	0.833625	7.147444	8.718143	0.003237	3.939421
846_s_at	2.127526	8.883141	8.685102	0.003263	3.909731
31830_s_at	0.90929	10.29754	8.585732	0.003501	3.819744
1801_at	-1.87408	8.139036	-8.56299	0.003501	3.799
40767_at	1.900989	8.605031	8.520022	0.003569	3.759664
38944_at	0.838444	8.059626	8.481077	0.003617	3.723838
349_g_at	-1.35414	7.530401	-8.46243	0.003617	3.706628
41641_at	0.752957	9.686598	8.382601	0.003777	3.632515

Πίνακας 3.1.

ID	M	A	t	P.Value	B
41400_at	-2.31225	10.04155	-19.9521	0.001164	9.823686
1824_s_at	-1.77799	9.238868	-17.3131	0.002027	8.959766
39755_at	-1.7223	12.13182	-15.9005	0.002803	8.40337
36134_at	-2.4666	8.275738	-15.4105	0.002803	8.19219
37043_at	1.496959	11.52942	13.98329	0.004946	7.51622
36833_at	-1.58154	8.709889	-13.7742	0.004946	7.408872
1854_at	-2.92426	8.532097	-13.4443	0.005225	7.234845
33252_at	-1.55986	8.000345	-12.8708	0.005581	6.91802
31798_at	-3.19872	12.11577	-12.7541	0.005581	6.851221
40425_at	1.376222	9.69109	12.72879	0.005581	6.836617
1505_at	-2.12252	8.764746	-12.6261	0.005581	6.777006
480_at	-1.0624	10.13993	-12.5302	0.005581	6.72075
1515_at	-1.95606	8.946283	-12.3313	0.00581	6.602286
2042_s_at	-2.11266	8.178136	-12.2486	0.00581	6.55234
947_at	-2.1023	10.97252	-11.9121	0.006877	6.344429
33255_at	-1.40414	8.367932	-11.7722	0.007129	6.255751
1884_s_at	-2.69304	9.034773	-11.5777	0.00773	6.13026
35312_at	-2.22932	10.12458	-11.3175	0.008852	5.958293
673_at	-1.47642	9.828868	-11.2073	0.009111	5.883954
1803_at	-1.33586	8.72651	-10.9644	0.010414	5.717035
1462_s_at	-0.9343	9.341027	-10.5255	0.013356	5.403896
39542_at	1.164624	10.0553	10.52498	0.013356	5.403522
41583_at	-2.21467	9.771343	-10.4626	0.013428	5.357783
40117_at	-2.41576	9.676526	-10.1615	0.015868	5.132482
1474_s_at	-1.15505	6.033744	-10.1541	0.015868	5.126826
1476_s_at	-0.99127	8.471458	-9.99964	0.015913	5.008176
38422_s_at	-1.54015	7.474208	-9.98855	0.015913	4.999582
37294_at	1.900328	8.194959	9.955109	0.015913	4.973588

37576_at	-1.28122	6.918036	-9.94924	0.015913	4.969015
1687_s_at	1.376881	8.996093	9.930995	0.015913	4.954785
910_at	-3.48441	9.660231	-9.82262	0.01687	4.869651
1516_g_at	-2.14961	9.685739	-9.59752	0.019795	4.689453
36499_at	-1.09687	9.216045	-9.56253	0.019795	4.661029
1376_at	-1.29333	8.990026	-9.47572	0.020717	4.590018
39059_at	-0.94874	11.28948	-9.4047	0.021414	4.531399
38551_at	2.013988	10.59207	9.12614	0.026662	4.296839
36813_at	-1.78936	8.606042	-9.05064	0.02777	4.231964
982_at	-1.6087	7.677224	-9.00745	0.028044	4.194598
967_g_at	-1.00433	8.180096	-8.97876	0.028044	4.169665
31831_at	0.769261	9.499375	8.954107	0.028044	4.148185
685_f_at	-0.90668	8.484465	-8.85896	0.029855	4.064688
1005_at	0.960637	8.623236	8.724912	0.032437	3.94549
36317_at	0.833625	7.147444	8.718143	0.032437	3.939421
846_s_at	2.127526	8.883141	8.685102	0.032694	3.909731
31830_s_at	0.90929	10.29754	8.585732	0.035083	3.819744
1801_at	-1.87408	8.139036	-8.56299	0.035083	3.799
40767_at	1.900989	8.605031	8.520022	0.035765	3.759664
38944_at	0.838444	8.059626	8.481077	0.036242	3.723838
349_g_at	-1.35414	7.530401	-8.46243	0.036242	3.706628
41641_at	0.752957	9.686598	8.382601	0.037848	3.632515

Πίνακας 3.2

ID	M	A	t	P.Value	B
41400_at	-2.31225	10.04155	-19.9521	0.000116	9.823686
1824_s_at	-1.77799	9.238868	-17.3131	0.000405	8.959766
39755_at	-1.7223	12.13182	-15.9005	0.000852	8.40337
36134_at	-2.4666	8.275738	-15.4105	0.001119	8.19219
37043_at	1.496959	11.52942	13.98329	0.002599	7.51622
36833_at	-1.58154	8.709889	-13.7742	0.00296	7.408872
1854_at	-2.92426	8.532097	-13.4443	0.003649	7.234845
33252_at	-1.55986	8.000345	-12.8708	0.005308	6.91802
31798_at	-3.19872	12.11577	-12.7541	0.005739	6.851221
40425_at	1.376222	9.69109	12.72879	0.005837	6.836617
1505_at	-2.12252	8.764746	-12.6261	0.006256	6.777006
480_at	-1.0624	10.13993	-12.5302	0.006678	6.72075
1515_at	-1.95606	8.946283	-12.3313	0.007657	6.602286
2042_s_at	-2.11266	8.178136	-12.2486	0.008109	6.55234
947_at	-2.1023	10.97252	-11.9121	0.010282	6.344429
33255_at	-1.40414	8.367932	-11.7722	0.011369	6.255751
1884_s_at	-2.69304	9.034773	-11.5777	0.013098	6.13026
35312_at	-2.22932	10.12458	-11.3175	0.01588	5.958293

673_at	-1.47642	9.828868	-11.2073	0.01725	5.883954
1803_at	-1.33586	8.72651	-10.9644	0.020754	5.717035
1462_s_at	-0.9343	9.341027	-10.5255	0.029263	5.403896
39542_at	1.164624	10.0553	10.52498	0.029273	5.403522
41583_at	-2.21467	9.771343	-10.4626	0.030767	5.357783
40117_at	-2.41576	9.676526	-10.1615	0.039276	5.132482
1474_s_at	-1.15505	6.033744	-10.1541	0.039513	5.126826
1476_s_at	-0.99127	8.471458	-9.99964	0.044896	5.008176
38422_s_at	-1.54015	7.474208	-9.98855	0.045309	4.999582
37294_at	1.900328	8.194959	9.955109	0.046589	4.973588
37576_at	-1.28122	6.918036	-9.94924	0.046814	4.969015
1687_s_at	1.376881	8.996093	9.930995	0.047531	4.954785
910_at	-3.48441	9.660231	-9.82262	0.052066	4.869651
1516_g_at	-2.14961	9.685739	-9.59752	0.063094	4.689453
36499_at	-1.09687	9.216045	-9.56253	0.065025	4.661029
1376_at	-1.29333	8.990026	-9.47572	0.070109	4.590018
39059_at	-0.94874	11.28948	-9.4047	0.074594	4.531399
38551_at	2.013988	10.59207	9.12614	0.09552	4.296839
36813_at	-1.78936	8.606042	-9.05064	0.102245	4.231964
982_at	-1.6087	7.677224	-9.00745	0.106321	4.194598
967_g_at	-1.00433	8.180096	-8.97876	0.109126	4.169665
31831_at	0.769261	9.499375	8.954107	0.111598	4.148185
685_f_at	-0.90668	8.484465	-8.85896	0.121766	4.064688
1005_at	0.960637	8.623236	8.724912	0.137865	3.94549
36317_at	0.833625	7.147444	8.718143	0.138728	3.939421
846_s_at	2.127526	8.883141	8.685102	0.143068	3.909731
31830_s_at	0.90929	10.29754	8.585732	0.157071	3.819744
1801_at	-1.87408	8.139036	-8.56299	0.160475	3.799
40767_at	1.900989	8.605031	8.520022	0.167138	3.759664
38944_at	0.838444	8.059626	8.481077	0.173441	3.723838
349_g_at	-1.35414	7.530401	-8.46243	0.176543	3.706628
41641_at	0.752957	9.686598	8.382601	0.190582	3.632515

Πίνακας 3.

Αν παρατηρήσουμε τις προσαρμοσμένες τιμές-p που δίνει η κάθε μία από τις τρεις διαδικασίες και θεωρήσουμε τον πολλαπλό έλεγχο υποθέσεων σε επίπεδο $a = 0.01$ διαπιστώνουμε τα εξής:

Σύμφωνα με τη διαδικασία των Benjamini και Hochberg (1995) για τη ρύθμιση της FDR, όλα τα γονίδια που φαίνονται στον πίνακα είναι διαφορετικά εκφρασμένα. Προφανώς ο αριθμός των υποθέσεων που απορρίπτονται σε επίπεδο $a = 0.01$ είναι μεγαλύτερος του 50. Η διαδικασία των Benjamini και Yekutieli (2001), η οποία αποτελεί διαδικασία ρύθμισης της

FDR για πολλαπλά εξαρτημένα τεστ, μας δίνει 19 απορρίψεις, ενώ η διαδικασία του Holm (1979) για τη ρύθμιση της FWER 14 απορρίψεις. Συγκρίνοντας τη διαδικασία του Holm (step-down διαδικασία για τη ρύθμιση της FWER) με αυτή των Benjamini και Hochberg (step-up διαδικασία για τη ρύθμιση της FDR) διαπιστώνουμε ότι και πάλι η πρώτη είναι πολύ πιο συντηρητική αφού οδηγεί σε πολύ λιγότερες απορρίψεις.

ΠΑΡΑΡΤΗΜΑ

Αποδείξεις Λημμάτων.

Απόδειξη του λήμματος 2.1. Η απόδειξη του λήμματος θα γίνει με τη χρήση της επαγωγικής μεθόδου στο *m*.

Για $m = 1$ είναι προφανές.

Δεχόμαστε ότι ισχύει για $m' \le m$ και θα δείξουμε ότι ισχύει για $m'+1$.

Αν $m_0 = 0$ τότε όλες οι μηδενικές υποθέσεις είναι ψευδείς, το Q είναι ίσο με μηδέν και θα ισχύει:

$$E\left(Q \mid P_1 = p_1, ..., P_{m'+1} = p_{m'+1} \right) = 0 \le \frac{m_0}{m'+1} q^*$$

Αν $m_0 > 0$, συμβολίζουμε με P_i' , $i = 1, 2, ..., m_0$, τις τυχαίες μεταβλητές για τις τιμές-*p* που αντιστοιχούν στις αληθείς μηδενικές υποθέσεις και με $P'_{(m_0)}$ τη μεγαλύτερη από αυτές. Επίσης διατάσσουμε τις m_1 τιμές-*p* που αντιστοιχούν στις ψευδείς μηδενικές υποθέσεις και τέλος ορίζουμε ως j_0 το μεγαλύτερο $j : 0 \le j \le m_1$ που ικανοποιεί την

$$p_j \le \frac{m_0 + j}{m'+1} q^* \qquad (\pi.1.1)$$

και με c το δεξιό μέρος της ανισότητας για $j = j_0$, δηλαδή $c = \frac{m_0 + j_0}{m'+1} q^*$.

Έτσι μπορούμε να γράψουμε:

$$E\left(Q \mid P_{m_0+1} = p_1, ..., P_m = p_{m_1} \right) = \int_0^c E\left(Q \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, ..., P_{m'} = p_{m_1} \right) f_{P'_{(m_0)}}(p)\, dp$$

$$+ \int_c^1 E\left(Q \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, ..., P_{m'} = p_{m_1} \right) f_{P'_{(m_0)}}(p)\, dp$$

με $f_{P'_{(m_0)}}(p) = m_0 p^{m_0-1}$. Στο πρώτο ολοκλήρωμα της παραπάνω σχέσης έχουμε $p \le c$ και έτσι όλες οι $m_0 + j_0$ υποθέσεις απορρίπτονται, ενώ το ποσοστό Q είναι ίσο με $\frac{m_0}{m_0 + j_0}$. Υπολογίζοντας πρώτα το ολοκλήρωμα και αντικαθιστώντας στην ανισότητα (π.1.1), έχουμε

$$\frac{m_0}{m_0 + j_0} c^{m_0} \le \frac{m_0}{m_0 + j_0} \frac{m_0 + j_0}{m' + 1} q^* c^{m_0 - 1} = \frac{m_0}{m' + 1} q^* c^{m_0 - 1} \qquad (\pi.1.2)$$

Για το δεύτερο ολοκλήρωμα, θεωρούμε ξεχωριστά κάθε $p_{j_0} < p_j \le P'_{(m_0)} = p < p_{j+1}$ και $p_{j_0} \le c \le P'_{(m_0)} = p < p_{j_0+1}$. Είναι σημαντικό να σημειωθεί ότι λόγω του ορισμού των j_0 και c, οι τιμές $p, p_{j+1}, p_{j+2}, ..., p_{m_1}$ δεν αρκούν για την λήψη απόφασης απόρριψης ή όχι των μηδενικών υποθέσεων. Έτσι, όταν θεωρούνται όλες οι υποθέσεις - αληθείς και ψευδείς - μαζί με τις αντίστοιχες διατεταγμένες τιμές-p, μία υπόθεση $H_{0(i)}$ μπορεί να απορριφθεί μόνον όταν υπάρχει ένα k, $i \le k \le m_0 + j - 1$ για το οποίο $p_{(k)} \le k(m' + 1) q^*$ ή ισοδύναμα

$$\frac{p_{(k)}}{p} \le \frac{k}{m_0 + j - 1} \frac{m_0 + j - 1}{(m' + 1) p} q^* \qquad (\pi.1.3)$$

Κατά τη δέσμευση στο $P'_{(m_0)} = p$, τα $\frac{P'_i}{p}, i = 1, 2, ..., m_0 - 1$, είναι ανεξάρτητες ομοιόμορφα κατανεμημένες τυχαίες μεταβλητές στο (0,1) και τα $\frac{p_i}{p}, i = 1, 2, ..., j$, είναι οι τιμές μεταξύ του 0 και του 1 που αντιστοιχούν στις ψευδείς μηδενικές υποθέσεις. Η χρήση της ανισότητας (π.1.3) για τον έλεγχο των $m_0 + j - 1 = m' \le m$ υποθέσεων είναι ισοδύναμη με την διαδικασία των Benjamini και Hochberg με τη σταθερά $\frac{m_0 + j - 1}{(m' + 1) p} q^*$ να παίρνει τη θέση του q^*. Εφαρμόζοντας τώρα την υπόθεση της επαγωγής, έχουμε

$$E\left(Q \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, ..., P_{m'} = p_{m_1} \right) \le \frac{m_0 - 1}{m_0 + j - 1} \frac{m_0 + j - 1}{(m' + 1) p} q^* = \frac{m_0 - 1}{(m' + 1) p} q^*.$$

Το παραπάνω όριο της ανισότητας εξαρτάται από το p, αλλά όχι από το διάστημα $p_j < p < p_{j+1}$ και έτσι

$$\int_c^1 E\left(Q \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, ..., P_{m'} = p_{m_1} \right) f_{P'_{(m_0)}}(p) dp \le \int_c^1 \frac{m_0 - 1}{(m' + 1) \cdot p} q^* \cdot m_0 p^{(m_0 - 1)} dp$$

$$= \frac{m_0 - 1}{(m'+1)\,p}\; q^* \int_c^1 (m_0 - 1)\, p^{(m_0-2)} dp = \frac{m_0}{m'+1} q^* \left\{1 - c^{m_0-1}\right\} \qquad (\pi.1.4)$$

Προσθέτοντας τα δεξιά μέλη των (π.1.2) και (π.1.4) καταλήγουμε στο ζητούμενο. □

Απόδειξη του λήμματος 2.2. Το ότι η (2.12) = (2.11) είναι προφανές. Για να αποδείξουμε ότι η (2.13) = (2.12) ορίζουμε την συνάρτηση

$$h(u) = \Pr\left(Y \le y \mid X = u\right)$$

έτσι ώστε

$$\Pr\left(Y \le y\,, X \le x\right) = \int_{-\infty}^{x} h(u)\, dF_X(u)$$

όπου F_x η περιθωριακή κατανομή της *X*. Κάτω από την υπόθεση ότι η *h* είναι φθίνουσα πρέπει να δείξουμε ότι

$$\int_{-\infty}^{x} h(u)\, dF_X(u) / \Pr\left(X \le x\right) \;\ge \int_{-\infty}^{x'} h(u)\, dF_X(u) / \Pr\left(X \le x'\right)$$

για $x \le x'$.

Παρατηρούμε ότι

$$\frac{\int_{-\infty}^{x'} h(u)\, dF_X(u)}{\Pr\left(X \le x'\right)} \le \frac{\int_{-\infty}^{x} h(u)\, dF_X(u) + h(x)\Pr\left(x < X \le x'\right)}{\Pr\left(X \le x\right) + \Pr\left(x < X \le x'\right)}$$

άρα αρκεί να δείξουμε ότι

$$\frac{\int_{-\infty}^{x} h(u)\, dF_X(u)}{\Pr\left(X \le x\right)} \ge \frac{\int_{-\infty}^{x} h(u)\, dF_X(u) + h(x)\Pr\left(x < X \le x'\right)}{\Pr\left(X \le x\right) + \Pr\left(x < X \le x'\right)}.$$

Αυτό είναι ισοδύναμο με την

$$\int_{-\infty}^{x} h(u)\, dF_X(u)\Pr\left(X \le x\right) + \int_{-\infty}^{x} h(u)\, dF_X(u)\Pr\left(x < X \le x'\right)$$

$$\ge \int_{-\infty}^{x} h(u)\, dF_X(u)\Pr\left(X \le x\right) + h(x)\Pr\left(X \le x\right)\Pr\left(x < X \le x'\right)$$

$$\Leftrightarrow \int_{-\infty}^{x} h(u)\,dF_X(u) \geq h(x)\Pr(X \leq x) \Leftrightarrow \frac{\int_{-\infty}^{x} h(u)\,dF_X(u)}{\Pr(X \leq x)} \geq h(x)$$

$\Leftrightarrow \Pr(Y \leq y \mid X \leq x) \geq \Pr(Y \leq y \mid X = x)$ και η οποία ισχύει. □

ΒΙΒΛΙΟΓΡΑΦΙΑ

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistic Societ, Series B*, **57**, 289-300.

Benjamini, Y. & Liu, W. (1999). A step down multiple hypothesis procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, **82**, 163-170.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165-1188.

Boos, D. D. (2003). Introduction to the Bootstrap World. *Statistical Science*, **18**, 168-174.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. , Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800-802.

Holm, S. (1979) A simple sequentially rejective multiple test procedure, *Scand. J. Statist*., **6**, 65-70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75**, 383-386.

Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika*, **79**, 624-625.

Karlin, S. & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. *I. J. Multivariate Anal*., **10**, 467-498.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, **37**, 1137-1153.

Marcus, R., Peritz, E. & Gabriel, K. R. (1976). On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. Biometrics, **63**, 655-660.

Sarkar, T. K. (1969). Some lower bounds of reliability. *Technical Report, 124*, Department of Statistics, Stanford University.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, **30**, 239-257.

Scholtens, D., Miron, A., Merchant, F. M., Miller, A., Miron, P. L., Iglehart, J. D., Gentleman, R. (2004). Analyzing Factorial Designed Microarray Experiments. *Journal of Multivariate Analysis*, **90**, 19-43.

Shaffer, J. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology*, **46**, 561-584.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751-754.

Storey, J. D. (2001). The false discovery rate: A Bayesian interpretation and the q-value. *Technical Report 2001-12*, Department of Statistics, Stanford University.

Storey, J. D. & Tibshirani, R. (2001). Estimating the positive false discovery rate under dependence, with applications to DNA microarrays. *Technical Report*, *2001-28*, Department of Statistics, Stanford University.

Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of Royal Statistic Society Series B*, **64**, 476-498.

Tamhame, A. C., Liu, W. and Dunnet, C. W. (1998). A generalized step-up-step-down multiple test procsdure. *Canad. J. Statist*, **26**, 353-363.

Westfall, P. H. & Young, S. S (1993). *Resampling-based Multiple Testing: Examples and methods for p-value adjustment*, John Wiley & Sons.