Πρόλογος

Η ανάπτυξη των υπολογιστών και η εξάπλωση τους ως εργαλεία στα χέρια στατιστικών έχει οδηγήσει στην ολοένα αυξανόμενη χρήση τους τόσο για την εφαρμογή παραδοσιακών μεθόδων όσο και για την ανάπτυξη νέων μεθόδων που χρησιμοποιούν την υπολογιστική ισχύ. Συνάμα, οι υπολογιστές προσφέρουν τη δυνατότητα για τη δημιουργία εξαιρετικών γραφημάτων για την οπτική παρουσίαση και μελέτη των δεδομένων.

Οι σημειώσεις που κρατάτε στα χέρια σας αποτελούν μια σύντομή εισαγωγή σε στατιστικές μεθόδους οι οποίες βασίζονται στην ευρεία χρήση του υπολογιστή. Είναι εξαιρετικά ενδιαφέρον πως αν και κάποιες από αυτές τις μεθόδους αναπτύχθηκαν πολλά χρόνια πριν, έγιναν εργαλεία στα χέρια των στατιστικών μόλις τα τελευταία χρόνια. Είναι επίσης πολύ σημαντικό να αναφερθεί πως οι μέθοδοι που θα περιγραφούν στη συνέχεια αποτελούν μερικά από τα σύγχρονα θέματα επιστημονικής έρευνας και οι εξελίξεις στα αντικείμενα αυτά είναι ραγδαίες.

Τα θέματα που αναπτύσσονται στις σημειώσεις αυτές είναι οργανωμένα σε ξεχωριστά κεφάλαια και πραγματεύονται διαφορετικά θέματα με κοινό παρονομαστή την ανάγκη χρήσης υπολογιστή. Στο πρώτο κεφάλαιο αναπτύσσεται η εκτίμηση συναρτήσεων πυκνότητας πιθανότητας με τη χρήση των kernels. Επίσης υπάρχει εκτενής συζήτηση για τη χρήση του ιστογράμματος, του πιο διαδεδομένου γραφήματος για απεικόνιση δεδομένων, καθώς και των μειονεκτημάτων του. Το δεύτερο κεφάλαιο αφορά ελέγχους τυχαιοποίησης. Είναι πολύ ενδιαφέρον πως αν και οι έλεγχοι αυτοί είναι γνωστοί από τον Fisher στη δεκαετία του 30, μόλις πρόσφατα έχουν ενσωματωθεί στα στατιστικά πακέτα. Το τρίτο κεφάλαιο περιέχει υλικό σχετικά με ελέγχους Monte Carlo. Σε αυτό το κεφάλαιο θέματα που αφορούν την προσομοίωση θεωρούνται γνωστά και επομένως ο αναγνώστης δεν θα βρει λεπτομέρειες σχετικά με την προσομοίωση. Το τέταρτο κεφάλαιο αφορά τη μέθοδο jackknife και τη μέθοδο cross-validation. Η μέθοδος jackknife είναι γνωστή από τα τέλη της δεκαετίας του 1940 ως μέθοδος που μειώνει τη μεροληψία των εκτιμητών. Η μέθοδος cross-validation έχει γίνει πολύτιμο εργαλείο για επιλογή μοντέλων όχι απαραίτητα γραμμικών. Το πέμπτο κεφάλαιο αποτελεί μια εισαγωγή σε ένα πολύ θερμό θέμα των τελευταίων χρόνων, τη μέθοδο bootstrap.

Αυτή η έκδοση των σημειώσεων περιέχει κι ένα νέο κεφάλαιο σχετικά με τον EM αλγόριθμο. Ο αλγόριθμος αυτός έχει βοηθήσει αρκετά την εκτίμηση παραμέτρων σε πολύπλοκα προβλήματα με τη χρήση της μεθόδου μεγίστης πιθανοφάνειας. Επίσης οι ομοιότητες του με άλλες μεθόδους τον κάνει ένα πολύ χρήσιμο εργαλείο για τη στατιστική συμπερασματολογία. Το έκτο κεφάλαιο περιέχει μια εισαγωγή στον αλγόριθμο αυτό.

Σε όλα τα κεφάλαια η προσέγγιση είναι κυρίως οι εφαρμογές και όχι η θεωρία που υπάρχει. Σε αρκετά σημεία όμως παρουσιάζεται και η θεωρία που αποτελεί το βασικό στοιχείο κατανόησης των μεθόδων. Στην έκδοση αυτή έχουν προστεθεί ασκήσεις στο τέλος κάθε κεφαλαίου.

Δυστυχώς προς το παρόν δεν υπάρχει ελληνική βιβλιογραφία στα θέματα τα οποία κάλυπτουν αυτές οι σημειώσεις επομένως οι ενδιαφερόμενοι θα πρέπει να χρησιμοποιήσουν ξενόγλωσση (κυρίως αγγλική) βιβλιογραφία. Μετά από κάθε κεφάλαιο ακολουθεί μια λίστα με βιβλία σχετικά με το θέμα με ένα μικρό σχολιασμό

με το περιεχόμενο τους και το βαθμό δυσκολίας τους. Η λίστα περιέχει μόνο κάποια βασικά βιβλία και άρθρα και σε καμιά περίπτωση δεν είναι πλήρης.

Χρειάστηκαν 4 χρόνια για να πάρουν οι σημειώσεις μια ολοκληρωμένη μορφή και να διορθωθούν πολλά από τα τυπογραφικά λάθη των προηγούμενων εκδόσεων ώστε να προκύψει η πρώτη εκδοση αυτών των σημειώσεων το Δεκέμβρη του 2004. Συνεπώς θα πρέπει να ευχαριστήσω τους φοιτητές του τμήματος Στατιστικής (προπτυχιακούς και μεταπτυχιακούς) που με ποικίλα σχόλια βοήθησαν τη διαμόρφωση των σημειώσεων αυτών. Θερμές ευχαριστίες στους συναδέλφους μου για την πολύτιμη βοήθεια τους όλα αυτά τα χρόνια αλλά και πολλά σχόλια που πιστεύω πως βελτίωσαν την παρουσίαση.

Η 2η εκδοση που κρατάτε στα χέρια σας περιέχει κάποια περισσότερα πράγματα σε σχέση με την προηγούμενη με βάση σχόλια και παρατηρήσεις φοιτητών όλα αυτά τα χρόνια. Αρκετά από τα τυπογραφικά λάθη έχουν διορθωθέι κι ελπίζω αυτά που παραμένουν να ειναι ελάχιστα και να αποκαλυφθούν συντομα από τους αναγνώστες. Παρά την προσπάθεια είναι βέβαιο πως υπάρχουν ακόμα κάποια τυπογραφικά λάθη. Θα βοηθήσει σημαντικά, όποιος βρει τέτοια λάθη να με ενημερώσει στέλνοντας ένα email στη διεύθυνση `karlis@aueb.gr` ώστε σε επόμενη έκδοση να απαλειφθούν όσα λάθη παραμένουν.

Αθήνα, Σεπτέμβριος 2008

Δημήτρης Καρλής

Περιεχόμενα

Κατάλογος Πινάκων

Καταλογος Γραφημάτων

Κεφάλαιο 1

Εκτίμηση Πυκνότητας Πιθανότητας

1.1 Εισαγωγή

Τα τελευταία χρόνια με την εκρηκτική διάδοση των ηλεκτρονικών υπολογιστών και με τη κατασκευή υπολογιστών με τεράστιες δυνατότητες υπολογισμών, η στατιστική επιστήμη βρήκε μια διέξοδο για να αναπτύξει καινούριες μεθοδολογίες αλλά και να υποστηρίξει και να βελτιώσει ήδη υπάρχουσες. Αυτό είχε σαν αποτέλεσμα, πολλά προβλήματα προηγούμενων γενεών να φαίνονται τώρα πια απλές ασκήσεις για κάποιο προπτυχιακό μάθημα, αλλά και προβλήματα που δεν μπορούσαν να αντιμετωπιστούν παρά μόνο μέσα από απλοποιήσεις των αρχικών υποθέσεων και προσεγγιστικές μεθόδους να μπορούν τώρα πια να επιλυθούν εύκολα, γρήγορα και με ακριβή αποτελέσματα.

Κάποια τέτοια παραδείγματα είναι τα εξής:

- 'Ενα μεγάλο κομμάτι της στατιστικής επιστήμης έχει να κάνει με την εκτίμηση των παραμέτρων κάποιου μοντέλου. 'Εστω ότι χρησιμοποιούμε τη μέθοδος μεγίστης πιθανοφάνειας. Σε πολλές περιπτώσεις δεν μπορούμε να βρούμε τις εκτιμήτριες σε κλειστή μορφή, δηλαδή να γράψουμε έναν τύπο που τις υπολογίζει και άρα αριθμητικές μέθοδοι πρέπει να χρησιμοποιηθούν. Φανταστείτε λοιπόν ένα μοντέλο με πολλές παραμέτρους. Είναι σαφές ότι για να είναι εφικτή η χρήση αριθμητικών μεθόδων χρειάζεται ένας δυνατός υπολογιστής για να φέρει σε πέρας τον τεράστιο όγκο υπολογισμών που απαιτούνται.

- Η περιγραφική στατιστική αποτελεί το πιο βασικό κομμάτι κάθε στατιστικής έρευνας και κάθε ανάλυσης, καθώς μας επιτρέπει να αποκομίσουμε μια πρώτη εικόνα των δεδομένων και να ανακαλύψουμε αφενός την ύπαρξη κάποιων σχέσεων που πρέπει να ερευνηθούν στη συνέχεια αλλά και να δούμε ποιες μέθοδοι ταιριάζουν και μπορούν να χρησιμοποιηθούν. Το να προσπαθήσει κάποιος να δημιουργήσει έστω και ένα απλό ιστόγραμμα για ένα

μεγάλο σετ δεδομένων με χιλιάδες παρατηρήσεις είναι ιδιαίτερα επίπονο και χρονοβόρο, πόσο μάλλον να χρησιμοποιήσει κάποια άλλη μέθοδο εκτίμησης της συνάρτησης πυκνότητας πιθανότητας του πληθυσμού που έχει μεγαλύτερη ακρίβεια από ότι ένα ιστόγραμμα.

- Η προσομοίωση αποτελεί πια ένα ζωτικό κομμάτι της στατιστικής επιστήμης κυρίως γιατί μας επιτρέπει πολύ εύκολα και με χαμηλό κόστος (σε χρόνο, μνήμη κ.ο.κ) να αναπαράγουμε κάποιες διαδικασίες δημιουργίας των δεδομένων και επομένως να εξετάσουμε ποικίλα χαρακτηριστικά τους. Ιδίως σε περιπτώσεις που τα θεωρητικά αποτελέσματα είναι δύσκολο να προκύψουν, η προσομοίωση αποτελεί ένα πολύτιμο βέλος στη φαρέτρα του στατιστικού.

- Οι περισσότερες στατιστικές διαδικασίες ξεκινούν και βασίζονται για την ακρίβεια των αποτελεσμάτων τους σε κάποιες υποθέσεις. Για παράδειγμα για πολλά χρόνια όλη σχεδόν η στατιστική ήταν βασισμένη σε υποθέσεις περί κανονικότητας του πληθυσμού. Τι μπορεί όμως να πει κάποιος για ένα δειγματικό συντελεστή συσχέτισης μεταξύ δύο μεταβλητών, όταν ο πληθυσμός ξέρουμε πως δεν είναι κανονικός αλλά ακολουθεί την εκθετική κατανομή; Τέτοια προβλήματα λύνονται πια σχετικά εύκολα με τη χρήση υπολογιστή και μεθόδων Monte Carlo ή bootstrap για τις οποίες θα μιλήσουμε στη συνέχεια.

Τα παραπάνω αποτελούν λίγα μόνο από τα πολλά παραδείγματα χρήσης των υπολογιστών για την επίλυση στατιστικών προβλημάτων. Η γενικευμένη χρήση εύχρηστων και ευέλικτων στατιστικών πακέτων είναι ένα απλό παράδειγμα για να καταλάβει κανείς τη μεγάλη εξάρτηση της στατιστικής επιστήμης από τους υπολογιστές.

Οι μέθοδοι που θα χρησιμοποιήσουμε στη συνέχεια βασίζονται και χρειάζονται τη χρήση υπολογιστών, κυρίως για να μπορέσει κάποιος να αντεπεξέλθει στον μεγάλο όγκο υπολογισμών που απαιτούνται. Είναι επίσης πολύ ενδιαφέρον να αναφέρουμε πως η θεωρητική ανάπτυξη των μεθόδων αυτών είναι αρκετά παλιά. Για παράδειγμα, η εκτίμηση με τη χρήση kernels αναπτύχθηκε στα τέλη της δεκαετίας του ’50, αλλά άρχισε να χρησιμοποιείται ευρέως στα μέσα της δεκαετίας του ’80 όταν η χρήση υπολογιστών άρχισε να γενικεύεται. Η μέθοδος jackknife αναπτύχθηκε στα τέλη της δεκαετίας του ’40, οι μέθοδοι cross-validation και bootstrap στη δεκαετία του ’70, ενώ οι έλεγχοι τυχαιοποίησης εμφανίστηκαν το 1935 από τον Fisher και τον ακριβή του έλεγχο για πίνακες συνάφειας και στην προσεγγιστική τους μορφή στα τέλη της δεκαετίας του ’60. Γίνεται άμεσα αντιληπτό πως έχει ήδη συσσωρευτεί ένας μεγάλος όγκος πληροφοριών για τις μεθόδους αυτές οι οποίες ολοένα και περισσότερο προσφέρονται από στατιστικά πακέτα που κυκλοφορούν στο εμπόριο.

1.2 Εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας

’Ενα από τα πιο συνηθισμένα προβλήματα που καλείται να αντιμετωπίσει ένας στατιστικός είναι να εκτιμήσει κάποια ποσότητα του πληθυσμού από ένα δείγμα

που έχει στα χέρια του. Μέχρι τώρα είχατε αντιμετωπίσει την εκτίμηση μέσων τιμών, διακυμάνσεων, ποσοστών και διαφόρων άλλων ποσοτήτων. Πολλές φορές το ενδιαφέρον εστιάζεται στην εκτίμηση της ίδιας της συνάρτησης πυκνότητας πιθανότητας (ή συνάρτησης πιθανότητας στη διακριτή περίπτωση) του πληθυσμού. Στο πρώτο λοιπόν μέρος θα μας απασχολήσει αυτό ακριβώς το πρόβλημα. Οι μεθοδολογίες που έχουν αναπτυχθεί και θα συζητηθούν στις σημειώσεις αυτές απαιτούν αρκετούς υπολογισμούς και επομένως η χρήση υπολογιστή κρίνεται απαραίτητη. Από την άλλη άποψη, δεδομένου ότι πολλές φορές αυτές οι μέθοδοι έχουν σαν σκοπό να οπτικοποιήσουν την εκτίμηση αυτή, κατασκευάζοντας το γράφημα, η χρήση υπολογιστών είναι επιβεβλημένη.

'Εστω ότι θέλουμε λοιπόν να εκτιμήσουμε μια συνάρτηση πυκνότητας πιθανότητας:

Μια πρώτη προσέγγιση θα ήταν να υποθέσουμε κάποια συγκεκριμένη μορφή για την υπό εξέταση συνάρτηση πυκνότητας πιθανότητας και στη συνέχεια να εκτιμήσουμε τις άγνωστες παραμέτρους. Για παράδειγμα αν υποθέσουμε πως η κανονική κατανομή είναι η κατανομή του πληθυσμού πρέπει να εκτιμήσουμε τη μέση τιμή και τη διακύμανση. Υποθέτοντας όμως μια συγκεκριμένη μορφή αυτό αυτόματα θέτει κάποιους περιορισμούς. Υποθέτοντας δηλαδή ότι ο πληθυσμός είναι κανονικός υποθέτουμε πως η κατανομή του πληθυσμού είναι συμμετρική.

Μια εντελώς διαφορετική προσέγγιση είναι να μην στηριχθούμε σε συγκεκριμένες υποθέσεις για την κατανομή του πληθυσμού. Από αυτή την άποψη, η προσέγγιση αυτή μπορεί να θεωρηθεί ως μη-παραμετρική μέθοδος. Στην πραγματικότητα υπάρχει μια πλειάδα τέτοιων μεθόδων οι οποίες πάντως έχουν κάποια κοινά χαρακτηριστικά. Οι πιο διαδεδομένες τέτοιες μέθοδοι είναι

- Το ιστόγραμμα
- Το πολύγωνο συχνοτήτων
- Ο απλοϊκός εκτιμητής
- Εκτιμητές με τη χρήση kernels

Εκτός από αυτές τις μεθόδους που θα μας απασχολήσουν στη συνεχεία υπάρχουν και άλλες μέθοδοι όπως η χρήση splines, ορθογώνιων σειρών, η χρήση πολυωνύμων και άλλες που δεν θα αναλυθούν εδώ.

1.3 Μέτρα σύγκρισης εκτιμητών

Στη συνέχεια θα χρησιμοποιήσουμε το συμβολισμό $\hat{f}(x)$ για την εκτίμηση της πραγματικής (και άγνωστης) συνάρτησης πυκνότητας πιθανότητας του πληθυσμού. Επειδή η ποσότητα $\hat{f}(x)$ είναι μια τυχαία μεταβλητή θα θέλαμε να εξετάσουμε τις ιδιότητές της. Είναι ευνόητο ότι αν πάρουμε διαφορετικά δείγματα από τον ίδιο πληθυσμό η τιμή που θα πάρουμε για την $\hat{f}(x)$ θα διαφέρει. Επομένως για να αξιολογήσουμε το πόσο καλή εκτιμήτρια είναι ως προς την πραγματική τιμή $f(x)$ του πληθυσμού χρειαζόμαστε κάποια μέτρα που να μας δείχνουν το πόσο κοντά είναι η $\hat{f}(x)$ στην πραγματική τιμή $f(x)$ αλλά και τη μεταβλητότητα της. Από την Εκτιμητική υπάρχουν κάποιες ποσότητες που χρησιμεύουν για την αξιολόγηση εκτιμητριών. Αυτές ήταν

η μεροληψία (Bias), η διακύμανση (Variance) και το μέσο τετραγωνικό σφάλμα (Mean Squared Error, MSE).

Πιο συγκεκριμένα για μια εκτιμήτρια $\hat{\theta}$ μιας άγνωστης παραμέτρου θ ορίζουμε τις εξής ποσότητες:

$$\begin{aligned} Bias(\hat{\theta}) &= E(\hat{\theta}) - \theta \\ Var(\hat{\theta}) &= E\left[(\hat{\theta} - E(\hat{\theta}))^2\right] \\ MSE(\hat{\theta}) &= E\left[(\hat{\theta} - theta)^2\right] \end{aligned}$$

Μπορεί κανείς να επιβεβαιώσει τη σχέση που συνδέει το MSE με τη μεροληψία και τη διακύμανση. Συγκεκριμένα

$$\begin{aligned} MSE(\hat{\theta}) &= E\left[(\hat{\theta} - \theta)^2\right] \\ &= E\left[\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2\right] \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \end{aligned}$$

Όμως

$$\begin{aligned} [Bias(\hat{\theta})]^2 + Var(\hat{\theta}) &= \left[E(\hat{\theta}) - \theta\right]^2 + E\left[(\hat{\theta} - E(\hat{\theta}))^2\right] \\ &= [E(\hat{\theta})]^2 - 2\theta E(\hat{\theta}) + \theta^2 + E(\hat{\theta}^2) - [E(\hat{\theta})]^2 \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \end{aligned}$$

κι επομένως η σχέση ισχύει

Η μεροληψία μας έδειχνε αν κατά μέσο όρο η εκτιμήτρια είναι η ίδια με την πραγματική τιμή (αμερόληπτη) ή αν διαφέρουν (μεροληπτική). Είναι αυτονόητο πως προτιμούνται εκτιμήτριες με όσο γίνεται μικρότερη μεροληψία σε απόλυτη τιμή.

Η διακύμανση μας έδειχνε πόση μεταβλητότητα έχει η εκτιμήτρια και μας ενδιαφέρει να έχει όσο γίνεται μικρότερη.

Επειδή όμως το να μικρύνει κανείς τη μεροληψία πολλές φορές οδηγεί σε αύξηση της διακύμανσης και το αντίθετο, το μέσο τετραγωνικό σφάλμα προσπαθεί να ισορροπήσει ανάμεσα στις δύο αυτές ποσότητες.

Για την περίπτωση της εκτίμησης μίας συνάρτησης πυκνότητας πιθανότητας έχουμε πως

$$\begin{aligned} Bias\left[\hat{f}(x)\right] &= E\left[\hat{f}(x)\right] - f(x) \\ Var\left[\hat{f}(x)\right] &= E\left[\left\{\hat{f}(x) - E\left[\hat{f}(x)\right]\right\}^2\right] \\ MSE\left[\hat{f}(x)\right] &= \left\{Bias\left[\hat{f}(x)\right]\right\}^2 + Var\left[\hat{f}(x)\right] \end{aligned}$$

Επειδή όμως το μέσο τετραγωνικό σφάλμα αφορά μια συγκεκριμένη τιμή του x, για να μπορέσουμε να αξιολογήσουμε μια εκτιμήτρια $\hat{f}(x)$ ενδιαφερόμαστε για το

πόσο καλή είναι σε όλον τον άξονα, άρα και για όλες τις τιμές του x. Για αυτό χρησιμοποιούμε το ολοκληρωμένο μέσο τετραγωνικό σφάλμα (Mean Integrated Squared Error) το οποίο υπολογίζεται ως

$$MISE\left[\hat{f}(x)\right] = E\left[\int_{-\infty}^{+\infty}\left\{\hat{f}(x) - f(x)\right\}^2 dx\right] =$$
$$= \int_{-\infty}^{+\infty} E\left\{\hat{f}(x) - f(x)\right\}^2 dx = \int_{-\infty}^{+\infty} MSE\left[\hat{f}(x)\right] dx$$

Δηλαδή, το ολοκληρωμένο μέσο τετραγωνικό σφάλμα, υπολογίζει για κάθε τιμή του x το μέσο τετραγωνικό σφάλμα και ολοκληρώνοντας ως προς όλες τις δυνατές τιμές του x υπολογίζει μια γενική ποσότητα.

Όλες οι αναμενόμενες τιμές υπολογίζονται με αναφορά την πραγματική συνάρτηση πιθανότητας $f(x)$, δηλαδή για παράδειγμα

$$E\left[\hat{f}(x)\right] = \int_{-\infty}^{+\infty} \hat{f}(x)f(x)dx$$

Όπως διαπιστώνει κανείς, συνήθως δεν είναι εύκολο να υπολογιστούν αυτά τα ολοκληρώματα.

Τα μέτρα που ορίστηκαν παραπάνω, είναι αυτά που θα χρησιμοποιηθούν σε αυτές τις σημειώσεις για να εξεταστούν κάποιες από τις ιδιότητες των εκτιμητριών μιας συνάρτησης πυκνότητας πιθανότητας. Επίσης τα μέτρα αυτά θα αποτελούν και κριτήρια για να εκτιμήσει κανείς κάποιες ποσότητες που όπως θα δούμε απαιτούνται. Σε κάθε περίπτωση, υπάρχουν πολλά άλλα μέτρα αξιολόγησης εκτιμητριών για τα οποία δεν θα μιλήσουμε.

1.4 Εκτίμηση με τη μέθοδο του ιστογράμματος

Το ιστόγραμμα αποτελεί την πιο απλή και γνωστή μέθοδο εκτίμησης μιας συνάρτησης πυκνότητας πιθανότητας, ή καλύτερα είναι γνωστό ως ένα σχετικά απλό γράφημα που επιτρέπει στον ερευνητή να αποκτήσει μια πρώτη εικόνα σχετικά με τα δεδομένα του και την άγνωστη συνάρτηση πυκνότητας πιθανότητας.

Όπως είναι γνωστό από την Περιγραφική Στατιστική για να κατασκευάσουμε ένα ιστόγραμμα βρίσκουμε τον αριθμό των παρατηρήσεων που ανήκουν στο κάθε κελί και για κάθε κελί φτιάχνουμε μια ράβδο με ύψος ίσο με τη συχνότητα των τιμών μέσα σε αυτό. Παρ' όλη την απλότητά του, πρέπει κάποιος να είναι ενημερωμένος για τον κίνδυνο να αποκτήσει μια εσφαλμένη εικόνα για τα δεδομένα απλά και μόνο επειδή το ιστόγραμμα που έφτιαξε (συνήθως αυτόματα με τη χρήση κάποιου στατιστικού πακέτου) δεν είναι από καμιά άποψη το καλύτερο.

Για να κατασκευάσουμε ένα ιστόγραμμα χρειάζεται να γνωρίζουμε

- Το πλάτος κάθε κελιού
- Τον αριθμό των κελιών (δηλαδή σε πόσα διαστήματα, όχι απαραίτητα ισομήκη, αν και αυτή είναι η συνηθισμένη πρακτική, θα χωρίσουμε τα δεδομένα μας)

- Την αριστερή άκρη του πρώτου κελιού.

Στην πραγματικότητα δεν χρειάζεται να ορίσουμε όλα τα παραπάνω καθώς αν ξέρουμε το πλάτος του κελιού και την αρχή μπορούμε εύκολα να βρούμε πόσα κελιά χρειαζόμαστε για το εύρος των παρατηρήσεων μας και ούτω καθεξής.

Μέχρι τώρα το ιστόγραμμα χρησιμοποιόταν απλά σαν μια γραφική απεικόνιση των δεδομένων. Ας δούμε λοιπόν πως μπορούμε να το χρησιμοποιήσουμε για να εκτιμήσουμε μια συνάρτηση πυκνότητας πιθανότητας.

Η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας στο σημείο x , δηλαδή η $\hat{f}(x)$ με τη μέθοδο του ιστογράμματος δίνεται ως

$$\hat{f}(x) = \frac{1}{n} \frac{\#\text{παρατηρήσεων στο ίδιο κελί με το } x}{\text{πλάτος κελιού που περιέχει το } x}$$

Παράδειγμα 1.1 : 'Εστω ότι έχουμε 10 παρατηρήσεις με τιμές 2, 6, 8, 11, 14, 3, 5, 13, 19, 5 και πως τα κελιά μας είναι τα $[0, 5)$, $[5, 10)$, $[10, 15)$ και $[15, 20)$. 'Εχουμε δηλαδή επιλέξει να χρησιμοποιήσουμε 4 κελιά. Τότε οι συχνότητες κάθε κελιού είναι 2, 4, 3 και 1 αντίστοιχα και συνεπώς η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας είναι η

$$\hat{f}(x) = \begin{cases} 0.04 & , \quad \text{x} \in [0,5) \\ 0.08 & , \quad \text{x} \in [5,10) \\ 0.06 & , \quad \text{x} \in [10,15) \\ 0.02 & , \quad \text{x} \in [15,20) \\ 0 & , \quad \text{x} < 0, \text{x} \geq 20 \end{cases}$$

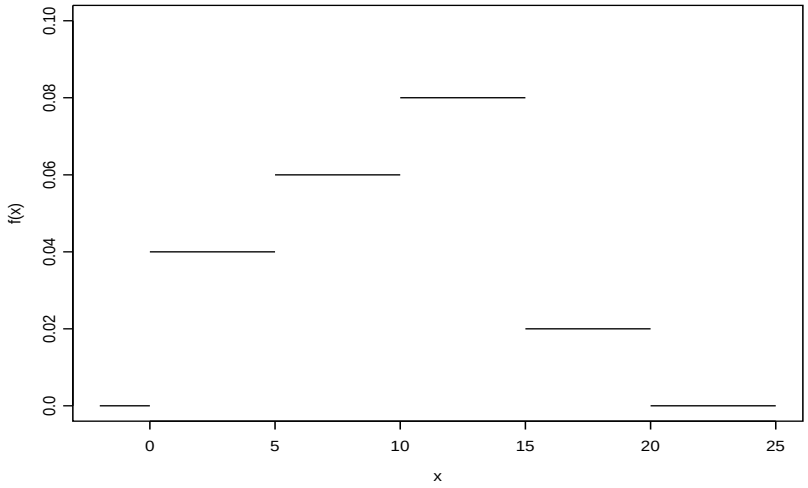
'Ετσι έχουμε πως $\hat{f}(3) = 0.04,\ \hat{f}(10) = 0.06,\ \hat{f}(14.3) = 0.06$, κλπ. Στο γράφημα 1.1 μπορεί κάποιος να δει την εκτίμηση αυτή και να παρατηρήσει τα "πηδηματάκια" που παρουσιάζει. Η εικόνα αυτή δεν είναι το ιστόγραμμα αλλά η απεικόνιση της εκτιμήτριας με τη μέθοδο του ιστογράμματος

Τα περισσότερα στατιστικά πακέτα κατασκευάζουν το ιστόγραμμα με έναν συγκεκριμένο τρόπο (δηλαδή με συγκεκριμένο αριθμό κελιών ή με τη χρήση κάποιου αλγορίθμου που αποφασίζει αυτόματα). Ο χρήστης, βέβαια, μπορεί να επέμβει και να αλλάξει τόσο τον αριθμό των κελιών όσο και το πλάτος τους ή την αρχή τους.

Το πλάτος του κελιού είναι βέλτιστο με την έννοια πως ελαχιστοποιεί το ολοκληρωμένο μέσο τετραγωνικό σφάλμα, και επομένως αν κάποιος χρησιμοποιήσει κάποιο άλλο κριτήριο σαφώς και θα καταλήξει σε κάποιο άλλο βέλτιστο πλάτος κελιού.

Αυτό που πρέπει να γίνει σαφές είναι πως ανάλογα με το πλάτος των κελιών μπορούμε να έχουμε πολύ μεγάλες διαφορές στην εικόνα που σχηματίζουν τα δεδομένα κι επομένως μπορούμε να οδηγηθούμε σε λανθασμένα συμπεράσματα για αυτά.

Παράδειγμα 1.2: Τα δεδομένα, τα οποία θα χρησιμοποιηθούν αρκετές φορές σε αυτό το κεφάλαιο, αφορούν πραγματικές παρατηρήσεις από ένα πείραμα όπου



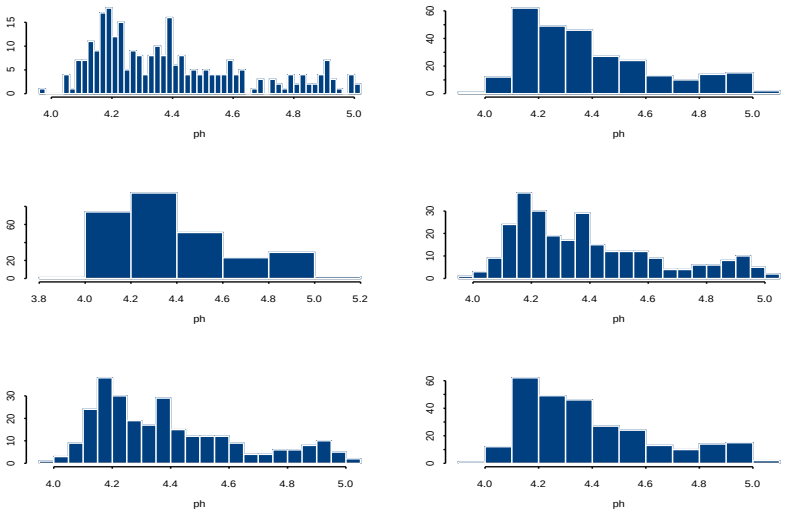
Γράφημα 1.1: Η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας για τα δεδομένα του παραδείγματος 1.1 με τη χρήση της μεθόδου του ιστογράμματος

300 γιαούρτια εξετάστηκαν για την ύπαρξη ή όχι κάποιων μικροοργανισμών. Οι 2 μεταβλητές που θα μας απασχολήσουν είναι το PH και η υγρασία. Αυτό που είναι ενδιαφέρον είναι πως ξέρουμε εκ των προτέρων πως καθώς υπάρχουν δύο είδη γιαουρτιών μέσα στο δείγμα (αγελαδινά και πρόβεια) η μορφή της κατανομής πρέπει, από τη θεωρία, να είναι ένα μείγμα 2 κατανομών και άρα είτε μια δίκορφη κατανομή είτε μια κατανομή με μεγάλη δεξιά ουρά. Κάτι τέτοιο όπως θα δούμε στη συνέχεια μάλλον επιβεβαιώνεται.

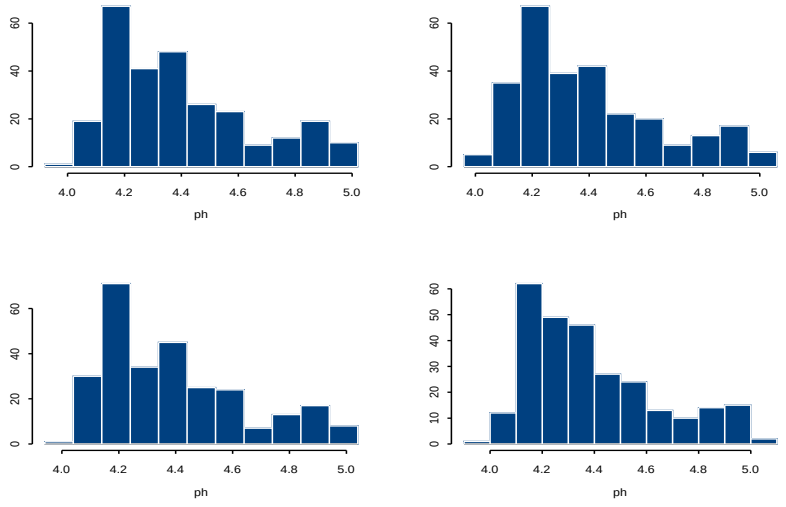
Τα γραφήματα που ακολουθούν αφορούν τα δεδομένα αυτά. Στο γράφημα 1.2 υπάρχουν 6 διαφορετικά ιστογράμματα για τα ίδια δεδομένα, με διαφορετικό αριθμό κελιών σε κάθε ιστόγραμμα. Στο γράφημα 1.3 αυτό που διαφέρει είναι η αρχή του γραφήματος.

Στο γράφημα 1.2 παρατηρείστε το πόσο αλλάζει η εικόνα ανάλογα με τον αριθμό των κελιών. Στην 3η εικόνα με 5 μόνο κελιά η εικόνα είναι πολύ γενική και τα ειδικά χαρακτηριστικά έχουν χαθεί. Στην 1η εικόνα με 30 κελιά η κατανομή φαίνεται να έχει πολλές κορυφές. Από όλα αυτά τα γραφήματα κανείς μπορεί να αποκτήσει μια πολύ γενική άποψη πως απλά η κατανομή είναι ασύμμετρη και από εκεί και πέρα τα υπόλοιπα εξαρτώνται από τον τρόπο που κανείς θα διαλέξει τον αριθμό των κελιών.

Στο γράφημα 1.3, βλέπουμε 4 διαφορετικά ιστογράμματα με διαφορετική αρχή.



Γράφημα 1.2: Διάφορα ιστογράμματα με διαφορετικό πλήθος κελιών για τα ίδια δεδομένα.



Γράφημα 1.3: Διάφορα ιστογράμματα με διαφορετική αρχή και ίδιο πλήθος κελιών.

Στο συγκεκριμένο παράδειγμα, επειδή το μέγεθος του δείγματος είναι αρκετά μεγάλο, οι διαφοροποιήσεις στην εικόνα δεν είναι ιδιαίτερα μεγάλες. Σε μικρά όμως δείγματα η επιλογή της αρχής μπορεί να είναι πολύ κρίσιμη και άρα χρειάζεται ιδιαίτερη προσοχή.

Στη συνέχεια θα δούμε τις ιδιότητες του ιστογράμματος με σκοπό να καθορίσουμε μια επιλογή βέλτιστου παραθύρου έτσι ώστε να αποκομίσουμε μια όσο το δυνατόν πιο αντικειμενική εικόνα από τα δεδομένα μας. Για να γίνει αυτό χρειάζεται να ορίσουμε κάποιες ποσότητες. Έστω λοιπόν πως η αρχή των κελιών, δηλαδή το αριστερό άκρο του πρώτου κελιού συμβολίζεται με b_0 ενώ το j κελί είναι το $[b_{j-1}, b_j)$. Η συχνότητα του j κελιού είναι n_j και $h = b_j - b_{j-1}$ είναι το πλάτος του κελιού που πολλές φορές το ονομάζουμε και παράθυρο. Τέλος υποθέτουμε πως έχουμε k κελιά, επομένως το δεξιό άκρο του ιστογράμματος είναι το b_k. Με αυτούς τους ορισμούς μπορούμε να γράψουμε την εκτιμήτρια με τη μέθοδο του ιστογράμματος ως

$$\hat{f}(x) = \frac{1}{n}\frac{n_j}{h}, \quad x \in [b_{j-1}, b_j) \tag{1.1}$$

Ο συμβολισμός αυτός προσφέρει ευκολία στην απόδειξη των ιδιοτήτων του ιστογράμματος που θα περιγράψουμε αμέσως τώρα.

1.4.1 Ιδιότητες ιστογράμματος

Χρησιμοποιώντας τον συμβολισμό (1.1) μπορεί κάποιος να παρατηρήσει πως μόνο η ποσότητα n_j είναι τυχαία μεταβλητή. Επομένως

$$E(\hat{f}(x)) = E\left(\frac{n_j}{nh}\right) = \frac{1}{nh}E(n_j) = \frac{n\left[F(b_j) - F(b_{j-1})\right]}{nh} = \frac{\left[F(b_j) - F(b_{j-1})\right]}{h},$$

όπου $F(x)$ είναι η συνάρτηση κατανομής του πληθυσμού (ισχύει δηλαδή $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$). Επομένως ο αριθμητής στην αναμενόμενη τιμή δεν είναι παρά η πιθανότητα να πάρουμε μια παρατήρηση μέσα στο συγκεκριμένο κελί στο οποίο ανήκει το x σύμφωνα με την πραγματική συνάρτηση πυκνότητας πιθανότητας του πληθυσμού. Παρατηρείστε πως για όλες τις τιμές του x μέσα στο κελί η αναμενόμενη τιμή είναι η ίδια. Παρατηρούμε επομένως πως η εκτιμήτρια του ιστογράμματος είναι μεροληπτική καθώς γενικά διαφέρει από την $f(x)$. Η μεροληψία της είναι περίπου ίση με

$$Bias\left[\hat{f}(x)\right] \approx \frac{1}{2}f'(x)\left[h - 2(x - b_{j-1})\right],$$

όπου 'περίπου ίση' σημαίνει ότι υπάρχει μια ακόμα πολύ μικρή αμελητέα ποσότητα την οποία δεν αναφέρουμε για να απλοποιήσουμε το συμβολισμό. Καλό είναι να έχετε στο μυαλό σας πως αυτό σημαίνει ότι τα αποτελέσματα είναι προσεγγιστικά, δηλαδή δεν ισχύει αυστηρά η ισότητα αλλά υπάρχει και μια πολύ μικρή ποσότητα που για λόγους ευκολίας την αγνοούμε. Αυτό που βλέπουμε λοιπόν από τη μεροληψία της εκτιμήτριας με τη μέθοδο του ιστογράμματος είναι πως:

- θα πρέπει $f'(x) = 0$, ώστε να είναι αμερόληπτη η εκτιμήτρια. Αυτό σημαίνει πως η $f(x)$ πρέπει να είναι σταθερή στο διάστημα που μελετάμε (π.χ. η ομοιόμορφη κατανομή).

- η μεροληψία δεν εξαρτάται άμεσα από το μέγεθος του δείγματος και άρα ακόμα και για πολύ μεγάλα μεγέθη δείγματος συνεχίζει να είναι μεροληπτική η εκτιμήτρια μας.

Επίσης μπορεί να δειχθεί πως

$$Var\left[\hat{f}(x)\right] \approx \frac{f(x)}{nh},$$

και επομένως συνδυάζοντας τη διακύμανση με τη μεροληψία παίρνουμε πως

$$MSE\left[\hat{f}(x)\right] \approx \frac{1}{4}\left[f'(x)\right]^2\left[h - 2(x - b_{j-1})\right]^2 + \frac{f(x)}{nh}.$$

Αυτό που μπορεί να παρατηρήσει κανείς είναι πως όσο μεγαλώνει η τιμή του h, μικραίνει η διακύμανση αλλά μεγαλώνει η μεροληψία. Επομένως αν κάποιος θέλει να βρει κάποιο βέλτιστο h πρέπει να το κάνει με τέτοιο τρόπο ώστε να ισορροπήσει ανάμεσα στη μείωση της διακύμανσης και την αύξηση της μεροληψίας.

1.4.2 Εύρεση βέλτιστου πλάτους κελιού

Δεδομένου, ότι μας ενδιαφέρει όλος ο άξονας των x και όχι κάποια συγκεκριμένη τιμή, μια μέθοδος για να βρούμε ένα βέλτιστο h είναι να χρησιμοποιήσουμε το ολοκληρωμένο μέσο τετραγωνικό σφάλμα και να διαλέξουμε την τιμή του h που το ελαχιστοποιεί. Το ολοκληρωμένο μέσο τετραγωνικό σφάλμα δίνεται από τον τύπο

$$MISE\left[\hat{f}(x)\right] \approx \frac{R(f')h^2}{12} + \frac{1}{nh},$$

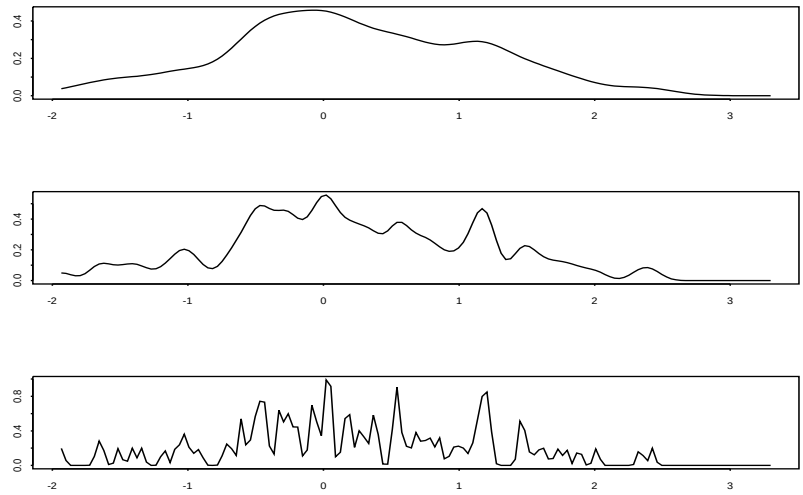
όπου $R(f') = \int \left[f'(u)\right]^2 du$.

Παίρνοντας την πρώτη παράγωγο του MISE ως προς h και εξισώνοντας την με 0 βρίσκουμε πως το βέλτιστο πλάτος κελιού h, που συμβολίζουμε με h_{opt} , είναι το

$$h_{opt} = \left[\frac{6}{R(f')}\right]^{1/3} n^{-1/3}.$$

Η ποσότητα $R(f')$ που εμφανίζεται στον παρονομαστή έχει να κάνει με τη μορφή της πραγματικής συνάρτησης πυκνότητας πιθανότητας, δηλαδή πόσο λεία είναι ή αν έχει μεγαλύτερα 'σκαμπανεβάσματα'. Στο γράφημα 1.4 που ακολουθεί μπορεί να δει κανείς 3 διαφορετικά γραφήματα με διάφορες συναρτήσεις f.

'Οπως βλέπει κανείς από πάνω προς τα κάτω η μορφή της f γίνεται ολοένα και λιγότερο λεία. Για την μορφή της f που είναι πάνω στο γράφημα 1.4, η τιμή της $R(f')$ είναι η μικρότερη. 'Αρα το βέλτιστο πλάτος κελιού, που εξαρτάται από την τιμή αυτή, θα είναι μεγαλύτερο σχετικά με τα άλλα. Στα δύο επόμενα γραφήματα βλέπουμε πως η $R(f')$ είναι μεγαλύτερη, άρα και τα σκαμπανεβάσματα της πραγματικής f είναι μεγαλύτερα, οπότε χρειαζόμαστε μικρότερο πλάτος κελιού, ώστε να μπορέσουμε να 'δούμε' καλύτερα τα σκαμπανεβάσματα της. Δηλαδή όταν η f είναι λεία (πχ μια μονοκόρυφη κατανομή) δεν χρειάζεται να πάρουμε μικρό παράθυρο αφού η γενική εικόνα της μπορεί να αναπαρασταθεί με παράθυρο μεγαλύτερου μεγέθους.



Γράφημα 1.4: Διάφορες δυνατές μορφές της $f(x)$. 'Οσο πιο λεία είναι η μορφή της τόσο μικρότερη είναι η τιμή της $R(f')$

Στην πράξη δουλεύουμε ως εξής:

Αν ξέρουμε πως η κατανομή του πληθυσμού είναι η κανονική, κάνοντας τις πράξεις βρίσκουμε πως το βέλτιστο παράθυρο είναι

$$h_{opt} = 3.491\sigma\ n^{-1/3},$$

όπου σ είναι η τυπική απόκλιση του πληθυσμού. 'Αρα στην περίπτωση που ξέρουμε πως ο πληθυσμός είναι κανονικός μπορούμε να χρησιμοποιήσουμε την δειγματική τυπική απόκλιση s και να βρούμε το βέλτιστο παράθυρο ως

$$h_{opt} = 3.491sn^{-1/3}.$$

Επειδή για δεδομένα από κανονικό πληθυσμό ισχύει πως

$$IQR = 1.345s,$$

όπου *IQR* είναι το ενδοτεταρτημοριακό εύρος, δηλαδή η διαφορά του τρίτου τεταρτημορίου από το πρώτο, μπορούμε εναλλακτικά να εκτιμήσουμε το βέλτιστο παράθυρο ως

$$h_{opt} = 3.491\frac{IQR}{1.345}n^{-1/3} = 2.595IQRn^{-1/3}.$$

Η χρήση του ενδοτεταρτημοριακού εύρους έχει το πλεονέκτημα ότι επειδή είναι πιο ανθεκτική (robust) σε αποκλίσεις από την κανονικότητα μπορεί να δώσει μια

καλύτερη εκτίμηση στην περίπτωση που η υπόθεση πως ο πληθυσμός είναι κανονικός δεν ισχύει.

Μια συμβιβαστική λύση για τις περιπτώσεις που ο πληθυσμός δεν είναι κανονικός αλλά η απόκλιση δεν είναι και πολύ μεγάλη (πχ η κατανομή t-student είναι μια συμμετρική στο 0 κατανομή, όπως και η κανονική, αλλά με κάπως παχύτερες ουρές, επομένως μπορεί κάποιος να ισχυριστεί πως μοιάζει και είναι κοντά στην κανονική κατανομή) είναι η χρήση της

$$\bar{s} = \min\left(\frac{IQR}{1.345}, s\right),$$

οπότε το βέλτιστο πλάτος κελιού είναι το

$$h_{opt} = 3.491\bar{s}n^{-1/3}.$$

Στην περίπτωση που η μορφή της κατανομής του πληθυσμού δεν μοιάζει καθόλου με την κανονική κατανομή, μια λύση είναι να χρησιμοποιήσουμε κάποια f που να μοιάζει με την κατανομή αυτή, να βρούμε την $R(f')$ και να υπολογίσουμε το βέλτιστο h. Για παράδειγμα αν τα δεδομένα φαίνεται να έχουν μια J-κλίση τότε η εκθετική μοιάζει μια επιλογή ενώ αν τα δεδομένα είναι δίκορφα μπορεί κάποιος να χρησιμοποιήσει μείγματα της κανονικής κατανομής. Είναι ξεκάθαρο ότι χρειάζεται αρκετή πείρα για να αποφασίσει κάποιος ποια κατανομή θα μπορούσε να μοιάζει με τα δεδομένα.

Η προσέγγισή μας στηρίχτηκε στην εύρεση της βέλτιστης τιμής του πλάτους του κελιού. Κάποιες άλλες προσεγγίσεις έχουν υπολογίσει ότι για την περίπτωση κανονικού πληθυσμού ο βέλτιστος αριθμός κελιών για το ιστόγραμμα πρέπει να είναι μεγαλύτερος από $(2n)^{1/3}$ το οποίο αντιστοιχεί σε βέλτιστο παράθυρο $h_{opt} \leq 3.55\sigma n^{-1/3}$ δηλαδή πολύ κοντά στο βέλτιστο πλάτος του κελιού που βρήκαμε ελαχιστοποιώντας το μέσο ολοκληρωμένο τετραγωνικό σφάλμα.

1.4.3 Εύρεση βέλτιστης αρχής για το ιστόγραμμα

Η επιλογή της αρχής του ιστογράμματος είναι κρίσιμη μόνο όταν το μέγεθος του δείγματος είναι μικρό. Επίσης επειδή έχει δειχτεί ότι δεν είναι τόσο σημαντική η επιλογή του ως προς το μέσο ολοκληρωμένο τετραγωνικό σφάλμα μια εμπειρική επιλογή είναι να ξεκινήσουμε να φτιάχνουμε τα κελιά που χρειαζόμαστε από το $\min(x_i) - h/2$.

'Ετσι για παράδειγμα αν η μικρότερη παρατήρηση είναι 10, η μεγαλύτερη είναι 200 και το βέλτιστο πλάτος κελιού είναι το 10, τότε το πρώτο διάστημα είναι το $(5, 15]$ και ακολούθως τα κελιά θα είναι $(15, 25], (25, 35], \ldots, (185, 195]$ και το τελευταίο $(195, 205]$.

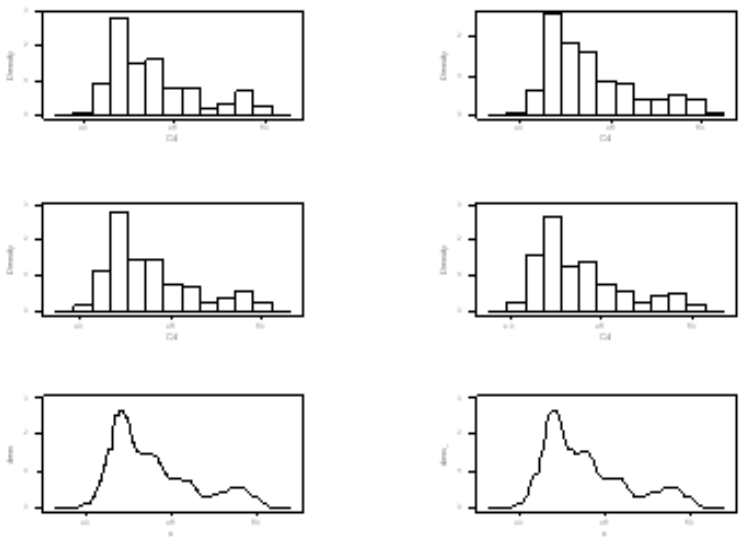
'Οπως θα δούμε σε λίγο, με σκοπό να περιορίσουμε την όποια επίδραση της επιλογής της αρχής μπορούμε να χρησιμοποιήσουμε διαφορετικές τιμές ως αρχή του ιστογράμματος και στη συνέχεια να πάρουμε το μέσο όρο τους ως εκτιμήτρια. Με αυτό τον τρόπο η εκτιμήτρια που προκύπτει είναι πιο λεία.

1.4.4 Η μέθοδος averaging Histograms

'Ενας τρόπος για να ξεπεράσουμε τα προβλήματα που δημιουργεί η επιλογή της αρχής του ιστογράμματος είναι να ξεκινήσουμε, κρατώντας το ίδιο πλάτος κελιού,

από διαφορετικές αρχές και στη συνέχεια να χρησιμοποιήσουμε ως εκτιμήτρια τη μέση τιμή που θα πάρουμε από τις διαφορετικές εκτιμήσεις με κάθε αρχή. Με αυτό τον τρόπο εξαφανίζουμε την οποιαδήποτε επίδραση από την επιλογή της αρχής και συγχρόνως καταλήγουμε σε μια πιο λεία μορφή της εκτιμήτριας με το ιστόγραμμα. Για να γίνει αυτό κατανοητό δημιουργήσαμε τις εκτιμήτριες με τη μέθοδο του ιστογράμματος για τα δεδομένα με τα γιαούρτια. Συγκεκριμένα θεωρήσαμε σταθερό το πλάτος των κελιών και ίσο με 0.10 αλλά η αρχή τέθηκε διαδοχικά $3.80, 3.82, 3.84$ και 3.86. Τα 4 πρώτα ιστογράμματα του γραφήματος 1.5 αντικατοπτρίζουν την εικόνα αυτή. Παρατηρείστε πως μεταξύ τους παρουσιάζουν οπτικές διαφορές που αντιστοιχούν σε διαφορετικές εκτιμήσεις αλλά και σε διαφορετική εικόνα που παίρνουμε κάθε φορά. Στη συνέχεια (κάτω αριστερά) πήραμε τη μέση τιμή αυτών των τεσσάρων εκτιμήσεων. Μπορεί κανείς να δει πως η εικόνα είναι αρκετά πιο καθαρή πια, αλλά ακόμα υπάρχουν πηδηματάκια της εκτιμήτριας. Είναι ξεκάθαρη πια η ύπαρξη δύο κορυφών κάτι που με κάποιες επιλογές αρχής του ιστογράμματος δεν ήταν προφανές. Στο τελευταίο γράφημα κάτω δεξιά έχουμε πάρει τη μέση τιμή 15 διαφορετικών ιστογραμμάτων με ίδιο παράθυρο αλλά διαφορετικές αρχές. Η εικόνα είναι ακόμα πιο λεία και τελικά εξαφανίζει δυο από τα προβλήματα του ιστογράμματος, αυτό της επιλογής παραθύρου αλλά και αυτό της όχι λείας εκτιμήτριας.

Η μέθοδος αυτή ονομάζεται στη βιβλιογραφία ως averaging histograms και εκτός από τις προφανείς ιδιότητές της που συζητήσαμε μειώνει τη μεροληψία και τη διακύμανση της εκτιμήτριας.



Γράφημα 1.5: Ιστόγραμμα με διαφορετική αρχή και ίδιο πλάτος κελιών και στην κάτω σειρά average ιστόγραμμα.

1.4.5 Συμπεράσματα

Συνοψίζοντας, τα πλεονεκτήματα του ιστογράμματος είναι

- Δεν χρειάζεται να κάνουμε καμιά σοβαρή υπόθεση για να το κατασκευάσουμε, παρά μόνο να έχουμε μια γενική εικόνα για τα δεδομένα μας
- Είναι σχετικά εύκολο και δεν χρειάζεται ειδικά προγράμματα για να το κατασκευάσουμε. 'Ολα τα στατιστικά πακέτα το κατασκευάζουν πολύ εύκολα και γρήγορα και μπορεί κάποιος πολύ εύκολα να το περιγράψει και να το εξηγήσει χωρίς απαραίτητα να έχει μεγάλες γνώσεις στατιστικής.
- Τα παραπάνω σχόλια αφορούν κυρίως το περιγραφικό μέρος του ιστογράμματος ως μια εικόνα που μας περιγράφει τα δεδομένα.

Από την άλλη κάποια σημαντικά μειονεκτήματα του είναι πως

- Δεν δίνει μια λεία εικόνα καθώς μέσα σε κάθε κελί η συνάρτηση πυκνότητας πιθανότητα είναι η ίδια και γενικά η εικόνα που παίρνουμε έχει 'πηδηματάκια'. Επειδή λοιπόν σκοπός της εκτίμησης μιας πυκνότητας πιθανότητας είναι να χρησιμοποιήσουμε την παράγωγό της η μέθοδος του ιστογράμματος αποτυγχάνει
- Δύσκολα γενικεύεται σε περισσότερες από μια διαστάσεις
- Παρατηρείστε πως τα μειονεκτήματα έχουν να κάνουν κυρίως με τη στατιστική χρήση του ιστογράμματος πέρα από την απλή περιγραφή των δεδομένων.

1.5 Εκτίμηση με τη μέθοδο του πολυγώνου

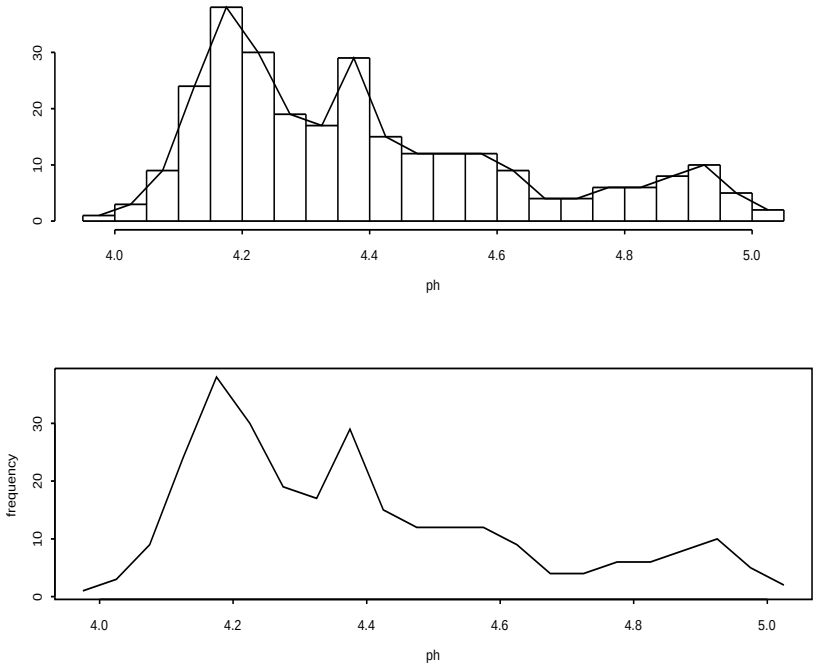
Το πολύγωνο συχνοτήτων προκύπτει αν ενώσουμε με μια ευθεία γραμμή τις κορυφές των ράβδων του ιστογράμματος ακριβώς στο μέσο του κάθε κελιού. Στο γράφημα 1.6 που ακολουθεί μπορείτε να δείτε πρώτα ένα ιστόγραμμα μαζί με το πολύγωνο συχνοτήτων και δίπλα μόνο το πολύγωνο συχνοτήτων για τα δεδομένα του παραδείγματος 1.2 που αφορούν τα γιαούρτια.

Το πολύγωνο συχνοτήτων είναι γνωστό κυρίως ως εναλλακτικός τρόπος να πάρει κανείς κάποια εικόνα από τα δεδομένα του, αλλά η χρήση του είναι περιορισμένη.

Η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας με τη μέθοδο του πολυγώνου είναι η

$$\tilde{f}(x) = \frac{1}{nh^2}\left[n_{j-1}c_j - n_j c_{j-1} + (n_j - n_{j-1})x\right], \quad x \in [c_{j-1}, c_j] \qquad (1.2)$$

όπου c_j είναι το κεντρικό σημείου του κελιού j. Δηλαδή έχουμε πως



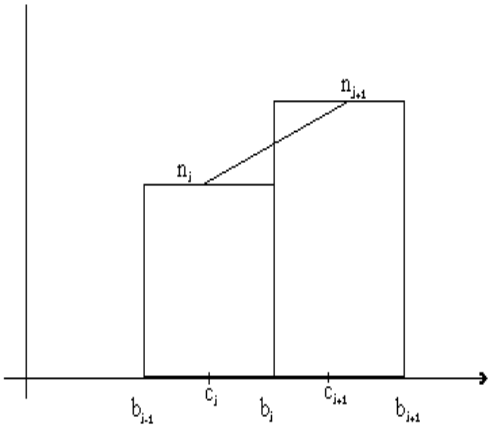
Γράφημα 1.6: Πολύγωνο συχνοτήτων για τα δεδομένα που αφορούν τα γιαούρτια.

	Διάστημα	Συχνότητα	Κεντρικό σημείο κελιού
1ο κελί	$(b_0, b_1]$	n_1	c_1
2ο κελί	$(b_1, b_2]$	n_2	c_2
	$\dots$	$\dots$	$\dots$
j κελί	$(b_{j-1}, b_j]$	n_j	c_j
	$\dots$	$\dots$	$\dots$
k κελί (τελευταίο κελί)	$(b_{k-1}, b_k]$	n_k	c_k

Μπορεί κάποιος να δει πως

$$\frac{n_{j-1}c_j - n_j c_{j-1} + (n_j - n_{j-1})x}{h}$$

είναι η εξίσωση ευθείας που συνδέει τα σημεία c_{j-1} και c_j. Σχηματικά αυτό περιγράφεται στο γράφημα 1.7 που ακολουθεί



Γράφημα 1.7: Σχηματική απεικόνιση της εκτίμησης με τη μέθοδο του πολυγώνου. Η εκτιμήτρια είναι απλά η εξίσωση της ευθείας που ενώνει τα δύο σημεία.

Με όμοιο τρόπο όπως και για το ιστόγραμμα μπορούμε να εξετάσουμε τις ιδιότητες και αυτού του εκτιμητή. Κάτι τέτοιο όμως δεν θα περιγραφεί εδώ. Αυτό που είναι χρήσιμο είναι πως ελαχιστοποιώντας το μέσο ολοκληρωμένο τετραγωνικό σφάλμα μπορεί να δειχτεί πως το βέλτιστο μήκος κελιού είναι το

$$h_{opt} = 2\left[\frac{15}{49R(f'')}\right]^{1/5} n^{-1/5}$$

Χρησιμοποιώντας πάλι την υπόθεση ότι ο πληθυσμός είναι κανονικός το βέλτιστο μήκος κελιού είναι το $h_{opt} = 2.15\sigma n^{-1/5}$ και επειδή η διακύμανση του πληθυσμού είναι άγνωστη χρησιμοποιούμε τη δειγματική τυπική απόκλιση s ή το ενδοτεταρτημοριακό εύρος κατάλληλα πολλαπλασιασμένο. 'Ετσι,

αν η κατανομή είναι κανονική, υπολογίζουμε $h_{opt} = 2.15\ s\ n^{-1/5}$ ενώ

αν αποκλίνει από την κανονική παίρνουμε $h_{opt} = 2.15\ \bar{s}\ n^{-1/5}$ όπου $\bar{s} = \min\left(\frac{IQR}{1.345}, s\right)$.

Συμπερασματικά μπορούμε να πούμε για το πολύγωνο συχνοτήτων ότι

- Η εκτίμηση μιας πυκνότητας με το πολύγωνο συχνοτήτων αν και απλή μπορεί να δώσει μια πολύ γενική εικόνα για τα δεδομένα
- 'Εχει αποδειχθεί θεωρητικά ότι, όσο το μέγεθος του δείγματος αυξάνει το πολύγωνο συχνοτήτων είναι προτιμότερο του ιστογράμματος (με κριτήριο το γεγονός ότι το μέσο ολοκληρωμένο τετραγωνικό σφάλμα είναι μικρότερο). Για μικρότερα δείγματα και οι δύο μέθοδοι δίνουν παρόμοια αποτελέσματα
- Επομένως το πολύγωνο συχνοτήτων σε καμιά περίπτωση δεν υπολείπεται του ιστογράμματος το οποίο χρησιμοποιείται πολύ περισσότερο στην πράξη
- Δεν μπορούμε να χρησιμοποιήσουμε το πολύγωνο για να βρούμε την κορυφή ή τις κορυφές των δεδομένων μας.
- Και πάλι (όπως και στην περίπτωση του ιστογράμματος) η αρχή από όπου θα ξεκινά το πρώτο κελί είναι σημαντική μόνο για μικρά δείγματα.

1.6 Απλοϊκός Εκτιμητής

Από τον ορισμό της συνάρτησης πυκνότητας πιθανότητας ως το όριο της συνάρτησης κατανομής προκύπτει πως

$$f(x) = \lim_{h \to 0} \frac{P(x-h < X < x+h)}{2h}.$$

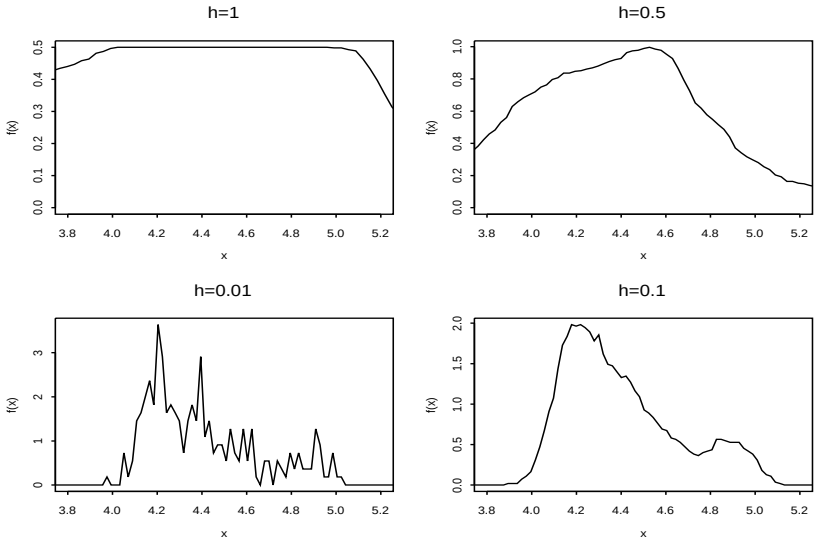
Επομένως φαίνεται λογικό να εκτιμήσουμε μια συνάρτηση πυκνότητας πιθανότητας χρησιμοποιώντας το δειγματικό ανάλογο, δηλαδή να ορίσουμε ένα μικρό διάστημα και να μετρήσουμε τον αριθμό των παρατηρήσεων μέσα σε αυτό. 'Ετσι ένας λογικός εκτιμητής της συνάρτησης πυκνότητας πιθανότητας είναι ο

$$\hat{f}(x) = \frac{1}{n} \frac{\#\text{παρατηρήσεων στο διάστημα} \quad (x-h, x+h)}{2h}$$

όπου h είναι ένα παράθυρο που καθορίζει πόσο μικρό ή μεγάλο είναι το διάστημα αυτό.

Ο εκτιμητής αυτός ονομάζεται απλοϊκός (naive) επειδή προκύπτει από αυτή την απλή μετάβαση από τον ορισμό της συνάρτησης πυκνότητας πιθανότητας στο δειγματικό αντίστοιχό της.

Στο Γράφημα 1.8 παρατηρείστε την επίδραση και τη σπουδαιότητα της επιλογής του παραθύρου στο οπτικό αποτέλεσμα. Βλέπουμε λοιπόν 4 διαφορετικές επιλογές παραθύρου h για τον απλοϊκό εκτιμητή. Τα δεδομένα που περιγράφουν είναι πάλι τα δεδομένα που αφορούν τα γιαούρτια. Στην πρώτη περίπτωση $h = 1$ και η εικόνα είναι μια σχεδόν ομοιόμορφη κατανομή σε όλο το διάστημα. Το παράθυρο ήταν πολύ μεγάλο οπότε σχεδόν όλες οι παρατηρήσεις έπεφταν μέσα στο παράθυρο.



Γράφημα 1.8: Εκτίμηση της συνάρτησης πυκνότητας πιθανότητας με τη χρήση του απλοϊκού εκτιμητή για 4 διαφορετικές επιλογές του h.

Στην εικόνα 3 με το μικρότερο παράθυρο $h = 0.01$ έχουμε μια πολύ 'τρεμουλιαστή' εικόνα με πολλά τοπικά ακρότατα. Αν αναλογιστεί κανείς πως τα δεδομένα μας είχαν μόνο 2 δεκαδικά ψηφία καταλαβαίνει κανείς γιατί η εικόνα έχει αυτή τη μορφή. Στην εικόνα 4 έχουμε κατά κάποιον τρόπο την πιο καθαρή εικόνα για τα δεδομένα μας.

Η διαφορά από το ιστόγραμμα είναι πως τώρα πια δεν έχουμε συγκεκριμένα κελιά και μετράμε πόσες παρατηρήσεις πέφτουν μέσα σε αυτά αλλά μετράμε πόσες παρατηρήσεις είναι μέσα σε συγκεκριμένη απόσταση από την τιμή μας. Δηλαδή τοποθετούμε ένα παραλληλόγραμμο μήκους $2h$ με κέντρο κάθε τιμή x και μετράμε πόσες παρατηρήσεις είναι μέσα σε αυτό το παραλληλόγραμμο.

Εναλλακτικά ο απλοϊκός εκτιμητής μπορεί να περιγραφεί ως εξής: βάζουμε σε κάθε μια παρατήρηση ένα παραλληλόγραμμο με πλάτος $2h$ και για κάθε τιμή x μετράμε πόσα παραλληλόγραμμα φτάνουν μέχρι αυτή.

Αυτό μπορούμε να το εκφράσουμε μαθηματικά ως εξής

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left(\frac{x - x_i}{h}\right) \tag{1.3}$$

όπου

$$w(x) = \begin{cases} 1/2 & \text{αν } |x| < 1 \\ 0 & \text{αλλού} \end{cases}$$

Μπορεί κανείς να παρατηρήσει πως η $w(x)$ είναι η ομοιόμορφη κατανομή στο διάστημα (-1,1) για αυτό και χρησιμοποιήσαμε τον όρο κουτάκια (η συνάρτηση πυκνότητας πιθανότητας της ομοιόμορφης κατανομής μοιάζει με κουτάκι).

Ο απλοϊκός εκτιμητής έχει το μειονέκτημα πως η εκτίμηση που δίνει έχει πηδηματάκια, όπως και στο ιστόγραμμα, κάτι που σημαίνει πως δεν υπάρχουν οι παράγωγοι σε όλα τα σημεία. Αυτά τα πηδηματάκια οφείλονται στη κορυφή της $w(x)$ και επομένως αν διαλέξουμε μια άλλη μορφή μπορούμε να τα ξεπεράσουμε. Η ανάγκη δηλαδή να γενικεύσουμε είναι προφανής. Αυτό αποτελεί και τη βάση για τη χρήση των kernels που θα δούμε αμέσως.

1.7 Εκτίμηση με τη χρήση Kernel

1.7.1 Εισαγωγή

Η κακή συμπεριφορά του απλοϊκού εκτιμητή οφείλεται κυρίως στην επιλογή της ομοιόμορφης κατανομής. Η μορφή της ομοιόμορφης είναι ένα παραλληλόγραμμο. Αν λοιπόν διαλέξουμε μια άλλη μορφή για τη συνάρτηση $w(x)$ μπορούμε να πάρουμε μια πολύ καλύτερη εικόνα. Η μέθοδος των kernels χρησιμοποιεί ένα kernel $K(x)$ με όμοιο τρόπο όπως χρησιμοποιούσε ο απλοϊκός εκτιμητής τη συνάρτηση $w(x)$. Επομένως η εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας με τη χρήση του kernel $K(x)$ δίνεται ως

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\,K\left(\frac{x-x_i}{h}\right) \tag{1.4}$$

όπου $K(x)$ είναι μια συνάρτηση (συνήθως συνάρτηση πυκνότητας πιθανότητας) για την οποία ισχύει

1. $K(x) \geq 0$,
2. $\int_{-\infty}^{+\infty} K(x)dx = 1$,
3. $\int_{-\infty}^{+\infty} xK(x)dx = 0$ και
4. $0 < \int_{-\infty}^{+\infty} x^2K(x)dx < \infty$.

Χωρίς να είναι απαραίτητο, είναι όμως χρήσιμο, η συνάρτηση $K(x)$ είναι συμμετρική ως προς το 0.

Μπορεί εύκολα κάποιος να δει πως, με βάση τον παραπάνω ορισμό, η εκτίμηση της συνάρτησης πυκνότητας είναι όντως μια συνάρτηση πυκνότητας πιθανότητας. Αυτό γιατί

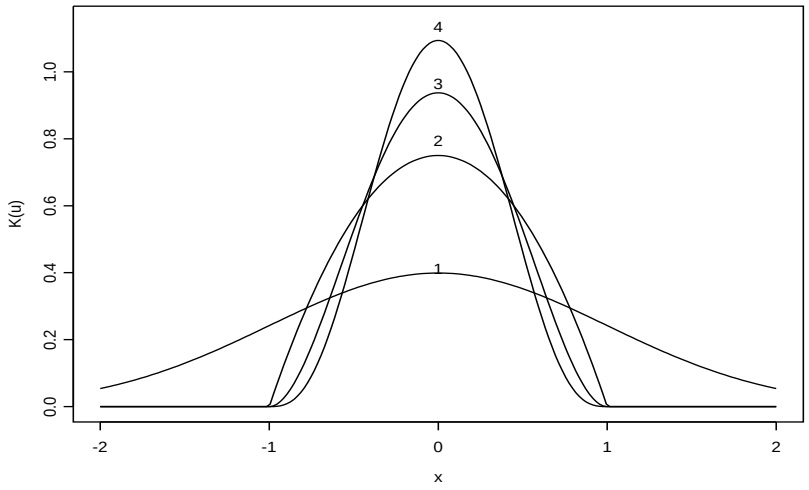
- είναι σίγουρα μη αρνητική αφού όλες οι ποσότητες είναι θετικές και

- ισχύει πως

$$\begin{aligned}\int_{-\infty}^{+\infty} \hat{f}(x)dx &= \int_{-\infty}^{+\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\, K\left(\frac{x-x_i}{h}\right)dx \\ &= \frac{1}{n}\frac{1}{h}\sum_{i=1}^{n} \int_{-\infty}^{+\infty} K\left(\frac{x-x_i}{h}\right)dx\end{aligned}$$

και αντικαθιστώντας $u = \frac{x-x_i}{h}$, προκύπτει πως

$$\int_{-\infty}^{+\infty} \hat{f}(x)dx = \frac{1}{n}\frac{1}{h}\sum_{i=1}^{n} h \int_{-\infty}^{+\infty} K\,(u)du = \frac{n}{n} = 1$$



Γράφημα 1.9: Διάφορα kernels. 1- Κανονικό, 2 -Epanenchnikov, 3-Biweight, 4-Triweight. Παρατηρείστε ότι τελικά οι διαφορές είναι αρκετά μεγάλες, όπως επίσης και πως το κανονικό kernel έχει πολύ μεγαλύτερες ουρές

Στο γράφημα 1.9 μπορεί να δει κανείς μερικές από τις πιθανές επιλογές kernel. Παρατηρείστε πως οι διαφορές δεν είναι αμελητέες και έχουν να κάνουν με την κύρτωση. Στον πίνακα 1.1 μπορεί να δει κανείς τον μαθηματικό ορισμό αυτών των kernels. Το κανονικό (Gaussian) kernel είναι το μόνο που δίνει θετικό βάρος έξω από το διάστημα $(-1, 1)$, έχει δηλαδή μεγαλύτερες ουρές. Αναγνωρίζει κανείς πως πρόκειται απλά για την συνάρτηση πυκνότητας πιθανότητας της τυποποιημένης κανονικής κατανομής.

Kernel	K(u)
Epanenchnikov	$\frac{3}{4}(1-u^2), \quad \lvert u\rvert \leq 1$
Biweight	$\frac{15}{16}(1-u^2)^2, \quad \lvert u\rvert \leq 1$
Triweight	$\frac{35}{32}(1-u^2)^3, \quad \lvert u\rvert \leq 1$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right), \quad u \in \Re$
Uniform	$\frac{1}{2} \quad \lvert u\rvert \leq 1$

Πίνακας 1.1: Μερικά από τα πιο διαδεδομένα kernels και η μορφή τους

Καταλαβαίνει κανείς πως στην περίπτωση της εκτίμησης με kernels, υπάρχουν δύο σημεία επιλογής του ερευνητή:

Το πρώτο αφορά την επιλογή του παραθύρου. Διαλέγοντας μεγάλο παράθυρο h καταλήγουμε στο να πάρουμε μια πολύ λεία εκτίμηση ενώ αν το παράθυρο h είναι μικρό η εικόνα της εκτίμησης είναι όλο "πηδηματάκια". Επομένως κανείς πρέπει να βρει ένα βέλτιστο παράθυρο το οποίο να ισορροπήσει αυτές τις δύο ακραίες περιπτώσεις, δηλαδή να είναι μικρό αλλά όχι πολύ μικρό.

Το δεύτερο αφορά την επιλογή του kernel. 'Οπως θα δούμε παρακάτω, η επιλογή του kernel δεν είναι τόσο καθοριστική καθώς σχεδόν όλα δίνουν παρόμοια εικόνα (εκτός ίσως από κάποιες εξαιρέσεις) δεδομένου ότι χρησιμοποιούμε ένα βέλτιστο κάθε φορά παράθυρο.

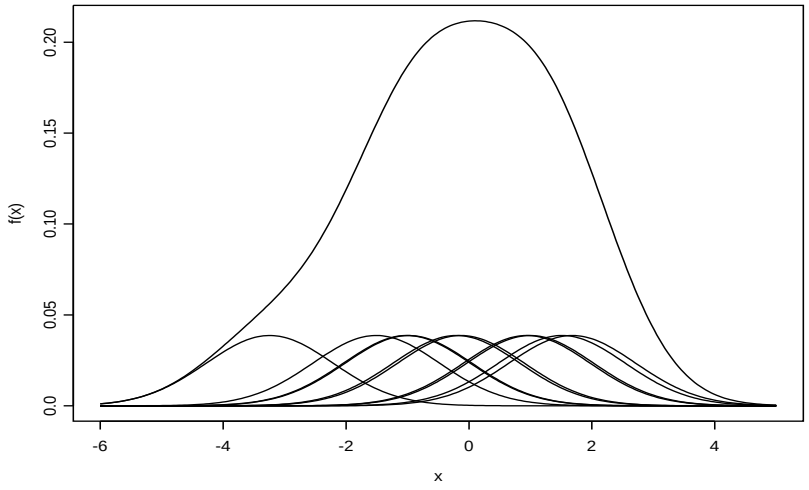
Πριν προχωρήσουμε λοιπόν με τα προβλήματα αυτά και πριν δούμε τις ιδιότητες της εκτιμήτριας με τη χρήση των kernel, ας ρίξουμε μια ματιά πως στην πραγματικότητα χρησιμοποιούμε τα kernel.

Σε κάθε παρατήρηση που έχουμε στο δείγμα τοποθετούμε ένα kernel με κέντρο την παρατήρηση αυτή. Τα δεδομένα μας αφορούν 10 μόνο παρατηρήσεις. Στο παράδειγμα χρησιμοποιούμε το κανονικό kernel αλλά προφανώς η ερμηνεία είναι η ίδια και για τα υπόλοιπα kernels. Δεδομένου πως κάθε kernel είναι συμμετρικό και έχει κορυφή στο 0 (αν και όπως είπαμε κάτι τέτοιο δεν είναι αναγκαίο, αλλά σε κάθε περίπτωση είναι λογικό) αυτό σημαίνει πως δίνουμε μεγαλύτερο βάρος ακριβώς σε αυτό το σημείο και συμμετρικά όσο απομακρυνόμαστε από κάθε παρατήρηση λιγότερο βάρος στα γειτονικά σημεία. Απομακρυσμένα σημεία δεν έχουν καθόλου ή έχουν αμελητέο βάρος. Η εκτίμηση λοιπόν σε κάθε σημείο είναι το άθροισμα αυτών των βαρών που προκύπτουν από την "τοποθέτηση" των kernels με κέντρο τις παρατηρήσεις. Αυτό μπορείτε να το δείτε στο γράφημα 1.10 όπου υπάρχει αφενός η εκτιμήτρια αλλά και οι επιμέρους όροι του αθροίσματος.

1.7.2 Ιδιότητες της εκτιμήτριας με τη χρήση των kernels

Η αναμενόμενη τιμή της εκτιμήτριας δίνεται από

$$E\left[\hat{f}(x)\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)\right] =$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}E\left[K\left(\frac{x-x_i}{h}\right)\right]$$



Γράφημα 1.10: Σχηματική απεικόνιση της μεθόδου των kernels

Επειδή όμως όλα τα X_i είναι ισόνομες και ανεξάρτητες μεταβλητές (έχουν την ίδια μέση τιμή) χρησιμοποιώντας για όλα τα x_i το συμβολισμό y βρίσκουμε πως

$$E\left[\hat{f}(x)\right] = \frac{1}{h}E\left[K\left(\frac{x-x_1}{h}\right)\right] = \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{x-y}{h}\right)f(y)dy$$

Κάνοντας αλλαγή μεταβλητής $\frac{x-y}{h} = u$ μπορεί να βρεθεί πως

$$E\left[\hat{f}(x)\right] = \int_{-\infty}^{+\infty} K(u)f(x+hu)du$$

Αυτό λοιπόν που βλέπουμε είναι πως:

- η αναμενόμενη τιμή είναι μια ομαλοποιημένη (smoothed) έκδοση της f και όχι γενικά η ίδια η f. Αυτό σημαίνει πως η εκτιμήτρια με τη χρήση kernels είναι μεροληπτική.

- η μεροληψία δεν εξαρτάται από το μέγεθος του δείγματος (παρατηρείστε πως το μέγεθος του δείγματος δεν υπάρχει πουθενά στον τύπο της αναμενόμενης τιμής). Δηλαδή ακόμα και για πολύ μεγάλα δείγματα μπορεί να υπάρχει μεροληψία.

Χρησιμοποιώντας προσεγγιστικά αποτελέσματα μπορεί να δειχτεί πως

$$Bias\left[\hat{f}(x)\right] \approx \tfrac{1}{2}h^2 f''(x)\sigma_k^2$$
$$Var\left[\hat{f}(x)\right] \approx \tfrac{f(x)R(K)}{nh}$$

όπου σ_k^2 είναι η διακύμανση του kernel που χρησιμοποιήσαμε και $f(x)$ είναι η πραγματική συνάρτηση πυκνότητας πιθανότητας την οποία συνήθως δεν γνωρίζουμε.

Αυτό που είναι πολύ σημαντικό να παρατηρήσει κανείς είναι πως όσο μεγαλώνει το παράθυρο h τόσο η μεν μεροληψία μεγαλώνει αλλά η διακύμανση μικραίνει. Επίσης ισχύει και το αντίστροφο δηλαδή αν μικρύνουμε το παράθυρο τότε μεγαλώνει η διακύμανση αλλά μικραίνει η μεροληψία.

Επομένως στην πραγματικότητα πρέπει να διαλέξει κανείς ένα μικρό παράθυρο αλλά όχι πολύ μικρό. Για την επιλογή του παραθύρου θα μιλήσουμε στη συνέχεια.

Είναι πολύ χρήσιμο να δούμε πως η εκτιμήτρια με τη μέθοδο των kernels είναι η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής $Z = X + Y$, όπου η τυχαία μεταβλητή X είναι μια διακριτή κατανομή με πιθανότητα $1/n$ σε καθένα από τα σημεία X_i και μηδέν σε κάθε άλλο σημείο ενώ η τυχαία μεταβλητή Y ακολουθεί την κατανομή $K(y)$, δηλαδή την κατανομή του kernel και X, Y είναι ανεξάρτητες τυχαίες μεταβλητές.

Παράδειγμα: Ας δούμε ένα παράδειγμα για να κατανοήσουμε καλύτερα τη μεροληψία της $\hat{f}(x)$. 'Οπως είδαμε πριν η αναμενόμενη τιμή μπορεί να γραφτεί ως

$$E\left[\hat{f}(x)\right] = \int_{-\infty}^{+\infty} \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)du$$

Ας υποθέσουμε πως γνωρίζουμε την κατανομή του πληθυσμού, δηλαδή την f και πως αυτή είναι η τυποποιημένη κανονική. Μπορεί κανείς να παρατηρήσει πως σε αυτή την περίπτωση η αναμενόμενη τιμή είναι η κατανομή του αθροίσματος (convolution) δύο τυχαίων μεταβλητών, η πρώτη ακολουθεί την fενώ η δεύτερη ακολουθεί την κατανομή της τυχαίας μεταβλητής $Y = hu$ όπου u είναι μια τυχαία μεταβλητή από την κατανομή του kernel. Ας υποθέσουμε για ευκολία πως το kernel είναι το κανονικό, επομένως με ότι είδαμε η τυχαία μεταβλητή θα ακολουθεί μια $N(0, h^2)$ κατανομή. Συνεπώς η αναμενόμενη τιμή θα είναι το άθροισμα δύο κανονικών τυχαίων μεταβλητών, εκ των οποίων η μια είναι τυποποιημένη και η άλλη ακολουθεί την $N(0, h^2)$κατανομή. 'Οπως ξέρουμε η κατανομή αυτού του αθροίσματος είναι η $N(0, h^2 + 1)$ κατανομή και επομένως για αυτήν την περίπτωση

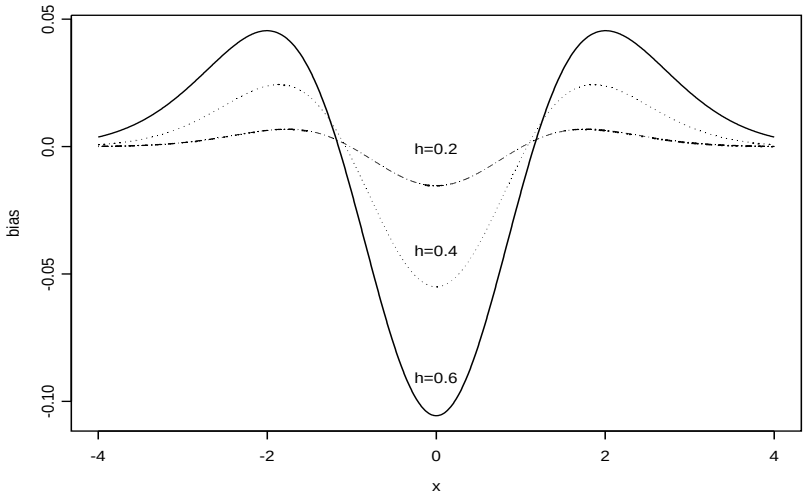
$$E\left[\hat{f}(x)\right] = \frac{1}{\sqrt{2\pi}\sqrt{h^2+1}} \exp\left(-\frac{x^2}{h^2+1}\right)$$

και επομένως αν με $\phi(x)$ συμβολίσουμε τη συνάρτηση πυκνότητας πιθανότητας της τυποποιημένης κανονικής κατανομής η μεροληψία θα είναι

$$Bias\left[\hat{f}(x)\right] = \frac{1}{\sqrt{h^2+1}}\phi\left(\frac{x}{\sqrt{h^2+1}}\right) - \phi(x)$$

Αν η τιμή του h είναι 0, τότε η εκτιμήτρια είναι αμερόληπτη αλλά μηδενική τιμή υπονοεί πως θα πάρουμε ως εκτίμηση την εμπειρική σχετική συχνότητα, που δεν είναι χρήσιμη αφού θα δίνει πιθανότητα μόνο στα σημεία που έχουμε παρατηρήσεις.

Το γράφημα 1.11 αναπαριστά τη μεροληψία για το παράδειγμα μας για τρεις διαφορετικές επιλογές παραθύρου, h=0.2, 0.4 και 0.6.



Γράφημα 1.11: Η μεροληψία της εκτιμήτριας με kernels για 3 διαφορετικές επιλογές παραθύρου. Η κατανομή του πληθυσμού είναι η τυποποιημένη κανονική κατανομή και το kernel που χρησιμοποιήσαμε είναι το κανονικό kernel.

Από το γράφημα 1.11 μπορεί κανείς να δει πως η μεροληψία μεγαλώνει γύρω από το 0, γύρω από την κορυφή της κατανομής. Επίσης παρατηρείστε πως έχουμε και σημεία με θετική αλλά και με αρνητική μεροληψία, και πως το παράθυρο παίζει σημαντικό ρόλο ως προς τη μεροληψία. Ουσιαστικά αυτό το οποίο δείξαμε είναι πως για να πάρουμε αμερόληπτη εκτιμήτρια θα πρέπει η κατανομή του πληθυσμού να είναι τέτοια ώστε η συνέλιξη της (convolution) με κάποια άλλη κατανομή (kernel) να είναι η ίδια η κατανομή του πληθυσμού, κάτι που δεν μπορεί να συμβεί εκτός από πολύ ειδικές περιπτώσεις. Παρατηρήστε επίσης πως για μεγαλύτερο h η μεροληψία (σε απόλυτη τιμή) είναι μεγαλύτερη σε κάθε περίπτωση.

1.7.3 Ροπές και άλλα χαρακτηριστικά της εκτιμήτριας

Ας δούμε ποιά ειναι η αναμενόμενη τιμή $E(X)$ και η διακύμανση $Var(X)$ της τμ X που ακολουθεί την $\hat{f}(x)$, δηλαδή την εκτιμήτρια της σππ από ένα δείγμα.

Θα ισχύει πως

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\hat{f}(x)dx \\ Var(X) &= E(X^2) - E(X)^2 \end{aligned}$$

Ας βρούμε πρώτα την $E(X)$. Θα ισχύει πως

$$E(X) = \int_{-\infty}^{\infty} x\hat{f}(x)dx = \int_{-\infty}^{\infty} x\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)dx$$

Αλλάζοντας τη σειρά ολοκληρώματος και αθροίσματος

$$E(X) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x\frac{1}{h}K\left(\frac{x-x_i}{h}\right)dx$$

Θέτοντας $u = \frac{x-x_i}{h}$ προκύπτει πως $dx\frac{1}{h} = du$ και συνεπώς το ολοκλήρωμα παίρνει τη μορφή

$$E(X) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(x_i + uh)K(u)du =$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\int_{-\infty}^{\infty} x_iK(u)du + \int_{-\infty}^{\infty} uhK(u)du\right)$$

επειδή όμως εξ' ορισμού το kernel έχει αναμενόμενη τιμή 0, προκύπτει πως

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

δηλαδή ταυτίζεται με τη μέση τιμή του δείγματος.

Για τον υπολογισμό της διακύμανσης είναι πιο απλό να βρούμε πρώτα τη δεύτερη ροπή. Χρησιμοποιώντας παρόμοια βήματα προκύπτει πως

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2\hat{f}(x)dx = \int_{-\infty}^{\infty} x^2\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x^2\frac{1}{h}K\left(\frac{x-x_i}{h}\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(x_i+uh)^2K(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(x_i^2+2uhx_i+u^2h^2)K(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\int_{-\infty}^{\infty} x_i^2K(u)du + 2hx_i\int_{-\infty}^{\infty} uK(u)du + h^2\int_{-\infty}^{\infty} u^2K(u)du\right) \end{aligned}$$

Όμως $\int_{-\infty}^{\infty} u^2 hK(u)du$ είναι η διακύμανση σ_K^2 του Kernel (επειδή $E(u) = 0$) και συνεπώς

$$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 + h^2\sigma_K^2$$

Κι επομένως για τη διακύμανση βρίσκουμε πως

$$Var(X) = E(X^2) - E(X)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 + h^2\sigma_K^2 - \bar{x}^2 = s^2 + h^2\sigma_K^2$$

και άρα η διακύμανση δεν ταυτίζεται με την αντίστοιχη δειγματική αλλά είναι μεγαλύτερη. Αυτό μπορεί να το δεί κανείς εύκολα καθώς προστίθεται ένας όρος που προέρχεται από το kernel κι έχει την ερμηνεία ότι στα πραγματικά δεδομένα προσθέτουμε κάποιο θόρυβο - αβεβαιότητα με τη χρήση των kernels, με αποτέλεσμα να μεγαλώνει η διακύμανση.

Επίσης ενδιαφέρον ειναι να δούμε ποιά είναι η συνάρτηση κατανομής που προκύπτει από την εκτιμηθείσα σππ. Σχετικά με την συνάρτηση κατανομής μπορεί κανείς να παρατηρήσει τα εξής:

Θα ισχύει πως

$$\begin{aligned}
\hat{F}(x) &= \int_{-\infty}^{x} \hat{f}(y)dy \\
&= \int_{-\infty}^{x} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{y - x_i}{h}\right) dy \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{x} \frac{1}{h} K\left(\frac{y - x_i}{h}\right) dy
\end{aligned}$$

Και χρησιμοποιώντας πάλι την ίδια αλλαγή μεταβλητής θέτοντας $u = \frac{y - x_i}{h}$ προκύπτει πως

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{(\frac{x - x_i}{h})} K(u)du$$

Μπορεί όμως να παρατηρήσει κανείς ότι η ποσότητα

$$\int_{-\infty}^{(\frac{x - x_i}{h})} K(u)du$$

είναι απλά η συνάρτηση κατανομής του kernel, έστω $\tilde{K}(u) = \int_{-\infty}^{u} K(t)dt$ και συνεπώς βρίσκουμε πως

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \tilde{K}\left(\frac{x - x_i}{h}\right)$$

δηλαδή αρκετή αναλογία σχετικά με την εκτιμήτρια σππ απλά τώρα μέσα στο άθροισμα υπάρχει η συνάρτηση κατανομής του kernel.

Ο παραπάνω τύπος έχει ενδιαφέρουσα ερμηνεία. Παρατηρήσεις που είναι μικρότερες από το x που μας ενδιαφέρει θα έχουν συνεισφορά μεγάλη κοντά στο 1 ειδικά αν είναι κατά πολύ μικρότερες. Αντίθετα παρατηρήσεις μεγαλύτερες θα έχουν συνεισφορά μικρή κοντά στο 0. Η εκτιμήτρια αυτή θυμίζει αρκετά την εμπειρική συνάρτηση κατανομής όπου

$$S(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

δηλαδή η σχετική συχνότητα των παρατηρήσεων που είναι μικρότερες από το x. Είναι προφανές πως όταν $h \to 0$ τότε η $\hat{F}(x)$ θα τείνει στην $S(x)$ αφού όλες οι μικρότερες παρατηρήσεις θα έχουν συνεισφορά 1 και όλες οι μεγαλύτερες 0.

1.7.4 Επιλογή βέλτιστου παραθύρου

Για να δούμε καλύτερα λοιπόν τις ιδιότητες χρειαζόμαστε το μέσο τετραγωνικό σφάλμα και το ολοκληρωμένο μέσο τετραγωνικό σφάλμα που δίνονται ως

$$MSE\left[\hat{f}(x)\right] \approx \frac{f(x)R(K)}{nh} + \frac{1}{4}h^4\left[f''(x)\right]^2\sigma_k^4,$$
$$MISE\left[\hat{f}(x)\right] \approx \frac{R(K)}{nh} + \frac{1}{4}h^4R(f'')\sigma_k^4$$

Επομένως χρησιμοποιώντας ως κριτήριο την ελαχιστοποίηση του μέσου ολοκληρωμένου τετραγωνικού σφάλματος προκύπτει πως το βέλτιστο παράθυρο εξαρτάται από το kernel που χρησιμοποιήσαμε και είναι ίσο με

$$h_{opt} = \left[\frac{R(K)}{\sigma_k^4 R(f'')}\right]^{1/5} n^{-1/5}.$$

Για το βέλτιστο παράθυρο, το ελαχιστοποιημένο ολοκληρωμένο μέσο τετραγωνικό σφάλμα είναι

$$MISE_{opt} = \frac{5}{4}\left[\sigma_K R(K)\right]^{4/5}\left[R(f'')\right]^{1/5} n^{-4/5}.$$

Ο όρος $R(f'')$ επειδή εξαρτάται από την πραγματική συνάρτηση πυκνότητας που είναι άγνωστη στον ερευνητή, δεν μπορεί να αντιμετωπιστεί. Όμως ο όρος $\left[\sigma_K R(K)\right]^{4/5}$ εξαρτάται μόνο από το kernel και επομένως μπορούμε να διαλέξουμε το κατάλληλο kernel .

Το kernel λοιπόν που ελαχιστοποιεί το ολοκληρωμένο μέσο τετραγωνικό σφάλμα είναι το Epanechnikov kernel (δες πίνακα 1.1). Επειδή για το kernel αυτό είναι $\sigma_K R(K) = 3/\left(5\sqrt{5}\right)$ η ποσότητα $\frac{\sigma_K R(K)}{3/\left(5\sqrt{5}\right)}$ είναι ένα μέτρο του κατά πόσο η επιλογή ενός άλλου kernel αυξάνει το ολοκληρωμένο μέσο τετραγωνικό σφάλμα και την ονομάζουμε αναποτελεσματικότητα (inefficiency).

Όπως βλέπετε στον πίνακα 1.2 που ακολουθεί, η αναποτελεσματικότητα είναι πολύ μικρή ακόμα και για την επιλογή του ομοιόμορφου kernel για το οποίο το

σφάλμα αυξάνεται μόλις 7%. (Θυμηθείτε πως η χρήση ομοιόμορφου kernel αντιστοιχεί στον απλοϊκό εκτιμητή που όπως είπαμε έχει κακές ιδιότητες). Συμπεραίνουμε λοιπόν πως

- Το Epanenchnikov kernel είναι το καλύτερο αλλά
- Η επιλογή του kernel δεν είναι καθόλου σημαντική. Σε λίγο θα δούμε κάποια γραφήματα που μαρτυρούν πως όποιο kernel και να πάρει κανείς η διαφορά είναι αμελητέα.

Kernel	Αναποτελεσματικότητα
Epanenchnikov	1
Biweight	1.0061
Triweight	1.0135
Gaussian	1.0513
Uniform	1.0758

Πίνακας 1.2: Η αναποτελεσματικότητα διαφόρων kernels

Ας δούμε τώρα πως μπορεί κανείς πρακτικά να υπολογίσει την τιμή του βέλτιστου παραθύρου. Για τον υπολογισμό του βέλτιστου παραθύρου θα χρησιμοποιήσουμε την υπόθεση πως το kernel είναι κανονικό. Αν το kernel δεν είναι κανονικό πρέπει κανείς να χρησιμοποιήσει κάποιους πολλαπλασιαστές (δίνονται στον πίνακα 1.3) ώστε να βρει το βέλτιστο παράθυρο για το συγκεκριμένο kernel.

Kernel	Πολλαπλασιαστής c_i
Epanenchnikov	2.214
Biweight	2.623
Triweight	2.978
Gaussian	1
Uniform	1.740

Πίνακας 1.3: Οι πολλαπλασιαστές που απαιτούνται για την εύρεση των βέλτιστων παραθύρων για διαφόρα kernels

'Ετσι υποθέτοντας ότι η πραγματική συνάρτηση πυκνότητας πιθανότητας είναι η κανονική κατανομή μπορεί να δειχτεί πως για κανονικό kernel το βέλτιστο παράθυρο είναι το

$$h = 1.059\sigma n^{-1/5}.$$

Στην πράξη χρειαζόμαστε κάποια εκτίμηση της τυπικής απόκλισης του πληθυσμού που δεν την ξέρουμε και μια λύση είναι η χρήση της δειγματικής τυπικής απόκλισης ή κάποιας ανθεκτικής (robust) εναλλακτικής εκτίμησής της (όπως για παράδειγμα το ενδοτεταρτημοριακό εύρος κατάλληλα πολλαπλασιασμένο).

	Χρήση κανονικού kernel	Χρήση άλλου kernel (εκτός του κανονικού)
Ο πληθυσμός είναι κανονικός	$h = 1.059sn^{-1/5}$ χρησιμοποιώ τη δειγματική τυπική απόκλιση s	$h = c1.059sn^{-1/5}$ όπου c είναι ένας πολλαπλασιαστής που εξαρτάται από το kernel που θα χρησιμοποιήσω.
Ο πληθυσμός είναι περίπου κανονικός	$h = 1.059\bar{s}n^{-1/5}$ αλλά χρησιμοποιώ $\bar{s} = \min\left(\frac{IQR}{1.345}, s\right)$ επειδή πια το s δεν είναι από μόνο του μια καλή εκτίμηση	$h = c\ 1.059\bar{s}n^{-1/5}$ Χρησιμοποιώ πάλι τον πολλαπλασιαστή μαζί με το $\bar{s} = \min\left(\frac{IQR}{1.345}, s\right)$
Ο πληθυσμός δεν μοιάζει να είναι καθόλου κανονικός	$h_{opt} = \left[\frac{R(K)}{\sigma_k^4 R(f'')}\right]^{1/5} n^{-1/5}$ χρησιμοποιώντας τη μορφή της f . Επειδή για το κανονικό kernel $\frac{R(K)}{\sigma_k^4} = \frac{1}{2\sqrt{\pi}}$, χρειάζεται να υπολογίσουμε μόνο το $R(f'')$. Ανάλογα με μια πρώτη ματιά των δεδομένων διαλέγουμε μια μορφή για την fπου να ταιριάζει.	$h = ch_{opt}$ όπου c είναι ένας πολλαπλασιαστής που εξαρτάται από το kernel που θα χρησιμοποιήσω και h_{opt} είναι το παράθυρο που αντιστοιχεί στο κανονικό kernel και υπολογίζεται ακριβώς στο διπλανό κελί

Πίνακας 1.4: Υπολογισμός βέλτιστου παραθύρου στην πράξη

Επομένως για να βρούμε το βέλτιστο παράθυρο για το Epanechnikov kernel πρέπει να πολλαπλασιάσουμε με τον κατάλληλο πολλαπλασιαστή c_i (δες πίνακα 1.3) και βρίσκουμε πως

$$h = 2.214 \times 1.059\sigma n^{-1/5} = 2.344\sigma n^{-1/5}.$$

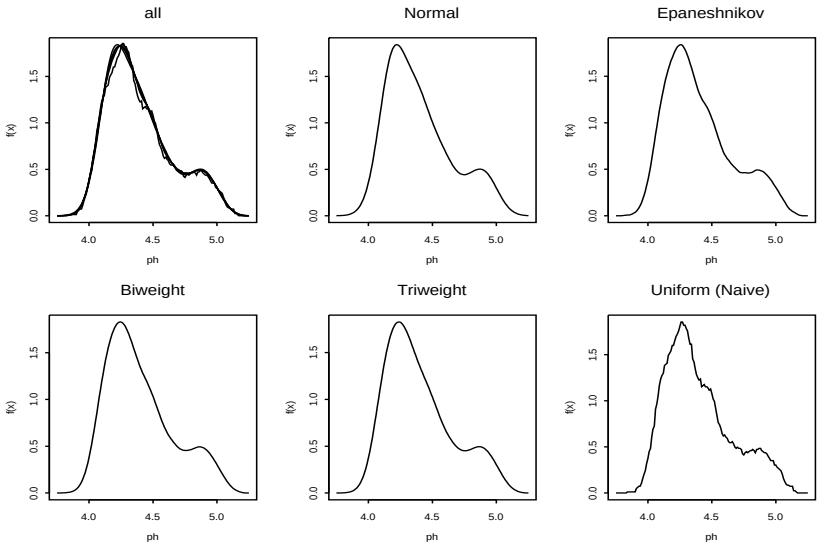
Ομοίως μπορούμε να δουλέψουμε και για τα υπόλοιπα kernels.

'Ενας γενικός οδηγός επιλογής του κατάλληλου kernel δίνεται στον πίνακα 1.4.

Στο γράφημα 1.12 που ακολουθεί μπορεί να δει κανείς την εκτίμηση με τη χρήση διάφορων kernels για τα δεδομένα με τα γιαούρτια. Για κάθε kernel έχουμε χρησιμοποιήσει το αντίστοιχο βέλτιστο παράθυρο. Μπορείτε να παρατηρήσετε πως οι εκτιμήσεις είναι πολύ όμοιες και σχεδόν δεν μπορεί κάποιος να τις ξεχωρίσει, εκτός από την περίπτωση του ομοιόμορφου kernel όπου εξαιτίας της "άγριας" μορφής της είναι εύκολα αναγνωρίσιμη. Στην πάνω δεξιά εικόνα έχουμε βάλει στο ίδιο γράφημα και τις 5 εκτιμήτριες. Μπορεί λοιπόν να συμπεράνει κανείς πως η επιλογή του kernel δεν είναι ιδιαίτερα σημαντική.

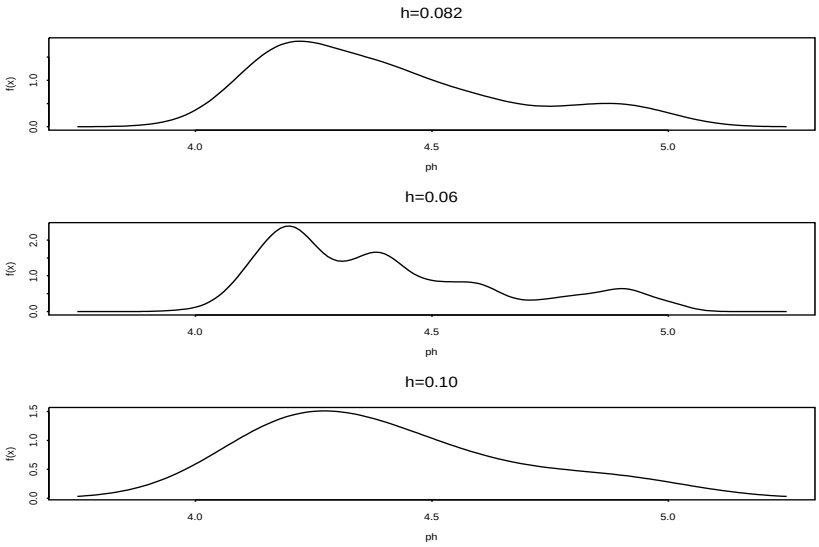
Δεν συμβαίνει όμως το ίδιο και με το παράθυρο. Η τιμή του βέλτιστου παραθύρου για το κανονικό kernel βρέθηκε να είναι 0.0862. Στο γράφημα 1.13 μπορεί κανείς να δει την εκτίμηση με το κανονικό kernel και με 3 διαφορετικά παράθυρα 0.04,

0.0862 (το βέλτιστο) και 0.15. Το βέλτιστο είναι στην κορυφή (h=0.0862). Στα άλλα δύο γραφήματα έχουμε χρησιμοποιήσει h=0.04 και h=0.10 αντίστοιχα. Παρατηρείστε πως όταν μικρύνουμε το παράθυρο η κατανομή εμφανίζει 4 κορυφές, ενώ με μεγάλο παράθυρο βλέπουμε πως η δεξιά κορυφή τείνει να απαλειφθεί. Είναι προφανές πως η επιλογή του παραθύρου είναι ιδιαίτερα κρίσιμη για να πάρουμε μια αντικειμενική εικόνα. Επομένως αυτό που συμπεραίνει κανείς είναι πως η επιλογή του παραθύρου είναι η κρίσιμη παράμετρος η οποία πρέπει να αντιμετωπιστεί με προσοχή.



Γράφημα 1.12: Εκτίμηση με τη χρήση διάφορων kernels για τα δεδομένα που αφορούν τα γιαούρτια. Κάθε γράφημα έχει στον y άξονα την ονομασία του kernel που χρησιμοποιήθηκε. Πάνω δεξιά παρουσιάζονται και οι 5 εκτιμήσεις.

Πριν προχωρήσουμε σε άλλες διαφορετικές επιλογές της τιμής του παραθύρου ας σταθούμε λίγο στην επιλογή της μορφής της κατανομής του πληθυσμού f. Από τον τύπο του υπολογισμού του βέλτιστου παραθύρου είναι σαφές πως βασικά μας ενδιαφέρει περισσότερο η f'' και όχι η ίδια η f . 'Οπως είπαμε και πριν η δεύτερη παράγωγος καθορίζει τη μορφή της κατανομής, δηλαδή αν έχει μια ή περισσότερες κορυφές, πόσο λεία είναι κλπ. Στη συνέχεια υψώνουμε τη δεύτερη παράγωγο στο τετράγωνο γιατί δεν μας ενδιαφέρει το πρόσημο αλλά πόσο λεία είναι και ολοκληρώνουμε πια την $R(f'')$ για να έχουμε ένα συνολικό μέτρο σε όλο τον άξονα των x. Επομένως αυτό που μας ενδιαφέρει είναι η f''. Για να πάρουμε μια εικόνα



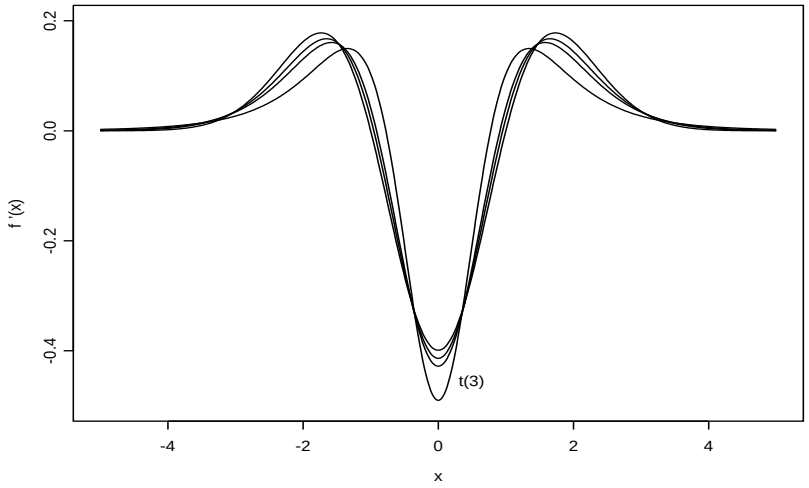
Γράφημα 1.13: Εκτίμηση με τη χρήση κανονικού kernel για τα δεδομένα μας με τη χρήση 3 διαφορετικών παραθύρων.

συγκρίναμε δύο κατανομές, την τυποποιημένη κανονική κατανομή και την κατανομή $t-$student. Γνωρίζουμε πως και οι δύο κατανομές είναι συμμετρικές αλλά η κατανομή t έχει μεγαλύτερες ουρές. Φυσικά είναι γνωστό πως η κανονική κατανομή προσεγγίζει πολύ καλά την κατανομή t για μεγάλους βαθμούς ελευθερίας. Συγκεκριμένα οι συναρτήσεις πυκνότητας πιθανότητας και οι δεύτερες παράγωγοι τους είναι:

Κανονική κατανομή	t-κατανομή
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),$	$f(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\sqrt{r\pi}} \left(1+\frac{x^2}{r}\right)^{-\frac{r+1}{2}},$
$\sigma^2>0,\ -\infty < x, \mu < +\infty$	$r > 0 - \infty < x < +\infty$
$f''(x) = f(x)\left(\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\right),$	$f''(x) = f(x)\left(\frac{r+1}{r+x^2}\right)\left(\frac{(r+3)x^2}{r+x^2} - 1\right)$

Στο γράφημα 1.14 μπορεί κανείς να δει τις δεύτερες παραγώγους για την κατανομή t με 3, 10 και 20 βαθμούς ελευθερίας καθώς και τη δεύτερη παράγωγο της τυποποιημένης κανονικής. Οι διαφορές είναι πολύ μικρές. Μόνο για την t με 3 βαθμούς ελευθερίας μπορούμε να διακρίνουμε διαφορές καθώς οι υπόλοιπες σχεδόν ταυτίζονται και επομένως τα βέλτιστα παράθυρα θα είναι σχεδόν τα ίδια. Βλέπει λοιπόν κανείς πως αν απλά γνωρίζουμε μια γενική μορφή της κατανομής αρκεί για να βρούμε το βέλτιστο παράθυρο και σε καμιά περίπτωση δεν χρειαζόμαστε πλήρη γνώση της κατανομής του πληθυσμού.

Για παράδειγμα μπορεί κανείς να χρησιμοποιήσει αντί για την f'' την $\hat{f}''$. Δηλαδή να πάρει μια αρχική εκτίμηση υποθέτοντας αυθαίρετα κανονικότητα του



Γράφημα 1.14: Η δεύτερη παράγωγος της συνάρτησης πυκνότητας πιθανότητας της τυποποιημένης κανονικής κατανομής και της κατανομής t με βαθμούς ελευθερίας 3,10,20. Παρατηρείστε πως μόνο για 3 βαθμούς ελευθερίας έχουμε διαφορές κοντά στο 0.

πληθυσμού και στη συνέχεια να βρει την $\hat{f}''$. Μπορεί πολύ εύκολα να δει κανείς πως

$$\hat{f}''(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K''\left(\frac{x-x_i}{h}\right)$$

'Ετσι αν κανείς έχει χρησιμοποιήσει κανονικό kernel προκύπτει πως

$$\hat{f}''(x) = \frac{1}{nh}\sum_{i=1}^{n}\left(\frac{x^2}{h^2} - \frac{1}{h}\right)K\left(\frac{x-x_i}{h}\right)$$

Με τη χρήση αυτής της παραγώγου μπορεί κανείς να βρει το $R(\hat{f}'')$ και να προχωρήσει στους υπόλοιπους υπολογισμούς. Βέβαια αυτή η ολοκλήρωση δεν μπορεί να γίνει εύκολα και χρειάζονται αριθμητικές μέθοδοι.

1.7.5 'Αλλες επιλογές βέλτιστου παραθύρου

Θα πρέπει να τονιστεί ότι όλη η συζήτηση που προηγήθηκε, σχετικά με την επιλογή του βέλτιστου παραθύρου, βασίστηκε στο κριτήριο της ελαχιστοποίησης του μέσου ολοκληρωμένου τετραγωνικού σφάλματος. Ανάλογα λοιπόν με το κριτήριο που χρησιμοποιεί κανείς μπορεί να βρει ένα άλλο βέλτιστο παράθυρο. Τα κριτήρια λοιπόν επιλογής είναι πάρα πολλά και είναι αλήθεια πως αυτό που περιγράψαμε είναι απλά μόνο ένα από τα πολλά εναλλακτικά μεταξύ των οποίων μπορεί κανείς

να διαλέξει. Επίσης πρέπει να τονιστεί ότι έχουν προταθεί στη βιβλιογραφία άλλες μέθοδοι προσδιορισμού του βέλτιστου παραθύρου που έχουν καλύτερες ιδιότητες.

'Ετσι για να διαλέξει κανείς το παράθυρο μπορεί να στηριχτεί σε

- Υποκειμενικά κριτήρια σχετικά με το πόσο εξομάλυνση θέλει στα δεδομένα του. Αν για παράδειγμα ο ερευνητής θεωρεί και πιστεύει ότι η κατανομή του έχει 2 κορυφές μπορεί να επιλέξει ένα παράθυρο που οδηγεί σε μια εκτίμηση με δύο κορυφές

- Κάποιο κριτήριο βελτιστοποίησης όπως αυτό που χρησιμοποιήσαμε. Για παράδειγμα το MISE έχει το μειονέκτημα πως επειδή υψώνουμε στο τετράγωνο την απόσταση ανάμεσα στην πραγματική και την εκτιμηθείσα πυκνότητα, κάποιες ακραίες τιμές μπορούν να μας οδηγήσουν σε λάθος αποτελέσματα καθώς η διαφορά σε εκείνα τα σημεία γίνεται πιο έντονη υψώνοντας στο τετράγωνο. Για αυτό το σκοπό μπορεί κανείς να χρησιμοποιήσει άλλα κριτήρια. 'Ενα τέτοιο κριτήριο είναι το MIAE (mean integrated absolute error) που έχει τη μορφή

  $$MIAE\left[\hat{f}(x)\right] = E\left[\int_{-\infty}^{+\infty}\left|\hat{f}(x) - f(x)\right|dx\right] = \int_{-\infty}^{+\infty} E\left[\left|\hat{f}(x) - f(x)\right|\right]dx$$

  και το οποίο επειδή χρησιμοποιεί τις απόλυτες διαφορές έχει καλύτερες ιδιότητες στις περιπτώσεις που υπάρχουν ακραίες τιμές. Βέβαια για να βρούμε το παράθυρο που το βελτιστοποιεί είναι πολύ πιο δύσκολο.

- Μεθόδους cross-validation και (ή) bootstrap. Αυτές οι μέθοδοι χρησιμοποιούν υπολογιστικές τεχνικές για τις οποίες θα μιλήσουμε στη συνέχεια και είναι ιδιαίτερα δημοφιλείς στη σύγχρονη στατιστική επιστήμη.

- Μεθόδους με τη χρήση πιθανοφάνειας ή πιθανοφάνειας με ποινή (penalized likelihood). Αυτές οι μέθοδοι χρησιμοποιούν μια ποινή ώστε να αποφεύγεται η χρήση πολύ μικρού παραθύρου.

- Γραφικές μεθόδους.

Στη συνέχεια θα περιγράψουμε τη μέθοδο cross-validation για να έχει ο αναγνώστης μια διαφορετική προσέγγιση στο πρόβλημα επιλογής παραθύρου.

1.7.6 Η μέθοδος cross-validation

Μια άλλη ιδέα για να βρούμε το βέλτιστο παράθυρο είναι να χρησιμοποιήσουμε μεθόδους μεγίστης πιθανοφάνειας. Γνωρίζουμε πως αν δώσουμε σε κάθε παρατήρηση πιθανότητα $1/n$ αυτή είναι η εκτιμήτρια μεγίστης πιθανοφάνειας και επομένως η μέθοδος αυτή καθεαυτή δεν πρόκειται να δουλέψει. Αντί αυτής μπορούμε να χρησιμοποιήσουμε μια εναλλακτική μορφή στην οποία θα χρησιμοποιηθεί επίσης η ιδέα του cross-validation.

Η μέθοδος cross-validation χρησιμοποιείται πολύ στην στατιστική σε προβλήματα που θέλουμε να εξετάσουμε το πόσο καλό είναι ένα στατιστικό μοντέλο.

Η ιδέα είναι σχετικά απλή αν και συνήθως χρειάζονται αυξημένες υπολογιστικές απαιτήσεις. Έτσι λοιπόν προκειμένου να εξετάσουμε αν ένα στατιστικό μοντέλο είναι καλό, αφήνουμε μια παρατήρηση έξω, προσαρμόζουμε το μοντέλο στις υπόλοιπες παρατηρήσεις και στη συνέχεια τσεκάρουμε πόσο καλά δουλεύει το μοντέλο για την παρατήρηση που αφήσαμε έξω. Αν επαναλάβουμε την ίδια προσέγγιση αφήνοντας έξω κάθε φορά μια διαφορετική τιμή του δείγματος μπορούμε να πάρουμε ένα συνολικό σκορ για το πόσο καλό ήταν το μοντέλο και επομένως να διαλέξουμε ανάμεσα σε διαφορετικά μοντέλα.

Στην περίπτωση της εκτίμησης με τη μέθοδο των kernels ουσιαστικά τα διαφορετικά μοντέλα είναι οι διαφορετικές εκτιμήσεις με διαφορετικά παράθυρα. Για ένα συγκεκριμένο δείγμα και για δοθέν παράθυρο h η πιθανοφάνεια του δείγματος δίνεται από το γινόμενο $L(h) = \prod_{i=1}^{n} \hat{f}_h(x_i)$. Όπως είπαμε πριν η συνάρτηση $L(h)$ μεγιστοποιείται για $h = 0$. Ας συμβολίσουμε με $\hat{f}_{h,i}(x)$ την εκτιμήτρια στο σημείο x αν αφαιρέσουμε την i παρατήρηση. Τότε η cross-validated πιθανοφάνεια θα δίνεται από τον τύπο

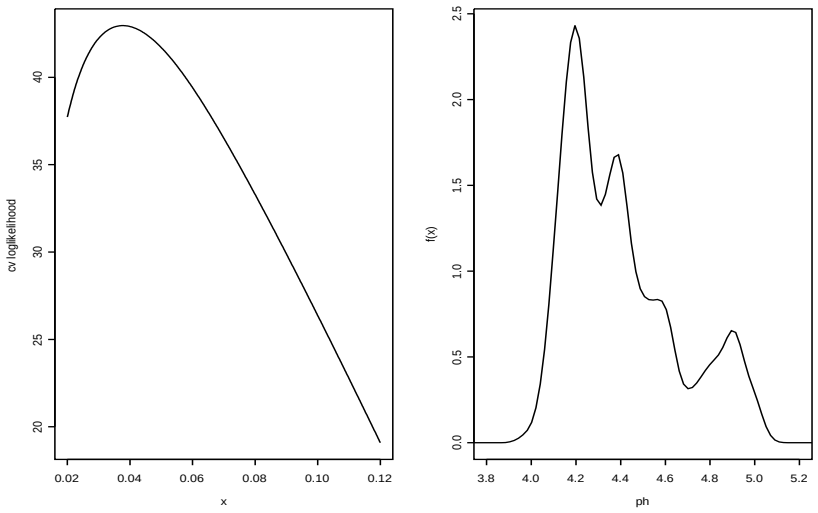
$$L(h, i) = \prod_{i=1}^{n} \hat{f}_{h,i}(x_i)$$

Δηλαδή η ιδέα είναι να αγνοήσουμε την i παρατήρηση, και να εκτιμήσουμε την πυκνότητα στην παρατήρηση που αφήσαμε έξω. Επαναλαμβάνοντας αυτή τη διαδικασία για κάθε παρατήρηση εκτιμάμε για δοθέν παράθυρο την πιθανοφάνεια. Η τιμή του h που μεγιστοποιεί την $L(h, i)$ είναι η επιλογή μας για το παράθυρο. Η μέθοδος αυτή ονομάζεται Maximum Likelihood Cross-Validation και έχει αποδειχθεί στην πράξη πως έχει καλές ιδιότητες. Εναλλακτικά μπορεί κανείς αντί να χρησιμοποιήσει ως κριτήριο τη μεγιστοποίηση μιας πιθανοφάνειας, να ελαχιστοποιήσει κάποιο άθροισμα τετραγώνων ή κάποια άλλη συνάρτηση.

Στο γράφημα 1.15 μπορεί να δει κανείς την cross-validated λογαριθμική πιθανοφάνεια για τα δεδομένα με τα γιαούρτια. Το μέγιστο παρατηρείται στην τιμή 0.0377 και αυτή η τιμή είναι η επιλογή του βέλτιστου παραθύρου με αυτό το κριτήριο. Παρατηρείστε πως το βέλτιστο παράθυρο είναι πολύ μικρότερο από αυτά που είχαμε βρει μέχρι τώρα. Η τιμή της cross-validated λογαριθμικής πιθανοφάνειας σε αυτό το σημείο είναι 42.9565. Στην πράξη αυτό που κάνουμε είναι πως διαλέγουμε πολλές τιμές του h σε κάποιο διάστημα που πιστεύουμε πως υπάρχει το μέγιστο και επαναλμβάνουμε τη διαδικασία για όλες αυτές τις τιμές. Στο παράδειγμά μας διαλέξαμε 400 τιμές του h στο διάστημα $0.02, 0.12$. Το βέλτιστο παράθυρο που βρήκαμε είναι αρκετά μικρότερο από αυτό που είχαμε χρησιμοποιήσει προηγουμένως με αποτέλεσμα η κατανομή τώρα να παρουσιάζει πολλές κορυφές.

1.7.7 Διαστήματα εμπιστοσύνης

Όπως τονίσαμε και στην αρχή, σκοπός μας είναι η εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας. Επομένως για τη $\hat{f}(x)$, η οποία είναι τυχαία μεταβλητή μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης. Αν τώρα κατασκευάσουμε το διάστημα εμπιστοσύνης για κάθε τιμή του x, παίρνουμε μια ζώνη μέσα στην οποία περιμένουμε πως κινείται η πραγματική συνάρτηση πυκνότητας πιθανότητας.



Γράφημα 1.15: Η cross-validated λογαριθμική πιθανοφάνεια και η εκτιμήτρια με τη χρήση του παραθύρου που την μεγιστοποιεί για τα δεδομένα με τα γιαούρτια

Προσοχή όμως στο γεγονός ότι αν κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για κάθε τιμή του x τότε η συνολική ζώνη έχει πολύ μικρότερη στατιστική ακρίβεια και για αυτό δεν μπορούμε να ισχυριστούμε πως είναι μια 95% ζώνη. Συνήθως η δημιουργία μιας ζώνης εμπιστοσύνης χρησιμεύει για να δούμε οπτικά το πόσο μεταβλητότητα έχει η εκτίμηση.

Για κάθε σημείο x ξεχωριστά μπορούμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης ως εξής. Συνήθως τα διαστήματα εμπιστοσύνης είναι της μορφής $(\hat{\theta} - z_{1-a/2}s_\theta, \hat{\theta} + z_{1-a/2}s_\theta)$, όπου $\hat{\theta}, s_\theta$ είναι η εκτίμηση για την παράμετρο θ και το τυπικό σφάλμα της αντίστοιχα και z_a είναι κάποιο ποσοστιαίο σημείο της κατανομής που η εκτιμήτρια ξέρουμε πως ακολουθεί (και τις περισσότερες φορές μπορεί να δειχτεί χρησιμοποιώντας ασυμπτωτικά αποτελέσματα πως είναι η κανονική κατανομή).

Στην περίπτωση της $\hat{f}(x)$, ενώ έχει δειχτεί ότι η κανονικότητα είναι συνήθως μια λογική υπόθεση, υπάρχει πρόβλημα λόγω της μεροληψίας της εκτιμήτριας. Λόγω της μεροληψίας αυτής τα διαστήματα εμπιστοσύνης δεν θα είναι συμμετρικά. Για αυτό ένα 95% διάστημα εμπιστοσύνης για την $\hat{f}(x)$ είναι το

$$\hat{f}(x) - \frac{h^2 f''(x)\sigma_K^2}{2}\Big) - 1.96 \left(\frac{f(x)R(K)}{nh}\right)^{1/2}, \quad \hat{f}(x) - \frac{h^2 f''(x)\sigma_K^2}{2}\Big) + 1.96 \left(\frac{f(x)R(K)}{nh}\right)^{1/2}$$

Μπορεί κάποιος να παρατηρήσει ότι έχουμε αφαιρέσει την μεροληψία (ο δεύτερος όρος της παρένθεσης) ενώ χρησιμοποιούμε το 97.5% ποσοστιαίο σημείο της κανονικής κατανομής (1.96) αφού έχει δειχτεί πως η εκτιμήτρια είναι ασυμπτωτικά κανονική. Τέλος ο τελευταίος όρος είναι η τυπική απόκλιση της $\hat{f}(x)$. Το παραπάνω

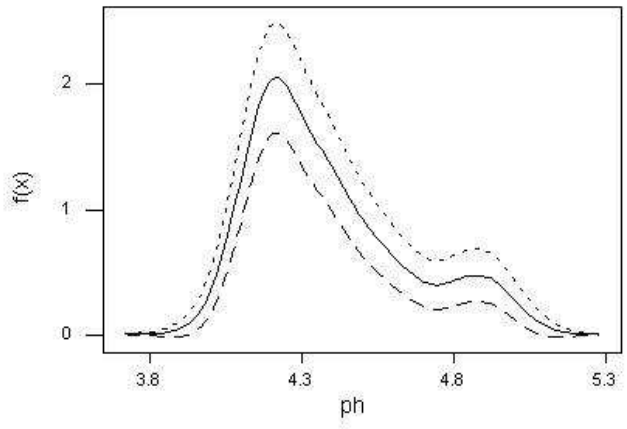
διάστημα εμπιστοσύνης δεν είναι το καλύτερο αλλά είναι μια πρώτη προσέγγιση. 'Ενα καλύτερο διάστημα εμπιστοσύνης είναι το

$$\left(\hat{f}(x) - \frac{c^2 f''(x)\sigma_K^2}{2n^{2/5}} + d_a\right), \left(\hat{f}(x) - \frac{c^2 f''(x)\sigma_K^2}{2n^{2/5}} - d_a\right),$$

όπου $d_a = z_{1-a/2}\sqrt{\frac{1}{c}f(x)R(K)}$ και $c = \frac{R(K)}{R(f'')\sigma_k^4}$.

Επίσης στην περίπτωση που η $f(x)$ είναι άγνωστη μπορούμε να χρησιμοποιήσουμε στη θέση της την $\hat{f}(x)$, αυτή δηλαδή που εκτιμήσαμε. Αν και ένα τέτοιο διάστημα εμπιστοσύνης είναι μάλλον απλοϊκό μπορεί να μας δώσει την εικόνα που θέλουμε και τελικά ενδιαφερόμαστε να πάρουμε σχετικά με την μεταβλητότητα της εκτίμησής μας.

Στο γράφημα 1.16 μπορεί κανείς να δει μια ζώνη 95% για την εκτίμηση της πυκνότητας πιθανότητας για τα δεδομένα με τα γιαούρτια. Παρατηρείστε πως η ζώνη είναι πιο πλατιά κοντά στις κορυφές επειδή υπάρχει αβεβαιότητα σχετικά με την πραγματική θέση τους. Επίσης παρατηρείστε πως χρησιμοποιώντας τη ζώνη εμπιστοσύνης η κορυφή στα δεξιά μπορεί να απαλειφθεί, δείχνοντας δηλαδή την αβεβαιότητα για το αν πραγματικά η κατανομή είναι δίκορφη ή όχι.



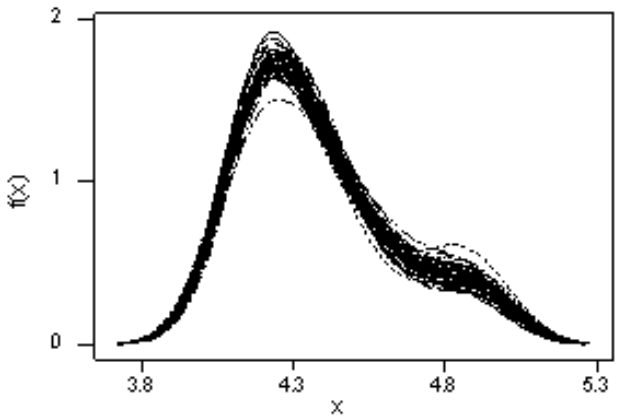
Γράφημα 1.16: Ζώνη εμπιστοσύνης γύρω από την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας για τα δεδομένα με τα γιαούρτια (κανονικό kernel, χρήση βέλτιστου παραθύρου).

1.7.8 Διαστήματα εμπιστοσύνης με τη μέθοδο bootstrap

'Ενας άλλος τρόπος κατασκευής διαστημάτων εμπιστοσύνης για μια συνάρτηση πυκνότητας πιθανότητας είναι η χρήση της μεθόδου bootstrap. Αν και για τη μέθοδο

και τη χρήση της για την κατασκευή διαστημάτων εμπιστοσύνης θα μιλήσουμε αργότερα ας δούμε με συντομία πως μπορούμε να κατασκευάσουμε μια ζώνη εμπιστοσύνης για την εκτιμήτρια μας.

Η μέθοδος bootstrap στηρίζεται στη δειγματοληψία με επανάθεση από το δείγμα που έχουμε στα χέρια μας. Η διαδικασία απαιτεί να δημιουργήσουμε πολλά δείγματα από το αρχικό μας δείγμα έτσι ώστε η μεταβλητότητα που υπάρχει στο δείγμα μας να μελετηθεί μέσα από τα πολλά δείγματα που θα δημιουργήσουμε. Στο γράφημα 1.17 μπορεί κανείς να δει 50 διαφορετικές εκτιμήσεις της συνάρτησης πυκνότητας πιθανότητας που αφορούν 50 δείγματα bootstrap. Στην πράξη η δημιουργία διαστημάτων εμπιστοσύνης είναι αρκετά πιο περίπλοκη αλλά αυτή θα περιγραφεί σε επόμενο κεφάλαιο. Αυτό που έχει σημασία είναι πως από το γράφημα μπορεί κανείς να δει ξεκάθαρα τα σημεία της εκτιμήτριας όπου υπάρχει μεγαλύτερη μεταβλητότητα και ειδικά το σημείο στη δεξιά πλευρά όπου στην πλειοψηφία των περιπτώσεων δεν υπάρχει δεύτερη κορυφή. Περισσότερα για αυτή τη μέθοδο κατασκευής ζώνης εμπιστοσύνης θα δούμε στο κεφάλαιο 5.



Γράφημα 1.17: 50 επαναλήψεις της εκτίμησης της συνάρτησης πυκνότητας πιθανότητας σε Bootstrap δείγματα. Το πάχος της γραμμής σε κάθε σημείο μας δίνει μια εικόνα της μεταβλητότητας στο σημείο αυτό.

1.7.9 Kernel με μεταβλητό παράθυρο

Μέχρι τώρα χρησιμοποιούσαμε για όλα τα σημεία που είχαμε το ίδιο παράθυρο. Αυτό όμως οδηγούσε στο να ομαλοποιούμε το ίδιο όλα τα σημεία ενώ είναι πιθανόν σε κάποια σημεία να υπάρχει περισσότερη δομή. Θυμηθείτε πως για κατανομές όχι

ιδιαίτερα λείες το παράθυρο που χρειαζόμαστε είναι μικρό, ενώ για κατανομές με μάλλον λεία συμπεριφορά χρησιμοποιούσαμε μεγαλύτερο παράθυρο. Επομένως είναι πολλές φορές λογικό να χρησιμοποιήσουμε σε κάθε σημείο διαφορετικό παράθυρο ανάλογα με την πυκνότητα των παρατηρήσεων σε εκείνο το σημείο.

Αν είχαμε απομακρυσμένες τιμές, αυτές κατέληγαν να δημιουργούν ένα λοφάκι (bump) στην άκρης της κατανομής, δίνοντας μια επιπλέον κορυφή η οποία δεν υπάρχει στην πραγματικότητα και απλά η πραγματική κατανομή έχει μεγάλη ουρά.

Από τα παραπάνω φαίνεται η ανάγκη να πάρουμε εκτιμήτριες με μεταβλητό παράθυρο, δηλαδή της μορφής

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(x_i)}K\left(\frac{x-x_i}{h(x_i)}\right)$$

Βλέπουμε λοιπόν πως τώρα πια το παράθυρο είναι συνάρτηση των παρατηρήσεων και διαφέρει από παρατήρηση σε παρατήρηση.

Στο γράφημα 1.18 που ακολουθεί δείτε τη βασική ιδέα. Σε κάθε παρατήρηση βάζουμε ένα kernel ανάλογα με την πυκνότητα σε αυτό το σημείο. Στην εικόνα δεξιά έχουμε το ίδιο παράθυρο για όλα τα σημεία ενώ αριστερά η απομακρυσμένη παρατήρηση έχει ένα kernel με μεγαλύτερη διασπορά και για αυτό αλλάζει και η εικόνα που παίρνουμε για την συνάρτηση πυκνότητας πιθανότητας.

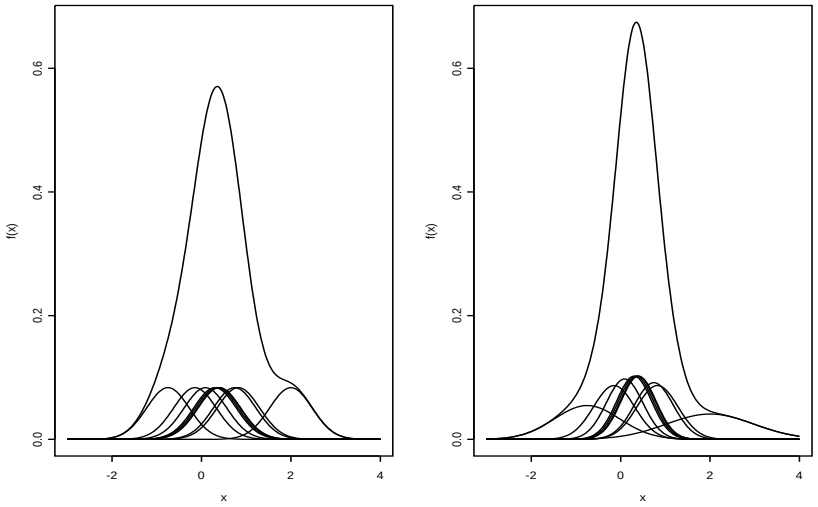
Μια σύγκριση των δύο εκτιμητριών που προκύπτουν μπορεί να δει κανείς στο γράφημα 1.19. Η πλήρης γραμμή αφορά την εκτιμήτρια με μεταβλητό παράθυρο ενώ η διακεκομμένη την εκτιμήτρια με σταθερό παράθυρο. Στην περίπτωση του σταθερού παραθύρου η κατανομή παρουσιάζει ένα σαμαράκι (bump) στη δεξιά ουρά που οφείλεται στην παρατήρηση που απέχει αρκετά από τις άλλες. Κάτι τέτοιο εξαφανίζεται στην περίπτωση του μεταβλητού παραθύρου, όπου όμως οι ουρές της κατανομής είναι παχύτερες.

Κάποιες επιλογές για τη μορφή του μεταβλητού παραθύρου είναι οι

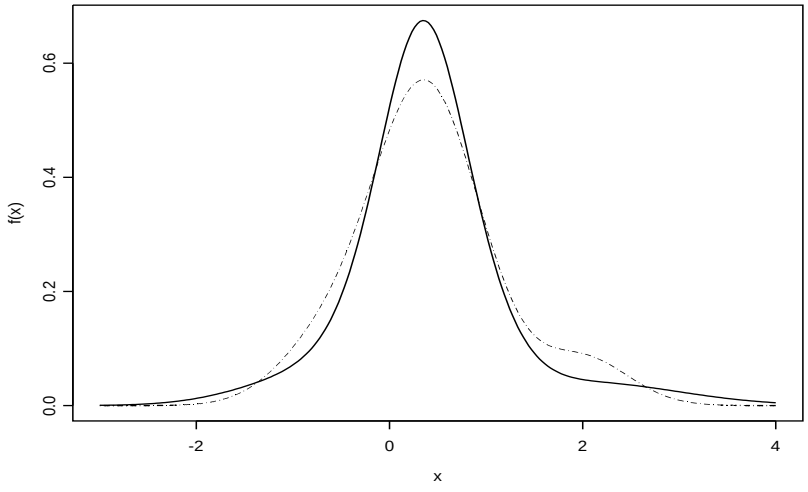
- $h(x_i) = \frac{h}{\sqrt{f(x_i)}}$όπου $f(x)$ είναι η συνάρτηση πυκνότητας πιθανότητας και
- $h(x_i) = d_{ij}, \quad d_{ij}$είναι η απόσταση της τιμής x_i από την πιο κοντινή της.

Επομένως όπου η απόσταση είναι μικρή, δηλαδή υπάρχουν πολλά σημεία μαζεμένα, δεν χρειαζόμαστε μεγάλη τιμή του παραθύρου, ενώ σε απομακρυσμένες τιμές συμβαίνει το αντίθετο. Παρατηρείστε πως η $f(x)$ συνήθως είναι άγνωστη. Σε αυτή την περίπτωση μπορούμε να χρησιμοποιήσουμε σε πρώτο στάδιο μια απλή εκτιμήτρια με κάποιο σταθερό παράθυρο και σε δεύτερο στάδιο να χρησιμοποιήσουμε μεταβλητό παράθυρο. Αλγοριθμικά, θα δουλέψουμε ως εξής

- Βήμα 1^o: Βρες το βέλτιστο παράθυρο h για την περίπτωση σταθερού παραθύρου
- Βήμα 2^o : Βρες την $\hat{f}(x)$ με το σταθερό παράθυρο



Γράφημα 1.18: Kernel με σταθερό (δεξιά) και μεταβλητό παράθυρο (αριστερά). Παρατηρείστε για την μεγαλύτερη παρατήρηση πόσο διαφορετικό είναι το kernel που έχουμε χρησιμοποιήσει (αριστερή εικόνα)



Γράφημα 1.19: Οι δύο εκτιμήτριες που προκύπτουν με μεταβλητό παράθυρο (πλήρης γραμμή) και με σταθερό παράθυρο (διακεκομμένη γραμμή).

- Βήμα 3^o : Βρες τα μεταβλητά παράθυρα για κάθε παρατήρηση από τον τύπο

$$h(x_i) = \frac{h}{\sqrt{\hat{f}(x_i)}}$$

- Βήμα 4^o : Χρησιμοποίησε αυτά τα παράθυρα για να εκτιμήσεις με τη μέθοδο των kernels αλλά με μεταβλητό πια παράθυρο.

'Ενας εναλλακτικός τρόπος να βρούμε τα μεταβλητά παράθυρα είναι ο εξής

- Βήμα 1^o: Εκτίμησε την $\hat{f}(x)$ χρησιμοποιώντας σταθερό παράθυρο h στα σημεία x_i που έχουμε παρατηρήσεις
- Βήμα 2^o: Βρες το γεωμετρικό μέσο των $\hat{f}(x_i)$, δηλαδή $G = \left(\prod_{i=1}^{n} \hat{f}(x_i)\right)^{1/n}$.
- Βήμα 3^o: Υπολόγισε τα $\lambda_i = \sqrt{\frac{G}{\hat{f}(x_i)}}$
- Βήμα 4^o: Υπολόγισε τα μεταβλητά παράθυρα $h_i = h\lambda_i$, $i = 1, ..., n$

Σε αυτή την περίπτωση η επιλογή του παραθύρου, στο πρώτο βήμα, δεν είναι ιδιαίτερα σημαντική αφού το μεταβλητό παράθυρο καθορίζεται περισσότερο από τα δεδομένα.

1.7.10 Μετασχηματισμοί

Πολλές φορές το ενδιαφέρον δεν είναι να βρούμε τη συνάρτηση πυκνότητας πιθανότητας του x αλλά μιας συνάρτησης του x, έστω της $g(x)$. Μια πρώτη προσέγγιση θα ήταν να μετασχηματίσουμε τα δεδομένα και να δουλέψουμε με αυτά. Πολλές φορές όμως συμφέρει να δουλέψουμε με τις πραγματικές τιμές και να κάνουμε αργότερα τον μετασχηματισμό.

Από τη θεωρία η συνάρτηση πυκνότητας πιθανότητας της μεταβλητής $Y = g(X)$ προκύπτει ως (οι υποδείκτες δείχνουν την τυχαία μεταβλητή)

$$f_X(x) = f_Y[g(x)]\,|g'(x)|$$

και επομένως μπορούμε να χρησιμοποιήσουμε μια ανάλογη σχέση ως

$$\hat{f}_X(x) = \frac{|g'(x)|}{nh_Y} \sum_{i=1}^{n} K\left(\frac{g(x) - g(x_i)}{h_Y}\right)$$

όπου το παράθυρο h_Y έχει υπολογιστεί από τις τιμές της y. Δηλαδή στην ουσία βρίσκουμε μια εκτίμηση για την $f_Y(x)$ την οποία στη συνεχεία μετασχηματίζουμε. Τέτοιοι μετασχηματισμοί είναι ιδιαίτερα χρήσιμοι όταν τα δεδομένα είναι ορισμένα σε ένα περιορισμένο διάστημα (πχ στο $(0, +\infty)$) ή όταν τα δεδομένα έχουν μεγάλες ουρές οπότε ο λογαριθμικός μετασχηματισμός μπορεί να μικρύνει κάπως την ουρά.

Παράδειγμα: Αν υποθέσουμε πως $Y = \log X$ τότε έχουμε για την κατανομή της ότι

$$\hat{f}_X(x) = \frac{1}{nxh_Y} \sum_{i=1}^{n} K\left(\frac{\log(x) - \log(x_i)}{h_Y}\right).$$

1.8 Δεδομένα ορισμένα σε διαστήματα

Μέχρι τώρα, ασχοληθήκαμε με δεδομένα ορισμένα σε όλο τον άξονα των πραγματικών αριθμών κι επομένως δεν υπήρχε κανένας περιορισμός για τις περιοχές στις οποίες η συνάρτηση πυκνότητας πιθανότητας είναι μηδενική ή δεν ορίζεται. Πολλές φορές όμως δεν συμβαίνει αυτό. Για παράδειγμα αν η τυχαία μεταβλητή που εξετάζουμε αντιπροσωπεύει το χρόνο επιβίωσης, αυτή η μεταβλητή θα πρέπει αναγκαστικά να είναι θετική, δηλαδή η συνάρτηση πυκνότητας πιθανότητας δεν μπορεί να είναι αρνητική. Καταλαβαίνει κανείς ότι επειδή τα kernels είναι συμμετρικά, για τιμές πολύ κοντά στο 0, κάποιο μικρό έστω κομμάτι μπορεί να περάσει στους αρνητικούς αριθμούς κι έτσι να πάρουμε μη μηδενική πιθανότητα σε αρνητική τιμή που είναι λάθος.

Έχει επίσης αποδειχτεί ότι σε αυτές τις περιπτώσεις υπάρχει σημαντική μεροληψία για τις τιμές της συνάρτησης πυκνότητας πιθανότητας κοντά στο όριο (το 0 στην συγκεκριμένη περίπτωση).

Για να επιτύχουμε λοιπόν η εκτίμηση της συνάρτησης πιθανότητας να είναι μέσα στα επιτρεπτά πλαίσια υπάρχουν διάφορες προσεγγίσεις. Αυτές είναι:

1. Να χρησιμοποιήσουμε κάποιο μετασχηματισμό των δεδομένων, όπως περιγράψαμε προηγουμένως, ούτως ώστε τα νέα δεδομένα να ορίζονται σε όλη την ευθεία των πραγματικών αριθμών, να εκτιμήσουμε τη συνάρτηση πυκνότητας πιθανότητας για αυτή την περίπτωση και στη συνεχεία με τη χρήση του αντίστροφου μετασχηματισμού να επιστρέψουμε στην εκτίμηση της συνάρτησης πιθανότητας για τη μεταβλητή που μας ενδιαφέρει.

2. Να εκτιμήσουμε κανονικά τη συνάρτηση πυκνότητας πιθανότητας, αλλά στις αρνητικές τιμές να τις αντανακλάσουμε στις αντίστοιχες θετικές, προσθέτοντας το μέρος της συνάρτησης πυκνότητας πιθανότητας που έχουμε στους αρνητικούς αριθμούς.

3. Να χρησιμοποιήσουμε κατάλληλα kernels που δεν μας επιτρέπουν να περάσουμε το συγκεκριμένο όριο. Δυστυχώς σε αυτή την περίπτωση η μεροληψία κοντά στο όριο παραμένει.

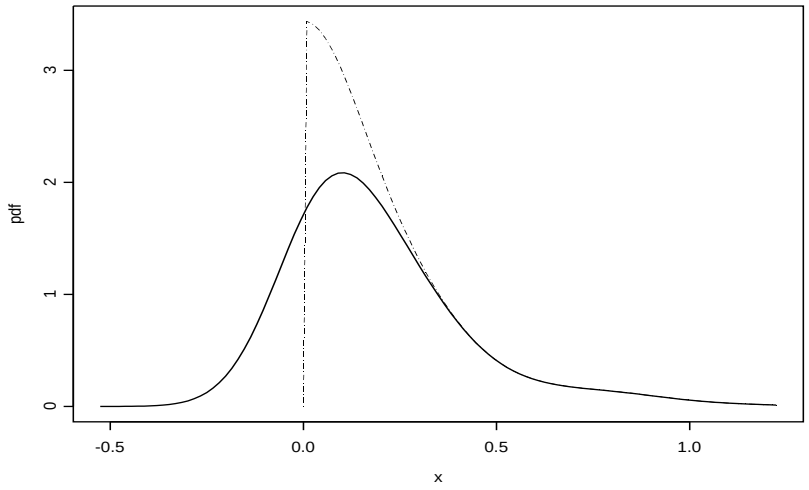
4. Να διπλασιάσουμε το μέγεθος του δείγματος δημιουργώντας άλλο ένα δείγμα με τις ίδιες παρατηρήσεις αλλά αλλάζοντας το πρόσημο. Στη συνέχεια προχωράμε κανονικά σε εκτίμηση με kernels και απλά κρατάμε το κομμάτι που είναι στο θετικό μέρος.

Παράδειγμα 1.3. 150 μηχανές συμμετείχαν σε ένα πείραμα και ο χρόνος σε εβδομάδες μέχρι να παρουσιάσουν κάποια βλάβη καταγράφηκε. Να εκτιμήσετε τη συνάρτηση πυκνότητας πιθανότητας του χρόνου λειτουργίας.

Επειδή τα δεδομένα αφορούν διάρκεια ζωής είναι απαραίτητα θετικά. Μια λογική υπόθεση είναι να υποθέσουμε ότι η πραγματική κατανομή του πληθυσμού είναι η εκθετική, κάτι που επιβεβαιώθηκε από τη J μορφή που είχαν τα δεδομένα. Για να βρει κανείς το βέλτιστο παράθυρο για Epanechnikov kernel αρκεί να παρατηρήσει πως για αυτό ισχύουν $R(K) = \frac{3\sigma_K^{-1}}{5\sqrt{5}}$ και για την εκθετική κατανομή $R(f'') =$

$\frac{\theta^5}{2}$. Επομένως το βέλτιστο παράθυρο είναι το $h_{opt} = \left[\frac{R(K)}{\sigma_k^4 R(f'')}\right]^{1/5} n^{-1/5} = \left[\frac{30}{n}\right]^{1/5} \theta$ και επειδή η εκτιμήτρια μεγίστης πιθανοφάνειας της παραμέτρου είναι η $\hat{\theta} = \bar{x}$ βρίσκουμε πως $h_{opt} = \left[\frac{30}{n}\right]^{1/5} \bar{x}$. Το γράφημα 1.20 δείχνει την εκτίμηση.

Επειδή όμως ένα μεγάλο μέρος της περνάει το μηδέν προς τη μεριά των αρνητικών αριθμών, που δεν είναι λογικό να χρησιμοποιήσουμε, στο γράφημα 1.20 έχουμε αντιστρέψει όλο το μέρος της συνάρτησης πυκνότητας πιθανότητας που ήταν στους αρνητικούς αριθμούς στους αντίστοιχους θετικούς. Τώρα η εικόνα μοιάζει πολύ περισσότερο με την εκθετική κατανομή.



Γράφημα 1.20: Η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας του χρόνου λειτουργίας. Η εκτίμηση με την εφαρμογή της κλασικής μεθόδου των kernels και η εκτίμηση αφού καθρεπτίσαμε την εκτίμηση στους αρνητικούς αριθμούς στους αντίστοιχους θετικούς

1.9 Εκτίμηση πολυμεταβλητής από κοινού συνάρτησης πυκνότητας πιθανότητας

Μέχρι τώρα ασχοληθήκαμε με την εκτίμηση συνάρτησης πυκνότητας πιθανότητας στη μονομεταβλητή περίπτωση. Η μεθοδολογία μπορεί να επεκταθεί και στην πολυμεταβλητή περίπτωση, όπου έχουμε περισσότερες από μια μεταβλητές. Μιλώντας γενικά ας υποθέσουμε πως έχουμε d μεταβλητές, επομένως μιλάμε για εκτίμηση της d-μεταβλητής από κοινού συνάρτησης πυκνότητας πιθανότητας.

Και πάλι η πιο απλή περίπτωση είναι το ιστόγραμμα. Τώρα τα κελιά δεν είναι διαστήματα αλλά πολυεπίπεδα, δηλαδή παραλληλόγραμμα για τη διμεταβλητή περίπτωση κλπ. Χρησιμοποιώντας και πάλι τη λέξη κελί για κάθε τέτοια περιοχή, και υποθέτοντας ότι έχουμε k κελιά η εκτιμήτρια της από κοινού συνάρτησης πυκνότητας πιθανότητας με τη μέθοδο του ιστογράμματος είναι

$$\hat{f}(x_1, x_{2,}..., x_d) = \hat{f}(\mathbf{x}) = \frac{n_k}{n h_1 h_2 ... h_d},\ \mathbf{x} \in k \quad \text{κελί}$$

όπου $h_1, h_2, ..., h_d$ είναι το μήκος του κελιού ως προς την i μεταβλητή.

Για να γίνει πιο κατανοητό έστω ότι έχουμε δύο μεταβλητές x και y. Και έστω ότι έχουμε χωρίσει όλο το χώρο σε κελιά, τα οποία δεν πρέπει να είναι αναγκαστικά τετράγωνα, αλλά έχουν διαστάσεις $h_x \times h_y$ όπου h_x, h_y είναι τα μήκη των κελιών ως προς την μεταβλητή x και y αντίστοιχα. Τότε η εκτίμηση της από κοινού συνάρτησης πυκνότητας πιθανότητας με τη χρήση του ιστογράμματος είναι η

$$\hat{f}(x, y) = \frac{n_k}{n h_x h_y},\ \mathbf{x} \in k \ \text{ κελί}$$

Η επιλογή των παραθύρων μπορεί να γίνει και πάλι με βάση κάποια κριτήρια αλλά κάτι τέτοιο δεν θα μας απασχολήσει σε αυτές τις σημειώσεις.

Κατά αντιστοιχία μπορούμε να ορίσουμε και να βρούμε την εκτιμήτρια της από κοινού συνάρτησης πυκνότητας πιθανότητας και με τη μέθοδο των kernels χρησιμοποιώντας τώρα πολυμεταβλητά kernel. Η εκτιμήτρια θα είναι η

$$\hat{f}(x_1, x_{2,}..., x_d) = \hat{f}(\mathbf{x}) = \frac{1}{n\,|H|} \sum_{i=1}^{n} K_d \left[\mathbf{H^{-1}}(\mathbf{x} - \mathbf{x}_i)\right],$$

όπου τώρα $\mathbf{x}_i$ είναι οι παρατηρήσεις μας που είναι διανύσματα και $\mathbf{H}$ είναι ο πίνακας που περιέχει τα παράθυρα προς όλες τις διαστάσεις (μεταβλητές) που χρησιμοποιούμε. 'Οπως καταλαβαίνετε ο παραπάνω τύπος είναι σε μια πολύ γενική μορφή. Αυτά που πρέπει να καθορίσουμε είναι ο πίνακας $\mathbf{H}$ και η μορφή του kernel.

Ως προς τον πίνακα $\mathbf{H}$ μπορούμε να σημειώσουμε τα εξής:

Χρησιμοποιούμε πίνακα, αντί να χρησιμοποιήσουμε κάποιο παράθυρο προς κάθε κατεύθυνση, για να έχουμε τη δυνατότητα να καθορίσουμε την ομαλοποίηση (smoothing) τόσο προς κάθε διάσταση χωριστά αλλά και από κοινού (κατά αναλογία με τη συνδιακύμανση που χρησιμοποιούμε σε πολυμεταβλητά δεδομένα). Επομένως μερικές ειδικές μορφές είναι οι εξής

- $\mathbf{H} = \begin{bmatrix} h & 0 & ... & 0 \\ 0 & h & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & h \end{bmatrix}$, ο πίνακας είναι διαγώνιος και χρησιμοποιούμε σε όλες τις διαστάσεις το ίδιο παράθυρο
- $\mathbf{H} = \begin{bmatrix} h_1 & 0 & ... & 0 \\ 0 & h_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & h_d \end{bmatrix}$, ο πίνακας είναι διαγώνιος αλλά για κάθε διάσταση χρησιμοποιούμε διαφορετικό παράθυρο. Σε αυτή την περίπτωση $|H| = h_1 h_2 ... h_d$. Τέλος

- $\mathbf{H} = \begin{bmatrix} h_{11} & h_{21} & ... & h_{d1} \\ h_{21} & h_2 & ... & h_{d2} \\ ... & ... & ... & ... \\ h_{d1} & h_{d2} & ... & h_{dd} \end{bmatrix}$ όπου πλέον ο πίνακας είναι συμμετρικός και λαμβάνουμε υπόψη παράθυρα με μια αίσθηση συνδιακυμάνσης, δηλαδή το πόσο θα κάνουμε ομαλοποίηση σε κάθε διάσταση εξαρτάται και από τις υπόλοιπες διαστάσεις.

Από την άλλη πλευρά η επιλογή των kernels μπορεί να στηριχτεί είτε σε γινόμενο ανεξάρτητων προς κάθε διάσταση kernels ή σε κατάλληλα ορισμένα στις d διαστάσεις kernels. Τα πιο διαδεδομένα kernels είναι μια γενίκευση του Epanechnikov με μορφή

$$K_d(\mathbf{x}) = \begin{cases} \left[\frac{d(d+2)}{4}\right] \Gamma(d/2)\pi^{-d/2}(1-\mathbf{x}'\mathbf{x}) & \mathbf{x}'\mathbf{x} \leq 1 \\ \\ 0 & \text{αλλού} \end{cases}$$

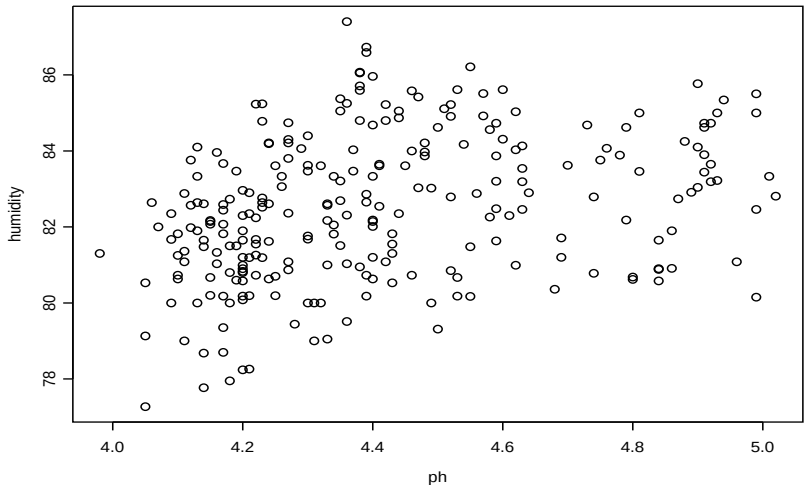
καθώς και μια πολυμεταβλητή κανονική κατανομή (κατά αντιστοιχία με τη χρήση της απλής κανονικής κατανομής για μονοδιάστατα δεδομένα).

Και πάλι η επιλογή των βέλτιστων παραθύρων στηρίζεται σε βελτιστοποίηση κάποιων κριτηρίων τα οποία είναι πια ιδιαίτερα πολύπλοκα και δεν θα μας απασχολήσουν.

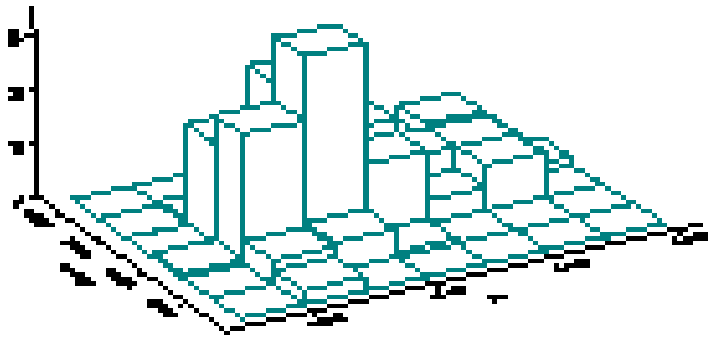
Παράδειγμα 1.2 (συνέχεια). Στο παράδειγμα με τα γιαούρτια θα χρησιμοποιήσουμε τώρα μια δεύτερη μεταβλητή την υγρασία (humidity). Από το διάγραμμα νέφους σημείων (γράφημα 1.21) μπορεί κανείς να δει πως κατανέμονται στο χώρο τα σημεία. Στη συνέχεια έχουμε κατασκευάσει ένα διμεταβλητό ιστόγραμμα καθώς και μια εκτίμηση της από κοινού συνάρτησης πυκνότητας πιθανότητας με τη χρήση ενός διμεταβλητού Epanechnikov kernel (γράφηματα 1.22 και 1.23). Αντικαθιστώντας $d = 2$ μπορεί κανείς να γράψει πως το διμεταβλητό Epanechnikov kernel έχει τη μορφή

$$K_2(x, y) = \begin{cases} \frac{3}{\pi}(1-(x^2+y^2)) & x^2+y^2 \leq 1 \\ \\ 0 & \text{αλλού} \end{cases}$$

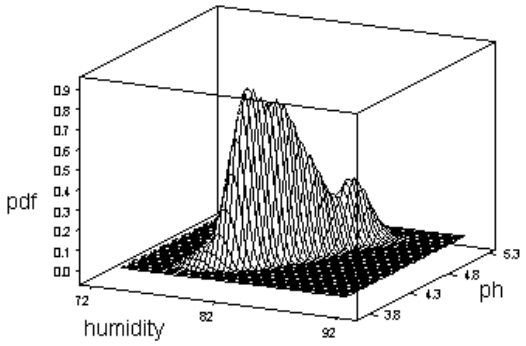
Στα γραφήματα που βλέπετε δεν έχει γίνει προσπάθεια να βελτιστοποιηθούν τα παράθυρα. Για την εκτίμηση με τη μέθοδο των kernels χρησιμοποιήθηκε ένας διαγώνιος πίνακας και τα παράθυρα υπολογίστηκαν ως τα βέλτιστα στη μονομεταβλητή περίπτωση. Είναι ξεκάθαρο ότι τα kernels δίνουν μια κατά πολύ πιο όμορφη εικόνα από ότι το ιστόγραμμα



Γράφημα 1.21: Διάγραμμα νέφους σημείων για την υγρασία και το PH. Παρατηρείστε πως υπάρχουν σημεία όπου μαζεύονται περισσότερα σημεία



Γράφημα 1.22: Διμεταβλητό ιστόγραμμα για τα δεδομένα



Γράφημα 1.23: Διμεταβλητή εκτίμηση με τη χρήση κατάλληλων kernels.

1.10 Κατηγορικά δεδομένα

Μέχρι τώρα μιλήσαμε μόνο για συνεχή δεδομένα. Παρόμοιες μέθοδοι εκτίμησης συνάρτησης πιθανότητας (τώρα πια δεν υπάρχει λόγος να μιλάμε για συνάρτηση πυκνότητας πιθανότητας όπως στη συνεχή περίπτωση) έχουν αναπτυχθεί και για κατηγορικά ή διακριτά δεδομένα. Θα πρέπει να τονισθεί ότι η ανάγκη για παρόμοιες μεθόδους σε κατηγορικά ή διακριτά δεδομένα δεν είναι τόσο μεγάλη καθώς οι σχετικές συχνότητες είναι αμερόληπτες εκτιμήτριες των πραγματικών πιθανοτήτων του πληθυσμού και επομένως χρησιμοποιούνται ευρύτατα. Η ανάγκη προκύπτει κυρίως όταν έχουμε να κάνουμε με μικρά σετ δεδομένων και κυρίως όταν κάποιες κατηγορίες έχουν μικρή πιθανότητα εμφάνισης και επομένως σε μικρά δείγματα τείνουμε να εκτιμάμε την πιθανότητα αυτών των κατηγοριών ως μηδενική.

Σε αυτή την ενότητα μιλάμε για ονομαστικά κατηγορικά δεδομένα (δηλαδή οι παρατηρήσεις μας είναι σε ονομαστική (nominal) κλίμακα, όπως για παράδειγμα, το χρώμα των ματιών, το κόμμα που θα ψηφίσει κάποιος κλπ).

'Εστω πως έχουμε k κατηγορίες και έστω πως από ένα συνολικό δείγμα μεγέθους n, η συχνότητα κάθε κατηγορίας είναι n_j, $j = 1, ..., k$.. 'Οπως είναι γνωστό μια αμερόληπτη εκτιμήτρια της πιθανότητας στον πληθυσμό είναι η $p_j = n_j/n$.

Επειδή όμως αυτή η εκτιμήτρια τείνει να μην είναι ιδιαίτερα λεία (smooth) για μικρά δείγματα και κατηγορίες με μικρή σχετικά πιθανότητα (μιλώντας στατιστικά, έχει μεγάλο μέσο τετραγωνικό σφάλμα) έχει προταθεί η χρήση της εκτιμήτριας

$$\hat{p}_j = \frac{n_j + \alpha}{n + \alpha k}$$

Το α παίζει το ρόλο μιας παραμέτρου εξομάλυνσης που οδηγεί σε μη μηδενικές

εκτιμήσεις ακόμα και για κατηγορίες που είχαν μηδενική συχνότητα στο δείγμα. Από την άλλη για πολύ μεγάλα μεγέθη δείγματος το α (που όπως θα δούμε δεν μπορεί να ξεπεράσει την τιμή 1) παίζει μικρό ρόλο και άρα οι σχετικές συχνότητες τείνουν να καθορίζουν την εκτίμηση.

Η τιμή του α έχει προταθεί να καθορίζεται ως

$$\alpha = \begin{cases} z^{-1} & z \geq 1 \\ \\ 1 & z < 1 \end{cases}$$

όπου

$$z = \frac{1}{k} \sum_{j=1}^{k} \frac{\left(n_j - \frac{n}{k}\right)^2}{\frac{n}{k}}.$$

Η ποσότητα z μπορεί να αναγνωριστεί ως η τιμή της στατιστικής συνάρτησης χ^2 του Pearson για έλεγχο της υπόθεσης ότι όλες οι κατηγορίες είναι ισοπίθανες, διαιρεμένης με το πλήθος των κατηγοριών. Επομένως μεγάλη τιμή της σημαίνει πως η υπόθεση ότι οι κατηγορίες είναι ισοπίθανες είναι εσφαλμένη. Παρατηρείστε πως όταν το α πάρει μικρή τιμή τότε και η διαφορά ανάμεσα στον εκτιμητή $\hat{p}_j$ και p_j θα είναι μικρή. 'Ομως μικρή τιμή του α σημαίνει πως τα δεδομένα δεν δείχνουν να είναι ισοπίθανα (μεγάλη τιμή του z).

Μπορεί επίσης να δειχτεί ότι ο παραπάνω εκτιμητής γράφεται ως

$$\hat{p}_j = \frac{\varepsilon}{k} + (1 - \varepsilon)\frac{n_j}{n},$$

όπου

$$\varepsilon = \frac{\alpha k}{n + \alpha k}$$

Μπορεί κάποιος να παρατηρήσει πως καθώς $1/k$ είναι η πιθανότητα κάθε κατηγορίας στην περίπτωση που όλες οι κατηγορίες είναι ισοπίθανες ο εκτιμητής αυτός είναι ένας σταθμικός μέσος των σχετικών συχνοτήτων και της περίπτωσης ισοπίθανων ενδεχομένων με βάρη $(1 - \varepsilon)$ και ε αντίστοιχα. 'Οταν η τιμή του α είναι μικρή, κάτι που από τον τρόπο υπολογισμού του α σημαίνει πως τα δεδομένα δεν μοιάζει να είναι ισοπίθανα, το βάρος του πρώτου όρου είναι μικρό και άρα οι σχετικές συχνότητες καθορίζουν την εκτίμηση.

Το πιο ενδιαφέρον είναι πως ο εκτιμητής αυτός μπορεί να συνδεθεί εύκολα και με τη μέθοδο των kernels, καθώς μπορεί να γραφτεί ως

$$\hat{p}_i = \sum_{j=1}^{k} \frac{n_j}{n} W_j(i, \lambda)$$

όπου το W παίζει το ρόλο του kernel και ορίζεται ως

$$W_j(i, \lambda) = \begin{cases} \lambda & j = i \\ \\ \frac{1-\lambda}{k-1} & j \neq i \end{cases}$$

Στην ουσία το kernel μας λέει πως αν έχουμε παρατηρήσει την τιμή j τότε η πιθανότητα να είναι σωστή η τιμή (σωστή με την έννοια ότι μας παρέχει πληροφορία για την τιμή αυτή) είναι λ ενώ όλες οι άλλες κατηγορίες έχουν μια πιθανότητα $\frac{1-\lambda}{k-1}$ η καθεμία. Το λ συνδέεται με τα α και ε που χρησιμοποιήσαμε προηγουμένως ως εξής: $\varepsilon = \frac{k(1-\lambda)}{k-1}$ και άρα μπορούμε να υπολογίσουμε και τη σχέση του α με το λ.

Παράδειγμα 1.4 Σε μια έρευνα ένα μικρό δείγμα 50 ατόμων δήλωσε ποιο κόμμα θα ψηφίσει στις εκλογές. Οι συχνότητες για τα 5 κόμματα (Α,Β,Γ,Δ,Ε) ήταν 20, 18, 7, 5, 0 αντίστοιχα. Να εκτιμήσετε την πιθανότητα του κόμματος Ε.

Αν χρησιμοποιήσουμε τις σχετικές συχνότητες παίρνουμε πως το κόμμα Ε έχει μηδενικό ποσοστό στον πληθυσμό κάτι που δεν είναι ρεαλιστικό. Χρησιμοποιώντας τον εξομαλυσμένο εκτιμητή βρίσκουμε

$$z = \frac{1}{5}\left[\left(\frac{20-\frac{50}{5}}{10}\right)^2 + ... + \left(\frac{0-\frac{50}{5}}{10}\right)^2\right] = 5.96$$

κι επομένως $\alpha = 1/5.96 = 0.167785$.

Κόμμα	n_j	p_j	$\hat{p}_j$
Α	20	0.40	0.3967
Β	18	0.36	0.3573
Γ	7	0.14	0.1410
Δ	5	0.10	0.1016
Ε	0	0	0.0033

Πίνακας 1.5: Εκτίμηση με τη χρήση του kernel εκτιμητή

Χρησιμοποιώντας αυτή την τιμή εκτιμάμε τις πιθανότητες που βλέπετε στον πίνακα 1.5. Τώρα πια το κόμμα Ε έχει μη μηδενική πιθανότητα. Παρατηρείστε πως οι αλλαγές είναι πολύ μικρές στα υπόλοιπα κόμματα κυρίως γιατί η υπόθεση των ισοπίθανων κατηγοριών δεν είναι ιδιαίτερα ισχυρή. Μπορεί κάποιος να υπολογίσει πως $\varepsilon = 0.0165$ και $\lambda = 0.9868$.

1.11 Διακριτά Δεδομένα

Ας δούμε τώρα την περίπτωση που τα δεδομένα μας αφορούν διακριτά δεδομένα όπου υπάρχει κάποια σχέση κατάταξης (ordinal data). Τέτοιες περιπτώσεις είναι για παράδειγμα ο αριθμός των ατυχημάτων που συνέβησαν σε ένα άτομο, ο αριθμός των γκολ που σκόραρε κάποια ομάδα και πολλά άλλα παραδείγματα.

Σε αυτή την περίπτωση η χρήση της γενικής μεθόδου για κατηγορικά δεδομένα, αν και μπορεί να εφαρμοσθεί, δεν είναι ιδιαίτερα σωστή καθώς η κατανομή τέτοιων δεδομένων συνήθως έχει θετική ασυμμετρία και οι μικρότερες τιμές κοντά στο μηδέν είναι πιο πιθανές από τις πιο απομακρυσμένες. Για αυτό χρησιμοποιούμε ένα λίγο τροποποιημένο kernel της μορφής

$$W_j(i,\lambda) = \begin{cases} \lambda & \text{αν } \ j = i \\ \frac{1-\lambda}{2^{|i-j|+1}} & \text{αν } \ 0 < |i-j| \leq j \\ \frac{1-\lambda}{2^{|i-j|}} & \text{αν } \ |i-j| > j \end{cases}$$

Η επιλογή του λ μπορεί και πάλι να γίνει με τη χρήση διαφόρων κριτηρίων και δεν θα μας απασχολήσει.

Παράδειγμα 1.5: Τα δεδομένα που ακολουθούν αφορούν τον αριθμό των γκολ που πέτυχε μια ομάδα στα τελευταία 20 παιχνίδια. Αν ένα παίκτης του Στοιχήματος θέλει να υπολογίσει κάποιες πιθανότητες για να παίξει στοίχημα (γιατί θέλει να παίξει έξυπνα και όχι με βάση άλλες εμπειρικές και αμφίβολες λογικές), επιθυμεί να δει πως περίπου είναι η κατανομή των γκολ της ομάδας.

Αν χρησιμοποιήσει απλά τις παρατηρηθείσες συχνότητες, αφενός θα έχει ίσως κάποια κενά (όπως στην τιμή 2 στο παράδειγμα μας όπου η συχνότητα είναι 0), αφετέρου δεν έχει κατάλληλη πληροφορία για το τι γίνεται στην ουρά της κατανομής (αφού ποτέ δεν έχει πετύχει η ομάδα του πάνω από 3 γκολ).

Με τη χρήση των kernels η εικόνα που παίρνει είναι πιο ομαλή και άρα πιο αξιόπιστη για να τη χρησιμοποιήσει για πρόβλεψη. Στον πίνακα 1.6 έχουμε χρησιμοποιήσει 4 διαφορετικές τιμές για το λ. Παρατηρείστε πως όσο αυξάνει το λ τόσο η κατανομή των γκολ πλησιάζει την παρατηρηθείσα συχνότητα αλλά σε κάθε περίπτωση είναι πιο ομαλή και δίνει μη μηδενικές πιθανότητες στα ενδεχόμενα να επιτύχει η ομάδα περισσότερα από 3 γκολ.

		Σχετική	Εκτίμηση με τη χρήση kernel			
x	Συχνότητα	συχνότητα	$\lambda = 0.25$	$\lambda = 0.45$	$\lambda = 0.65$	$\lambda = 0.85$
0	8	0.4	0.205	0.257	0.309	0.361
1	11	0.55	0.292	0.361	0.430	0.498
2	0	0	0.188	0.138	0.088	0.038
3	1	0.05	0.153	0.126	0.098	0.071
4	0	0	0.080	0.058	0.037	0.016
5	0	0	0.040	0.029	0.019	0.008
6	0	0	0.020	0.015	0.009	0.004
7	0	0	0.011	0.008	0.005	0.002
8	0	0	0.006	0.004	0.003	0.001
9	0	0	0.003	0.002	0.001	0.001
10	0	0	0.001	0.001	0.001	0.000
11	0	0	0.001	0.001	0.000	0.000

Πίνακας 1.6: Εκτίμηση με τη χρήση διακριτού kernel

1.12 Εφαρμογές της μεθόδου των kernels

Στο κεφάλαιο αυτό μιλήσαμε για εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας. Η μέθοδος των kernels είναι αυτή που έχει τις καλύτερες ιδιότητες και που χρησιμοποιείται στην πράξη για διάφορες εφαρμογές πέραν της απλής κατασκευής ενός πληροφοριακού γραφήματος. Μερικές από τις εφαρμογές της είναι:

- Παρουσίαση Δεδομένων. Πολλές φορές όταν τα δεδομένα πρέπει να παρουσιαστούν με κάποιον τρόπο η χρήση του ιστογράμματος δεν είναι αρκετά καλή και μπορεί να οδηγήσει σε λανθασμένη εκτίμηση για την πραγματική κατανομή του πληθυσμού. Παρουσιάζοντας μια εκτίμηση με τη χρήση των kernels η εικόνα παρέχει μεγαλύτερη πληροφορία ειδικά αν έχει βελτιστοποιηθεί κατά κάποιον τρόπο με τη χρήση βέλτιστου παραθύρου. Για παράδειγμα, στη σύγχρονη στατιστική μεθοδολογία και ιδίως στην Μπευζιανή προσέγγιση, ολοένα και περισσότερο χρησιμοποιείται η ιδέα πως δεν χρειάζεται κανείς να γνωρίζει τη συνάρτηση πυκνότητας πιθανότητας μιας κατανομής αρκεί να μπορεί να προσομοιώσει τιμές από αυτήν. Αν έχουμε ένα αρκετά μεγάλο δείγμα τιμών από μια κατανομή μπορούμε να πάρουμε μια εικόνα της απλά ζωγραφίζοντας την εκτιμήτρια της με τη μέθοδο των kernels.

- Περιγραφή των δεδομένων. Σε συνέχεια της παρουσίασης των δεδομένων τα συμπεράσματα που μπορεί να βγάλει κανείς είναι πολύ πιο αξιόπιστα για κάποια περιγραφικά μέτρα. Έτσι η συμμετρία ή όχι της κατανομής μπορεί πιο εύκολα να ανιχνευτεί, όπως και η κορυφή, η ύπαρξη ή όχι δύο κορυφών και άλλα τέτοια περιγραφικά συμπεράσματα.

- Μη παραμετρικές τεχνικές. Μια από τις πιο σημαντικές εφαρμογές των kernels είναι στη συμπερασματολογία χωρίς χρήση κάποιου παραμετρικού μοντέλου. Δηλαδή ενώ πολλές στατιστικές τεχνικές είναι βασισμένες πάνω στην υπόθεση πως τα δεδομένα προέρχονται από κάποια συγκεκριμένη κατανομή (συνήθως την κανονική), μπορεί κανείς να προχωρήσει στην ανάλυση χρησιμοποιώντας την συνάρτηση πυκνότητας πιθανότητας που εκτίμησε με τα kernels. Μια τέτοια προσέγγιση είναι μη παραμετρική και επομένως δεν χρειαζόμαστε υποθέσεις. Τέτοιες προσεγγίσεις έχουν αναπτυχθεί για μια σειρά από στατιστικές τεχνικές γνωστές στους περισσότερους, οι οποίες βασίζονται στην υπόθεση της κανονικότητας. Αγνοώντας λοιπόν την υπόθεση της κανονικότητας και προχωρώντας με την εκτιμηθείσα συνάρτηση πυκνότητας πιθανότητας μπορεί κανείς να μιλήσει για:

  1. Μη παραμετρική διακριτική ανάλυση (τεχνική που προσπαθεί να βρει κανόνες με τους οποίους να διακρίνει ανάμεσα σε διαφορετικούς πληθυσμούς)
  2. Μη παραμετρική ανάλυση κατά συστάδες (τεχνική που προσπαθεί να βρει συστάδες παρατηρήσεων με όμοια χαρακτηριστικά για να δημιουργήσει ομοιογενείς ομάδες)
  3. Μη παραμετρική παλινδρόμηση. Γενικά το γραμμικό μοντέλο στηρίζεται πάνω σε υποθέσεις περί κανονικότητας των σφαλμάτων, αλλά και πάνω

στην ύπαρξη μιας γραμμικής σχέσης. Η μη παραμετρική παλινδρόμηση δεν είναι απαραίτητα γραμμική (ή καλύτερα *ποτέ* δεν είναι) και δεν βασίζεται σε υποθέσεις περί κανονικότητας. Θα περιγράψουμε στην επόμενη ενότητα τη μέθοδο αυτή.

- Τέλος κάποιες άλλες εφαρμογές της μεθόδου θα περιγραφούν στη συνέχεια. Αυτές είναι η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας όταν τα δεδομένα δεν είναι προϊόν απλής τυχαίας δειγματοληψίας καθώς και η περίπτωση εκτίμησης με τη μέθοδο των ελαχίστων αποστάσεων. Το ενδιαφέρον στην τελευταία περίπτωση είναι πως χρησιμοποιούμε τη μέθοδο των kernels όχι για καθαρά πρακτικούς σκοπούς αλλά για τη θεωρητική ανάπτυξη στατιστικών εννοιών και μεθόδων με απώτερο σκοπό την πρακτική τους εφαρμογή.

1.13 Μη Παραμετρική Παλινδρόμηση

1.13.1 Εκτίμηση του μοντέλου

Η μέθοδος των kernels βρίσκει εφαρμογή στη μέθοδο της μη παραμετρικής παλινδρόμησης. Η λέξη μη παραμετρική παλινδρόμηση προδιαθέτει για μια μεθοδολογία όπου δεν γίνεται καμιά υπόθεση σχετικά με την κατανομή του πληθυσμού από όπου προήλθαν τα δεδομένα αλλά όπως θα δούμε στη συνέχεια, τελικά αφορά ένα ακόμα πιο ευέλικτο μοντέλο όπου δεν υπάρχει και καμιά υπόθεση σχετικά με τη σχέση ανάμεσα στις δύο μεταβλητές.

Αν ξεκινήσουμε από το κλασσικό μοντέλο της απλής γραμμικής παλινδρόμησης, στην ουσία ξεκινάμε από την υπόθεση ότι η δεσμευμένη αναμενόμενη τιμή της τμ Y για δοθείσα τιμή της τμ $X = x$ είναι γραμμική δηλαδή πως

$$m(x) = E(Y \mid X = x) = \alpha + \beta x$$

Γενικά στη στατιστική όταν αναφερόμαστε στην παλινδρόμηση της τμ Y ως προς την τμ X εννοούμε τη δεσμευμένη αναμενόμενη τιμή. Το μοντέλο της απλής γραμμικής παλινδρόμησης εκτός από την υπόθεση πως η παλινδρόμηση είναι γραμμική χρησιμοποιεί και διάφορες άλλες υποθέσεις, όπως ομοσκεδαστικότητα, κανονικότητα για την κατανομή του πληθυσμού κλπ.

'Εστω δύο τμ X και Y με από κοινού κατανομή $f(x, y)$. Τότε χρησιμοποιώντας απλές ιδιότητες από τις πιθανότητες προκύπτει πως

$$\begin{aligned} m(x) = E(Y \mid X = x) &= \int y f(y \mid x) dy \\ &= \int y \frac{f(x, y)}{f_x(x)} dy \end{aligned}$$

'Εστω πως έχουμε δύο εκτιμήτριες της από κοινού κατανομής και της περιθώριας κατανομής του X χρησιμοποιώντας εκτιμήτριες με τη μέθοδο των kernels. Οι εκτι-

μήτριες αντίστοιχα είναι

$$\begin{aligned}\hat{f}(x,y) &= \frac{1}{nh_xh_y}\sum_{i=1}^{n}K_x\left(\frac{x-x_i}{h_x}\right)K_y\left(\frac{y-y_i}{h_y}\right)\\ \hat{f}_x(x) &= \frac{1}{nh_x}\sum_{i=1}^{n}K_x\left(\frac{x-x_i}{h_x}\right)\end{aligned}$$

όπου ο δείκτης είτε στο kernel είτε στο παράθυρο αναφέρεται στη μεταβλητή που χρησιμοποιούμε. Παρατηρείστε πως στην περίπτωση της από κοινού μεταβλητής έχουμε χρησιμοποιήσει το γινόμενο δύο ανεξαρτήτων kernels, ένα για κάθε μεταβλητή. Θα μπορούσαμε να χρησιμοποιήσουμε πιο πολύπλοκα kernels αλλά με κόστος να γίνει πιο πολύπλοκη η κατάσταση.

Αντικαθιστώντας λοιπόν στη δεσμευμένη αναμενόμενη τιμή προκύπτει πως

$$\begin{aligned}\hat{m}(x) &= \int \frac{y}{\hat{f}_x(x)}\frac{1}{nh_xh_y}\sum_{i=1}^{n}K_x\left(\frac{x-x_i}{h_x}\right)K_y\left(\frac{y-y_i}{h_y}\right)dy =\\ &= \frac{1}{\hat{f}_x(x)}\sum_{i=1}^{n}\frac{1}{nh_x}K_x\left(\frac{x-x_i}{h_x}\right)\int\frac{y}{h_y}K_y\left(\frac{y-y_i}{h_y}\right)dy =\\ &= \frac{1}{\hat{f}_x(x)}\sum_{i=1}^{n}\frac{1}{nh_x}K_x\left(\frac{x-x_i}{h_x}\right)\int\left[uh_y+y_i\right]K_y\left(u\right)du\end{aligned}$$

και παρατηρώντας πως από τον ορισμό των kernels $\int K(u)du = 1$ και $\int uK(u)du = 0$ προκύπτει πως

$$\hat{m}_{NW}(x) = \sum_{i=1}^{n}\frac{1}{nh_x}\frac{K_x\left(\frac{x-x_i}{h_x}\right)y_i}{\hat{f}_x(x)} = \sum_{i=1}^{n}w_iy_i$$

Ο παραπάνω εκτιμητής ονομάζεται Nadaraya-Watson εκτιμητής. Πρέπει να σημειωθούν τα εξής:

Προφανώς τα παραπάνω ισχύουν για οποιαδήποτε επιλογή kernel.

Ο παραπάνω εκτιμητής στην ουσία εκτιμά την παλινδρόμηση για κάθε δοθέν τιμή x χρησιμοποιώντας τα δεδομένα. Δεν έχουμε υποθέσει καμιά μορφή για την παλινδρόμηση και για αυτό, όπως θα δούμε και παρακάτω, η μορφή της είναι ελεύθερη χωρίς περιορισμούς. Στην ουσία δεν έχουμε κάνει καμιά υπόθεση ούτε για την κατανομή του πληθυσμού αλλά ούτε και για τη μορφή της παλινδρόμησης

Και πάλι το παράθυρο καθορίζει το πόσο λεία ή όχι θα είναι η εκτιμήτρια που θα πάρουμε. Κριτήρια επιλογής βέλτιστου παραθύρου υπάρχουν πολλά. Αν η τιμή του παραθύρου τείνει στο 0, τότε η εκτιμήτρια θα είναι 0 για σημεία όπου δεν έχουμε παρατηρήσεις και θα ταυτίζεται με την παρατήρηση στα υπόλοιπα σημεία. Από την άλλη αν το παράθυρο είναι μεγάλο, θα πάρουμε ως εκτίμηση μια ευθεία σταθερή στη μέση τιμή των y.

Καταλαβαίνει κανείς πως για να βρούμε την παλινδρόμηση για δοθέν x λαμβάνουμε υπόψη μας μόνο τα γειτονικά σημεία, όπως το εύρος του παραθύρου ορίζει, δίνοντας περισσότερο βάρος σε παρατηρήσεις κοντά στο x.

Είναι πολύ ενδιαφέρον να παρατηρήσει κανείς πως η εκτιμήτρια μπορεί να γραφτεί ως ένας γραμμικός συνδυασμός των παρατηρήσεων y_i με βάρη w_i, όπου

$$w_i = \frac{1}{nh_x} \frac{K_x\left(\frac{x-x_i}{h_x}\right)}{\hat{f}_x(x)}$$

Στην απλή παλινδρόμηση υπάρχει μια ουσιώδης υπόθεση την οποία συνήθως ξεχνούμε για λόγους πρακτικότητας. Υποτίθεται δηλαδή πως τα X_i είναι σταθερά και προεπιλεγμένα με κάποια κριτήρια. Σε αυτή την περίπτωση δεν είναι πια τυχαίες μεταβλητές όπως η εκτιμήτρια Nadaraya-Watson υποθέτει και είναι λογικό να αντικαταστήσουμε την εκτιμήτρια τους $\hat{f}_x(x)$ με την πραγματική συνάρτηση πυκνότητας πιθανότητας $f_x(x)$. Σε αυτή την περίπτωση η εκτιμήτρια γίνεται

$$w_i = \frac{1}{nh} \frac{K_x\left(\frac{x-x_i}{h_x}\right)}{f_x(x)}$$

Τέλος σε πολλές εφαρμογές οι τιμές των X_i είναι συγκεκριμένοι αριθμοί που ισαπέχουν μεταξύ τους. Για παράδειγμα έστω ότι κάποιος έχει μετρήσεις ανά τακτικά χρονικά διαστήματα και θέλει να δει πως εξελίσσεται το φαινόμενο ως προς το χρόνο. Σε αυτή την περίπτωση η κατανομή του x μπορεί να εκτιμηθεί ως

$$\hat{f}(x) = \frac{1}{n(x_i - x_{i-1})}$$

και η εκτιμήτρια της παλινδρόμησης θα προκύψει με αντίστοιχη αντικατάσταση στα βάρη w_i. Επομένως η εκτιμήτρια της παλινδρόμησης θα είναι

$$\hat{m}(x) = h^{-1} \sum_{i=1}^{n} (x_i - x_{i-1}) K_x\left(\frac{x - x_i}{h}\right) y_i$$

1.13.2 Ερμηνεία της μη παραμετρικής παλινδρόμησης

Μια πολύ ενδιαφέρουσα ερμηνεία της μη παραμετρικής παλινδρόμησης είναι πως ελαχιστοποιεί την παρακάτω συνάρτηση

$$\sum_{i=1}^{n} (y_i - b_0)^2 K\left(\frac{x - x_i}{h}\right)$$

η οποία μπορεί εύκολα να αναγνωριστεί σαν ένα σταθμισμένο άθροισμα τετραγωνικών καταλοίπων, όπου οι σταθμίσεις δίνονται από τα $K\left(\frac{x-x_i}{h}\right)$, δηλαδή τις τιμές των kernels. Στην ουσία ο τύπος αυτός μας λέει πως εκτιμάμε την παλινδρόμηση τοπικά ως μια σταθερή συνάρτηση. Το επόμενο βήμα θα ήταν να εκτιμήσουμε την παλινδρόμηση τοπικά ως ένα πολυώνυμο ως προς x καθώς αυτό θα μας έδινε μεγαλύτερη ευελιξία, δηλαδή να ελαχιστοποιήσουμε μια ποσότητα όπως η

$$\sum_{i=1}^{n} \left[y_i - b_0 - b_1(x - x_i) - ... - b_p(x - x_i)^p\right]^2 K\left(\frac{x - x_i}{h}\right)$$

Η μέθοδος αυτή ονομάζεται local polynomial regression estimation και όπως μπορεί να δει κανείς έχει καλύτερες ιδιότητες (και μεγαλύτερη όμως πολυπλοκότητα) από την απλή μέθοδο που είδαμε.

1.13.3 Διάφορα θέματα και παραδείγματα

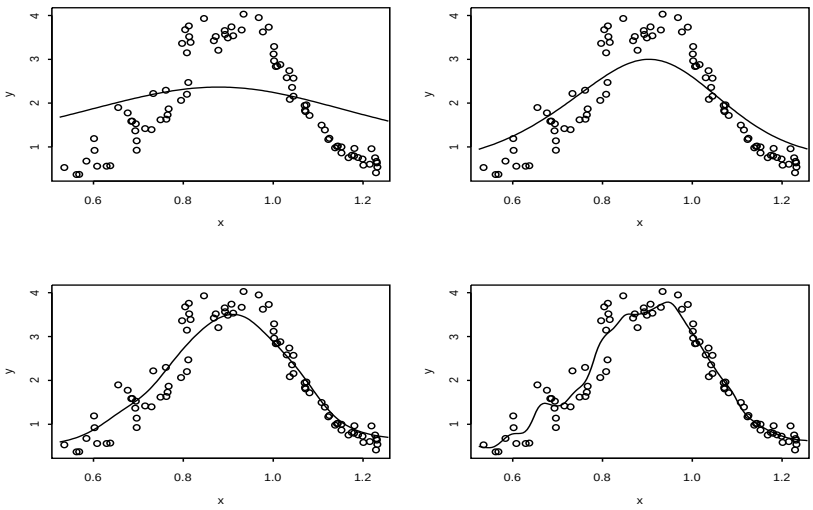
Και στην περίπτωση της μη παραμετρικής παλινδρόμησης η επιλογή της τιμής του παραθύρου είναι σημαντική για το πόσο πολύ θα εξομαλύνουμε ή όχι τα δεδομένα μας. Οι μέθοδοι που χρησιμοποιούσαμε πριν στην περίπτωση που σκοπός ήταν η εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας αν και μπορούν να χρησιμοποιηθούν δεν είναι πια οι καλύτεροι καθώς ο σκοπός είναι διαφορετικός. Για να βρεθεί το βέλτιστο παράθυρο υπάρχουν πολλές μέθοδοι που ελαχιστοποιούν κάποιο κριτήριο, όπως και προηγουμένως, καθώς και μέθοδοι που εισάγουν κάποια ποινή (penalty) ως προς το πόσο λεία είναι η εκτιμήτρια που παίρνουμε (penalized methods). Τέλος και πάλι μπορούν να χρησιμοποιηθούν μέθοδοι cross-validation. Δεν θα επεκταθούμε περισσότερο στη βέλτιστη επιλογή παραθύρου.

Πολλές φορές είναι ενδιαφέρον όχι απλά να εκτιμήσει κανείς την παλινδρόμηση αλλά και να εκτιμήσει τη μεταβλητότητα γύρω από αυτή. Επομένως μπορεί κανείς να κατασκευάσει διαστήματα εμπιστοσύνης γύρω από την εκτιμηθείσα παλινδρόμηση, είτε χρησιμοποιώντας ασυμπτωτικά αποτελέσματα, είτε χρησιμοποιώντας τη μέθοδο bootstrap.

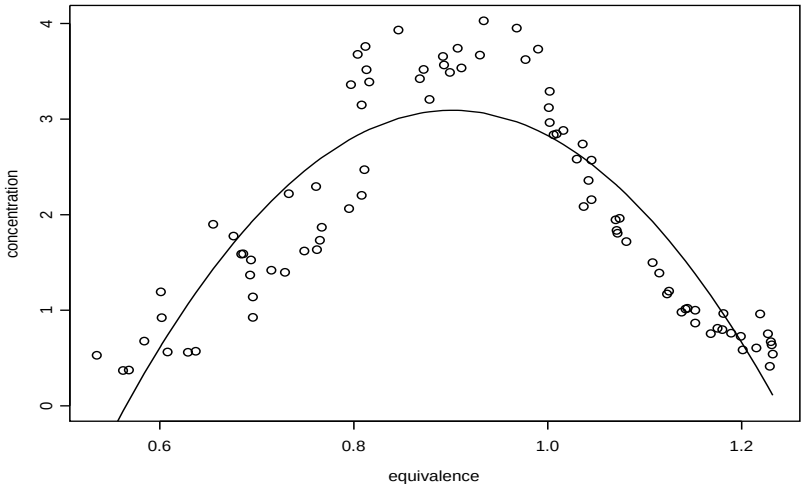
Μπορεί κανείς εύκολα να δει πως η ιδέα της μη παραμετρικής παλινδρόμησης μπορεί να εφαρμοστεί και στην περίπτωση που οι επεξηγηματικές μεταβλητές είναι περισσότερες από μια. Σε αυτή την περίπτωση πρέπει κανείς να εκτιμήσει πολυμεταβλητές από κοινού συναρτήσεις πυκνότητας πιθανότητας και άρα όλο το εγχείρημα απαιτεί ολοένα και περισσότερους υπολογισμούς.

Παράδειγμα 1.6: Το παράδειγμα που ακολουθεί αφορά τη συγκέντρωση νιτρικών οξειδίων κατά τη καύση σε μηχανές αυτοκινήτου. Από μηχανολογική άποψη αυτό που έχει σημασία είναι ο λόγος αέρα και αιθανόλης κατά την καύση. Στο γράφημα 1.24 έχουν προσαρμοσθεί μια σειρά από εκτιμήσεις με τη μέθοδο της μη παραμετρικής παλινδρόμησης. Από το γράφημα μπορεί να δει κανείς πως η συγκέντρωση νιτρικών οξειδίων αυξάνεται μέχρι ο λόγος αέρα και αιθανόλης να γίνει 1 και μετά μειώνεται. Μια πρώτη ιδέα θα ήταν να προσαρμόσει κανείς ένα πολυώνυμο δευτέρου βαθμού αλλά όπως μπορεί κανείς να διακρίνει από το γράφημα 1.25 η προσαρμογή του μοντέλου είναι πολύ κακή.

Στο γράφημα 1.24 έχουν προσαρμοσθεί μια σειρά από μη παραμετρικές καμπύλες με διάφορες τιμές του παραθύρου και συγκεκριμένα $h = 0.02, 0.05, 0.10$ και 0.2 αντίστοιχα. 'Οταν το παράθυρο πάρει τη μέγιστη τιμή (0.2) στην ουσία η παλινδρόμηση τείνει να είναι μια ευθεία γραμμή παράλληλη με τον άξονα, ενώ στην περίπτωση πολύ μικρού παραθύρου (0.02) η παλινδρόμηση φαίνεται να ακολουθεί πιστά τα σημεία. Αντίθετα μέτριες τιμές (0.05 και 0.10) φαίνεται να περιγράφουν πολύ καλά τα δεδομένα χωρίς όμως να χάνουν την ομαλότητα τους.



Γράφημα 1.24: Μη παραμετρική παλινδρόμηση με τη χρήση διαφόρων τιμών για το παράθυρο.



Γράφημα 1.25: Προσαρμογή δευτεροβάθμιου πολυωνύμου στα δεδομένα

1.14 'Αλλες εφαρμογές

1.14.1 Bias Length models

Αν και τις περισσότερες φορές στη στατιστική ξεκινάμε με δεδομένα που είτε προέρχονται από τυχαία δειγματοληψία, είτε εμείς υποθέτουμε πως προέρχονται από τυχαία δειγματοληψία υπάρχουν φαινόμενα όπου είναι ευνόητο πως η δειγματοληψία δεν είναι τυχαία και πως κάθε μονάδα του δείγματος έχει διαφορετική πιθανότητα να παρατηρηθεί. Σε αυτή την περίπτωση υπάρχουν αντικειμενικές δυσκολίες και πρέπει να χρησιμοποιηθούν διαφορετικές μέθοδοι για την οποιαδήποτε στατιστική συμπερασματολογία.

'Εστω ότι η κατανομή του πληθυσμού μας είναι η $f(x)$ αλλά κάθε μονάδα έχει πιθανότητα να παρατηρηθεί η οποία εξαρτάται από την τιμή της, δηλαδή το x και ας υποθέσουμε πως μια παρατήρηση με τιμή x μπορούμε να την παρατηρήσουμε με πιθανότητα ανάλογη της συνάρτησης $w(x)$. Σε αυτή την περίπτωση μπορεί κανείς να δει εύκολα πως η κατανομή από όπου εμείς παίρνουμε δείγμα δεν είναι πια η $f(x)$ αλλά η

$$g^*(x) = \frac{w(x)f(x)}{E(W(x))}$$

Μια ενδιαφέρουσα επιλογή είναι η $w(x) = x$ για την οποία υποθέτουμε πως όσο πιο μεγάλη είναι η τιμή της παρατήρησης τόσο πιο μεγάλη είναι η πιθανότητα να την συμπεριλάβουμε στο δείγμα. Υπάρχουν πολλά παραδείγματα όπου αυτό το μοντέλο είναι λογικό όπως τα εξής:

- θέλουμε να μελετήσουμε κοπάδια πληθυσμών κάποιου ζώου. 'Οσο μεγαλύτερο είναι το κοπάδι τόσο πιο πιθανό είναι να το παρατηρήσουμε.
- θέλουμε να μελετήσουμε τη διάρκεια ζωής κάποιου ανταλλακτικού. Αν το ανταλλακτικό έχει μικρή διάρκεια ζωής η πιθανότητα να το παρατηρήσουμε είναι πολύ μικρότερη από την περίπτωση που το ανταλλακτικό έχει μεγάλη διάρκεια ζωής.

Στην περίπτωση αυτή συνήθως λέμε πως έχουμε bias length sampling και η κατανομή του πληθυσμού από όπου παίρνουμε τελικά το δείγμα δεν είναι η πραγματική αλλά η

$$g(x) = \frac{xf(x)}{\mu}$$

όπου $\mu = E(x)$.

Το πρόβλημα που καλούμαστε να λύσουμε είναι να εκτιμήσουμε τη συνάρτηση πυκνότητας πιθανότητας του πληθυσμού $f(x)$ δοθέντος πως εμείς έχουμε δείγμα από τον πληθυσμό $g(x)$.

Με απλή αντικατάσταση προκύπτει πως

$$f(x) = \frac{g(x)\mu}{x}$$

και επομένως μια εύλογη εκτιμήτρια θα μπορούσε να είναι η

$$\hat{f}(x) = \frac{\hat{g}(x)\hat{\mu}}{x}$$

όπου $\hat{g}(x)$ είναι μια εκτιμήτρια της $g(x)$ από όπου έχουμε δείγμα και $\hat{\mu}$ αντίστοιχα μια εκτιμήτρια του μ. Επειδή οι παρατηρήσεις μας προέρχονται από τη $g(x)$ μπορούμε να την εκτιμήσουμε με τη χρήση της μεθόδου των kernels (ή οποιαδήποτε άλλη μέθοδο). Ομοίως μπορούμε να παρατηρήσουμε πως

$$E_g(x^{-1}) = 1/\mu$$

όπου ο υποδείκτης δηλώνει πως η αναμενόμενη τιμή είναι ως προς τη συνάρτηση πυκνότητας πιθανότητας $g(x)$.Συνεπώς μια εύλογη εκτιμήτρια του μ είναι η

$$\hat{\mu} = \frac{n}{\sum_{i=1}^{n} x_i^{-1}}$$

και συνεπώς είμαστε σε θέση να εκτιμήσουμε την κατανομή του πληθυσμού αν και έχουμε ένα μεροληπτικό δείγμα από αυτόν.

Παράδειγμα 1.7:

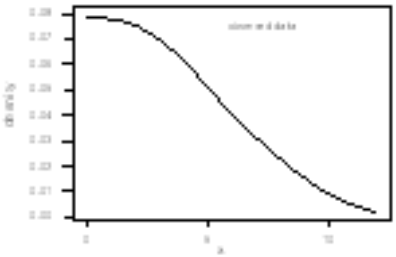
Οι 15 παρατηρήσεις που θα χρησιμοποιήσουμε προέρχονται από ένα μηχανολογικό εργαστήριο που μελετά την αντοχή μεταλλικών συρμάτων όταν σε αυτά εξασκηθεί δύναμη. Μοιάζει λογικό να υποθέσουμε ότι επειδή δεν υπάρχει κάποιο συγκεκριμένο σχέδιο δειγματοληψίας, όπου τα σύρματα υπόκεινται σε συγκεκριμένες διαδικασίες αλλά απλά ο ερευνητής καταγράφει για συγκεκριμένα σύρματα την αντοχή τους. Επομένως, τα σύρματα με μεγαλύτερη αντοχή είναι πιο πιθανό να περιληφθούν στο δείγμα, αφού αν είχαν μικρή αντοχή θα περίμενε κανείς να έχουν καταστραφεί προηγουμένως και επομένως να μην προλάβει ο ερευνητής να τα παρατηρήσει.

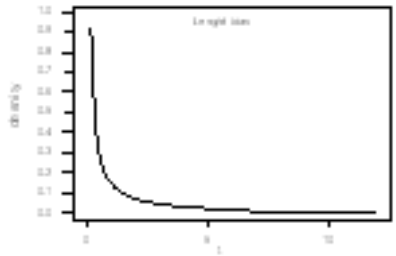
Στα δύο γραφήματα που ακολουθούν μπορεί να δει κανείς την εκτιμήτρια της συνάρτησης πυκνότητας πιθανότητας, τόσο από τα πραγματικά δεδομένα όσο και για την περίπτωση του length bias μοντέλου. Χρησιμοποιήθηκε κανονικό kernel και για να περιορίσουμε την εκτιμήτρια στους θετικούς αριθμούς χρησιμοποιήσαμε τη μέθοδο του διπλασιασμού του δείγματος. Παρατηρείστε πόσο διαφορετικές είναι οι εκτιμήτριες. Τα δεδομένα είχαν σχετικά μεγάλη μέση τιμή, αλλά αν αναλογιστεί κανείς πως μεγάλες τιμές είναι πολύ πιο πιθανό να παρατηρηθούν καταλήγουμε στην εκτιμήτρια κάτω που μας δείχνει πως ο χρόνος αντοχής είναι μια κατανομή που μοιάζει πολύ με την εκθετική.

Γράφημα 1.26 Εκτιμήσεις της συνάρτησης πυκνότητας πιθανότητας των παρατηρηθέντων δεδομένων αλλά και της κατανομής του πληθυσμού υποθέτοντας bias length δειγματοληψία

1.14.2 Διακριτική Ανάλυση

Στη διακριτική ανάλυση σκοπός μας είναι να δημιουργήσουμε κανόνες, από τα δεδομένα που έχουμε και γνωρίζουμε σε ποιο υποπληθυσμό ανήκουν, έτσι ώστε





να μπορούμε με βάση αυτούς τους κανόνες να κατατάξουμε μελλοντικές παρατηρήσεις στους επιμέρους υποπληθυσμούς. Είναι ευνόητο πως για αυτόν τον σκοπό είναι βασικό να ξέρουμε την κατανομή κάθε υποπληθυσμού, αφού μόνο τότε μπορούμε να χρησιμοποιήσουμε πιθανοθεωρητικά κριτήρια κατάταξης. Αν και η διακριτική ανάλυση έχει αναπτυχθεί πάνω στην ιδέα της πολυμεταβλητής κανονικότητας, δηλαδή συνήθως υποθέτουμε πως τα δεδομένα μας προέρχονται από μια πολυμεταβλητή κανονική κατανομή, είναι ξεκάθαρο πως κανείς μπορεί να λειτουργήσει μη παραμετρικά χωρίς να κάνει κάποια παραμετρική υπόθεση για τους υποπληθυσμούς και να χρησιμοποιήσει απλά κάποιες εκτιμήτριες της συνάρτησης πυκνότητας πιθανότητας.

1.14.3 Εκτίμηση με τη χρήση αποστάσεων

Η χρήση των μεθόδων εκτίμησης με kernels έχει επίσης χρησιμοποιηθεί και σε θέματα τα οποία δεν είναι απλές εφαρμογές αλλά αποσκοπούν σε θεωρητικά θέματα, όπως η εκτίμηση παραμέτρων με τη χρήση αποστάσεων. Ας υποθέσουμε πως επιλέγουμε τις παραμέτρους με κριτήριο τα δεδομένα μας να προσαρμόζουν όσο γίνεται πιο καλά στην κατανομή που έχουμε υποθέσει. Αν τα δεδομένα μας είναι διακριτά για παράδειγμα, οι περισσότεροι έχουμε υπόψη μας τον έλεγχο με τη χρήση της ελεγχοσυνάρτησης χ^2 του Pearson. Κάτι το οποίο συνήθως δεν αναφέρεται είναι πως για να έχει νόημα η χρήση της ελεγχοσυνάρτησης αυτής πρέπει και οι παράμετροι να έχουν εκτιμηθεί ώστε να την ελαχιστοποιούν. Δηλαδή όλα τα αποτελέσματα που ισχύουν για τον έλεγχο με την ελεγχοσυνάρτηση χ^2 του

Pearson, ουσιαστικά (χωρίς να το κάνουμε δυστυχώς) στηρίζονται στην ιδέα πως αυτή η απόσταση είναι η μικρότερη δυνατή ανάμεσα στα δεδομένα και τη θεωρητική κατανομή που έχουμε υποθέσει. Αυτό σημαίνει πως η παράμετρος, για παράδειγμα, της κατανομής Poisson έχει υπολογιστεί έτσι ώστε να ελαχιστοποιεί την ελεγχοσυνάρτηση χ^2. Μια τέτοια εκτιμήτρια δεν είναι η εκτιμήτρια μεγίστης πιθανοφάνειας αν και στην πράξη την χρησιμοποιούμε χωρίς ενδοιασμούς.

Στην περίπτωση των διακριτών δεδομένων έχουμε τις παρατηρούμενες συχνότητες που γενικά είναι μια καλή εκτίμηση της πιθανότητας για κάθε τιμή. Ανάλογα με την απόσταση που θα χρησιμοποιήσουμε οδηγούμαστε σε μια μέθοδο εκτίμησης ελαχίστων αποστάσεων (minimum distance estimation) και μπορεί να δειχθεί πολύ εύκολα πως και η μέθοδος μεγίστης πιθανοφάνειας ανήκει σε αυτή την κατηγορία, δηλαδή ελαχιστοποιεί κάποια συγκεκριμένη απόσταση ανάμεσα στα δεδομένα και τη θεωρητική κατανομή.

Για παράδειγμα ας υποθέσουμε πως έχουμε παρατηρήσεις από την κατανομή Poisson και συμβολίσουμε με $d(x)$ την παρατηρηθείσα σχετική συχνότητα της τιμή x και με $\hat{d}(x,\theta)$ την αντίστοιχη αναμενόμενη με βάση την κατανομή Poisson. Χρησιμοποιούμε στον συμβολισμό την παράμετρο θ για να δείξουμε πως η εκτίμηση των αναμενόμενων σχετικών συχνοτήτων εξαρτάται από την τιμή της παραμέτρου της κατανομής που υποθέσαμε. Η εκτιμήτρια ελάχιστης απόστασης θα έχει τη μορφή της ελαχιστοποίησης της ποσότητας $\rho(d(\cdot),\hat{d}(\cdot,\theta))$ ως προς θ. Για παράδειγμα υποθέτοντας μια απόσταση ελαχίστων τετραγώνων ουσιαστικά θα πρέπει να βρούμε το θ το οποίο ελαχιστοποιεί την ποσότητα

$$\sum_{x=0}^{\infty}\Big[d(x)-\hat{d}(x,\theta)\Big]^2$$

Συνήθως αυτό δεν είναι κάτι εύκολο και χρειάζεται αρκετή προσπάθεια. Η βιβλιογραφία είναι γεμάτη από διάφορες αποστάσεις που έχει αποδειχτεί πως έχουν καλές ιδιότητες. Σκοπός μας δεν είναι όμως να μιλήσουμε για την εκτίμηση με ελάχιστες αποστάσεις .

Τι γίνεται όμως στην περίπτωση που έχουμε συνεχή δεδομένα; Η απευθείας χρήση της εμπειρικής συνάρτησης πυκνότητας πιθανότητας δεν διευκολύνει καθόλου καθώς θα είχε πολλές τιμές με μηδενική πυκνότητα. Αντί αυτής η ιδέα είναι να χρησιμοποιήσουμε μια εκτιμήτρια της πυκνότητας και στη συνέχεια να ελαχιστοποιήσουμε την απόσταση από την εκτιμήτρια αυτή ως προς τη θεωρητική κατανομή. Δηλαδή για το παράδειγμα της τετραγωνικής απόστασης ελαχιστοποιούμε

$$\int_{-\infty}^{+\infty}\Big[f(x;\theta)-\hat{f}(x)\Big]^2dx$$

όπου $f(x;\theta)$ είναι η θεωρητική κατανομή με παράμετρο(υς) θ που θέλουμε να εκτιμήσουμε και $\hat{f}(x)$ μια εκτιμήτρια με τη χρήση kernel της συνάρτησης πυκνότητας πιθανότητας που προκύπτει προφανώς από το δείγμα. Έχει αποδειχθεί πως αυτή η προσέγγιση έχει πολύ ενδιαφέρουσες ιδιότητες.

1.15 Προσομοίωση από μια εκτιμήτρια με τη μέθοδο των kernels

'Εστω πως έχουμε μια εκτιμήτρια $\hat{f}(x)$ της συνάρτησης πυκνότητας του πληθυσμού που προέκυψε με τη χρήση του kernel (u). Μερικές φορές είναι χρήσιμο να προσομοιώσουμε από αυτή την κατανομή. Ευτυχώς αυτό είναι πολύ εύκολο αν θυμηθούμε πως η λογική της εκτίμησης με kernel ήταν πως σε κάθε παρατήρηση τοποθετούμε ένα kernel, δηλαδή προσθέτουμε έναν ακόμα όρο μεταβλητότητας. Με αυτή τη λογική μπορούμε να προσομοιώσουμε από την $\hat{f}(x)$ ως εξής

- Βήμα 1ο : Διαλέγουμε τυχαία μια παρατήρηση από το δείγμα. Κάθε παρατήρηση έχει πιθανότητα $1/n$ να επιλεγεί. 'Εστω λοιπόν πως έχει την τιμή X
- Βήμα 2ο : Στη συνέχεια προσομοιώνουμε μια τυχαία μεταβλητή Y από την κατανομή του kernel.
- Βήμα 3ο : Η τυχαία μεταβλητή $Z = X + Y$ είναι μια τυχαία μεταβλητή από την $\hat{f}(x)$

Επομένως το πρόβλημα είναι να προσομοιώσουμε από την κατανομή του kernel κάτι το οποίο γενικά είναι εύκολο. Για παράδειγμα αν το kernel είναι το κανονικό απλά χρειάζεται να προσομοιώσουμε από την κανονική κατανομή κάτι που όλα τα στατιστικά πακέτα μας προσφέρουν.

1.16 Kernel εκτιμήτριες και μίξεις κατανομών

'Εστω πως ένας πληθυσμός αποτελείται από k υποπληθυσμούς, καθένας από τους οποίους είναι ομοιογενής και επομένως όλα τα μέλη του υποπληθυσμού ακολουθούν μια συγκεκριμένη κατανομή, έστω f_j όπου ο δείκτης δείχνει πως μιλάμε για την κατανομή του j υποπληθυσμού. Για να διευκολύνουμε την παρουσίαση ας υποθέσουμε πως όλοι οι υποπληθυσμοί ακολουθούν την ίδια κατανομή αλλά με διαφορετικές παραμέτρους. Π.χ. ας υποθέσουμε πως όλοι οι υποπληθυσμοί ακολουθούν μια κανονική κατανομή αλλά με διαφορετική μέση τιμή έστω $\mu_j, j = 1, \ldots k$. Επομένως συμβολίζουμε την κατανομή του j υποπληθυσμού ως $f(x \mid \theta_j)$. Αν πάρουμε τυχαία ένα άτομο από τον πληθυσμό αυτό και δεν γνωρίζουμε από ποιο υποπληθυσμό προέρχεται τότε από το θεώρημα ολικής πιθανότητας η κατανομή του θα είναι

$$f(x) = \sum_{j=1}^{k} p_j f(x|\theta_j)$$

όπου $0 < p_j < 1, \sum_{j=1}^{k} p_j = 1$ δηλώνει την πιθανότητα ένα τυχαίο άτομο να ανήκει στον υποπληθυσμό j. Παρατηρείστε πως κανείς μπορεί να γενικεύσει την ιδέα υποθέτοντας πως υπάρχουν άπειροι υποπληθυσμοί και να αντικαταστήσει το άθροισμα με ολοκλήρωμα.

Οι μίξεις κατανομών έχουν βρει πληθώρα από εφαρμογές σε μια μεγάλη ποικιλία από στατιστικές μεθόδους. Σκοπός μας είναι να δούμε πως σχετίζονται όμως με την εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας και για αυτό δεν θα επεκταθούμε σε περισσότερες λεπτομέρειες.

Ας περιοριστούμε στη συζήτηση μίξεων κανονικών κατανομών και ας υποθέσουμε για αρχή πως έχουμε $k = 2$, δηλαδή μόλις δύο υποπληθυσμούς. Αν

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

συμβολίσουμε τη συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής τότε ένα μείγμα από δύο κανονικές θα έχει συνάρτηση πυκνότητας πιθανότητας

$$\begin{aligned} f(x|\mu_1, \mu_2, \sigma^2, p_1) &= p_1 f(x|\mu_1, \sigma^2) + (1-p_1) f(x|\mu_2, \sigma^2) \\ &= \frac{p_1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right) + \frac{(1-p_1)}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right) \end{aligned}$$

Η χρησιμότητα της κατανομής αυτής είναι πως μπορεί να πάρει πολλά διαφορετικά σχήματα και επομένως προσφέρει μεγάλη ποικιλία διαφορετικών πληθυσμών που θα μπορούσε να περιγράψει. Για παράδειγμα η παραπάνω κατανομή μπορεί να πάρει διάφορες μορφές όπως να γίνει δικόρυφη, να έχει μεγάλη ουρά προς όποια κατεύθυνση κλπ.

Γενικεύοντας λοιπόν στην περίπτωση πολλών υποπληθυσμών καταλαβαίνει κανείς πως η ποικιλία σχημάτων είναι τεράστια και επομένως μπορεί κανείς να περιγράψει μια τεράστια ποικιλία διαφορετικών περιπτώσεων χρησιμοποιώντας μίξεις της κανονικής κατανομής.

Ας γυρίσουμε τώρα στην περίπτωση της εκτίμησης με τη μέθοδο των kernels. Ας υποθέσουμε, για λόγους ευκολίας, πως δουλεύουμε με το κανονικό kernel. Από τον ορισμό της $\hat{f}(x)$ μπορεί κανείς να δει πως η εκτιμήτρια δεν είναι παρά ένα μείγμα κανονικών κατανομών. Το πλήθος τους είναι ίσο με το μέγεθος του δείγματος, $p_i = \frac{1}{n}$, $i = 1, \ldots n$, ενώ η κατανομή κάθε υποπληθυσμού είναι η κανονική με μέση τιμή x_i και διακύμανση h^2, δηλαδή κάθε υποπληθυσμός έχει μέση τιμή μια παρατήρηση και διακύμανση το παράθυρο υψωμένο στο τετράγωνο.

Αυτό μας επιτρέπει μια ολότελα διαφορετική ματιά στην εκτίμηση συναρτήσεων πυκνότητας πιθανότητας. Για παράδειγμα αν η εκτίμηση για δεδομένα ορισμένα σε όλους τους πραγματικούς αριθμούς μπορεί να γίνει με μίξεις κανονικών, μπορεί κανείς να χρησιμοποιήσει μίξεις της κατανομής Poisson για διακριτά δεδομένα και μίξεις της Γάμμα κατανομής για δεδομένα ορισμένα στους θετικούς αριθμούς. Η περίπτωση του μεταβλητού παραθύρου αντιστοιχεί σε μίξεις της κανονικής κατανομής με διαφορετικές διακυμάνσεις, ενώ ο απλοϊκός εκτιμητής αντιστοιχεί σε μίξεις ομοιόμορφων κατανομών.

1.17 Η μέθοδος Warping

'Ενα πρόβλημα που παρουσιάζεται πολλές φορές στην εκτίμηση με τη χρήση των kernels έχει να κάνει με το μέγεθος του δείγματος. Για πολύ μεγάλα μεγέθη

δείγματος είναι ιδιαίτερα χρονοβόρος ο υπολογισμός καθώς πρέπει κανείς να αθροίσει ως προς όλες τις παρατηρήσεις. Με σκοπό να αποφευχθεί αυτό το πρόβλημα αναπτύχθηκε η μέθοδος WARP (Weighted Average of Rounded Points) η οποία βασίζεται στην απλή υπόθεση πως αντί να πάρουμε όλα τα σημεία τα χωρίζουμε σε κελιά, και για κάθε κελί χρησιμοποιούμε μόνο το μέσο του κελιού τόσες φορές όσες και οι παρατηρήσεις μέσα στο κελί. Δηλαδή αντικαθιστούμε κάθε παρατήρηση μέσα στο κελί με το μέσο του κελιού. Βέβαια το πλάτος του κελιού πρέπει να είναι αρκετά μικρό ώστε να μη δημιουργούμε πρόβλημα από την αντικατάσταση αυτή. Προτείνεται στη βιβλιογραφία το πλάτος αυτών των κελιών να είναι ίσο με $h/5$ ή εναλλακτικά με το 1/100 του εύρους των παρατηρήσεων. Για μεγάλα μεγέθη δείγματος το κέρδος σε υπολογισμούς είναι τεράστιο μπροστά στο μικρό κόστος της ακρίβειας στον υπολογισμό της εκτιμήτριας.

1.18 Ασκήσεις

1. Δίνονται οι παρακάτω 80 παρατηρήσεις. Να εκτιμήσετε τη συνάρτηση πυκνότητας πιθανότητας στο σημείο $x = 100$, χρησιμοποιώντας το ομοιόμορφο kernel και παράθυρο $h = 1.92$.

   Οι παρατηρήσεις σε αύξουσα σειρά είναι

   90,90,91,93,93,94,94,94,94,94,95,96,96,96,96,96,97,97,97,97,97,97,97,97,97,97

   98,98,98,98,98,99,99,99,99,99,99,99,99,99,100,100,100,100,100,100,100,100,100

   101,101,101,101,102,102,102,102,103,103,104,104,104,104,104,104,104,104,104

   105,105,106,106,107,107,107,108,109,109,110,114

2. 'Εστω οι ακόλουθες 50 παρατηρήσεις που αντιστοιχούν στη βαθμολογία 50 φοιτητών στο μάθημα Στατιστικής.

   30 32 32 33 34 38 38 40 42 43

   43 44 44 45 46 46 49 49 51 52

   52 54 56 56 56 59 61 61 63 65

   65 66 69 72 74 74 74 74 75 75

   76 80 81 81 82 83 83 85 85 89

   Εκτιμήστε, χρησιμοποιώντας τη μέθοδο των kernels και χρησιμοποιώντας ομοιόμορφο kernel την πιθανότητα $P(X > 50)$ (Χρησιμοποιήστε παράθυρο $h = 5$).

3. 'Ενας ερευνητής κατέγραψε την ικανοποίηση των πελατών μιας εταιρείας χρησιμοποιώντας ένα δείγμα μεγέθους 20. Οι απαντήσεις ήταν

Πάρα Πολύ	Πολύ	Μέτρια	Λίγο	Καθόλου
4	5	7	4	0

Α) Να εκτιμήσετε την πιθανότητα ένας τυχαίος πελάτης να είναι λίγο ή καθόλου ικανοποιημένος από την εταιρεία

Β) Πως θα μπορούσατε να βελτιώσετε την εκτίμηση σας;

4. Δίνονται 5 παρατηρήσεις με τιμές 10, 15, 25, 25, 30. Χρησιμοποιήστε το ομοιόμορφο kernel ($f(u) = 0.5, |u| \leq 1$) για να εκτιμήσετε τη συνάρτηση κατανομής χρησιμοποιώντας kernel με παράθυρο $h = 5$.

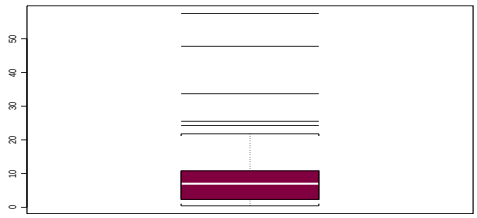
5. Δείξτε πως το μέσο τετραγωνικό σφάλμα (MSE) μιας εκτιμήτριας $\hat{\theta}$ υπολογίζεται ως $MSE(\hat{\theta}) = [Bias(\hat{\theta})]^2 + Var(\hat{\theta})$ όπου Bias($\hat{\theta}$) και Var($\hat{\theta}$) είναι η μεροληψία και η διακύμανση αντίστοιχα της εκτιμήτριας.

6. Δίνονται οι παρακάτω 25 παρατηρήσεις που αφορούν τη βαθμολογία 25 φοιτητών σε ένα τεστ. Να εκτιμήσετε την πιθανότητα $P(99 < \ < 102)$ χρησιμοποιώντας τη μέθοδο των Kernels με ομοιόμορφο kernel και παράθυρο $h = 1.89$.

   Οι παρατηρήσεις σε αύξουσα σειρά είναι

98	98	99	99	99	99	99	99	99	100	100	100	100
100	101	101	102	102	102	103	104	104	109	109	110	

7. Στο διάγραμμα boxplot (1.26 που ακολουθεί απεικονίζονται 50 τιμές διαφόρων ειδών σοκολάτας (σε Euro/ κιλό). Επίσης μπορείτε να δείτε κάποια περιγραφικά στατιστικά μέτρα των δεδομένων



Γράφημα 1.26: Τα δεδομένα της ασκησης

$Min.$	$1stQu.$	$Median$	$Mean$	$3rdQu.$	$Max.$	$stdev$
0.43123	2.41329	7.00666	9.70615	10.71178	57.49416	11.49223

Χρησιμοποιώντας αυτά τα δεδομένα προτείνετε την τιμή του παραθύρου h για να εκτιμήσετε τη συνάρτηση πυκνότητας πιθανότητας των δεδομένων με τη χρήση κανονικού kernel. Κοιτάζοντας το boxplot τι έχετε να προτείνετε για τη βελτίωση της εκτίμησης;

8. Για όλα τα kernels του πίνακα 1.1 υπολογίστε τη διακύμανση. Ποιό kernel έχει τη μεγαλύτερη διακύμανση;

9. Για τα δεδομένα της άσκησης 1 χρησιμοποιήστε μεταβλητό παράθυρο. Τι παρατηρείτε;

10. Δείξτε πως η εκτιμήτρια μιας σππ με τη μέθοδο του ιστογράμματος είναι όντως συνάρτηση πυκνότητας πιθανότητας. Στη συνέχεια βρείτε πια είναι η αναμενόμενη τιμή $E(X)$ για αυτή την εκτιμήτρια, δηλαδή βρείτε την

$$E(X) = \int x\hat{f}(x)dx$$

11. Για την εκτιμήτρια μιας σππ με τη μέθοδο του ιστογράμματος βρείτε την αντίστοιχη εκτιμήτρια συνάρτηση κατανομής $\hat{F}(x)$. Υπενθυμίζουμε πως $\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(t)dt$

12. Έστω η συνάρτηση

$$K(u) = \begin{cases} \frac{1}{\pi}(1-u^2)^{-1/2}, & |u| \leq 1 \\ 0, & \text{αλλού} \end{cases}$$

Χρησιμοποιείστε αυτό το kernel για να εκτιμήσετε τη συνάρτηση πυκνότητας πιθανότητας ενός τυχαίου δείγματος με 100 παρατηρήσεις από την κανονική κατανομή (στην `R` χρησιμοποιήστε την εντολή `rnorm(100)` για να τις προσομοιώσετε), χρησιμοποιώντας $h = 0.2, 0.5, 1$. Πιστεύετε ότι αυτό το kernel είναι κατάλληλο; (Κάντε ένα γράφημα για το kernel για να σας βοηθήσει στο συμπέρασμα σας)

13. Στο Γράφημα 1 υπάρχουν τέσσερα διαφορετικά kernels που προτείνει ένας ερευνητής να χρησιμοποιηθούν. Σχολιάστε για το καθένα την καταλληλότητα του. Αν απορρίπτετε κάποιο από αυτά εξηγήστε τους λόγους.

14. Δίνονται οι παρακάτω 6 παρατηρήσεις:

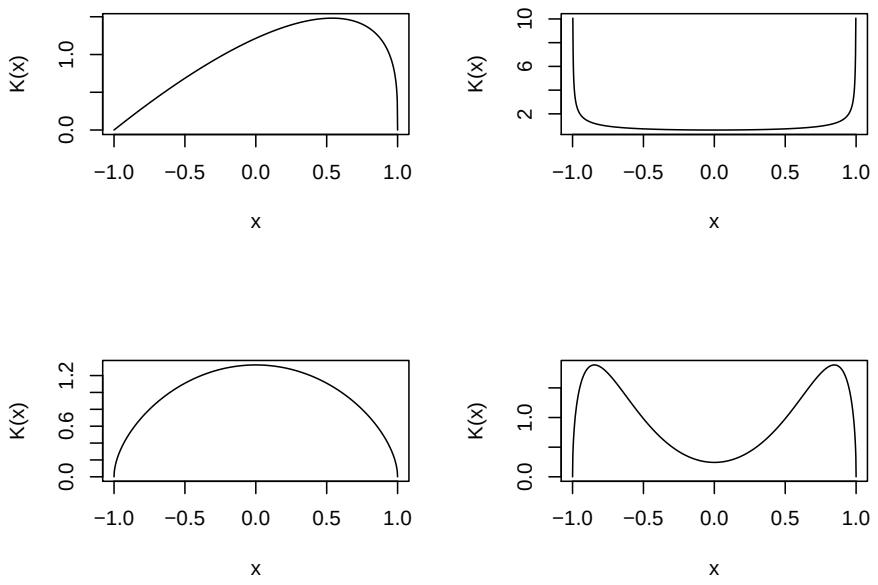
90, 91, 93, 96, 97, 99

Χρησιμοποιείστε αυτές τις 6 παρατηρήσεις, το Epaneshnikov kernel και υπολογίστε την cross-validated πιθανοφάνεια για τιμές του $h = 0.2, 0.5, 0.8, 1$.

Για ποια τιμή του h μεγιστοποιείται η cross-validated πιθανοφάνεια;

15. Για την εκτιμήτρια με τη μέθοδο των kernels βρείτε ποιά είναι η αντίστοιχη εκτιμήτρια της πρώτης και δεύτερης παραγώγου $\hat{f}'(x)$ και $\hat{f}''(x)$ υποθέτοντας κανονικό kernel. Στη συνέχεια προσομοιώστε 100 παρατηρήσεις από την κανονική κατανομή (χρησιμοποιήστε την εντολή `rnorm(100)`) και απεικονίστε τις $\hat{f}(x)$, $\hat{f}'(x)$,$\hat{f}''(x)$ χρησιμοποιώντας το βέλτιστο παράθυρο.

Στη συνέχεια επαναλάβετε την ίδια διαδικασία, αντικαθιστώντας το κανονικό kernel με το ομοιόμορφο kernel. Τι παρατηρείτε;



Γράφημα 1.27:

16. Υποθέτοντας τη χρήση του kernel του Epaneshnikov βρείτε τα σ_K^2 και $R(K)$ αντίστοιχα

17. Για την εκτιμήτρια με τη μέθοδο των kernels βρείτε ποιά είναι η αναμενόμενη τιμή $E()$ και η διακύμανση $Var(X)$ δηλαδή βρείτε την

$$E(X^r) = \int x^r \hat{f}(x)dx$$

$r = 1, 2$ και από αυτές τη διακύμανση $Var(X)$.

18. Για την εκτιμήτρια με τη μέθοδο των kernels βρείτε ποιά είναι η αντίστοιχη εκτιμήτρια συνάρτηση κατανομής $\hat{F}(x)$. Υπενθυμίζουμε πως $\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(t)dt$.

Τι παρατηρείτε; Τι θα συμβεί αν διαλέξετε ένα πολύ μικρό h; Ποιά η σχέση της εκτιμησης σας με την εμπειρική συνάρτηση κατανομής;

19. Δίνονται οι παρακάτω 80 παρατηρήσεις. Να εκτιμήσετε τη συνάρτηση πυκνότητας πιθανότητας στα σημεία $x = 95, 95.8, 97.9, 100.2$, χρησιμοποιώντας ομοιόμορφο kernel με $h = 1, 2, 5, 10, 20$.

Οι παρατηρήσεις σε αύξουσα σειρά είναι

90,90,91,93,93,94,94,94,94,94,95,96,96,96,96,96,97,97,97,97,97,97,97,97,97,97

98,98,98,98,98,99,99,99,99,99,99,99,99,99,100,100,100,100,100,100,100,100,100

101,101,101,101,102,102,102,102,103,103,104,104,104,104,104,104,104,104,104

105,105,106,106,107,107,107,108,109,109,110,114

(Υπόδειξη : Δεν χρειάζεται να χρησιμοποίησετε τον υπολογιστή.)

20. Για τα ίδια δεδομένα χρησιμοποιήστε το κανονικό kernel και το kernel και το kernel του Epaneshnikov διαλέγοντας 3 διαφορετικά h (όποια θέλετε). Δείτε συγκριτικά τα γραφήματα των εκτιμητριών σας.

21. Αν υποθέσουμε πως ο πληθυσμός ακολουθεί την εκθετική κατανομή με παράμετρο θ και σππ $f(x) = \theta exp(-x\theta)$ βρείτε το βέλτιστο παράθυρο για την εκτίμηση της σππ με kernels. Πως θα εκτιμήσετε το θ;

22. Ένας ερευνητής πρότεινε τη χρήση του εξής kernel για την εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας:

$$K(x) = \begin{cases} 6x(1-x) & x \in (0,1) \\ 0 & \text{αλλού} \end{cases}$$

Σχολιάστε την επιλογή του συγκεκριμένου kernel.

Βιβλιογραφία

Silverman, C (1986) *Density Estimation for statistics and Data Analysis*. Chapman and Hall.

Εισαγωγικό βιβλίο κατάλληλο για αρχάριους. Δυστυχώς έχει γραφτεί αρκετά παλιά με αποτέλεσμα να μην περιέχει αρκετές από τις νέες ιδέες στο θέμα

Simonoff, J. (1996) *Smoothing Methods in Statistics*. Springer

Πολύ χρήσιμο βιβλίο καθώς αν και περιέχει μια ποικιλία από θέματα σε αρκετά προχωρημένη μορφή δεν είναι δύσκολο να το παρακολουθήσει και κάποιος χωρίς πολλές γνώσεις στο αντικείμενο

Hardle, W (1991) *Smoothing Techniques with Implementation in S.* Springer-Verlag

'Οχι πολύ προχωρημένο βιβλίο αν και πυκνογραμμένο. 'Ολα τα παραδείγματα συνοδεύονται και από τις εντολές για την εκτέλεση τους στο στατιστικό πακέτο S-Plus

Κεφάλαιο 2

'Ελεγχοι Τυχαιοποίησης

2.1 Εισαγωγή

Οι έλεγχοι υποθέσεων αποτελούν ένα σημαντικό κομμάτι της στατιστικής επιστήμης και συμπερασματολογίας. Τα βασικά 'συστατικά' ενός ελέγχου υποθέσεων είναι:

- Η μηδενική και η εναλλακτική υπόθεση, όπου η μηδενική συνήθως περιέχει μια υπόθεση που θέλουμε να δούμε αν ισχύει και η εναλλακτική είναι η υπόθεση προς την οποία πρέπει να κινηθούμε αν δούμε ότι η μηδενική υπόθεση δεν ισχύει.

- Η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε για να κάνουμε τον έλεγχο. Η ελεγχοσυνάρτηση, η οποία είναι συνάρτηση των παρατηρήσεων, πρέπει να έχει την ικανότητα να διακρίνει ανάμεσα στις δύο υποθέσεις, δηλαδή να παίρνει διαφορετικές τιμές στην περίπτωση που ισχύει η μηδενική υπόθεση από την περίπτωση που δεν ισχύει.

- Το επίπεδο στατιστικής σημαντικότητας το οποίο είναι η πιθανότητα να απορρίψουμε εσφαλμένα την μηδενική υπόθεση. Συνήθως διαλέγουμε $\alpha = 5\%$.

- Μια κριτική τιμή από την κατανομή που ακολουθεί η ελεγχοσυνάρτηση και η οποία είναι συνάρτηση και του επιπέδου στατιστικής σημαντικότητας. Με αυτή την κριτική τιμή και συγκρίνοντας την με την παρατηρούμενη τιμή της ελεγχοσυνάρτησης μπορούμε να καταλήξουμε σε κάποια απόφαση σχετικά με το αν έχουμε αρκετά στοιχεία να απορρίψουμε τη μηδενική υπόθεση.

Τα δύο τελευταία συστατικά τείνουν τα τελευταία χρόνια να ξεπεραστούν. Παλιότερα οι έλεγχοι υποθέσεων κατέληγαν σε μια απόφαση δέχομαι ή απορρίπτω (ή καλύτερα απορρίπτω ή δεν απορρίπτω) τη μηδενική υπόθεση συγκρίνοντας την τιμή της ελεγχοσυνάρτησης με την κριτική (ή τις κριτικές μερικές φορές) τιμή. Τα τελευταία χρόνια ολοένα και περισσότερο χρησιμοποιείται μια άλλη λογική. Οι έλεγχοι υποθέσεων έγιναν πια έλεγχοι σημαντικότητας (significance testing) με την έννοια πως δεν καταλήγουμε σε μια απόφαση αλλά απαντούμε με μια πιθανότητα,

το p-value, που είναι η πιθανότητα να πάρει η ελεγχοσυνάρτησή μας μια τιμή τόσο ακραία ή και περισσότερο ακραία από την παρατηρούμενη. Εμπειρικά μπορεί κανείς να πει πως το p-value μας δείχνει πόσο ισχυρή είναι η μηδενική υπόθεση.

Η σχέση ανάμεσα στους κλασσικούς ελέγχους υποθέσεων και τους ελέγχους σημαντικότητας είναι πως απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο στατιστικής σημαντικότητας α αν το p-value είναι μικρότερο από α. Για παράδειγμα αν $\alpha = 5\%$ και το p-value που βρήκαμε είναι 0.10 δεν απορρίπτουμε τη μηδενική υπόθεση.

Η χρήση του p-value έχει εξαπλωθεί κυρίως λόγω της χρήσης των ηλεκτρονικών υπολογιστών που μπορούν εύκολα και γρήγορα να το υπολογίσουν. Παλιότερα ο ερευνητής έπρεπε να έχει μια σειρά από πίνακες με τις κριτικές τιμές κάθε κατανομής ώστε να μπορεί να κάνει τον έλεγχο. Κάτι τέτοιο φαίνεται πια ξεπερασμένο.

Πριν λοιπόν προχωρήσουμε στην περιγραφή ελέγχων υποθέσεων με τη χρήση του υπολογιστή ας σχολιάσουμε ένα άλλο πρόβλημα που απασχολούσε παλιότερα τους ερευνητές. Αυτό είχε να κάνει με την επιλογή της ελεγχοσυνάρτησης. Η ελεγχοσυνάρτηση δεν έπρεπε μόνο να μπορεί να διακρίνει ανάμεσα στις δύο υποθέσεις αλλά από την άλλη έπρεπε η κατανομή της να είναι γνωστή ώστε να είναι δυνατό να πινακοποιηθούν οι κριτικές τιμές που χρειάζονται για να καταλήξει κανείς σε αποφάσεις δέχομαι ή απορρίπτω. Μερικές από τις πιο ευρέως γνωστές ελεγχοσυναρτήσεις είναι πολύπλοκες μόνο και μόνο για να μπορούν να πινακοποιηθούν εύκολα. Πολλές φορές μάλιστα η κατανομή της ελεγχοσυνάρτησης κάτω από τη μηδενική υπόθεση είναι γνωστή μόνο ασυμπτωτικά, όταν δηλαδή το μέγεθος του δείγματος είναι μεγάλο. Σε άλλες περιπτώσεις η κατανομή αυτή βασίζεται σε υποθέσεις που μερικές φορές δεν ισχύουν ή τουλάχιστον είναι δύσκολο να ελέγξουμε αν ισχύουν.

Στη συνέχεια θα παρουσιάσουμε τον τρόπο με τον οποίο μπορεί κανείς να χρησιμοποιήσει μεθόδους προσομοίωσης ώστε να κάνει ελέγχους υποθέσεων ή σημαντικότητας.

2.2 Ακριβής 'Ελεγχος Τυχαιοποίησης

Οι έλεγχοι τυχαιοποίησης αποτελούν μια ομάδα ελέγχων με τη χρήση υπολογιστή οι οποίοι μπορούν να βοηθήσουν σε πολλά προβλήματα. Είναι βασικό πως ο ερευνητής δεν χρειάζεται σχεδόν καθόλου υποθέσεις για να τους χρησιμοποιήσει, και επομένως μπορούν να χρησιμοποιηθούν εκεί που άλλοι κλασικοί έλεγχοι αποτυγχάνουν. Ας δούμε λοιπόν ένα παράδειγμα ενός ελέγχου τυχαιοποίησης.

Παράδειγμα 2.1: 5 ασθενείς υποβλήθηκαν σε δυο διαφορετικές θεραπείες Α και Β. Στην συνέχεια οι γιατροί με βάση κάποια ιατρική κλίμακα βαθμολόγησαν την πρόοδο των ασθενών. Τα δεδομένα είναι τα εξής:

Θεραπεία Α (2 ασθενείς): 1 ,2

Θεραπεία Β (3 ασθενείς): 3,5,9

Υπάρχει διαφορά ανάμεσα στις δύο θεραπείες;

Το πρόβλημα έχει να κάνει με το αν δύο μέσες τιμές διαφέρουν ή όχι, δηλαδή ο ερευνητής θέλει να ελέγξει την

$H_0 : \mu_{\mathrm{A}} = \mu_{\mathrm{B}}$ έναντι της εναλλακτικής

$H_1 : \mu_{\mathrm{A}} \neq \mu_{\mathrm{B}}$

Για να ελεγχθεί λοιπόν αυτή η υπόθεση ο ερευνητής θα μπορούσε να χρησιμοποιήσει το κλασσικό t-test το οποίο όμως βασίζεται σε κάποιες υποθέσεις περί κανονικότητας, οι οποίες δεν ξέρουμε αν ισχύουν στην περίπτωση του παραδείγματος. Εναλλακτικά, θα μπορούσε να χρησιμοποιήσει κάποιο μη παραμετρικό έλεγχο (non parametric test).

Στην πραγματικότητα αν η μηδενική υπόθεση ισχύει, αυτό σημαίνει πως οι δύο θεραπείες είναι ίδιες και επομένως δεν έχει σημασία ποια θεραπεία πήρε ο κάθε ασθενής. Δηλαδή, αν κάποιος 'ξαναμοίραζε' τυχαία τους ασθενείς στις 2 θεραπείες θα έπρεπε να πάρει παρόμοια αποτελέσματα και η όποια διαφορά στα αποτελέσματα είναι απλά τυχαιότητα. Επίσης αν ισχύει η μηδενική υπόθεση κάθε συνδυασμός των 5 ασθενών σε 2 ομάδες από 3 και 2 ασθενείς είναι ισοπίθανος.

Πριν προχωρήσουμε ας ορίσουμε την ελεγχοσυνάρτηση με την οποία θα δουλέψουμε. Αν $\bar{a}, \bar{b}$ είναι οι μέσες τιμές για τους ασθενείς των δύο θεραπειών, η ποσότητα $T = |\bar{a} - \bar{b}|$ είναι μια ελεγχοσυνάρτηση η οποία αν ισχύει η μηδενική υπόθεση παίρνει τιμές κοντά στο 0 ενώ αν δεν ισχύει, τιμές μακριά από το 0. Για τα δεδομένα μας βρίσκουμε πως η τιμή της ελεγχοσυνάρτησης είναι $t_{obs} = 4.51$.

Ο έλεγχος τυχαιοποίησης βασίζεται στο γεγονός πως αν πάρουμε όλους τους δυνατούς τρόπους ώστε να μοιράσουμε τους 5 ασθενείς σε 2 ομάδες από 2 και 3 ασθενείς αντίστοιχα (κάθε ασθενής κρατά μαζί του το βαθμό βελτίωσης που έχει) τότε βρίσκοντας για όλους αυτούς τους συνδυασμούς την τιμή της ελεγχοσυνάρτησης μπορούμε να κρίνουμε αν η τιμή που πήραμε είναι πραγματικά μεγάλη ή απλά οφείλεται στην τύχη. Για τα δεδομένα μας υπάρχουν $\begin{pmatrix} 5 \\ 3 \end{pmatrix} = 10$ διαφορετικοί συνδυασμοί, όπως μπορεί να δει κανείς στον πίνακα 2.1.

Θεραπεία Α	Θεραπεία Β	$\bar{a}$	$\bar{b}$	Τ
1,3	2,5,9	2	5.33	3.33
1,5	2,3,9	3	4.66	1.66
1,9	2,3,5	5	3.33	1.66
1,2	3,5,9	1.5	5.66	4.51
2,3	1,5,9	2.5	5	2.5
2,5	1,3,9	3.5	4.33	0.83
2,9	1,3,5	5.5	3	2.5
3,5	1,2,9	4	4	0
3,9	1,2,5	6	2.66	3.34
5,9	1,,2,3	7	2	5

Πίνακας 2.1: 'Ολοι οι δυνατοί συνδυασμοί ασθενών σε 2 ομάδες και η τιμή της ελεγχοσυνάρτησης για καθένα από αυτούς.

Παρατηρούμε επομένως πως η τιμή 4.51 που υπολογίσαμε για τα δεδομένα μας είναι η δεύτερη μεγαλύτερη από τις 10 δυνατές τιμές που μπορεί να πάρει η ελεγχοσυνάρτηση. Από τον ορισμό του p-value που είδαμε νωρίτερα έχουμε πως

$$\text{p} - \text{value} = P(T \geq t_{obs} | H_0)$$

και επομένως μπορούμε να υπολογίσουμε το p-value του ελέγχου τυχαιοποίησης ως

$$p - value = \frac{\text{αριθμός συνδυασμών με } \ T \geq t_{obs}}{\text{συνολικός αριθμός συνδυασμών}}.$$

΄Ετσι στο παράδειγμα μας βρίσκουμε πως p-value=0.2. Αυτό σημαίνει πως αν θέλουμε να κάνουμε κλασσικό έλεγχο υποθέσεων σε επίπεδο στατιστικής σημαντικότητας $\alpha = 10\%$ δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση

Πριν περιγράψουμε τον έλεγχο τυχαιοποίησης με βήματα ας παρατηρήσουμε το εξής. Στον ορισμό του p-value πιο πάνω χρησιμοποιήσαμε την ποσότητα $P(T \geq t_{obs}|H_0)$. Αυτό δεν είναι γενικά σωστό καθώς σε μερικές περιπτώσεις οι τιμές της ελεγχοσυνάρτησης που μας οδηγούν σε απόρριψη είναι οι μικρές, δηλαδή απορρίπτουμε όταν $T \leq t_{obs}$, και όχι οι μεγάλες όπως υπονοεί η παραπάνω ποσότητα. Από την άλλη καταλαβαίνει κανείς πως αλλάζοντας το πρόσημο ή κάνοντας κάποιον άλλο μετασχηματισμό, μπορούμε να κάνουμε την ελεγχοσυνάρτηση να δείχνει απόρριψη της μηδενικής υπόθεσης για μεγάλες τιμές. Για αυτό το λόγο στο εξής θα θεωρήσουμε πως απορρίπτουμε για μεγάλες τιμές κρατώντας στο νου μας πως απλά αρκεί να ορίσουμε κατάλληλα την ελεγχοσυνάρτηση.

’Αρα η γενική διαδικασία για να κάνουμε έναν ακριβή έλεγχο τυχαιοποίησης είναι η εξής

Ακριβής ’Ελεγχος Τυχαιοποίησης

- Βήμα 1ο: Θέσε τη μηδενική υπόθεση που δείχνει ότι δεν υπάρχει διαφορά (η εναλλακτική είναι ότι υπάρχει διαφορά). Γενικά η μηδενική υπόθεση σε ελέγχους τυχαιοποίησης δείχνει πως δεν υπάρχει κάποια δομή στα δεδομένα.

- Βήμα 2ο : Διάλεξε την ελεγχοσυνάρτηση, υπολόγισε την τιμή της t_{obs} για τα δεδομένα που έχεις. Για τους λόγους που εξηγήσαμε παραπάνω η ελεγχοσυνάρτηση πρέπει να δείχνει απόκλιση από τη μηδενική υπόθεση όταν παίρνει μεγάλες τιμές

- Βήμα 3ο: Δημιούργησε όλους τους δυνατούς συνδυασμούς δεδομένων και υπολόγισε την τιμή της ελεγχοσυνάρτησης για καθέναν από αυτούς. ’Ολες αυτές οι τιμές, δεδομένου πως αποτελούν το σύνολο των δυνατών τιμών της ελεγχοσυνάρτησης για το δείγμα μας, αποτελούν την ακριβή συνάρτηση κατανομής της ελεγχοσυνάρτησης.

- Βήμα 4ο: Υπολόγισε το p-value ως

$$p - value = \frac{\text{αριθμός συνδυασμών με } \ T \geq t_{obs}}{\text{συνολικός αριθμός συνδυασμών}}$$

Μερικά ενδιαφέροντα σημεία είναι τα εξής:

1. Η επιλογή της ελεγχοσυνάρτησης έχει το εξής χρήσιμο χαρακτηριστικό. Δεν μας ενδιαφέρει να διαλέξουμε μια ελεγχοσυνάρτηση για την οποία να ξέρουμε την κατανομή της όταν ισχύει η μηδενική υπόθεση. Για παράδειγμα ο έλεγχος

t-test για δύο δείγματα χρησιμοποιεί την ελεγχοσυνάρτηση $\frac{\bar{x}-\bar{y}}{\sqrt{s_p^2}}$ όπου s_p^2 είναι η εκτίμηση της κοινής διακύμανσης με σκοπό η κατανομή της ελεγχοσυνάρτησης να είναι η κατανομή *t-Student* με κάποιους βαθμούς ελευθερίας. Στον έλεγχο τυχαιοποίησης δεν μας ενδιαφέρει να βρούμε την κατανομή της ελεγχοσυνάρτησης θεωρητικά, αφού στην πράξη την κατασκευάζουμε εμείς. Προσέξτε πως έχουμε στον πίνακα όλες τις δυνατές τιμές της ελεγχοσυνάρτησης για τα δεδομένα μας.

2. Στο παράδειγμα μας θέλαμε να ελέγξουμε αν δύο μέσες τιμές διαφέρουν. Πολλά άλλα προβλήματα μπορούν να τεθούν κάτω από τη μορφή ενός ελέγχου τυχαιοποίησης, όπως αν υπάρχει σχέση ανάμεσα σε μεταβλητές, αν υπάρχει τυχαιότητα ή όχι στα δεδομένα κλπ.

Ο παραπάνω έλεγχος λέγεται ακριβής γιατί εξαντλήσαμε όλους τους δυνατούς συνδυασμούς που μπορούσαν να προκύψουν από τα δεδομένα μας. Λέγεται επίσης έλεγχος τυχαιοποίησης γιατί μοιράσαμε τυχαία τις παρατηρήσεις στις ομάδες για να δούμε αν υπάρχουν ή όχι διαφορές. Μερικές φορές στη βιβλιογραφία χρησιμοποιείται και η ονομασία *permutation test*. Στην πράξη λίγες φορές κάνουμε τέτοιους ακριβείς ελέγχους, για λόγους που θα δούμε αμέσως μετά. Αξίζει να σημειωθεί ότι, ο πιο γνωστός ίσως, και πιο διαδεδομένος ακριβής έλεγχος τυχαιοποίησης είναι ο ακριβής έλεγχος του Fisher για έλεγχο ανεξαρτησίας σε πίνακα συνάφειας.

Δυστυχώς δεν είναι πάντα εύκολο να δημιουργήσουμε όλα τα δυνατά δείγματα που προκύπτουν από την αναδιάταξη των δεδομένων. Σε μερικές όμως περιπτώσεις είμαστε σε θέση να βρούμε τον ακριβή αριθμό των δειγμάτων που θα έδιναν μια τιμή στην ελεγχοσυνάρτηση πιο ακραία από αυτή που έχουμε. Για παράδειγμα ας υποθέσουμε πως έχουμε δύο ομάδες με 20 και 10 παρατηρήσεις αντίστοιχα που δινονται στον πίνακα 2.2:

A	3,4,4.3,4.5,4.8,5,6,5,7,8,9,10,12,13,14,15,16,17,22,23
B	20,24,26,28,29,31,32,34,34,35

Πίνακας 2.2: Τα δεδομένα του παραδείγματος μας

Θέλουμε να ελέγξουμε τη μηδενική υπόθεση ότι οι δύο μέσοι είναι ίσοι έναντι της εναλλακτικής ότι η ομάδα Β έχει μεγαλύτερη μέση τιμή. 'Οπως θα δούμε και αργότερα μια χρήσιμη ελεγχοσυνάρτηση είναι η $T = \sum_{i=1}^{10} X_i^B$, δηλαδή το άθροισμα των τιμών της ομάδας Β. Η παρατηρούμενη τιμή είναι 293. Ο συνολικός αριθμός δειγμάτων είναι $\begin{pmatrix} 30 \\ 10 \end{pmatrix} = 30045015$, δηλαδή περισσότερες από 30 εκατομμύρια περιπτώσεις. Ο αριθμός είναι απαγορευτικά μεγάλος και ίσως κάποιος να σήκωνε τα χέρια ψηλά και να προσπαθούσε να απαντήσει με διαφορετικό τρόπο. Μπορεί όμως να παρατηρήσει κανείς πως ο μόνος τρόπος για να επιτευχθεί μεγαλύτερη τιμή στην ελεγχοσυνάρτηση είναι να υπάρχουν στην ομάδα Β, οι 8 τελευταίες παρατηρήσεις δηλαδή 26,28,29,31,32,34,34 και 35 και οι υπόλοιπες 2 να είναι τα ζεύγη (22,24), (22,23) και (23,24) και αυτό βέβαια που έχουμε παρατηρήσει, το (20,24).

Σε κάθε άλλη περίπτωση η τιμή της ελεγχοσυνάρτησης θα είναι μικρότερη από την παρατηρηθείσα. Συνεπώς, χωρίς να δημιουργήσουμε όλα τα δείγματα μπορούμε να βρούμε πως υπάρχουν μόλις 4 από τα 30045015 δυνατά δείγματα και άρα το p-value είναι ο λόγος 4/30045015, που πρακτικά είναι 0.

2.3 Προσεγγιστικοί έλεγχοι τυχαιοποίησης

'Ενα σημαντικό πρόβλημα που έχουν οι ακριβείς έλεγχοι τυχαιοποίησης είναι το εξής. Στο παράδειγμα που μόλις είδαμε έχουμε πάνω από 30 εκατομμύρια περιπτώσεις. Μπορεί βέβαια τα δεδομένα να ήταν βολικά στην περίπτωση που εξετάσαμε αλλά γενικά σε μια άλλη περίπτωση θα μπορούσε τα πράγματα να ήταν περισσότερο πολύπλοκα. Και αν τα 30 εκατομμύρια μπορούν να αντιμετωπιστούν, έστω με τεράστιο κόπο από τη σημερινή υπολογιστική δύναμη υπάρχουν ακόμα πιο δύσκολες περιπτώσεις.

'Εστω ότι είχαμε 30 και 20 ασθενείς αντίστοιχα στις 2 θεραπείες. Για να κάνουμε τον έλεγχο όπως είδαμε προηγουμένως θα έπρεπε να φτιάξουμε όλους τους δυνατούς συνδυασμούς που τώρα είναι $\begin{pmatrix} 50 \\ 20 \end{pmatrix}$=84832582039128, ένα μάλλον απαγορευτικό μέγεθος ακόμα και για πολύ ικανούς υπολογιστές.

Επειδή λοιπόν πλήρης υπολογισμός όλων των συνδυασμών δεν είναι εφικτός, μια λύση είναι να δημιουργήσουμε όχι όλους αλλά έναν αριθμό από τους δυνατούς συνδυασμούς με τυχαίο τρόπο και να χρησιμοποιήσουμε αυτές τις τιμές της ελεγχοσυνάρτησης ως μια εκτίμηση, πλέον, της κατανομής της ελεγχοσυνάρτησης για να κάνουμε τον έλεγχο.

'Ετσι ο προσεγγιστικός έλεγχος τυχαιοποίησης διαφέρει από τον ακριβή μόνο στο ότι αντί να πάρουμε όλους τους συνδυασμούς παίρνουμε ένα τυχαίο δείγμα από αυτούς. Δηλαδή έχουμε

Προσεγγιστικός 'Ελεγχος Τυχαιοποίησης

- Βήμα 1ο: Θέσε τη μηδενική υπόθεση που δείχνει ότι δεν υπάρχει διαφορά
- Βήμα 2ο: Διάλεξε την ελεγχοσυνάρτηση, υπολόγισε την τιμή της t_{obs} για τα δεδομένα που έχεις.
- Βήμα 3ο: Δημιούργησε k συνδυασμούς δεδομένων αντί για όλους τους δυνατούς και υπολόγισε την τιμή της ελεγχοσυνάρτησης για καθέναν από αυτούς
- Βήμα 4ο: Εκτίμησε το p-value ως

$$p - value = \frac{m+1}{\mathrm{k}+1} \quad ,$$

όπου $m =$ αριθμός συνδυασμών με$T \geq t_{obs}$

Από τα παραπάνω υπάρχουν κάποια σημεία που χρειάζονται εξηγήσεις. Αυτά είναι

Πόσους συνδυασμούς να πάρουμε, δηλαδή ποια είναι η τιμή του k;

Γιατί εκτιμούμε έτσι το p-value;

Σε αυτά τα ερωτήματα θα απαντήσουμε σε λίγο. Πριν όμως γίνει αυτό πρέπει να τονιστούν τα εξής:

- Επειδή παίρνουμε απλά ένα τυχαίο δείγμα από όλους τους δυνατούς συνδυασμούς, αυτό σημαίνει πως 2 ερευνητές με τα ίδια δεδομένα μπορεί να καταλήξουν σε διαφορετικά αποτελέσματα (διαφορετικά p-values). Η διαφορά έχει να κάνει και με την τιμή του k, πόσους από τους δυνατούς συνδυασμούς θα πάρουμε αλλά και με άλλα θέματα.
- Από την άλλη αν τα p-values που βρουν οι ερευνητές διαφέρουν μεν αλλά στην ουσία δείχνουν προς την ίδια κατεύθυνση ως προς την ισχύ της μηδενικής υπόθεσης δεν υπάρχει πρόβλημα. 'Ενα p-value με τιμή 0.80 διαφέρει από ένα άλλο με τιμή 0.82 αλλά και τα 2 δείχνουν πως η μηδενική υπόθεση είναι αρκετά ισχυρή. Το πρόβλημα υπάρχει αν η τιμή του p-value είναι κοντά στο 0.05, για παράδειγμα, το οποίο σε πολλές περιπτώσεις είναι η τιμή που αποφασίζω να δεχτώ ή όχι τη μηδενική υπόθεση. Εκεί χρειάζεται προσοχή. (Αν κάνουμε έλεγχο σημαντικότητας και άρα δεν αποφασίζουμε δέχομαι ή απορρίπτω πάλι δεν υπάρχει πρόβλημα)
- Επειδή το p-value λοιπόν είναι και αυτό με τη σειρά του μια τυχαία μεταβλητή, μπορούμε να κάνουμε στατιστική συμπερασματολογία χρησιμοποιώντας στατιστικά εργαλεία όπως θα δούμε σε λίγο.

Παράδειγμα 2.2: 20 φοιτητές και 20 φοιτήτριες έγραψαν ένα διαγώνισμα στα Μαθηματικά και βαθμολογήθηκαν με άριστα το 60. Τα δεδομένα δίνονται στον πίνακα 2.3 . Υπάρχει διαφορά στους μέσους όρους των δύο φύλων;

Αγόρια	39, 44, 43, 47, 39, 46, 43, 43, 47, 41, 53, 60, 46, 45, 47, 53, 53, 57, 50, 46
Κορίτσια	53, 40, 53, 33, 52, 60, 49, 51, 44, 36, 44, 49, 39, 53, 51, 37, 48, 47, 42, 37

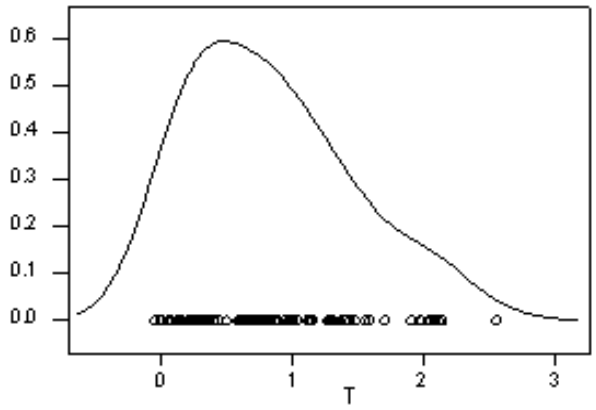
Πίνακας 2.3: Τα δεδομένα του παραδείγματος 2.2

Το πρόβλημα είναι όμοιο με αυτό που είδαμε προηγουμένως. Θέλουμε να συγκρίνουμε δύο μέσες τιμές. Από το δείγμα βρίσκουμε πως οι δειγματικοί μέσοι είναι 47.1 για τα αγόρια και 45.9 για τα κορίτσια. Η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε είναι πάλι η $T = |\bar{x}_A - \bar{x}_K|$ και η τιμή για το δείγμα μας είναι 1.2. Αποτελεί αυτή η τιμή ένδειξη για να απορρίψουμε τη μηδενική υπόθεση; Επειδή όλοι οι δυνατοί συνδυασμοί είναι $\begin{pmatrix} 40 \\ 20 \end{pmatrix}$ =137846528820, επομένως αποφασίζουμε να μη κάνουμε ακριβή έλεγχο αλλά προσεγγιστικό. 'Αρα δημιουργούμε $k = 99$ συνδυασμούς (δηλαδή μοιράζουμε τις 40 παρατηρήσεις τυχαία σε 2 ομάδες από 20 παρατηρήσεις) και υπολογίζουμε την τιμή της ελεγχοσυνάρτησης. Οι 99 τιμές της ελεγχοσυνάρτησης που βρήκαμε είναι (σε αύξουσα σειρά)

0.00, 0.00, 0.00, 0.00, 0.05, 0.05, 0.05, 0.10, 0.10, 0.15, 0.15,

0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.25, 0.25, 0.25, 0.25, 0.25,
0.25, 0.30, 0.30, 0.30, 0.35, 0.35, 0.35, 0.35, 0.35, 0.40, 0.40,
0.40, 0.40, 0.45, 0.45, 0.45, 0.55, 0.55, 0.55, 0.60, 0.65, 0.70,
0.70, 0.70, 0.70, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.85,
0.85, 0.85, 0.85, 0.85, 0.90, 0.90, 0.90, 0.95, 0.95, 0.95, 0.95,
1.00, 1.05, 1.05, 1.05, 1.10, 1.10, 1.15, 1.15, 1.15, 1.25, 1.30,
1.30, 1.35, 1.35, 1.35, 1.40, 1.45, 1.45, 1.50, 1.50, 1.50, 1.55,
1.55, 1.70, 1.95, 2.00, 2.00, 2.00, 2.05, 2.05, 2.05, 2.15, 2.55

και επομένως μπορούμε να βρούμε ότι 24 φορές η τιμή της ελεγχοσυνάρτησης ήταν μεγαλύτερη ή ίση από το 1.2 που πήραμε για τα δεδομένα μας, άρα $m = 24$. Επομένως βρίσκουμε πως η εκτίμηση του p-value είναι $\tilde{p} = \frac{24+1}{99+1} = 0.25$. 'Αρα αν θέσουμε α=5% δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση πως δεν υπάρχει διαφορά στη μέση τιμή των αγοριών και των κοριτσιών.



Γράφημα 2.1: Εκτίμηση της συνάρτησης πυκνότητας πιθανότητας της ελεγχοσυνάρτησης .

Παρατηρείστε πως οι 99 τιμές της ελεγχοσυνάρτησης επιτρέπουν μια προσέγγιση της κατανομής της. Στο γράφημα 2.1 μπορείτε να δείτε μια εκτίμηση για τη συνάρτηση πυκνότητας πιθανότητας της ελεγχοσυνάρτησης, με τη μέθοδο των kernels. Παρατηρούμε πως η κατανομή της ελεγχοσυνάρτησης κάτω από τη μηδενική

υπόθεση είναι ασύμμετρη προς τα αριστερά. Η δεξιά ουρά οφείλεται στην απομακρυσμένη τιμή 2.55.

Και πάλι πρέπει να σημειώσουμε πως αν κάποιος άλλος ερευνητής πάρει 99 συνδυασμούς, πιθανότατα θα εκτιμήσει ένα λίγο διαφορετικό p-value, αφού μάλλον θα πάρει διαφορετικούς συνδυασμούς. Αυτό δεν αποτελεί μειονέκτημα της μεθόδου καθώς τη γενική εικόνα σχετικά με το αν ισχύει ή όχι η μηδενική υπόθεση μπορεί κανείς να τη σχηματίσει.

Τέλος είναι ευνόητο ότι όσο αυξάνει η τιμή του k τόσο αυξάνει και η σιγουριά μας για το συμπέρασμα. Το κόστος είναι πως πιθανότατα ο υπολογιστικός χρόνος που χρειάζεται είναι μεγάλος.

2.4 Εκτίμηση του p-value

Ας προσπαθήσουμε τώρα να εξηγήσουμε γιατί εκτιμούμε το p-value ως $\tilde{p} = \frac{m+1}{k+1}$. Κατά αρχάς δεδομένου πως έχουμε k τιμές της ελεγχοσυνάρτησης αυτό σημαίνει πως όλη την ευθεία των πραγματικών αριθμών την έχουμε μοιράσει σε $k+1$ διαστήματα και θέλουμε να δούμε μέσα σε ποιο από αυτά τα διαστήματα ανήκει η παρατηρούμενη τιμή της ελεγχοσυνάρτησης. Επίσης αν ισχύει η μηδενική υπόθεση η τιμή της ελεγχοσυνάρτησης που παρατηρήσαμε μπορεί να πέσει μέσα σε οποιοδήποτε διάστημα με ίδια πιθανότητα και επομένως με πιθανότητα $1/(k+1)$. Αυτό εξηγεί τον παρονομαστή.

Κάθε μια τιμή είτε θα είναι μικρότερη από την παρατηρούμενη είτε μεγαλύτερη. Επίσης εξ ορισμού $P(T \geq t_{obs}) = p_{true}$ δηλαδή το πραγματικό (αλλά άγνωστο) p-value. Συνεπώς η τυχαία μεταβλητή m ακολουθεί τη διωνυμική κατανομή για k επαναλήψεις και πιθανότητα επιτυχίας p_{true}. Επομένως $E(m) = kp_{true}$ και άρα

$$E(\tilde{p}) = E\left(\frac{m+1}{k+1}\right) = \frac{kp_{true}+1}{k+1} > p_{true}$$

Επομένως παρατηρούμε πως η ποσότητα $\tilde{p}$ είναι μεροληπτική και πως υπερεκτιμά το πραγματικό p-value. Από την άλλη αν είχαμε χρησιμοποιήσει για p-value την ποσότητα $\frac{m}{k+1}$ θα ήταν μεροληπτική προς τα κάτω δηλαδή θα υποεκτιμούσαμε το p-value. Αν σκεφτεί κανείς πως το p-value το χρησιμοποιούμε για να παίρνουμε αποφάσεις, είναι προτιμότερο να το υπερεκτιμούμε οπότε θα απορρίπτουμε λιγότερο συχνά από ότι πρέπει τη μηδενική υπόθεση, ενώ αν το υποεκτιμούσαμε θα απορρίπταμε πιο συχνά.

Από την άλλη είναι ενδιαφέρον να δούμε πόσο μεγάλη είναι η μεροληψία που έχουμε. 'Ετσι

$$Bias(\tilde{p}) = E(\tilde{p}) - p_{true} = \frac{kp_{true}+1}{k+1} - p_{true} = \frac{1-p_{true}}{k+1}$$

Ο αριθμητής είναι σε κάθε περίπτωση μικρότερος ή ίσος της μονάδας οπότε αν επιλέξουμε $k = 99$ η μεροληψία είναι, κατά μέσο όρο, μικρότερη από 0.01. Επομένως ακόμα και για 99 επαναλήψεις το σφάλμα είναι στο δεύτερο δεκαδικό ψηφίο και άρα είμαστε αρκετά σίγουροι για το πόσο περίπου είναι η πραγματική τιμή του p-value.

Τέλος μπορεί κάποιος να κατασκευάσει εύκολα και γρήγορα προσεγγιστικά διαστήματα εμπιστοσύνης για το πραγματικό p-value χρησιμοποιώντας την ασυμπτωτική κανονικότητα του $\tilde{p}$ (θυμηθείτε πως η διωνυμική κατανομή προσεγγίζει σχετικά γρήγορα την κανονική κατανομή). Έτσι ένα προσεγγιστικό 95% διάστημα εμπιστοσύνης είναι το

$$\tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{k}}.$$

Προσοχή: Είναι σωστό να παρατηρήσει κανείς πως η προσέγγιση της διωνυμικής από την κανονική κατανομή δεν είναι σωστή αν για παράδειγμα το p-value είναι πολύ κοντά στο 0. Όμως σε αυτή την περίπτωση ο ερευνητής είναι σχετικά αρκετά βέβαιος για την απόρριψη της μηδενικής του υπόθεσης οπότε δεν έχει νόημα να φτιάξουμε διάστημα εμπιστοσύνης. Γενικά τα διαστήματα εμπιστοσύνης μας βοηθάνε στις περιπτώσεις που είμαστε κοντά στο επίπεδο στατιστικής σημαντικότητας, οπότε θέλουμε να δούμε κατά πόσο δεχόμαστε ή απορρίπτουμε τη μηδενική υπόθεση.

Επίσης μια άλλη ερμηνεία για τον τρόπο που υπολογίζουμε το p-value είναι πως αν έχουμε δημιουργήσει k δείγματα και με αυτά θέλουμε να προσεγγίσουμε την κατανομή της ελεγχοσυνάρτησης, υπάρχει και μια ακόμα τιμή διαθέσιμη από την κατανομή της ελεγχοσυνάρτησης: η παρατηρηθείσα τιμή. Επομένως έχουμε διαθέσιμες $k+1$ τιμές. Ως προς τον αριθμητή ξέρουμε σίγουρα πως αυτή η $k+1$ τιμή είναι ίση με την παρατηρηθείσα αφού είναι η παρατηρηθείσα και έτσι προκύπτει ο αριθμητής.

Καταλήγοντας για το πόσες επαναλήψεις (πόσους συνδυασμούς) χρειαζόμαστε ένας εμπειρικός κανόνας είναι ο εξής:

Αν το k είναι πολύ μικρό (πχ 20) και η ελεγχοσυνάρτηση μας έχει μεγάλη μεταβλητότητα αυτό αυξάνει την πιθανότητα σφάλματος. Άρα μια τιμή του k κοντά στο 100 είναι ικανοποιητική εφόσον το πρόβλημα επιτρέπει τόσες επαναλήψεις από άποψη χρόνου. (Παρατήρηση: συνήθως διαλέγουμε όχι 100 αλλά 99 επαναλήψεις, απλά ώστε ο παρονομαστής στην εκτίμηση του p-value να είναι στρογγυλός αριθμός και άρα η εκτίμηση να μην έχει πολλά δεκαδικά ψηφία. Το ίδιο ισχύει και για άλλες τιμές του k που χρησιμοποιούνται στην πράξη).

Αν το p-value που παίρνουμε είναι κοντά στο επίπεδο στατιστικής σημαντικότητας, αύξησε το k μέχρι το 95% διάστημα εμπιστοσύνης να είναι ξεκάθαρο ως προς την απόφαση που θα πάρεις, δηλαδή να μην περιέχει την τιμή του επιπέδου στατιστικής σημαντικότητας.

Επειδή οι υπολογιστές είναι πια αρκετά γρήγοροι 999 επαναλήψεις είναι τις περισσότερες φορές εύκολο να γίνουν επομένως αυτή είναι μια καλή επιλογή. Για παράδειγμα κάποια στατιστικά πακέτα που προσφέρουν ελέγχους τυχαιοποίησης χρησιμοποιούν ακόμα μεγαλύτερες τιμές, όπως για παράδειγμα το SPSS for WINDOWS που η αρχική επιλογή του είναι 10000 δείγματα.

Σε κάθε περίπτωση ο ερευνητής πρέπει να λάβει υπόψη του το κόστος σε υπολογιστικό χρόνο.

Τέλος θα πρέπει να σημειωθεί, πως αν το k είναι σχετικά μεγάλο μπορεί κανείς να χρησιμοποιήσει ως εκτίμηση του p-value απλά το λόγο m/k ο οποίος είναι απλοϊκός μεν, αλλά μπορεί να μας δείξει αν ισχύει ή όχι η μηδενική υπόθεση.

Μερικοί συγγραφείς χρησιμοποιούν αυτό για να εκτιμούν το p-value. Είναι ξεκάθαρο ότι για μεγάλες τιμές του k ουσιαστικά η διαφορά είναι αμελητέα. Προσέξτε επίσης ότι με τη χρήση του $\tilde{p}$ δεν υπάρχει περίπτωση να εκτιμήσουμε το p-value ως 0, κάτι που μοιάζει λογικό ακόμα και αν έχουμε μεγάλο k, αφού πάντα υπάρχει μια μικρή πιθανότητα να πάρουμε μια ακόμα πιο ακραία τιμή.

2.5 Παραδείγματα

2.5.1 Αμερικάνικες Εκλογές

Το σύστημα των εκλογών στις ΗΠΑ για την ανάδειξη προέδρου είναι πλειοψηφικό, δηλαδή αυτός που θα κερδίσει σε μια πολιτεία κερδίζει όλους τους εκλέκτορες αυτής της πολιτείας. Επομένως μερικοί πολιτικοί αναλυτές υποστηρίζουν πως όσο το αποτέλεσμα μοιάζει να είναι πιο αμφίρροπο τόσο μεγαλύτερη είναι η συμμετοχή των ψηφοφόρων στις εκλογές, προφανώς γιατί θεωρούν πως είναι πιο σημαντικό να συμμετάσχουν στις εκλογές αφού η ψήφος τους είναι πολύτιμη. Τα δεδομένα που βλέπετε στον πίνακα 2.4 , αφορούν τις αμερικάνικες εκλογές του 1844 όπου οι 2 υποψήφιοι είχαν τελικά πολύ μικρή διαφορά ψήφων. Το ερώτημα που θέλουμε να εξετάσουμε είναι αν αληθεύει ο ισχυρισμός των αναλυτών .

Ας αρχίσουμε προσπαθώντας να ορίσουμε το πρόβλημα με στατιστική ορολογία. Η υπόθεση πως όσο μικρότερη είναι η διαφορά τόσο μεγαλύτερη είναι η συμμετοχή, σημαίνει πως η συσχέτιση μεταξύ των δύο μεταβλητών είναι αρνητική. Επομένως θέλουμε να ελέγξουμε την

H_0: δεν υπάρχει σχέση ανάμεσα στη συμμετοχή και τη διαφορά

H_1: υπάρχει αρνητική σχέση

Προσέξτε πως

- η μηδενική υπόθεση πρέπει να δηλώνει την ανυπαρξία κάποιας δομής στα δεδομένα.

- η εναλλακτική υπόθεση είναι μονόπλευρη

Αφού λοιπόν θέσαμε τις υποθέσεις, πρέπει να διαλέξουμε την ελεγχοσυνάρτηση. Από τις υποθέσεις βλέπουμε πως ο συντελεστής συσχέτισης του Pearson είναι υποψήφια ελεγχοσυνάρτηση, αφού περιμένουμε να παίρνει την τιμή 0 αν ισχύει η μηδενική υπόθεση και αρνητικές τιμές αν ισχύει η εναλλακτική. Προφανώς δεν είναι η μόνη υποψήφια ελεγχοσυνάρτηση. Μπορεί κάποιος να πει πως ο συντελεστής β της γραμμικής παλινδρόμησης είναι και αυτός μια εναλλακτική ιδέα, καθώς και διάφορες άλλες ελεγχοσυναρτήσεις. Διαλέγουμε να δουλέψουμε με το συντελεστή συσχέτισης.

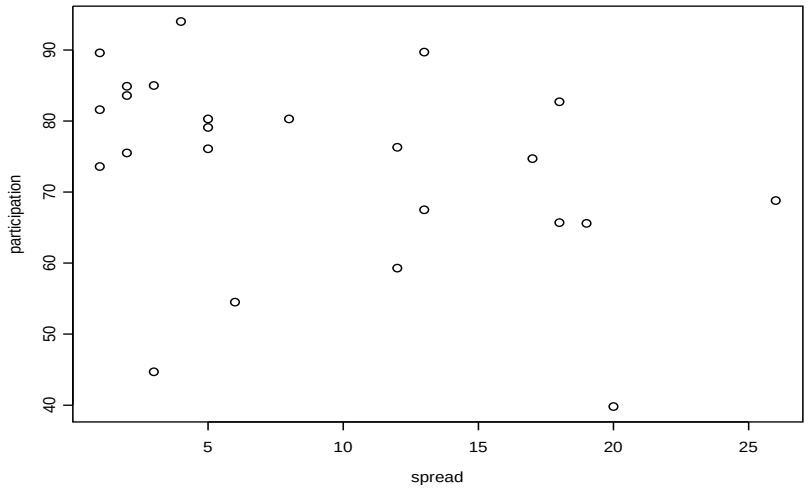
Είπαμε προηγουμένως πως για καθαρά τεχνικούς λόγους θέλουμε να απορρίπτουμε τη μηδενική υπόθεση για μεγάλες τιμές της ελεγχοσυνάρτησης. Επομένως επειδή στην περίπτωση μας απορρίπτουμε για μικρές τιμές (αρνητικός συντελεστής συσχέτισης),

Πολιτεία	Συμμετοχή %	Απόλυτη διαφορά %
Maine	67.5	13
New Hampshire	65.6	19
Vermont	65.7	18
Masschusetts	59.3	12
Rhode Island	39.8	20
Connecticut	76.1	5
New York	73.6	1
New Jersey	81.6	1
Pensylvania	75.5	2
Maryland	80.3	3
Virginia	54.5	6
North Carolina	79.1	5
Georgia	94.0	4
Kentucky	80.3	8
Tennessee	89.6	1
Louisianna	44.7	3
Alabama	82.7	18
Missisipi	89.7	13
Ohio	83.6	2
Indiana	84.9	2
Illinois	76.3	12
Missouri	74.7	17
Arkansas	68.8	26
Michigan	79.3	6

Πίνακας 2.4: Τα αποτελέσματα των αμερικάνικων προεδρικών εκλογών του 1844 σε 24 πολιτείες. Στον πίνακα βλέπετε το ποσοστό της συμμετοχής και την απόλυτη διαφορά ανάμεσα στους δύο υποψηφίους.

αρκεί να αλλάξουμε το πρόσημο και να διαλέξουμε για ελεγχοσυνάρτηση την

$$T = -corr(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}.$$



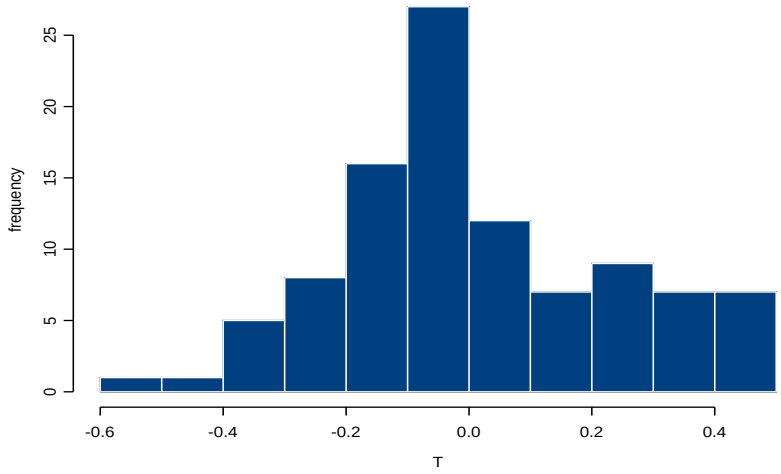
Γράφημα 2.2: Διάγραμμα σημείων για τα δεδομένα των αμερικάνικων εκλογών

Υπολογίζουμε από τα δεδομένα μας πως $t_{obs} = 0.36$, δηλαδή η αντίθετη τιμή του συντελεστή συσχέτισης των δεδομένων μας. Θέλουμε να δούμε κατά πόσο αυτή η τιμή είναι λογική αν ισχύει η μηδενική υπόθεση.

Δείτε στο γράφημα 2.2 το διάγραμμα σημείων για τις 24 παρατηρήσεις μας. Υπάρχει μια μάλλον αμυδρή αρνητική συσχέτιση, αλλά σίγουρα δεν φαίνεται κάτι προφανές.

Στην περίπτωση μας η τυχαιοποίηση υπονοεί ότι θα πρέπει να κατασκευάσουμε όλες τις δυνατές 24άδες ζευγών διαφορών και συμμετοχής. Δεδομένου ότι για να κατασκευάσουμε όλες αυτές τις περιπτώσεις είναι ισοδύναμο με το να κρατήσουμε τη μια μεταβλητή σταθερή και μετά για την άλλη στήλη, να πάρουμε όλες τις δυνατές διατάξεις, είναι κατανοητό ότι ένας ακριβής έλεγχος τυχαιοπόιησης δεν είναι εφικτός. Επομένως θα κάνουμε έναν προσεγγιστικό έλεγχο. Διαλέγουμε να πάρουμε $k = 99$ διαφορετικές 24άδες ζευγών τυχαία, και να υπολογίσουμε την ελεγχοσυνάρτηση.

Στο γράφημα 2.3 μπορείτε να δείτε το ιστόγραμμα των 99 τιμών που πήραμε.



Γράφημα 2.3: Ιστόγραμμα των τιμών της ελεγχοσυνάρτησης για το παράδειγμα των αμερικάνικων εκλογών.

Παρατηρούμε πως η τιμή 0.36 που έχουμε παρατηρήσει είναι στη δεξιά ουρά της κατανομής εκεί δηλαδή που απορρίπτουμε τη μηδενική υπόθεση. Αν εκτιμήσουμε το p-value όπως είδαμε πριν και επειδή $m =$ αριθμός συνδυασμών με $T \geq t_{obs}$=3, βρίσκουμε πως $\tilde{p} = \frac{m+1}{\text{k+1}} = \frac{4}{100} = 0.04$. Βλέπουμε πως το εκτιμημένο p-value είναι πολύ κοντά στο 5% το οποίο είναι το συνηθισμένο επίπεδο στατιστικής σημαντικότητας, επομένως δεν μπορεί κάποιος να νιώθει σίγουρος για την απόφασή του. 'Ενα 95% διάστημα εμπιστοσύνης για το πραγματικό p-value είναι το $(0.002, 0.078)$.

Σε αυτό το σημείο πρέπει να τονιστεί ότι πιθανότατα αν κάποιος ξανατρέξει για τα ίδια δεδομένα άλλες 99 επαναλήψεις ενδέχεται να πάρει διαφορετικό p-value (θα δούμε σε λίγο την κατανομή του p-value όπως εκτιμήθηκε με 100 επαναλήψεις της διαδικασίας).

Για να είμαστε πιο σίγουροι πρέπει να αυξήσουμε τον αριθμό των επαναλήψεων. Κάνοντας τώρα 999 επαναλήψεις βρίσκουμε $m = 37$, οπότε $\tilde{p} = 0.038$. Τώρα ένα 95% διάστημα εμπιστοσύνης είναι $(0.027, 0.049)$. Τώρα είμαστε κάπως πιο βέβαιοι καθώς φαίνεται πως το p-value είναι μάλλον μικρότερο από 5% και άρα απορρίπτουμε τη μηδενική υπόθεση, δηλαδή υπάρχει αρνητική σχέση στις 2 μεταβλητές.

Είναι πολύ σημαντικό να παρατηρήσει κανείς πως αν και στην πραγματικότητα κάνουμε έλεγχο για έναν συντελεστή συσχέτισης, δεν έχουμε κάνει καμία υπόθεση για την κατανομή του πληθυσμού. 'Εχοντας κάνει την υπόθεση πως η από κοινού κατανομή των δύο μεταβλητών ήταν η διμεταβλητή κανονική κατανομή, θα μπορούσε

κάποιος να χρησιμοποιήσει υπάρχοντες ελέγχους υποθέσεων για το συντελεστή συσχέτισης. Θυμηθείτε πως τέτοιοι έλεγχοι στηρίζονται σε ασυμπτωτικά αποτελέσματα και πως η κατανομή του δειγματικού συντελεστή συσχέτισης από κανονικούς πληθυσμούς είναι μια περίπλοκη κατανομή που δύσκολα μπορεί κανείς να χρησιμοποιήσει. Μάλιστα για τα δεδομένα μας μπορεί κάποιος να επαληθεύσει πως η υπόθεση της κανονικότητας του πληθυσμού δεν φαίνεται να ισχύει.

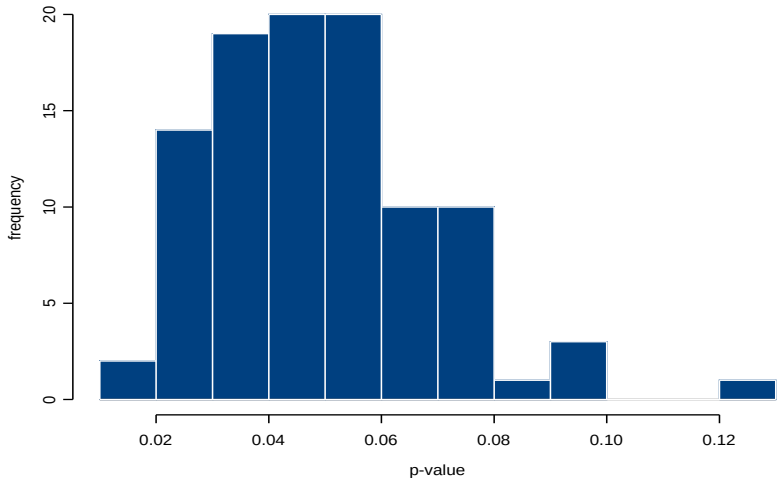
Εναλλακτικά θα μπορούσε κανείς να χρησιμοποιήσει για ελεγχοσυνάρτηση το συντελεστή β του γραμμικού μοντέλου. Σε αυτή την περίπτωση οι υποθέσεις θα ήταν:

$H_0 : \beta = 0$

$H_1 : \beta < 0$

'Αρα θα κάναμε έναν έλεγχο για τη στατιστική σημαντικότητα του συντελεστή β του γραμμικού μοντέλου χωρίς να έχουμε κάνει καμιά υπόθεση για την κανονικότητα του πληθυσμού. Είναι μάλιστα γνωστό πως ο εκτιμητής ελαχίστων τετραγώνων του β και ο δειγματικός συντελεστής συσχέτισης έχουν μια απλή συναρτησιακή σχέση και επομένως η ισοδυναμία των δύο ελέγχων μπορεί να αποδειχτεί και θεωρητικά.

Αξίζει εδώ να αναφέρουμε το εξής. Αν επαναλάβουμε τον έλεγχο 100 φορές θα καταλήξουμε σε 100 τιμές για το p-value. Στην πραγματικότητα αυτές οι τιμές θα αποτελούν μια καλή προσέγγιση της κατανομής του p-value. Με αυτόν τον τρόπο κατασκευάσαμε το ιστόγραμμα που ακολουθεί. Δεδομένου πως $k = 99$ τα p-value που πήραμε είχαν κάθε φορά ακρίβεια 2 δεκαδικών ψηφίων.



Γράφημα 2.4: Ιστόγραμμα συχνοτήτων για το p-value βασισμένο σε 100 επαναλήψεις του ελέγχου.

Θεραπεία Α	13 ασθενείς	$\bar{x}_A = 3$
Θεραπεία Β	24 ασθενείς	$\bar{x}_B = 6$
Θεραπεία Γ	17 ασθενείς	$\bar{x}_\Gamma = 2$
Θεραπεία Δ	8 ασθενείς	$\bar{x}_\Delta = 1$

Πίνακας 2.5: Αποτελέσματα για τις 4 διαφορετικές θεραπείες

Από το γράφημα 2.4 μπορεί κανείς να παρατηρήσεις πως οι περισσότερες τιμές είναι μικρότερες ή ίσες του 5% (για την ακρίβεια 75 στις 100). Η μέση τιμή είναι 0.0427 ενώ αν κανείς χρησιμοποιήσει όλες τις επαναλήψεις τότε θα βρει πως από τις 9900 τιμές της ελεγχοσυνάρτησης το εκτιμηθέν p-value είναι 0.0431 ενώ το 95% διάστημα εμπιστοσύνης γίνεται $(0.039214, 0.047320)$ και μας οδηγεί εκ του ασφαλούς να απορρίψουμε τη μηδενική υπόθεση. Αυτό που είναι ενδιαφέρον είναι να παρατηρήσει κανείς πως μπορεί να εκτιμήσει όχι μόνο την συνάρτηση πυκνότητας πιθανότητας της ελεγχοσυνάρτησης αλλά με επανάληψη της διαδικασίας μπορεί να εκτιμήσει και την κατανομή του p-value του ελέγχου. Μάλιστα, παρατηρείστε πως για να κατασκευάσουμε το 95% διάστημα εμπιστοσύνης επικαλεστήκαμε την κανονικότητα της κατανομής του $\tilde{p}$. Από το ιστόγραμμα μπορεί κανείς να παρατηρήσει πως η κατανομή του δεν φαίνεται να είναι κανονική και επομένως και τα διαστήματα εμπιστοσύνης είναι αμφιβόλου ποιότητας. Επειδή έχουμε χρησιμοποιήσει $k = 99$, οι τιμές που μπορεί να πάρει το $\tilde{p}$ είναι διακριτές (διακριτές θα ήταν σε κάθε περίπτωση!) αλλά και λίγες στον αριθμό. Αυτός είναι και ο λόγος που η προσέγγιση της κατανομής από την κανονική δεν είναι πολύ καλή.

2.5.2 Ανάλυση Διακύμανσης

Ας δούμε ένα ακόμα παράδειγμα που αφορά τον έλεγχο ισότητας πολλών μέσων τιμών.

62 ασθενείς υποβλήθηκαν σε 4 διαφορετικές θεραπείες και για καθένα μετρήθηκε η βελτίωση που παρουσίασε σε κάποιο συστατικό του αίματος. Τα αποτελέσματα παρουσιάζονται στον πίνακα 2.5

Η υπόθεση που θέλουμε να ερευνήσουμε είναι αν υπάρχει διαφορά στις θεραπείες.

Το πρόβλημα είναι ένα κλασικό πρόβλημα ελέγχου διαφόρων πολλών μέσων τιμών. Για κανονικούς πληθυσμούς μπορούμε να το αντιμετωπίσουμε χρησιμοποιώντας Ανάλυση Διακύμανσης (ANOVA). Επειδή όμως δεν έχουμε εδώ καμιά τέτοια πληροφορία, προχωράμε με έναν έλεγχο τυχαιότητας. Οι υποθέσεις μας είναι

H_0: όλες οι μέσες τιμές είναι ίδιες

H_1: τουλάχιστον 2 διαφέρουν μεταξύ τους

Στην κλασική ανάλυση διακύμανσης χρησιμοποιούμε την ελεγχοσυνάρτηση F. Ο κύριος λόγος είναι πως αυτή η ελεγχοσυνάρτηση έχει γνωστή κατανομή όταν ισχύει η μηδενική υπόθεση. Η βασική ιδέα της ανάλυσης διακύμανσης είναι πως η συνολική διακύμανση μπορεί να διασπαστεί σε δύο μέρη, το ένα αφορά τη διασπορά των μέσων τιμών των γκρουπ μεταξύ τους και το άλλο τις αποκλίσεις μέσα στα γκρουπ. Αν όλες οι μέσες τιμές ήταν ίδιες αυτό σήμαινε πως οι τετραγωνικές αποκλίσεις των 4 μέσων από το συνολικό μέσο θα ήταν μικρές. Επομένως μια

λογική ελεγχοσυνάρτηση είναι η $T = \sum_{i=A,B,\Gamma,\Delta} (\bar{x}_i - \bar{x})^2$.

Για να κάνουμε τον έλεγχο τυχαιοποίησης θα πρέπει να μοιράσουμε τους 62 ασθενείς στις 4 θεραπείες έτσι ώστε κάθε θεραπεία να έχει το σωστό αριθμό ασθενών και να υπολογίσουμε την ελεγχοσυνάρτηση. Αυτό πρέπει να επαναληφθεί k φορές και στη συνέχεια να συγκρίνουμε την παρατηρούμενη τιμή με αυτές. Συνήθως τα δεδομένα μας θα είναι δύο στήλες (μεταβλητές), η μια εκ των οποίων θα περιέχει τις τιμές των ασθενών και η άλλη την ομάδα στην οποία ανήκει κάθε ασθενής. Συνεπώς, υπολογιστικά η τυχαιοποίηση μπορεί να επιτευχθεί κρατώντας μια από τις δύο μεταβλητές που έχουμε σταθερή και ανακατεύοντας την άλλη.

Για τα δεδομένα μας έχουμε $t_{obs} = 14$, και χρησιμοποιώντας $k = 99$ βρήκαμε $m = 17$, άρα $\tilde{p} = \frac{17+1}{99+1} = 0.18$, άρα δεν απορρίπτουμε τη μηδενική υπόθεση.

2.6 Επιλογή Ελεγχοσυνάρτησης

Ας δούμε λίγο το θέμα της επιλογής της ελεγχοσυνάρτησης και πως μπορούμε να απλοποιήσουμε την επιλογή. Για τη διεξαγωγή του ελέγχου ισότητας δύο μέσων τιμών χρησιμοποιούμε την ελεγχοσυνάρτηση $t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}}$. Ο παρονομαστής στην ουσία εξυπηρετεί μόνο το γεγονός πως η κατανομή της ελεγχοσυνάρτησης είναι γνωστή και τίποτα άλλο. Επομένως στην περίπτωση μας μπορούμε να την αγνοήσουμε και να χρησιμοποιήσουμε την ελεγχοσυνάρτηση $t = \bar{x} - \bar{y}$. 'Ομως θα ισχύει

$$t = \bar{x} - \bar{y} = \sum_{i=1}^{n_x} X_i \left(\frac{1}{n_x} + \frac{1}{n_y}\right) - \left(\sum_{i=1}^{n_x} X_i + \sum_{i=1}^{n_y} Y_i\right) \frac{1}{n_y}$$

'Ομως σε αυτή τη μορφή οι ποσότητες $\left(\frac{1}{n_x} + \frac{1}{n_y}\right)$ και $\left(\sum_{i=1}^{n_x} X_i + \sum_{i=1}^{n_y} Y_i\right)$ είναι γνωστές αφού η μεν πρώτη έχει να κάνει με τα μεγέθη των επιμέρους δειγμάτων, η δε δεύτερη με το άθροισμα όλων των παρατηρήσεων που είναι σταθερό για κάθε διάταξη του δείγματος. Επομένως μοιάζει λογικό να χρησιμοποιήσει κανείς μόνο την ελεγχοσυνάρτηση $t' = \sum_{i=1}^{n_x} X_i$.

Δηλαδή στους ελέγχους τυχαιοποίησης μπορούμε να χρησιμοποιήσουμε απλές ελεγχοσυναρτήσεις οι οποίες είναι ισοδύναμες με άλλες πιο πολύπλοκες που χρησιμοποιούνται στους κλασσικούς ελέγχους αποκλειστικά και μόνο για να μπορούμε να βρούμε την κατανομή τους. Στους ελέγχους τυχαιοποίησης κάτι τέτοιο δεν είναι πρόβλημα αφού εκτιμάμε την κατανομή της ελεγχοσυνάρτησης και άρα δεν χρειάζεται να την ξέρουμε.

Στο παράδειγμα των αμερικάνικων εκλογών μπορεί κάποιος να παρατηρήσει πως ο παρονομαστής της ελεγχοσυνάρτησης είναι πάντα ο ίδιος αφού αφορά τα αθροίσματα τετραγώνων κάθε μεταβλητής και παραμένει ο ίδιος σε κάθε επανάληψη. Η συνδιακύμανση, δηλαδή ο αριθμητής μπορεί να γραφτεί ως $\sum_{i=1}^{n} X_i Y_i / n - \bar{X}\bar{Y}$ όπου και πάλι ο δεύτερος όρος είναι σταθερός σε κάθε επανάληψη. Συνεπώς

μια εντελώς ισοδύναμη ελεγχοσυνάρτηση είναι η

$$T_2 = -\sum_{i=1}^{n} X_i Y_i$$

που είναι απλός γραμμικός συνδυασμός της προηγούμενης ελεγχοσυνάρτησης. Συνεπώς τα αποτελέσματα που θα πάρει κάποιος θα είναι ακριβώς τα ίδια (δηλαδή θα βρεί ίδιο p-value). Ο υπολογισμός της T_2 είναι σαφώς πιο απλός.

Είναι ευνόητο πως για έναν συγκεκριμένο έλεγχο υπάρχουν περισσότερες από μια ελεγχοσυναρτήσεις που θα μπορούσαν να χρησιμοποιηθούν, συνήθως με παρόμοια αποτελέσματα. Παρόλα αυτά υπάρχει κίνδυνος μερικές φορές οι διάφορες ελεγχοσυναρτήσεις να οδηγούν σε διαφορετικά αποτελέσματα σχετικά με την απόρριψη ή όχι της μηδενικής υπόθεσης. Πως θα διαλέξουμε λοιπόν ανάμεσα σε πιθανές ελεγχοσυναρτήσεις;

Η απάντηση δεν είναι απλή. Είναι πάντα χρήσιμο να προτιμούμε ελεγχοσυναρτήσεις γνωστές και δοκιμασμένες από το να ορίζουμε ελεγχοσυναρτήσεις από την αρχή. Η βασική αρχή είναι πως η ελεγχοσυνάρτηση θα πρέπει να μπορεί να διακρίνει καθαρά ανάμεσα στη μηδενική και την εναλλακτική υπόθεση, δηλαδή να παίρνει διαφορετικές τιμές όταν ισχύει η μηδενική υπόθεση από όταν δεν ισχύει.

Θεωρητικά αυτό που πρέπει να εξετάζει κανείς είναι η ισχύς του ελέγχου. Υπενθυμίζουμε, κάποια βασικά πράγματα σχετικά με τους ελέγχους υποθέσεων. Σε κάθε έλεγχο υποθέσεων υπάρχουν δύο διαφορετικά σφάλματα που μπορούμε να κάνουμε. Το σφάλμα τύπου Ι αναφέρεται στο να απορρίψουμε τη μηδενική υπόθεση ενώ είναι αληθινή ενώ το σφάλμα τύπου ΙΙ αναφέρεται στο να δεχτούμε τη μηδενική υπόθεση ενώ η εναλλακτική είναι σωστή. Η πιθανότητα σφάλματος τύπου Ι, συνήθως την ονομάζουμε επίπεδο στατιστικής σημαντικότητας ενώ το συμπληρωματικό της πιθανότητας σφάλματος τύπου ΙΙ αποτελεί την ισχύ του ελέγχου. Δηλαδή έχουμε

$\alpha = P(\text{σφάλμα τύπου I})$

$\beta = P(\text{σφάλμα τύπου II})$ και

$\gamma = 1 - \beta$

Αυτό λοιπόν που έχει σημασία είναι να διαλέξουμε μια ελεγχοσυνάρτηση που να έχει μεγάλη ισχύ, αφού συνήθως στους ελέγχους υποθέσεων το α είναι προεπιλεγμένο και άρα μόνο το β μένει να εξαρτάται από την ελεγχοσυνάρτηση.

Κάτι που περιπλέκει ακόμα περισσότερο τα πράγματα έχει να κάνει με τη μορφή της εναλλακτικής καθώς κάποιες ελεγχοσυναρτήσεις έχουν μεγάλη ισχύ ως προς συγκεκριμένες εναλλακτικές υποθέσεις και μικρή ισχύ προς κάποιες άλλες. Υπάρχουν όμως ελεγχοσυναρτήσεις που να είναι βέλτιστες έναντι σε κάθε εναλλακτική υπόθεση; Η απάντηση είναι πως γενικά δεν υπάρχουν, εκτός από ειδικές περιπτώσεις όπου κανείς μπορεί να δείξει τις απαιτούμενες ιδιότητες, όπως το πολύ γνωστό μας t-test . Συνεπώς θα πρέπει κανείς να μελετήσει την ισχύ του ελέγχου για να αποφανθεί αν ο έλεγχος είναι ο καλύτερος από μια σειρά προτεινόμενων ελέγχων.

Δεν θα επεκταθούμε περισσότερο σε αυτό το πρόβλημα. Ως μια τελευταία συμβουλή όμως αξίζει να παρατηρήσουμε πως θα πρέπει να προτιμώνται έλεγχοι με γνωστές ιδιότητες ως προς την ισχύ τους, ή στην περίπτωση ελέγχων τυχαιοποίησης έλεγχοι που να είναι ισοδύναμοι με αυτούς.

2.7 Ο ακριβής έλεγχος του Fisher

Ο ακριβής έλεγχος του Fisher αποτελεί έναν από τους πλέον γνωστούς ελέγχους τυχαιοποίησης. Συγκεκριμένα ο έλεγχος αρχικά αναπτύχθηκε για πίνακες συνάφειας 2×2 (δηλαδή πίνακες συνάφειας με 2 γραμμές και 2 στήλες) κυρίως για λόγους υπολογιστικούς, αλλά στη συνέχεια με την πάροδο των χρόνων και την ύπαρξη ολοένα και πιο ισχυρών υπολογιστών γενικεύτηκε αρκετά.

Ας θεωρήσουμε έναν πίνακα συνάφειας με r γραμμές και c στήλες που γενικά έχει τη μορφή που βλέπουμε στον πίνακα 2.6

X_{11}	X_{12}	X_{1c}	m_1
X_{21}	X_{22}	X_{2c}	
X_{r1}	X_{r2}	X_{rc}	m_r
n_1	n_2	n_c	N

Πίνακας 2.6: Υπόδειγμα πίνακα συνάφειας

όπου τα X_{ij} είναι η συχνότητα του ij κελιού και αντίστοιχα m_i και n_jείναι τα αθροίσματα γραμμών και στηλών του πίνακα. Αν θέλουμε να ελέγξουμε την ανεξαρτησία γραμμών και στηλών (και άρα την ανεξαρτησία δύο μεταβλητών), ο πιο διαδεδομένος έλεγχος είναι αυτός με τη χρήση της ελεγχοσυνάρτησης χ^2 του Pearson που δίνεται από την

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - \hat{x}_{ij})^2}{\hat{x}_{ij}}$$

όπου $\hat{x}_{ij} = \frac{m_i n_j}{N}$ είναι οι αναμενόμενες συχνότητες αν ίσχυε η υπόθεση της ανεξαρτησίας. Ο έλεγχος χ^2 είναι πολύ διαδεδομένος δυσανάλογα με την ακρίβεια του. Για τους σκοπούς του ελέγχου η κατανομή της ελεγχοσυνάρτησης θεωρούμε πως ασυμπτωτικά είναι η χ^2 κατανομή με $(r-1)\times(c-1)$ βαθμούς ελευθερίας αλλά αυτό το ασυμπτωτικό αποτέλεσμα είναι σε πολλές περιπτώσεις χωρίς αντίκρισμα καθώς χρειάζεται να υπάρχουν κάποιες προϋποθέσεις ως προς το μέγεθος του δείγματος αλλά και τις αναμενόμενες τιμές ενός κελιού. Επομένως ο έλεγχος δεν είναι αξιόπιστος αν:

- το μέγεθος του δείγματος είναι μικρό,
- οι αναμενόμενες τιμές των κελιών είναι μικρές,
- ο πίνακας είναι αραιός, δηλαδή σε κάποια κελιά έχουμε πολύ λίγες παρατηρήσεις.

Ο ακριβής έλεγχος του Fisher αλλά και οι προσεγγιστικοί έλεγχοι που προκύπτουν έχουν σκοπό να ξεπεράσουν αυτά τα προβλήματα.

Θα περιγράψουμε τον έλεγχο με ένα παράδειγμα:

8 άτομα, από τα οποία 3 άνδρες και 5 γυναίκες ρωτήθηκαν για το αν είδαν ένα τηλεοπτικό πρόγραμμα. Από τους 3 άνδρες οι 2 είδαν το πρόγραμμα, ενώ από τις

5 γυναίκες μόνο 1 είδε το πρόγραμμα. Τα δεδομένα μπορούν να συνοψιστούν στον πίνακα 2.7.

	Είδαν	Δεν είδαν	
'Ανδρες	2	1	3
Γυναίκες	1	4	5
	3	5	8

Πίνακας 2.7: Τα δεδομένα του παραδείγματος της τηλεθέασης

Πιστεύετε πως τα ποσοστά αυτών που είδαν το πρόγραμμα διαφέρουν ανάμεσα στα δύο φύλα;

Είναι ξεκάθαρο πως θέλουμε να ελέγξουμε αν οι μεταβλητές 'είδαν το πρόγραμμα' και 'φύλο' είναι ανεξάρτητες ή όχι. Ο έλεγχος χ^2 δεν φαίνεται ικανός να απαντήσει το ερώτημα καθώς το μέγεθος του δείγματος είναι απαγορευτικά μικρό. Παρόλα αυτά μπορεί κανείς να υπολογίσει την τιμή 1.742 που δίνει ως p-value (χρησιμοποιώντας το ασυμπτωτικό αποτέλεσμα) την τιμή 0.187.

Σύμφωνα με αυτά που είπαμε στους ελέγχους τυχαιοποίησης, αν ισχύει η μηδενική υπόθεση τότε όλοι οι δυνατοί συνδυασμοί απαντήσεων θα μπορούσαν να έχουν προκύψει και επομένως θέλουμε να δούμε αν αυτός που παρατηρήσαμε εμείς είναι ακραίος ή όχι. Τα δεδομένα θα μπορούσαν να αναπαρασταθούν αντί σε πίνακα συνάφειας ως δύο στήλες (μεταβλητές) (για λόγους οικονομίας χώρου τις έχουμε παρουσιάσει ως γραμμές) όπως φαίνεται στον πίνακα 2.8.

Φύλο (Α: άνδρας, Γ: γυναίκα)	Α	Α	Α	Γ	Γ	Γ	Γ	Γ
Τηλεθέαση (1: Είδε 0: όχι)	1	1	0	0	1	0	0	0

Πίνακας 2.8: Διαφορετική μορφή για τα δεδομένα του παραδείγματος της τηλεθέασης.

Για να προχωρήσουμε λοιπόν στον έλεγχο τυχαιοποίησης, με ότι έχουμε πει μέχρι τώρα, θα έπρεπε να αναδιατάξουμε τα στοιχεία της στήλης αν είδε ή όχι το πρόγραμμα και να πάρουμε όλους τους δυνατούς συνδυασμούς. Για παράδειγμα μια τέτοια αναδιάταξη είναι και η

A	A	A	Γ	Γ	Γ	Γ	Γ
1	0	0	1	0	0	0	1

η οποία οδηγεί στον πίνακα συνάφειας που βλέπετε στον πίνακα 2.7

Είναι προφανές ότι τα περιθώρια αθροίσματα θα παραμείνουν σταθερά.

Επομένως ο ακριβής έλεγχος τυχαιοποίησης θα στηριχτεί σε όλα τα δυνατά δείγματα που μπορούμε να πάρουμε μοιράζοντας τις 8 τιμές της τηλεθέασης σε δύο ομάδες από 3 και 5 άτομα αντίστοιχα, άρα ο συνολικός αριθμός δειγμάτων είναι $\begin{pmatrix} 8 \\ 3 \end{pmatrix} = 56$ διαφορετικοί συνδυασμοί. Βέβαια επειδή έχουμε μόνο τιμές 0 και 1 κάποιοι συνδυασμοί θα επαναλαμβάνονται.

Αν το δούμε από την πλευρά των πινάκων συνάφειας και δεδομένου πως τα περιθώρια αθροίσματα είναι σταθερά οι δυνατοί πίνακες που μπορούμε να πάρουμε είναι οι

$$A = \begin{bmatrix} 0 & 3 \\ 3 & 2 \end{bmatrix}, B = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}, \Gamma = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}, \Delta = \begin{bmatrix} 3 & 0 \\ 0 & 5 \end{bmatrix}$$

Παρατηρείστε επίσης πως σε κάθε πίνακα αρκεί να ορίσω το X_{11} στοιχείο και αυτόματα προκύπτουν και τα υπόλοιπα

Στον πίνακα 2.10 μπορεί κανείς να δει όλα τα δυνατά δείγματα για το παράδειγμα μας καθώς και τον πίνακα συνάφειας που προκύπτει σε κάθε περίπτωση. Για να κάνουμε τον έλεγχο τυχαιοποίησης στην ουσία χρειαζόμαστε μια ελεγχοσυνάρτηση που να μας επιτρέπει να δούμε ανάμεσα στις δύο υποθέσεις που έχουμε και είναι

H_0 : τα δύο χαρακτηριστικά είναι ανεξάρτητα

H_1 : τα δύο χαρακτηριστικά δεν είναι ανεξάρτητα

Η ελεγχοσυνάρτηση του Pearson είναι μια υποψήφια αλλά παρατηρείστε πως για την περίπτωση μας αρκεί να διαπιστώσουμε ποιοι από τους 4 δυνατούς πίνακες συνάφειας είναι περισσότερο ακραίοι ως προς την υπόθεση που έχουμε. 'Ετσι αν ίσχυε η μηδενική υπόθεση θα περιμέναμε να έχουμε ποσοστά ανδρών και γυναικών που παρακολούθησαν το πρόγραμμα σχετικά κοντά. 'Εχουμε παρατηρήσει πως 2 στους 3 άνδρες (66.6%) και 1 στις 5 γυναίκες (20%) παρακολούθησαν. Επομένως πιο ακραίες περιπτώσεις είναι οι πίνακες Α και Δ. Από τον πίνακα 2.9 παρατηρούμε με ποιές συχνότητες οι πίνακες εμφανίζονται στα 56 δείγματα.

Πίνακας	Συχνότητα	Τιμή ελεγχοσυνάρτησης
Α	10	2.880
Β	30	0.036
Γ	15	1.742
Δ	1	8

Πίνακας 2.9: Οι συχνότητες εμφάνισης κάθε δυνατού πίνακα συνάφειας και η τιμή της ελεγχοσυνάρτησης του Pearson

'Αρα από τον πίνακα 2.9 βλέπουμε ότι είναι 11 οι πιο ακραίοι πίνακες που μπορούμε να πάρουμε από τους 56 δυνατούς πίνακες ενώ άλλοι 15 ταυτίζονται με τον πίνακα που παρατηρήσαμε. Επομένως σύμφωνα με αυτά που είπαμε για τους ακριβείς ελέγχους τυχαιοποίησης το p-value είναι 26/56 = 0.464 και επομένως η υπόθεση περί ανεξαρτησίας δεν μπορεί να απορριφθεί. Στον πίνακα 2.9 επίσης μπορεί κανείς να δει την τιμή της ελεγχοσυνάρτησης του Pearson. Παρατηρείστε πως αν και παίρνει μόνο 4 τιμές εμείς στην ουσία, με βάση το ασυμπτωτικό αποτέλεσμα, υποθέτουμε πως είναι μια συνεχής τυχαία μεταβλητή και πως μάλιστα ακολουθεί και την κατανομή χ^2 με έναν βαθμό ελευθερίας!.

Στο παράδειγμα μας βέβαια και το ασυμπτωτικό αποτέλεσμα δεν απέρριπτε τη μηδενική υπόθεση. Πολλές φορές όμως η χρήση του ασυμπτωτικού αποτελέσματος

	Α	Α	Α	Γ	Γ	Γ	Γ	Γ	Πίνακας συνάφειας
Δείγμα 1	1	1	1	0	0	0	0	0	Δ
Δείγμα 2	1	1	0	1	0	0	0	0	Γ
Δείγμα 3	1	1	0	0	1	0	0	0	Γ
Δείγμα 4	1	1	0	0	0	1	0	0	Γ
Δείγμα 5	1	1	0	0	0	0	1	0	Γ
Δείγμα 6	1	1	0	0	0	0	0	1	Γ
Δείγμα 7	1	0	1	1	0	0	0	0	Γ
Δείγμα 8	1	0	1	0	1	0	0	0	Γ
Δείγμα 9	1	0	1	0	0	1	0	0	Γ
Δείγμα 10	1	0	1	0	0	0	1	0	Γ
Δείγμα 11	1	0	1	0	0	0	0	1	Γ
Δείγμα 12	1	0	0	1	1	0	0	0	Β
Δείγμα 13	1	0	0	1	0	1	0	0	Β
Δείγμα 14	1	0	0	1	0	0	1	0	Β
Δείγμα 15	1	0	0	1	0	0	0	1	Β
Δείγμα 16	1	0	0	0	1	1	0	0	Β
Δείγμα 17	1	0	0	0	1	0	1	0	Β
Δείγμα 18	1	0	0	0	1	0	0	1	Β
Δείγμα 19	1	0	0	0	0	1	1	0	Β
Δείγμα 20	1	0	0	0	0	1	0	1	Β
Δείγμα 21	1	0	0	0	0	0	1	1	Β
Δείγμα 22	0	1	1	1	0	0	0	0	Γ
Δείγμα 23	0	1	1	0	1	0	0	0	Γ
Δείγμα 24	0	1	1	0	0	1	0	0	Γ
Δείγμα 25	0	1	1	0	0	0	1	0	Γ
Δείγμα 26	0	1	1	0	0	0	0	1	Γ
Δείγμα 27	0	1	0	1	1	0	0	0	Β
Δείγμα 28	0	1	0	1	0	1	0	0	Β
Δείγμα 29	0	1	0	1	0	0	1	0	Β
Δείγμα 30	0	1	0	1	0	0	0	1	Β
Δείγμα 31	0	1	0	0	1	1	0	0	Β
Δείγμα 32	0	1	0	0	1	0	1	0	Β
Δείγμα 33	0	1	0	0	1	0	0	1	Β
Δείγμα 34	0	1	0	0	0	1	1	0	Β
Δείγμα 35	0	1	0	0	0	1	0	1	Β
Δείγμα 36	0	1	0	0	0	0	1	1	Β
Δείγμα 37	0	0	1	1	1	0	0	0	Β
Δείγμα 38	0	0	1	1	0	1	0	0	Β
Δείγμα 39	0	0	1	1	0	0	1	0	Β
Δείγμα 40	0	0	1	1	0	0	0	1	Β

συνεχίζεται στην επόμενη σελίδα

	Α	Α	Α	Γ	Γ	Γ	Γ	Γ	Πίνακας συνάφειας
Δείγμα 41	0	0	1	0	1	1	0	0	Β
Δείγμα 42	0	0	1	0	1	0	1	0	Β
Δείγμα 43	0	0	1	0	1	0	0	1	Β
Δείγμα 44	0	0	1	0	0	1	1	0	Β
Δείγμα 45	0	0	1	0	0	1	0	1	Β
Δείγμα 46	0	0	1	0	0	0	1	1	Β
Δείγμα 47	0	0	0	1	1	1	0	0	Α
Δείγμα 48	0	0	0	1	1	0	1	0	Α
Δείγμα 49	0	0	0	1	1	0	0	1	Α
Δείγμα 50	0	0	0	1	0	1	1	0	Α
Δείγμα 51	0	0	0	1	0	1	0	1	Α
Δείγμα 52	0	0	0	1	0	0	1	1	Α
Δείγμα 53	0	0	0	0	1	1	1	0	Α
Δείγμα 54	0	0	0	0	1	1	0	1	Α
Δείγμα 55	0	0	0	0	1	0	1	1	Α
Δείγμα 56	0	0	0	0	0	1	1	1	Α

Πίνακας 2.10: 'Ολα τα πιθανά δείγματα και οι πίνακες συνάφειας που τους αντιστοιχούν.

μπορεί να οδηγήσει σε λανθασμένες αποφάσεις. Θα δούμε αργότερα τέτοια παραδείγματα. Ας προσπαθήσουμε να δούμε την περίπτωση πιο γενικά.

Ας δούμε, λοιπόν, στη γενική περίπτωση πως μπορούμε να υπολογίσουμε το p-value χωρίς να χρειαζόμαστε να κατασκευάσουμε και να μελετήσουμε όλους τους πίνακες

Στη γενική του μορφή ο 2×2 πίνακας συνάφειας έχει τη μορφή που μπορείτε να δείτε στον πίνακα 2.11.

x_{11}	x_{12}	m_1
x_{21}	x_{22}	m_2
n_1	n_2	N

Πίνακας 2.11: 2×2 πίνακας συνάφειας.

Επομένως σε αυτή την περίπτωση ο συνολικός αριθμός συνολικών δειγμάτων είναι $\begin{pmatrix} N \\ n_1 \end{pmatrix}$. 'Οπως είπαμε και πριν αφού τα περιθώρια αθροίσματα θα είναι σταθερά αρκεί κανείς να ορίσει ένα μόνο κελί και με βάση αυτό να υπολογίσει πόσες φορές εμφανίζεται κάθε πίνακας. Επίσης, μπορεί διαιρώντας προς το συνολικό αριθμό πινάκων να υπολογίσει την πιθανότητα εμφάνισης κάθε πίνακα.

Αυτή είναι

$$P(x_{11}) = \frac{\left(\begin{array}{c} m_1 \\ x_{11} \end{array} \right) \left(\begin{array}{c} m_2 \\ n_1 - x_{11} \end{array} \right)}{\left(\begin{array}{c} N \\ n_1 \end{array} \right)}$$

όπου x_{11} είναι το στοιχείο στην πρώτη γραμμή και στήλη, σύμφωνα με τα όσα είπαμε πριν. Η παραπάνω πιθανότητα είναι μια υπεργεωμετρική πιθανότητα που μας υπολογίζει την πιθανότητα από ένα σύνολο N πραγμάτων τα οποία χωρίζονται σε δύο κατηγορίες με m_1 και m_2 αντίστοιχα σε κάθε κατηγόρια να πάρουμε ακριβώς x_{11} πράγματα από την πρώτη κατηγορία. Για να υπολογίσουμε το p-value πρέπει να βρούμε όλους τους πίνακες συνάφειας που είναι πιο ακραίοι από ότι έχουμε παρατηρήσει και να προσθέσουμε την πιθανότητα κάθε πίνακα. Σε περίπτωση δίπλευρου ελέγχου, δηλαδή η εναλλακτική υπόθεση είναι 'όχι ανεξαρτησία', οι πίνακες που μας ενδιαφέρουν θα βρίσκονται και πως τις δύο πλευρές. Στην περίπτωση μονόπλευρου ελέγχου, με εναλλακτική της μορφής, π.χ. οι άνδρες είδαν σε μεγαλύτερο ποσοστό το πρόγραμμα, οι πίνακες θα βρίσκονται προς τη μια μεριά και άρα είναι πιο εύκολο να βρεθούν. Στην περίπτωση του μονόπλευρου ελέγχου η ελεγχοσυνάρτηση του Pearson δεν είναι κατάλληλη καθώς στην ουσία μετρά αποκλίσεις και προς τις δύο κατευθύνσεις. Επομένως το ακριβές p-value θα προκύψει αθροίζοντας τις υπεργεωμετρικές πιθανότητες των πινάκων που θεωρούνται πιο ακραίοι. Στο παράδειγμα μας, το p-value υπολογίστηκε ως

$$p - value = P(X_{11} = 0) + P(X_{11} = 2) + P(X_{11} = 3)$$

Συνεπώς χρειάζεται κανείς να υπολογίσει μια σειρά από υπεργεωμετρικές πιθανότητες και άρα να υλοποιήσει μια σειρά από παραγοντικά. Αυτό θα μπορούσε να είναι μεγάλο πρόβλημα καθώς τα παραγοντικά ακόμα και για μικρές σχετικά τιμές οδηγούν σε προβλήματα τον υπολογιστή. Ευτυχώς οι πιθανότητες μπορούν να υπολογιστούν χωρίς μεγάλη δυσκολία και συνήθως με τη χρήση επαναληπτικών αλγορίθμων. Για παράδειγμα αν συμβολίσουμε ως

$$P(x) = \frac{\left(\begin{array}{c} m_1 \\ x \end{array} \right) \left(\begin{array}{c} m_2 \\ n - x \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)},$$

μπορεί κανείς να χρησιμοποιήσει τον αναδρομικό τύπο

$$P(x+1) = P(x) \left(\frac{m_1 - x}{x+1} \right) \left(\frac{n-x}{m_2 - n + x + 1} \right), \quad x = 0, 1, 2, \ldots$$

με αρχική τιμή $P(0) = \frac{m_2!(N-n)!}{(m_2-n)!N!}$.

Στην πράξη δεν χρειάζεται κανείς να υλοποιήσει τα παραγοντικά καθώς και πάλι θα μπορούσε να κάνει επαναληπτικά τους πολλαπλασιασμούς συγχρόνως και στον αριθμητή αλλά και στον παρονομαστή ώστε να μην υπάρξουν προβλήματα με τους μεγάλους αριθμούς.

Αν προσπαθήσουμε να γενικεύσουμε για έναν πίνακα με r γραμμές και c στήλες τότε η πιθανότητα κάθε πίνακα X με στοιχεία x_{ij} δίνεται από τον τύπο

$$P(X) = \frac{\prod_{j=1}^{c} n_j! \prod_{i=1}^{r} m_i!}{N! \prod_{j=1}^{c} \prod_{i=1}^{r} x_{ij}!}$$

Υπάρχουν στη βιβλιογραφία έξυπνοι αλγόριθμοι με σκοπό αφενός το γρήγορο υπολογισμό των πιθανοτήτων, αλλά κυρίως τον υπολογισμό των πιθανοτήτων που χρειάζονται για το p-value.

Βέβαια σε πολλές περιπτώσεις που είτε οι διαστάσεις του πίνακα είναι μεγάλες, είτε το μέγεθος του δείγματος είναι μεγάλο οι υπολογισμοί, ακόμα και με τη χρήση κατάλληλων αλγορίθμων, είναι πρακτικά αδύνατοι.

Η λύση είναι να καταφύγει κανείς σε προσεγγιστικούς ελέγχους, δηλαδή να διαλέξει ένα τυχαίο δείγμα από όλους τους δυνατούς συνδυασμούς και να δουλέψει με αυτούς. Ουσιαστικά ισχύουν όλα όσα είπαμε και για τους απλούς ελέγχους τυχαιοποίησης. Βέβαια στην περίπτωση μεγαλύτερων πινάκων δεν είναι εύκολο κανείς να διακρίνει ποιοι είναι οι πίνακες που είναι πιο ακραίοι από τον παρατηρούμενο και επομένως καλό είναι να διαλέξει μια ελεγχοσυνάρτηση και να προχωρήσει σε έλεγχο τυχαιοποίησης με τον αλγόριθμο που περιγράψαμε στην αρχή του κεφαλαίου.

Παράδειγμα 2.3: Σε ένα διαγώνισμα εξετάστηκαν 20 φοιτητές από 3 διαφορετικά έτη. Η βαθμολογία τους φαίνεται στον πίνακα 2.12. Ο πρόεδρος του τμήματος ισχυρίζεται πως υπάρχει σχέση ανάμεσα στο βαθμό και το έτος στο οποίο βρίσκονται οι φοιτητές. Αληθεύει ο ισχυρισμός;

	2^o έτος	3^o έτος	4^o έτος	Μεγαλύτερο	
Κάτω από 5	5	2	2	0	9
5-7	0	1	0	1	2
8-10	0	2	3	4	9
	5	5	5	5	20

Πίνακας 2.12: Αποτελέσματα διαγωνίσματος, ως προς το βαθμό και το έτος του φοιτητή

Παρατηρείστε πως ο πίνακας είναι αραιός, έχει 20 παρατηρήσεις και 12 κελιά. Κάποια κελιά είναι άδεια και γενικά οι παρατηρούμενες συχνότητες είναι πολύ μικρές. Η τιμή της ελεγχοσυνάρτησης χ^2 είναι 11.556 και με βάση τα ασυμπτωτικά αποτελέσματα το p-value = 0.073 οπότε σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$ δεν απορρίπτουμε τη μηδενική υπόθεση. Ο συνολικός αριθμός διαφορετικών δειγμάτων που μπορούμε να πάρουμε είναι $5!2!2!2!3!4! = 13824$. Χρησιμοποιώντας τον ακριβή έλεγχο μπορεί να βρει κανείς πως το ακριβές p-value είναι 0.040. Βέβαια πρέπει κανείς να δημιουργήσει όλους τους πίνακες, κάτι που μπορεί να γίνει με τη χρήση των ειδικών αλγορίθμων που αναφέραμε. Παρατηρείστε πως ο

ακριβής έλεγχος οδήγησε σε διαφορετική απόφαση: απορρίπτουμε την υπόθεση της ανεξαρτησίας.

Για τον προσεγγιστικό έλεγχο δημιουργήσαμε 99 δείγματα. Και πάλι όπως βλέπετε στον πίνακα 2.7 η κατανομή της ελεγχοσυνάρτησης είναι διακριτή και όχι συνεχής (βέβαια παρατηρεί κανείς πως έχω μόλις 99 τιμές, αλλά είναι απλό να δείξει πως η ελεγχοσυνάρτηση μπορεί να πάρει διακριτές και μόνο τιμές)

Τιμή	Συχνότητα	Τιμή	Συχνότητα
2.6667	10	8.8889	3
3.5556	18	9.3333	6
4.4444	4	9.7778	4
5.3333	18	10.6667	2
6.2222	8	11.1111	5
7.1111	5	12.4444	1
7.5556	8	14.6667	1
8	6	Σύνολο	99

Πίνακας 2.13: Πίνακας συχνοτήτων για τις 99 τιμές της ελεγχοσυνάρτησης.

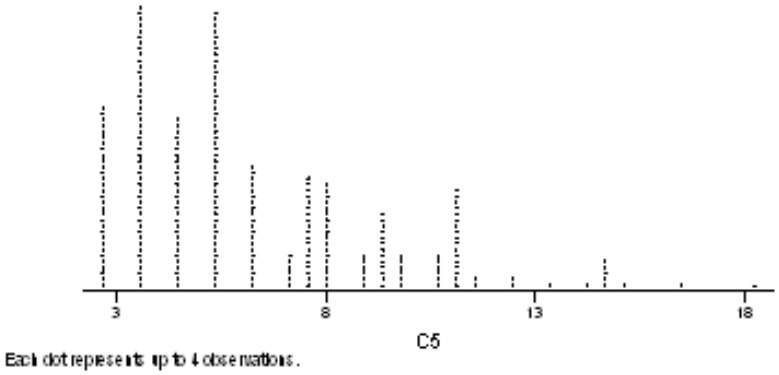
Με βάση όσα έχουμε πει το p-value εκτιμάται ως 0.03 και οδηγεί σε διαφορετική απάντηση (απορρίπτουμε τη μηδενική υπόθεση) από αυτή που θα παίρναμε με τη χρήση του ασυμπτωτικού αποτελέσματος. Βέβαια χρειάζεται να χρησιμοποιήσει κανείς περισσότερες επαναλήψεις ώστε να είναι σίγουρος (υποθέτουμε πως δεν έχουμε στα χέρια μας το ακριβές αποτέλεσμα). Επαναλαμβάνοντας το πείραμα 999 φορές το p-value που βρήκαμε ήταν 0.0398 κάτι που και αν δεν ξέραμε το ακριβές αποτέλεσμα θα βελτίωνε τη σιγουριά μας για να απορρίψουμε τη μηδενική υπόθεση. Παρατηρείστε πως όσο αυξάνουμε τον αριθμό των επαναλήψεων τόσο θα πλησιάζουμε το ακριβές αποτέλεσμα.

Αξίζει να παρατηρήσει κανείς πως αν ίσχυε το ασυμπτωτικό αποτέλεσμα τότε η μέση τιμή της ελεγχοσυνάρτησης όφειλε να είναι 6 και η διακύμανση 12 (θυμηθείτε πως για την χ^2 με k βαθμούς ελευθερίας η μέση τιμή είναι k και η διακύμανση $2k$). Για τις 999 τιμές που έχουμε βρήκαμε πως η μέση τιμή είναι 6.29 και η διακύμανση 8.98, επομένως η κατανομή αποκλίνει αρκετά από την ασυμπτωτική που συνήθως υποθέτουμε.

Στο γράφημα 2.5 μπορεί να δει κανείς το διάγραμμα σημείων για τις 999 τιμές της ελεγχοσυνάρτησης που προσομοιώσαμε. Η διακριτή φύση της κατανομής είναι ξεκάθαρη.

Αξίζει να αναφερθούν τα εξής

- Η προσέγγιση μπορεί να γενικευτεί και με πλήθος άλλες ελεγχοσυναρτήσεις που χρησιμοποιούμε για διάφορους ελέγχους σε πίνακες συνάφειας (πχ likelihood ratio test, odds ratio κλπ).



Γράφημα 2.5: Διάγραμμα σημείων για την κατανομή της ελεγχοσυνάρτησης

- Γενικά τα ασυμπτωτικά αποτελέσματα είναι πολύ λίγο αξιόπιστα και επομένως αν υπάρχει η δυνατότητα για ελέγχους τυχαιοποίησης (είτε ακριβείς είτε προσεγγιστικούς) αυτοί συστήνονται ανεπιφύλακτα.

- Εκτός από τους πίνακες συνάφειας υπάρχουν πολλοί άλλοι μη παραμετρικοί έλεγχοι οι οποίοι μπορούν να βελτιωθούν με τη χρήση της τυχαιοποίησης. Η μη παραμετρική στατιστική στηρίζεται στη λογική των ranks για τις οποίες υπάρχουν ασυμπτωτικά αποτελέσματα περί κανονικότητας. Οι περισσότερες ελεγχοσυναρτήσεις χρησιμοποιούν αυτό το αποτέλεσμα για να καταλήξουμε σε ασυμπτωτικά αποτελέσματα. Είναι ευνόητο πως κανείς μπορεί να βελτιώσει πολύ την συμπεριφορά τέτοιων ελέγχων με τη χρήση τυχαιοποίησης όπου η κατανομή της ελεγχοσυνάρτησης προσεγγίζεται από τις τιμές που προσομοιώσαμε από αυτήν.

2.8 Συμπεράσματα

Πριν τελειώσουμε τη περιγραφή των ελέγχων τυχαιοποίησης θα πρέπει να τονίσουμε τα εξής

- Μπορούμε να χρησιμοποιήσουμε απλές ελεγχοσυναρτήσεις χωρίς να ενδιαφερόμαστε αν οι κατανομές τους, όταν ισχύει η μηδενική υπόθεση, είναι γνωστές. Αρκεί μια απλή ελεγχοσυνάρτηση που να μπορεί να διακρίνει ανάμεσα στην H_0 και την H_1. Βέβαια η επιλογή κατάλληλης ελεγχοσυνάρτησης απαιτεί εμπειρία. Επίσης για τον ίδιο έλεγχο μπορεί πολλές ελεγχοσυναρτήσεις να κάνουν την ίδια δουλειά. 'Ομως εφαρμόζοντας στα ίδια δεδομένα

διαφορετικές ελεγχοσυναρτήσεις (προφανώς στα ίδια δείγματα) ενδέχεται να πάρει κάποιος διαφορετικές εκτιμήσεις για τα p-values. Αυτό έχει να κάνει με τη δυνατότητα κάθε ελεγχοσυνάρτησης να διακρίνει ανάμεσα στις δύο υποθέσεις. Δηλαδή έχει να κάνει όχι μόνο με τη στατιστική σημαντικότητα αλλά και το σφάλμα τύπου II .

- Δεν χρειάζεται να κάνουμε καμία υπόθεση σχετικά με τον πληθυσμό μας. Η μόνη υπόθεση που ουσιαστικά χρειαζόμαστε είναι πως κάτω από τη μηδενική υπόθεση μπορούμε να αλλάξουμε 'ταμπέλες' στις παρατηρήσεις μας, δηλαδή πως κάθε συνδυασμός δεδομένων είναι πιθανός κάτω από τη μηδενική υπόθεση. Δεν χρειάζεται να κάνουμε καμιά άλλη υπόθεση σχετικά με το μηχανισμό που δημιούργησε τα δεδομένα μας.

- Αυτό μπορεί βέβαια να είναι και μειονέκτημα καθώς η μηδενική υπόθεση πρέπει να είναι γενικά η απουσία κάθε δομής στα δεδομένα (πχ όλοι οι μέσοι είναι ίσοι, δεν υπάρχει συσχέτιση κλπ). 'Ετσι, για παράδειγμα, δεν μπορούμε να κάνουμε έλεγχο τυχαιοποίησης για μια πολύ συγκεκριμένη μηδενική υπόθεση όπως H_0: η μέση τιμή της θεραπείας Α είναι 3 φορές μεγαλύτερη από τις υπόλοιπες.

- 'Οπως είδαμε και στα παραδείγματα, προβλήματα που ανάγονται είτε σε γραμμική παλινδρόμηση είτε σε ανάλυση διακύμανσης αντιμετωπίζονται πολύ εύκολα με ελέγχους τυχαιοποίησης και έχουν ικανοποιητικά αποτελέσματα. Δεδομένου πως δεν χρειαζόμαστε σχεδόν καθόλου υποθέσεις, οι έλεγχοι τυχαιοποίησης (σχεδόν πάντα προσεγγιστικοί) είναι ενδιαφέρουσες εναλλακτικές μέθοδοι για τέτοια προβλήματα. Θα πρέπει πάντως να σημειωθεί ότι στις περιπτώσεις που η υπόθεση του γραμμικού μοντέλου που παραβιάζεται είναι η ομοσκεδαστικότητα ή η ανεξαρτησία, τότε και οι έλεγχοι τυχαιοποίησης έχουν μικρή ικανότητα, οπότε μπορούν να προσφέρουν λίγα πράγματα. 'Οσον αφορά την παραβίαση της ανεξαρτησίας γιατί καταρρέει όλη η στατιστική συμπερασματολογία που βασίζεται σε τυχαίο δείγμα. Ως προς την ομοσκεδαστικότητα, η μέθοδος ελαχίστων τετραγώνων παύει πια να δίνει αξιόπιστες εκτιμήσεις και επομένως πρέπει να χρησιμοποιηθούν πιο σύνθετες μέθοδοι εκτίμησης. Και η κλασσική παλινδρόμηση όμως παύει να είναι εφαρμόσιμη σε αυτή την περίπτωση. Αντίθετα αν παραβιάζεται η υπόθεση της κανονικότητας ο έλεγχος τυχαιοποίησης είναι εφαρμόσιμος.

- Οι έλεγχοι τυχαιοποίησης σε πολλές περιπτώσεις έχουν την ίδια ή και μεγαλύτερη ισχύ από υπάρχοντες παραμετρικούς ελέγχους και στις περισσότερες περιπτώσεις έχουν πολύ μεγαλύτερη ισχύ από τους κλασσικούς μη παραμετρικούς ελέγχους που βασίζονται στα ranks. Για παράδειγμα ο έλεγχος τυχαιοποίησης για την ανάλυση διακύμανσης έχει σχεδόν την ίδια ισχύ με τον κλασσικό έλεγχο ανάλυσης διακύμανσης για κανονικά δεδομένα όταν τα δεδομένα προέρχονται από κανονική κατανομή, είναι υποδεέστερος μόνο στην περίπτωση πολύ μικρών δειγμάτων (μικρότερα των 10 παρατηρήσεων). Στην περίπτωση που ο πληθυσμός δεν είναι κανονικός ο έλεγχος τυχαιοποίησης έχει πολύ μεγαλύτερη ισχύ και δεδομένου πως σχεδόν ποτέ δεν είμαστε σε θέση

να αποδείξουμε την υπόθεση της κανονικότητας, ο έλεγχος τυχαιοποίησης θα έπρεπε να προτιμάται από το κλασσικό F-test. Αυτό ισχύει ακόμα και όταν χρησιμοποιεί κανείς την ελεγχοσυνάρτηση F για τον έλεγχο τυχαιοποίησης (δηλαδή η ανωτερότητα οφείλεται όχι στην ελεγχοσυνάρτηση αλλά στο ότι δεν κάνουμε καμιά παραμετρική υπόθεση!). Σε κάθε περίπτωση ο έλεγχος τυχαιοποίησης είναι ανώτερος του μη παραμετρικού ελέγχου των Kruskal-Wallis.

Κεφάλαιο 3

'Ελεγχοι Monte Carlo

3.1 Εισαγωγή

Πολλές φορές η προσομοίωση αναφέρεται στη στατιστική ως μέθοδος Monte Carlo. Αυτό οφείλεται στο γεγονός πως κατά τη διάρκεια του Β' Παγκοσμίου Πολέμου, οπότε και η ανάγκη για καινούρια όπλα απαιτούσε εντατικά πειράματα αλλά με γρήγορα αποτελέσματα, χρησιμοποιήθηκαν κατά κόρον μέθοδοι προσομοίωσης με το κωδικό όνομα 'Monte Carlo experiment' για να μην γίνει αντιληπτό από τους εχθρούς. Επομένως όταν μιλάμε για μεθόδους Monte Carlo μιλάμε για μεθόδους που χρησιμοποιούν την προσομοίωση.

Η βασική ιδέα είναι η εξής. Ακόμα και αν δεν ξέρουμε την κατανομή μιας στατιστικής συνάρτησης, αν μπορούμε να προσομοιώσουμε τιμές αυτής της συνάρτησης, στην πραγματικότητα ξέρουμε πολλά πράγματα για αυτή. Για παράδειγμα, με τη χρήση κάποιας μεθόδου εκτίμησης της συνάρτησης πυκνότητας πιθανότητας μπορούμε να αναπαράγουμε έστω γραφικά μια εικόνα της κατανομής. Επίσης μπορούμε να υπολογίσουμε ποσότητες που μας ενδιαφέρουν όπως η μέση τιμή, ή η πιθανότητα η τιμή να ξεπερνά το 70.

Ας δούμε ένα παράδειγμα. 'Εστω ότι θέλουμε να κάνουμε έναν απλό έλεγχο για μια μέση τιμή, κι έστω ότι ξέρουμε τη διακύμανση. Μπορούμε να χρησιμοποιήσουμε την ελεγχοσυνάρτηση Z για την οποία ξέρουμε πως ακολουθεί την τυποποιημένη κανονική κατανομή. 'Ομως αυτό βασίζεται στην υπόθεση ότι τα δεδομένα προέρχονται από κανονικό πληθυσμό. Αν μας πουν πως τα δεδομένα προέρχονται, για παράδειγμα, από την εκθετική κατανομή, τότε μπορεί κάποιος να επικαλεστεί το κεντρικό οριακό θεώρημα και να χρησιμοποιήσει πάλι την Z. Αν όμως μας πουν πως το μέγεθος του δείγματος είναι 8, πολύ μικρό για να ισχύει το κεντρικό οριακό θεώρημα ,τι μπορούμε να πούμε για την κατανομή της ελεγχοσυνάρτησης Z; Προφανώς πρέπει να δουλέψουμε σκληρά με χαρτί και μολύβι για να βρούμε την κατανομή.

'Ισως το παράδειγμα με τη μέση τιμή να μην ήταν πειστικό, αφού υπάρχει αρκετή θεωρία που δικαιολογεί τη χρήση του κεντρικού οριακού θεωρήματος ακόμα και για μικρά δείγματα. Επίσης για την περίπτωση που ενδιαφερόμαστε για τον μέσο της εκθετικής μπορούμε να δείξουμε πως η κατανομή του είναι η Γάμμα

κατανομή. Πως όμως μπορεί να κάνει κανείς έλεγχο για τη διάμεσο ή για έναν περικομμένο μέσο (trimmed mean), όπου η θεωρία είναι περιορισμένη και δεν μπορούμε να βρούμε εύκολα την κατανομή της ελεγχοσυνάρτησης τους; Επίσης μπορεί κάποιος να αναφέρει λογής άλλες ελεγχοσυναρτήσεις των οποίων την κατανομή την ξέρουμε μόνο όταν πληρούνται κάποιες υποθέσεις. Για παράδειγμα την κατανομή της ελεγχοσυνάρτησης για τον έλεγχο ότι ο συντελεστής συσχέτισης είναι 0.6 δεν την ξέρουμε και χρειαζόμαστε την προσέγγιση του Fisher. Στην πραγματικότητα τις περισσότερες φορές δεν ξέρουμε την κατανομή της ελεγχοσυνάρτησης και βασιζόμαστε σε κάποια ασυμπτωτικά αποτελέσματα. Επίσης αν δούμε το θέμα πιο σφαιρικά, το ίδιο πρόβλημα παρουσιάζεται και για οποιαδήποτε συνάρτηση του δείγματος, όχι απαραίτητα μια ελεγχοσυνάρτηση, όπου δεν ξέρουμε για την κατανομή της.

Οι έλεγχοι Monte Carlo μπορούν να αντιμετωπίσουν αυτό το πρόβλημα ως εξής: Για το πρόβλημα ελέγχου μιας μέσης τιμής από εκθετικό πληθυσμό και μικρό δείγμα, αρκεί να προσομοιώσουμε πολλά (πόσο πολλά θα το δούμε αργότερα) δείγματα από τη μηδενική υπόθεση και για κάθε ένα δείγμα να υπολογίσουμε την τιμή της ελεγχοσυνάρτησης για κάθε δείγμα. Αυτές οι τιμές αποτελούν μια καλή προσέγγιση της κατανομής της ελεγχοσυνάρτησης και επομένως αν τις συγκρίνουμε με την παρατηρούμενη τιμή της ελεγχοσυνάρτησης μπορούμε να οδηγηθούμε σε συμπέρασμα. Το ίδιο γενικεύεται και στην περίπτωση οποιασδήποτε συνάρτησης του δείγματος, π.χ. trimmed μέση τιμή, όπου με την προσέγγιση αυτή επιτυγχάνουμε να κατασκευάσουμε την κατανομή τους.

3.2 Περιγραφή Ελέγχων

Σε σχέση με τους ελέγχους τυχαιοποίησης, η διαφορά είναι πως εδώ η μηδενική υπόθεση είναι πιο συγκεκριμένη. Στους ελέγχους τυχαιοποίησης η μηδενική υπόθεση υπέθετε την απουσία δομής και χρησιμοποιούσαμε αυτή την απουσία δομής για να φτιάξουμε πολλά δείγματα για να προσεγγίσουμε την κατανομή της ελεγχοσυνάρτησης. Στους ελέγχους Monte Carlo προσομοιώνουμε από την κατανομή που η μηδενική υπόθεση καθορίζει. Σε πολλές περιπτώσεις μάλιστα χρειάζεται να ξέρουμε την κατανομή του πληθυσμού για να μπορέσουμε να προχωρήσουμε.

'Ετσι τα βήματα για έναν έλεγχο Monte Carlo είναι τα εξής

'Ελεγχοι Monte Carlo

- Βήμα 1ο: Θέσε τη μηδενική και την εναλλακτική υπόθεση
- Βήμα 2ο: Διάλεξε την ελεγχοσυνάρτηση (με βάση τα κριτήρια που είπαμε και προηγουμένως). Για να μπορούμε να γενικεύσουμε τη μέθοδο υποθέτουμε πως η ελεγχοσυνάρτηση μπορεί να απορρίπτει τη μηδενική υπόθεση και στις δύο ουρές
- Βήμα 3α: Προσομοίωσε k δείγματα μεγέθους n (το μέγεθος του δικού μας πραγματικού δείγματος) από τη μηδενική υπόθεση (αυτό συνήθως σημαίνει από την κατανομή και τις παραμέτρους που υπονοεί η μηδενική υπόθεση)

- Βήμα 3β: Για κάθε δείγμα που δημιούργησες υπολόγισε την τιμή της ελεγχοσυνάρτησης.

- Βήμα 4ο: Συγκρίνοντας την τιμή της ελεγχοσυνάρτησης για τα πραγματικά δεδομένα με αυτές τις τιμές από τα δείγματα που προσομοίωσες, δες πόσο ακραία είναι και εκτίμησε το p-value ως

$$p - value = \frac{m+1}{\mathrm{k}+1} \quad ,$$

όπου m είναι τώρα ο αριθμός των φορών που η τιμή της ελεγχοσυνάρτησης από τα προσομοιωμένα δείγματα είναι πιο ακραία από αυτή που παρατηρήσαμε.

Μερικά ενδιαφέροντα σημεία είναι τα εξής:

'Οταν λέμε ότι η τιμή της ελεγχοσυνάρτησης από τα προσομοιωμένα δείγματα είναι πιο ακραία από αυτή που παρατηρήσαμε εξαρτάται από την εναλλακτική υπόθεση και τη μορφή της. Δηλαδή ποιες τιμές, σε ποια ουρά της κατανομής της ελεγχοσυνάρτησης οδηγούν σε απόρριψη της μηδενικής υπόθεσης.

Η ιδέα των ελέγχων αυτών είναι πως με την προσομοίωση βρίσκουμε τι τιμές θα έπαιρνε η ελεγχοσυνάρτηση αν πραγματικά ίσχυε η μηδενική υπόθεση. Αν αυτό που παρατηρήσαμε είναι πολύ μακριά σημαίνει πως μάλλον τα δεδομένα μας δεν έχουν προκύψει από τη μηδενική υπόθεση.

Μερικά παραδείγματα είναι τα εξής:

- 'Ελεγχος για έναν συντελεστή συσχέτισης όταν ξέρουμε πως οι πληθυσμοί ακολουθούν κατανομές Γάμμα. Τα δείγματα πρέπει να προσομοιωθούν από δύο Γάμμα κατανομές, πιθανότατα με συγκεκριμένη συσχέτιση όπως αυτή καθορίζεται στη μηδενική υπόθεση.

- 'Ελεγχος για μια μέση τιμή όταν η κατανομή είναι η Αρνητική Διωνυμική. Προσομοιώνουμε δείγματα του δοθέντος μεγέθους από την αρνητική διωνυμική κατανομή.

- 'Ελεγχος για την ισότητα διακυμάνσεων όταν οι δυο πληθυσμοί ακολουθούν την κατανομή Poisson. Προσομοιώνουμε δείγματα από δύο κατανομές Poisson με παραμέτρους που είτε δίνονται, είτε προκύπτουν από τη μηδενική υπόθεση

- 'Ελεγχος για το αν τα δεδομένα ακολουθούν τη κατανομή *t-student*. Προσομοιώνουμε δείγματα από την κατανομή που έχουμε υποθέσει και για κάθε δείγμα υπολογίζουμε την τιμή της ελεγχοσυνάρτησης.

Παρατήρηση: Το κοινό στοιχείο σε όλα τα παραπάνω είναι πως ξέρουμε την κατανομή του πληθυσμού είτε γιατί προκύπτει από τη μηδενική υπόθεση είτε γιατί έχουμε βάσιμα στοιχεία για να υποθέσουμε κάτι τέτοιο. Βέβαια σε αυτή την περίπτωση, όπου χρειάζεται να κάνουμε μια παραμετρική υπόθεση σχετικά με την κατανομή του πληθυσμού, το μόνο κέρδος σε σχέση με τους κλασσικούς ελέγχους είναι πως δεν χρειαζόμαστε αποκλειστικά την υπόθεση της κανονικότητας αλλά μπορούμε να δουλέψουμε και με πιο περίπλοκες επιλογές.

3.3 Παραδείγματα

Ας δούμε κάποια παραδείγματα με περισσότερη λεπτομέρεια.

3.3.1 'Ελεγχος για μια μέση τιμή από εκθετικό πληθυσμό

Σε ένα εργοστάσιο, οι χρόνοι σε μέρες που μεσολάβησαν ανάμεσα σε δύο διαδοχικές βλάβες ενός μηχανήματος καταγράφηκαν με σκοπό να μελετηθεί η συχνότητα εμφάνισης βλαβών. Οι 12 παρατηρήσεις που πάρθηκαν ήταν οι εξής

0.18, 0.89, 0.61, 0.23, 0.16, 0.37, 0.29, 0.03, 0.05, 0.47, 1.17, 0.24

Να ελεγχθεί η μηδενική υπόθεση ότι ο μέσος χρόνος ανάμεσα σε δύο διαδοχικές βλάβες είναι 1 μέρα ξέροντας πως η κατανομή του χρόνου είναι η εκθετική

Α. Αν η εναλλακτική είναι πως $\mu > 1$

Β. Αν η εναλλακτική είναι πως $\mu < 1$

Γ. Αν η εναλλακτική είναι πως $\mu \neq 1$

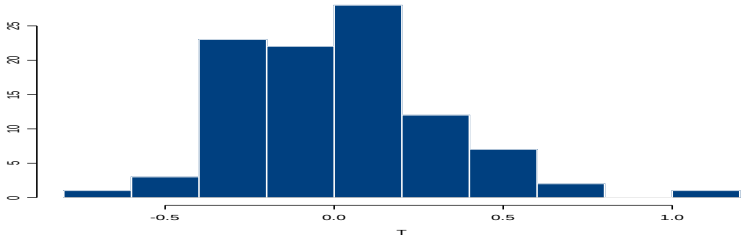
Αφού ξέρουμε πως η κατανομή είναι η εκθετική και δεδομένου ότι το μέγεθος του δείγματος είναι μικρό, και άρα το κεντρικό οριακό θεώρημα δεν μπορεί να εφαρμοσθεί, προχωράμε σε έναν έλεγχο Monte Carlo. Η ελεγχοσυνάρτηση που διαλέγουμε είναι η $T = \bar{x} - 1$. Αυτή η ελεγχοσυνάρτηση, αν ισχύει η μηδενική υπόθεση θα παίρνει την τιμή 0 ενώ μεγάλες τιμές μακριά από το 0 δείχνουν απόκλιση από τη μηδενική υπόθεση. Το πρόσημο των τιμών που δείχνουν υπέρ της εναλλακτικής σαφώς και εξαρτάται από την εναλλακτική. 'Ετσι για την περίπτωση Α απορρίπτουμε τη μηδενική υπόθεση όταν οι τιμές της Τ είναι μεγάλες (δεξιά ουρά), για την περίπτωση Β όταν είναι μικρές (αριστερή ουρά) ενώ για την περίπτωση Γ και στις δύο ουρές. Δηλαδή για την περίπτωση Γ θα μπορούσε κανείς να πάρει ως ελεγχοσυνάρτηση την $|T|$.

'Ο έλεγχος γίνεται ως εξής:

- Προσομοιώνουμε k, έστω 99, δείγματα από την εκθετική κατανομή με μέση τιμή 1. Κάθε τέτοιο δείγμα έχει 12 παρατηρήσεις.
- Για κάθε δείγμα υπολογίζουμε την ελεγχοσυνάρτηση (δηλαδή βρίσκουμε τη μέση τιμή και αφαιρούμε 1). Αυτές λοιπόν οι 99 τιμές που συγκεντρώνουμε αποτελούν μια εκτίμηση της κατανομής της ελεγχοσυνάρτησης (την οποία δεν ξέρουμε γιατί αν την ξέραμε δεν θα χρειαζόταν να κάνουμε έλεγχο Monte Carlo).
- Στη συνέχεια βρίσκουμε πόσες από τις 99 τιμές της ελεγχοσυνάρτησης είναι πιο ακραίες (θυμηθείτε ότι το ακραίες έχει να κάνει με την εναλλακτική υπόθεση) και εκτιμούμε το p-value

Για τα δεδομένα μας έχουμε $\bar{x} = 0.39$ οπότε $T = -0.61$. Στο γράφημα 3.1 μπορείτε να δείτε ένα ιστόγραμμα με τις 99 τιμές της ελεγχοσυνάρτησης.

Για να απαντήσουμε το Α ερώτημα πρέπει να δούμε πόσες παρατηρήσεις είναι μεγαλύτερες από -0.61 που βρήκαμε για το δείγμα μας. Βρίσκουμε λοιπόν πως 98 παρατηρήσεις είναι μεγαλύτερες από το $t_{obs} = 0.61$ και άρα $\tilde{p} = \frac{m+1}{k+1} = \frac{98+1}{99+1} =$



Γράφημα 3.1: Ιστόγραμμα για την ελεγχοσυνάρτηση βασισμένο σε 99 επαναλήψεις

0.99 είμαστε λοιπόν αρκετά σίγουροι και δεχόμαστε τη μηδενική υπόθεση. Αυτό μοιάζει πολύ λογικό. Η μέση τιμή που βρήκαμε είναι 0.39 και επομένως θα ήταν παράλογο να απορρίψουμε την υπόθεση πως $\mu = 1$ έναντι της $\mu > 1$.

Για το ερώτημα Β ενδιαφερόμαστε για τις παρατηρήσεις που είναι μικρότερες του --0.61. Μόνο μια παρατήρηση είναι μικρότερη επομένως τώρα έχουμε $\tilde{p} = \frac{1+1}{99+1} = 0.02$ επομένως σε $\alpha = 5\%$ απορρίπτουμε τη μηδενική υπόθεση.

Για την υπόθεση Γ ενδιαφερόμαστε και για τις 2 ουρές δηλαδή πόσες παρατηρήσεις είναι μικρότερες από --0.61 και πόσες μεγαλύτερες από 0.61. 'Εχουμε 4 παρατηρήσεις και επομένως $\tilde{p} = \frac{4+1}{99+1} = 0.05$ δηλαδή δεν είμαστε σε θέση να απορρίψουμε τη μηδενική υπόθεση, δηλαδή να πάρουμε μια απόφαση αφού το p-value είναι οριακό. Προφανώς θα πρέπει να πάρουμε περισσότερα δείγματα. Επαναλαμβάνοντας τον έλεγχο με $k = 999$ βρίσκουμε πως $\tilde{p} = 0.039$ οπότε και απορρίπτουμε τη μηδενική υπόθεση

Αξίζει εδώ να σημειωθεί ότι χρησιμοποιώντας το κεντρικό οριακό θεώρημα η κατανομή της ελεγχοσυνάρτησης πρέπει να είναι η κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση $1/\sqrt{12}$. Χρησιμοποιώντας αυτή την προσέγγιση βρίσκουμε για τους 3 ελέγχους τα εξής p-value 0.975, 0.025, 0.051 τα οποία είναι πολύ κοντά σε αυτά που βρήκαμε. Παρατηρείστε όμως πως στον τρίτο έλεγχο και για $\alpha = 5\%$ καταλήγουμε σε διαφορετική απόφαση.

3.3.2 'Ελεγχος καλής προσαρμογής

Σε μια αγωνιστική του πρωταθλήματος ποδοσφαίρου σημειώθηκαν συνολικά σε κάθε παιχνίδι τα γκολ που βλέπετε στον πίνακα 3.1. Θέλουμε να απαντήσουμε στο ερώτημα κατά πόσο τα δεδομένα ακολουθούν την κατανμή Poisson.

Ο έλεγχος που θέλουμε να κάνουμε είναι ένας έλεγχος καλής προσαρμογής. Επειδή δεν έχουμε καμιά πληροφορία σχετικά με την παράμετρο της κατανομής Poisson θα πρέπει να την εκτιμήσουμε. Επομένως η μηδενική υπόθεση είναι πως

H_0: η κατανομή είναι Poisson έναντι της

H_1: η κατανομή δεν είναι Poisson

Επειδή η εναλλακτική είναι πολύ γενική, για να διευκολύνουμε το παράδειγμα, θα υποθέσουμε πως η εναλλακτική είναι η

H_1: η κατανομή είναι Αρνητική Διωνυμική

Αριθμός Γκολ	Συχνότητα (Αριθμός ομάδων)
0	6
1	3
2	4
3	2
4	1

Πίνακας 3.1: Αριθμός γκολ που επιτεύχθηκαν συνολικά σε κάθε αγώνα

Ας βρούμε μια κατάλληλη ελεγχοσυνάρτηση. Γενικά για ελέγχους καλής προσαρμογής έχουμε τις εξής επιλογές:

Να διαλέξουμε ένα μέτρο που να μας δείχνει πόσο απέχουν τα δεδομένα μας από αυτά που θα έδινε η κατανομή που ελέγχουμε. Για να γίνει αυτό θα πρέπει να χρησιμοποιήσουμε κάποια απόσταση. Μερικοί από τους πιο γνωστούς (αλλά όχι απαραίτητα καλύτερους) ελέγχους καλής προσαρμογής χρησιμοποιούν αυτή την ιδέα, όπως ο γνωστός χ^2 έλεγχος καλής προσαρμογής ή ο έλεγχος Kolmogorov --Smirnov κατάλληλα διασκευασμένος για να δουλεύει με διακριτά δεδομένα.

Να βρούμε κάποια ποσότητα που χαρακτηρίζει την κατανομή σε σχέση με την εναλλακτική. Για παράδειγμα η κανονική κατανομή είναι η μόνη συνεχής κατανομή με ασυμμετρία 0 και κύρτωση 3. Για την περίπτωση της Poisson ξέρουμε πως έχει την ιδιότητα να έχει μέση τιμή και διακύμανση ίσες κάτι που δεν ισχύει για την αρνητική διωνυμική. Εδώ χρειάζεται πολύ προσοχή και θεωρητική γνώση καθώς η ιδιότητα που θα χρησιμοποιήσουμε πρέπει να χαρακτηρίζει την κατανομή, δηλαδή να μην την έχει καμιά άλλη κατανομή. Για παράδειγμα, αν θέλαμε να ελέγξουμε για την κατανομή t με 3 βαθμούς ελευθερίας θα μπορούσε να παρατηρήσει κανείς πως η μέση τιμή είναι 0 και η διακύμανση 3. Αυτή όμως η ιδιότητα δεν είναι καθόλου χαρακτηριστική για την κατανομή καθώς και μια κανονική κατανομή θα μπορούσε να έχει μέση τιμή 0 και διακύμανση 3, οπότε μια ελεγχοσυνάρτηση βασισμένη σε αυτή την ιδιότητα δεν θα μπορούσε ποτέ να ξεχωρίσει ανάμεσα στην κατανομή t και την κανονική κατανομή.

Επομένως, επιστρέφοντας στο παράδειγμα με την κατανομή Poisson, μια ελεγχοσυνάρτηση που μπορούμε να χρησιμοποιήσουμε είναι η $T = \frac{s^2}{\bar{x}}$ για την οποία περιμένουμε τιμές κοντά στο 1 αν τα δεδομένα είναι από την κατανομή Poisson ενώ αν προέρχονται από την αρνητική διωνυμική οι τιμές πρέπει να είναι μεγαλύτερες από τη μονάδα. 'Εχετε υπόψη ότι αυτός ο λόγος διακύμανσης προς τη μέση τιμή λέγεται και δείκτης διασποράς (index of dispersion) και χρησιμοποιείται στην πράξη για διακριτά δεδομένα σαν ένας δείκτης υπερδιασποράς (overdispersion). Καθώς υπάρχουν συγκεκριμένοι μηχανισμοί που οδηγούν σε υπερδιασπορά μπορεί κανείς να μελετήσει ένα φαινόμενο με τη χρήση παρόμοιων δεικτών.

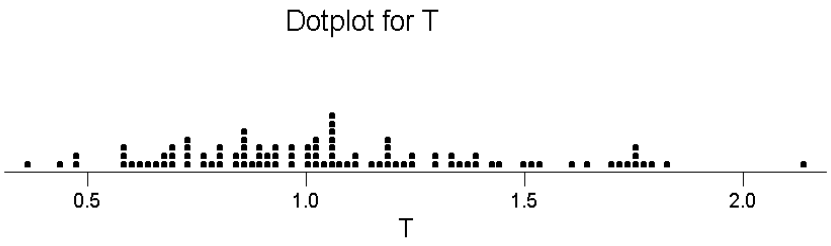
Επομένως ο έλεγχος έχει ως εξής:

- Προσομοιώνουμε k, έστω 99, δείγματα από την κατανομή Poisson με παράμετρο 1.31 (ο μέσος των τιμών του δείγματος που έχουμε). Κάθε τέτοιο δείγμα έχει 16 παρατηρήσεις. Επειδή η μηδενική υπόθεση δεν καθορίζει κάποια παράμετρο για την κατανομή Poisson θα πρέπει να την εκτιμήσουμε εμείς

από τα δεδομένα. Είναι γνωστό πως η εκτιμήτρια μεγίστης πιθανοφάνειας για την παράμετρο της κατανομής Poisson είναι η μέση τιμή του δείγματος.

- Για κάθε δείγμα υπολογίζουμε την ελεγχοσυνάρτηση (δηλαδή βρίσκουμε τη μέση τιμή και τη διακύμανση και παίρνουμε το λόγο τους).
- Στη συνέχεια βρίσκουμε πόσες από τις 99 τιμές της ελεγχοσυνάρτησης είναι πιο μεγάλες από αυτή του δείγματος μας.

Για το δείγμα μας βρίσκουμε $\bar{x} = 1.31, s^2 = 1.69, T = 1.29$. Οι 99 τιμές της ελεγχοσυνάρτησης φαίνονται στο γράφημα 3.2. Υπάρχουν 24 τιμές μεγαλύτερες από 1.29 επομένως $\tilde{p} = \frac{24+1}{99+1} = 0.25$ και άρα δεν απορρίπτουμε σε $\alpha = 5\%$ την υπόθεση ότι τα δεδομένα προέρχονται από την κατανομή Poisson. Με άλλα λόγια για ένα δείγμα μεγέθους 12, από την κατανομή Poisson με μέση τιμή 1.31 δεν είναι απίθανο να πάρουμε έναν δείκτη διασποράς ίσο με 1.29



Γράφημα 3.2: Διάγραμμα σημείων για την ελεγχοσυνάρτηση του παραδείγματος

Θα πρέπει σε αυτό το σημείο να τονίσουμε πως η ελεγχοσυνάρτηση που χρησιμοποιήσαμε θα ήταν καταδικασμένη να αποτύχει αν η εναλλακτική υπόθεση ήταν πολύ γενική, δηλαδή όπως είχαμε ορίσει στην αρχή, η εναλλακτική υπόθεση ήταν πως τα δεδομένα δεν ακολουθούν την κατανομή Poisson (κι επομένως ακολουθούν κάποια οποιαδήποτε κατανομή). Ο λόγος είναι πως ο λόγος της διακύμανσης προς τη μέση τιμή αν και είναι χαρακτηριστική ιδιότητα της κατανομής Poisson ως προς όλες τις κατανομές που ορίζονται στους ακέραιους αριθμούς $0, 1, \ldots$ δεν ισχύει το ίδιο για όλες τις διακριτές κατανομές. Δηλαδή υπάρχουν διακριτές κατανομές που ορίζονται σε ένα πεπερασμένο αριθμό ακέραιων τιμών κι έχουν αυτή την ιδιότητα. Σε αυτή την περίπτωση η ελεγχοσυνάρτηση δεν μπορεί να ξεχωρίσει ανάμεσα στις 2 υποθέσεις. Συνεπώς μια ελεγχοσυνάρτηση μπορεί να είναι καλή επιλογή για κάποια εναλλακτική υπόθεση ενώ για άλλες εναλλακτικές υποθέσεις να μην είναι καλή επιλογή.

3.4 Συμπεράσματα

Πριν τελειώσουμε την περιγραφή των ελέγχων Monte Carlo θα πρέπει να παρατηρήσουμε τα εξής:

- Σχετικά με την επιλογή της ελεγχοσυνάρτησης, τον αριθμό των επαναλήψεων και την εκτίμηση του p-value, ισχύουν όλα όσα είπαμε στους ελέγχους τυχαιοποίησης.

- Στους ελέγχους Monte Carlo χρησιμοποιούμε απλά την πρόσθετη πληροφορία που έχουμε σχετικά με την κατανομή του πληθυσμού.

- Οι έλεγχοι αυτοί είναι ιδιαίτερα χρήσιμοι για ελέγχους καλής προσαρμογής όπου η μηδενική υπόθεση μας κατευθύνει για την κατανομή από την οποία θα προσομοιώσουμε τα δείγματα μας.

Ας ξεφύγουμε όμως λίγο από την περίπτωση των ελέγχων υποθέσεων. Στην πραγματικότητα αυτό που κάναμε ήταν να εκτιμούμε τη συνάρτηση πυκνότητας πιθανότητας της ελεγχοσυνάρτησης και μετά να χρησιμοποιούμε αυτή την εκτίμηση, βασισμένη στην προσομοίωση, για να κάνουμε τον έλεγχο υποθέσεων. Από αυτή την οπτική, μπορεί κανείς να χρησιμοποιήσει παρόμοια τεχνική για να εκτιμήσει την κατανομή, όχι πια μιας ελεγχοσυνάρτησης με σκοπό κάποιον έλεγχο υπόθεσης, αλλά οποιασδήποτε άλλης συνάρτησης του δείγματος. Για παράδειγμα μπορεί κάποιος να εκτιμήσει την κατανομή της δειγματικής διαμέσου από ένα δείγμα μεγέθους πχ 43, όταν ξέρουμε πως ο πληθυσμός είναι εκθετικός. Ακόμα, μπορεί κανείς χρησιμοποιώντας Monte Carlo να εκτιμήσει διάφορες ποσότητες για αυτές τις στατιστικές συναρτήσεις, όπως τυπικά σφάλματα, πιθανότητες να ξεπεράσουν κάποια τιμή κλπ.

Η ιδέα αυτή θα αναπτυχθεί με περισσότερες λεπτομέρειες στη συνέχεια όπου θα δούμε πως η μέθοδος Monte Carlo αποτελεί ειδική περίπτωση μιας πολύ γενικότερης μεθοδολογίας που ονομάζεται bootstrap (κεφάλαιο 5). Η βασική ιδέα που πρέπει να κρατήσει ο αναγνώστης είναι πως αν δεν είμαστε σε θέση να βρούμε την κατανομή μιας στατιστικής συνάρτησης όταν αυτή προέρχεται από πληθυσμό με γνωστή κατανομή τότε μπορούμε να καταφύγουμε στη λύση της προσομοίωσης. Αν καταφέρουμε να πάρουμε δείγμα από την κατανομή του πληθυσμού τότε αυτόματα μπορούμε να έχουμε μια πολύ καλή εικόνα και από την κατανομή της συνάρτησης που μας ενδιαφέρει χρησιμοποιώντας τα χαρακτηριστικά αυτού του δείγματος. Συνεπώς το πρόβλημα της εύρεσης της κατανομής μιας στατιστικής συνάρτησης μετασχηματίζεται σε πρόβλημα προσομοίωσης από την κατανομή του πληθυσμού. Ο τρόπος που προσομοιώνουμε από διάφορες κατανομές δεν μας απασχολεί σε αυτές τις σημειώσεις αλλά υπάρχουν ποικίλες μέθοδοι που προσφέρονται από όλα τα στατιστικά πακέτα.

Για παράδειγμα αν και δεν γνωρίζουμε την κατανομή ης ποσότητας $V = \sigma^2/\bar{x}$ από ένα δείγμα μεγέθους n από κατανομή Poisson, εντούτοις όπως είδαμε στο παράδειγμα είμαστε σε θέση να γνωρίζουμε αρκετά στοιχεία για αυτή την κατανομή, όπως να εκτιμήσουμε τη μέση της τιμή, τη διακύμανση της, το ποσοστό τιμών μεγαλύτερων από 1, αλλά ακόμα και να εκτιμήσουμε τη συνάρτηση πυκνότητας πιθανότητας της.

Επομένως η μέθοδος Monte Carlo δεν αφορά μόνο ελέγχους υποθέσεων αλλά και άλλα αντικείμενα της στατιστικής συμπερασματολογίας. Περισσότερα στο κεφάλαιο 5.

3.5 Ασκήσεις κεφαλαίων 2 και 3

1. Ας υποθέσουμε πως τα δεδομένα μας αποτελούνται από ζευγάρια μετρήσεων (X_i, Y_i). 'Εστω ότι θέλουμε να ελέγξουμε την υπόθεση ότι ο συντελεστής συσχέτισης $\rho = -0.10$ έναντι της εναλλακτικής ότι είναι διάφορος του -0.10 χρησιμοποιώντας έναν έλεγχο τυχαιοποίησης. Ποια από τις τρεις στατιστικές συναρτήσεις T_1, T_2, T_3 θα προτιμήσετε και γιατί;

   Δίνονται:

$$T_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \quad T_2 = \sum_{i=1}^{n} X_i Y_i, \quad T_3 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

2. Στη φυσιολογία ο καρδιακός ρυθμός κάθε ανθρώπου έχει κάποια συνηθισμένη περιοδικότητα η οποία είναι διαφορετική για κάθε άνθρωπο. 'Εστω ότι είναι διαθέσιμες μετρήσεις για μια περίοδο 6 συνεχόμενων ημερών ανά 4 ώρες για 10 διαφορετικά άτομα και ότι θέλουμε να ελέγξουμε αν πράγματι υπάρχει αυτή η περιοδικότητα με τη χρήση ενός ελέγχου τυχαιοποίησης.

   Α. Θεωρείτε πως η στατιστική συνάρτηση $T = \sum_{i=1}^{10}\sum_{j=1}^{24}(x_{ij} - \bar{x}_j)^2$, όπου x_{ij} είναι η μέτρηση τη χρονική στιγμή j για το i άτομο και $\bar{x}_j$ είναι η μέση τιμή για τη χρονική στιγμή j για όλα τα άτομα, είναι κατάλληλη;

   Β. Περιγράψτε με βήματα πως θα γίνει ο έλεγχος τυχαιοποίησης (άσχετα αν χρησιμοποιήσετε την T ή κάποια άλλη ελεγχοσυνάρτηση).

3. 'Εστω 30 παρατηρήσεις από την κατανομή Γάμμα. Βρήκαμε πως η μέση απόλυτη απόκλιση του δείγματος είναι 12. Περιγράψτε πως θα ελέγξετε την υπόθεση πως η μέση απόλυτη απόκλιση στον πληθυσμό είναι 15, χρησιμοποιώντας όποια μέθοδο θέλετε. Αναφέρετε πως θα απαντούσατε το ερώτημα χρησιμοποιώντας κάποια διαφορετική μέθοδο.

   $(MAD = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}\right|)$.

4. 'Ενα δείγμα από 8 φοιτητές ρωτήθηκαν για την ποιότητα του μαθήματος που διδάχθηκαν. Οι απαντήσεις τους, λαμβάνοντας υπόψη και τη συχνότητα των παρακολουθήσεων τους ήταν οι εξής:

	Θετική γνώμη	Αρνητική γνώμη
Τακτική παρακολούθηση	3	1
'Οχι τακτική παρακολούθηση	1	3

   Ισχύει ο ισχυρισμός πως αυτοί που παρακολουθούσαν τακτικά έχουν πιο θετική γνώμη για το μάθημα;

5. Για να δοκιμαστεί μια καινούρια θεραπεία, 4 ασθενείς πήραν ένα καινούριο φάρμακο, ενώ 3 άλλοι πήραν το συνηθισμένο φάρμακο. Οι τιμές που παρατηρήσαμε ήταν 4,7,8,5 και 5,10, 4 αντίστοιχα για τις δύο ομάδες. Είναι αποτελεσματικότερη η καινούρια θεραπεία;

6. 'Ενα δείγμα από 108 τηλεθεατές ρωτήθηκαν για την ποιότητα ενός τηλεοπτικού προγράμματος. Οι απαντήσεις τους, λαμβάνοντας υπόψη και τη συχνότητα των παρακολουθήσεων τους ήταν οι εξής:

	Θετική γνώμη	Αρνητική γνώμη
Είδαν και τα 12 επεισόδια	34	17
Είδαν 6-11 επεισόδια	15	18
Είδαν λιγότερα από 6 επεισόδια	12	12

Χρησιμοποιείστε τον έλεγχο του Fisher (προσεγγιστικό) για να ελέγξετε αν ισχύει ο ισχυρισμός πως αυτοί που παρακολουθούσαν τακτικά έχουν διαφορετική γνώμη για το πρόγραμμα; Περιγράψτε πλήρως τη διαδικασία με αριθμό δειγμάτων 10. (Δηλαδή περιγράψτε πλήρως τους πίνακες που σχηματίσατε, την τιμή της ελεγχοσυνάρτησης κλπ)

7. Για να δοκιμαστεί ένα καινούριο υλικό, 24 συσκευές κατασκευάστηκαν με το καινούριο υλικό, ενώ 7 με το παλιό υλικό και στη συνέχεια υποβλήθηκαν σε τεστ αντοχής. Οι τιμές που παρατηρήσαμε σε ώρες ήταν

Θεραπεία	Παρατηρήσεις
A	$3, 2, 4, 5, 6, 7, 5, 6, 7, 4, 5, 6, 4, 8, 9, 12, 12, 7, 8, 12, 12, 14, 18, 16$
B	$17, 19, 21, 19, 21, 28, 25$

Χρησιμοποιώντας την ακριβή μέθοδο τυχαιοποίησης θεωρείτε πως είναι αποτελεσματικότερο το καινούριο υλικό;

8. Ο κύριος Χ πιστεύει πως μπορεί να ξεχωρίσει ανάμεσα σε 5 ποικιλίες τσαγιού απλά από το χρώμα τους και χωρίς να τα γευτεί. Για να το αποδείξει ζήτησε από κάποιον να βάλει 5 διαφορετικές ποικιλίες σε 5 φλυτζάνια και να υποδείξει ποιό φλυτζάνι έχει ποια ποικιλία. Τα δεδομένα είναι τα εξής:

Ποικιλία τσαγιού στο φλυτζάνι	A	B	Γ	Δ	E
Επιλογή ειδικού	E	B	Γ	Δ	A

Πιστεύετε πως είναι τελικά ειδικός ή απλά απάντησε στην τύχη;

9. Έστω πως έχουμε k διαφορετικά δείγματα και μας ενδιαφέρει ο έλεγχος της H_0 ότι οι μέσοι των δειγμάτων είναι ίσοι έναντι της H_1 πως οι μέσοι είναι σε μια συγκεκριμένη σειρά. Για παράδειγμα τα δεδομένα που ακολουθούν αφορούν τον αριθμό των μεταλλάξεων στα χρωμοσώματα σε ποντίκια όταν αυτά υποβληθούν σε συγκεκριμένες δόσεις ενός φαρμάκου

Δόση (mg/kg)	αριθμός μεταλλάξεων				
0	0	0	0	0	
5	1	1	1	4	5
20	0	0	0	4	
80	2	3	5	11	20

Κάποιος ερευνητής πρότεινε τη χρήση της ελεγχοσυνάρτησης

$$T = \sum_{j=1}^{k} g(j)x_j$$

όπου $g(j)$ είναι μια μονότονα άυξουσα συνάρτηση και x_j είναι το άθροισμα των παρατηρήσεων του j δείγματος

- Εξηγήστε γιατί η συγκεκριμένη ελεγχοσυνάρτηση είναι κατάλληλη για τον έλεγχο υποθέσεων που θέλουμε να κάνουμε
- Επιλέξτε 2 διαφορετικές συναρτήσεις $g(j)$ για τον έλεγχο και συγκρίνετε τα αποτελέσματα τους. Ποια θα προτιμήσετε και γιατί;
- Θα συμφωνούσατε με τη χρήση της συνάρτησης $g(j) = \log(dose_j + 1)$;

10. Στην ψυχολογία, η θεωρία της γνωσιακής παραφωνίας (cognitive dissonance) υποστηρίζει πως οι άνθρωποι όταν έχουν αλληλοσυγκρουόμενες απόψεις για ένα θέμα τείνουν να μειώνουν την παραφωνία αυτή για να νιώσουν καλύτερα. Σε ένα πείραμα που έγινε στην Αμερική κι έλαβαν μέρος φοιτητές, αυτοί έκαναν μια σειρά από απλές και βαρετές εργασίες και στη συνέχεια πληρώθηκαν 1 ή 20 δολάρια για να πουν ψέματα στους επόμενους φοιτητές που θα λάμβαναν μέρος στο πείραμα, ότι αυτό ήταν πολύ ενδιαφέρον. Στη συνέχεια ερωτήθηκαν κατά πόσο πίστευαν πως οι εργασίες που έκαναν ήταν ενδιαφέρουσες αξιολογώντας σε μια κλίμακα από 0 (πολύ βαρετό) εως 10 (πολύ συναρπαστικό). Αυτοί που έλαβαν 20 δολάρια έδωσαν βαθμολογίες 3, 1, 2, 4, 0, 5, 4, 5 ενώ αυτοί που πήραν μόνο 1 δολάριο έδωσαν βαθμολογίες 4, 8, 6, 9, 3, 6, 7, 10, 4 , 8. Είναι η διαφορά στατιστικά σημαντική; Επιβεβαιώνει τη θεωρία;

11. Μερικοί ερευνητές ακόμα και τώρα προτείνουν τη χρήση του ελέγχου χ^2 του Pearson για να ελέγξουν αν τα δεδομένα προέρχονται από μια κατανομή. Ας υποθέσουμε πως τα δεδομένα προέρχονται από κατανομή Poisson με μέση τιμή 1. Χρησιμοποιήστε Monte Carlo για να εκτιμήσετε τη συνάρτηση πυκνότητας πιθανότητας της ελεγχοσυνάρτησης του ελέγχου, δηλαδή την

$$\chi^2 = \sum_{j=0}^{m} \frac{(O_j - E_j)^2}{E_j}$$

όπου m είναι η μεγαλύτερη παρατηρηθείσα τιμή και O_j και E_j είναι η παρατηρηθείσα συχνότητα της τιμής j και η αναμενόμενη αντίστοιχα. Για λόγους ευκολίας υποθέστε ότι το μέγεθος δείγματος είναι 100 και πως $m = 5$,

τιμές μεγαλύτερες από 5 ομαδοποιούνται στην τελευταία ομάδα. Συγκρίνετε την κατανομή που κατασκευάσατε με το Monte Carlo με την θεωρητική κατανομή της ελεγχοσυνάρτησης. Τι παρατηρείτε;

12. Ο έλεγχος Kruskal-Wallis θεωρείται ο μη παραμετρικός αντίστοιχος έλεγχος της απλής ανάλυσης διακύμανσης. Για τον έλεγχο αυτό δεν χρειάζεται να ξέρουμε την κατανομή του πληθυσμού, υποθέτουμε όμως πως είναι συμμετρική. Ο έλεγχος στηρίζεται στην έννοια των ranks και η κατανομή της ελεγχοσυνάρτησης είναι η ίδια ανεξάρτητα από την κατανομή του πληθυσμού. Υποθέστε πως έχετε 3 ομάδες με ιδιο μέγεθος δείγματος $n = 20$ που έχουν με ίδια μέση τιμή στον πληθυσμό $\mu = 2$.

    - Προσομοιώστε την κατανομή της ελεγχοσυνάρτησης όταν ο πληθυσμός ακολουθεί $pN(\mu, 4) + (1 - pN(\mu, 2)$ κατανομή, δηλαδή μια μείξη από δύο κανονικές κατανομές που έχει μέση τιμή 2.
    - Προσομοιώστε την κατανομή της ελεγχοσυνάρτησης όταν ο πληθυσμός ακολουθεί κατανομή Γάμμα με μέση τιμή 2 και διακύμανση 4.
    - Συγκρίνετε τις κατανομές που προέκυψαν με αυτή που χρησιμοποιούμε στην πράξη

13. ΄Ενα ενδιαφέρον ερώτημα σχετικά με την πολιτική είναι η ύπαρξη ή όχι γεωγραφικής αυτοσυσχέτισης στα εκλογικά αποτελέσματα, με άλλα λόγια αν υπάρχει κάποια συσχέτιση ανάμεσα στα αποτελέσματα γειτονικών περιοχών - νομών. Βρείτε τα αποτελέσματα των τελευταίων εκλογών από την ιστοσελίδα του Υπουργείου Εσωτερικών (www.ypes.gr) και επιλέξτε ένα κόμμα. Στη συνέχεια χρησιμοποιείστε ως ελεγχοσυναρτήσεις τους δυο παρακάτω συντελεστές χωρικής αυτοσυσχέτισης (spatial autocorrelation):

    Το δείκτη του Moran

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\left(\sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} \right) \sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2}$$

    Το δείκτη του Geary

$$G = \frac{(N-1) \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij}(X_i - X_j)^2}{2 \left(\sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} \right) \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

    όπου X_i είναι οι παρατηρήσεις που έχουμε δηλαδή το ποσοστό του κόμματος στο νομό i, N είναι το πλήθος των νομών και W_{ij} είναι τα στοιχεία του πίνακα γειτνίασης W, διαστάσεων $N \times N$, με στοιχεία που δείχνουν αν οι νομοί γειτονεύουν. Για την περίπτωση που μας αφορά τα στοιχεία W_{ij} έχουν την

τιμή 1 αν οι δύο νομοί έχουν κοινό σύνορο και 0 αν δεν έχουν. Κάντε έναν έλεγχο τυχαιοποίησης για τα δεδομένα σας για να διαπιστώσετε αν υπάρχει ή όχι γεωγραφική αυτοσυσχέτιση.

14. Οι πόντοι που σημειωσαν 4 ομάδες μπάσκετ στα παιχνίδια της χρονιάς είναι οι ακόλουθοι (οι ομάδες δεν έδωσαν τον ίδιο αριθμό παιχνιδιών)

    Α: 63,63,80,70,64,69,89,75,83,66,84,80,95,74,62,59,72,44,71,74,75, 87

    Β: 63,86,104,55,93,58,99,91,84,41,90,87,77,90,99,92,90,101,96,65

    Γ: 81,61,72,80,82,53,87,94,73,76,86,45,53,80,40,72,91

    Δ: 48,60,91,65,92,54,95,52,68,74,83,58,62,73,63,71,81,91,69

    Από αυτά τα δεδομένα ο εθνικός προπονητής πιστεύει πως οι ομάδες διαφέρουν στην ικανότητα τους στο σκοράρισμα. Κάντε έναν έλεγχο τυχαιοποιησης για να δείτε κάτά πόσο ο προπονητής έχει δίκιο ή όχι. Ποιά ελεγχοσυνάρτηση θα χρησιμοποιήσετε; Περιγράψτε όλα τα βήματα για να κάνετε τον έλεγχο.

15. Ο κύριος Χ θεωρεί πως είναι ειδικός στο να αναγνωρίζει διαφορετικές ποικιλίες κρασιών. Για αυτό το λόγο έγινε το εξής πείραμα. Σε 10 ποτήρια τοποθετήθηκαν 10 διαφορετικά κρασιά (για τα οποία όμως γνωριζε ποιά 10 είναι) και ο κύριος Χ κλήθηκε να αναγνωρίσει τα κρασιά αυτά. Αναγνώρισε σωστά 6 από τα 10. Ελέγξτε την υπόθεση ότι ο κύριος Χ είναι όντως ειδικός ή όχι Περιγράψτε με λεπτομέρεια τα βήματα της διαδικασίας που θα χρησιμοποιήσετε. Υλοποιήστε τον έλεγχο με αριθμό επαναλήψεων 10.

16. Ένας ερευνητής έχει δοκιμάσει 4 τεχνικές κοπής μετάλλου και θέλει να ελέγξει αν η μέθοδος Α (που είναι καινούρια) μειώνει στο μισό το χρόνο κοπής. Για κάθε μέθοδο έχει 10 παρατηρήσεις που τις συμβολίζουμε ως X_{ij} το i συμβολίζει την παρατήρηση και το j τη μέθοδο, $i = 1, \ldots, 10, \;\; j = 1, \ldots, 4$. Τι έλεγχο θα του προτείνατε να κάνει και γιατί; Ποιά ελεγχοσυνάρτηση να χρησιμοποιήσει; Περιγράψτε τα βήματα για τον έλεγχο.

Βιβλιογραφία

Noreen, E.W. (1989) *Computer Intensive Methods for Testing Hypotheses: An Introduction.* Wiley

Ιδανικό εισαγωγικό βιβλίο. που απευθύνεται σε αρχάριους. Περιέχει πλήθος παραδειγμάτων.

Hope, A.C.A (1968) A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society, Series B*, 30, 582-598

Στην πραγματικότητα αυτό το άρθρο εισήγαγε για πρώτη φορά συστηματικά την ιδέα των ελέγχων Monte Carlo, αρκετά νωρίς. Επειδή πρόκειται για άρθρο και όχι για βιβλίο είναι πυκνογραμμένο και κάπως δυσκολονόητο για αρχάριους.

Westfall, P.H. and Young, S.S. (1993) *Resampling-Based Multiple Testing*. Wiley

Αν και το θέμα του είναι πολύ πιο γενικό, περιγράφονται πολλές από τις ιδέες, όπως και επίσης νέα αποτελέσματα σχετικά με τα θέματα αυτά. Επίσης δίνονται παραδείγματα με τη χρήση του στατιστικού πακέτου SAS.

Good, P. (1998). *Resampling Methods: A Practical Guide to Data Analysis*. Birkhauser, Boston

Εισαγωγικό βιβλίο ακόμα και για αυτούς χωρίς ιδιαίτερες γνώσεις στατιστικής. Οι ιδέες εισάγονται απλοκά και δίνονται πολλά παραδείγματα εφαρμογής των μεθόδων

Κεφάλαιο 4

Η μέθοδος Jackknife

4.1 Εισαγωγή

Η μέθοδος jackknife αναπτύχθηκε αρκετά νωρίς το 1950 κυρίως ως μέθοδος περιορισμού της μεροληψίας μερικών εκτιμητριών. Στην πραγματικότητα η μέθοδος οδηγεί σε τυπικά σφάλματα εκτιμητριών τα οποία είναι σχετικά εύκολο να υπολογιστούν. Πιο συγκεκριμένα αν θέλουμε να εκτιμήσουμε μια μέση τιμή ενός πληθυσμού, αυτό είναι συνήθως αρκετά εύκολο καθώς η δειγματική μέση τιμή είναι αμερόληπτη εκτιμήτρια και η διακύμανση της είναι γνωστή. Για πιο σύνθετες εκτιμήτριες όμως, όπως για παράδειγμα ένας περικομμένος μέσος, ο θεωρητικός υπολογισμός του τυπικού σφάλματος είναι αρκετά πολύπλοκος και επομένως υπάρχει η ανάγκη για μια εναλλακτική μεθοδολογία. Η μέθοδος jackknife μπορεί να βοηθήσει σε αυτή την κατεύθυνση. Επομένως η μέθοδος jackknife προσφέρει

1. Εκτιμήτριες συναρτήσεις με μικρότερη μεροληψία σε απόλυτη τιμή

2. Εύκολο υπολογισμό του τυπικού σφάλματος της εκτιμήτριας.

Πριν προχωρήσουμε στην περιγραφή της μεθόδου ας δούμε λίγο το συμβολισμό που θα χρησιμοποιήσουμε. 'Οπως πάντα έχουμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ μεγέθους n και μια εκτιμήτρια συνάρτηση $\hat{\theta} = T(X_1, X_2, ..., X_n)$ της πραγματικής παραμέτρου θ του πληθυσμού η οποία είναι μια συνάρτηση του δείγματος (πχ η μέση τιμή, η διακύμανση, η διάμεσος κλπ). Επίσης συμβολίζουμε με $\hat{\theta}_{(i)} = T(X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n)$ την τιμή της εκτιμήτριας όταν όμως έχουμε αφαιρέσει την i παρατήρηση από το δείγμα μας.

Η βασική ιδέα για τη μέθοδο jackknife είναι η εξής: Αφαιρώντας παρατηρήσεις από το αρχικό δείγμα και εκτιμώντας ξανά την παράμετρο που μας ενδιαφέρει, μπορούμε να πάρουμε πληροφορία σχετικά με τη σταθερότητα, και άρα τη μεταβλητότητα, της εκτιμήτριας. Επομένως αν αφαιρούμε κάθε φορά μια παρατήρηση εξετάζοντας πόσο αλλάζουν οι τιμές της εκτιμήτριας παίρνουμε μια εικόνα σχετικά με τη διακύμανση της εκτιμήτριας.

4.2 Η εκτιμήτρια jackknife

Η jackknife εκτιμήτρια $\hat{\theta}_J$ της απλής εκτιμήτριας $\hat{\theta}$ υπολογίζεται ως

$$\hat{\theta}_J = n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\bullet)} \quad ,$$

όπου $\bar{\hat{\theta}}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$

Επομένως, βλέπουμε πως από τον ορισμό της jackknife εκτιμήτριας $\hat{\theta}_J$ χρειάζεται να επαναλάβουμε n φορές τον υπολογισμό της εκτιμήτριας όπου κάθε φορά το δείγμα μας θα είναι το αρχικό έχοντας όμως αφαιρέσει μια παρατήρηση κάθε φορά. Δηλαδή την πρώτη φορά αφαιρούμε την πρώτη παρατήρηση X_1 υπολογίζουμε την εκτιμήτρια $\hat{\theta}_{(1)}$, μετά την επιστρέφουμε στο δείγμα και αφαιρούμε τη δεύτερη και ούτω καθεξής μέχρι να αφαιρέσουμε και την τελευταία παρατήρηση. Είναι σαφές πως σε όλες τις περιπτώσεις το πλήθος των παρατηρήσεων είναι $n-1$.

Μπορεί επίσης να παρατηρήσει κανείς από τον ορισμό της $\hat{\theta}_J$ πως απλά διορθώνει την εκτιμήτρια από το συνολικό δείγμα με τη χρήση των εκτιμητριών από τα δείγματα όπου έχουμε αφαιρέσει μια τιμή.

Εναλλακτικά η $\hat{\theta}_J$ μπορεί να υπολογιστεί με τη χρήση των ψευδοτιμών

$$p_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$$

ως $\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n} p_i$.

Παράδειγμα 4.1: 'Εστω οι εξής 6 παρατηρήσεις που αφορούν το χρόνο αναμονής σε λεπτά σε μια στάση λεωφορείου: 4 ,3 ,7 ,6 ,5 ,9. Να υπολογιστούν οι εκτιμήτριες jackknife για τη μέση τιμή και τη διάμεσο του πληθυσμού.

Θα χρησιμοποιήσουμε τη δειγματική μέση τιμή $\bar{x}$ και τη δειγματική διάμεσο για να εκτιμήσουμε τις αντίστοιχες ποσότητες του πληθυσμού. Επομένως βρίσκουμε πως $\bar{x} = 5.666$ και $M = 5.5$. Οι υπολογισμοί για τις αντίστοιχες jackknife εκτιμήτριες φαίνονται στον πίνακα 4.1.

Παρατηρήσεις	Τιμή της εκτιμήτριας όταν αφαιρέσουμε την i παρατήρηση		Ψευδοτιμές	
	Μέση τιμή	Διάμεσος	Μέση τιμή	Διάμεσος
4	6	6	4	3
3	6.2	6	3	3
7	5.4	5	7	8
6	5.6	5	6	8
5	5.8	6	5	3
9	5	5	9	8
34	34	33	34	33

Πίνακας 4.1: Jackknife εκτίμηση της μέσης τιμής και της διαμέσου

'Ετσι παρατηρούμε πως αν αφαιρέσουμε την πρώτη παρατήρηση η μέση τιμή των 5 παρατηρήσεων που απομένουν είναι (3 + 7 + 6 +5 +9) / 5 =6 και ομοίως η

διάμεσος των παρατηρήσεων 3, 7, 6, 5, 9, είναι το 6. Με τον ίδιο τρόπο υπολογίζουμε και τις υπόλοιπες τιμές των στηλών 2 και 3. Οι στήλες 4 και 5 προκύπτουν από τον ορισμό των ψευδοτιμών. 'Ετσι η πρώτη ψευδοτιμή για την μέση τιμή είναι

$$p_1 = n\bar{x} - (n-1)\bar{x}_{(1)} = 6 \times 5.666 - 5 \times 6 = 4.$$

Κάνοντας λοιπόν τις πράξεις βρίσκουμε πως η jackknife εκτιμήτρια για τη μέση τιμή είναι 5.666 και για τη διάμεσο 5.5. Στην περίπτωση μας δηλαδή βλέπουμε πως συμπίπτουν με τις εκτιμήσεις που είχαμε. Συμβαίνει αυτό πάντα; Θα το δούμε στην επόμενη ενότητα.

4.3 Διάφορες jackknife εκτιμήτριες

4.3.1 Μέση τιμή

Είδαμε στο παράδειγμα πως η εκτιμήτρια jackknife για τη μέση τιμή συμπίπτει με το δειγματικό μέσο. Συμβαίνει πάντα αυτό;

Αν συμβολίσουμε $\hat{\theta} = \bar{x}$ βλέπουμε πως οι ψευδοτιμές μας υπολογίζονται ως

$$p_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)} = n\hat{\theta} - (n-1)\frac{\sum_{j=1}^{n} X_j - X_i}{n-1} =$$
$$= n\hat{\theta} - (n-1)\frac{n\hat{\theta} - X_i}{n-1} = X_i$$

Η εκτιμήτρια jackknife επομένως θα είναι $\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n} p_i = \bar{x}$ άρα η ίδια με την απλή εκτιμήτρια που είχαμε πριν. Επομένως δεν υπάρχει λόγος να χρησιμοποιήσουμε jackknife για να εκτιμήσουμε μια μέση τιμή (αυτό ήταν αναμενόμενο αφού για τη μέση τιμή ξέρουμε την αμερόληπτη εκτιμήτρια για την οποία ο υπολογισμός του τυπικού της σφάλματος είναι εύκολος).

Με όμοιο τρόπο μπορεί να δειχτεί πως για οποιαδήποτε εκτιμήτρια της μορφής $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ η jackknife εκτιμήτρια ταυτίζεται με την απλή εκτιμήτρια. Παρατηρείστε πως αν $h(X_i) = X_i$ παίρνουμε τη μέση τιμή και πως αυτός ο συμβολισμός περιέχει όλες οι ροπές του δείγματος και πλήθος άλλων στατιστικών συναρτήσεων.

4.3.2 Διάμεσος

Για τη διάμεσο μπορεί να δει κανείς πως αν το πλήθος των παρατηρήσεων είναι άρτιος αριθμός τότε η jackknife εκτιμήτρια ταυτίζεται με τη δειγματική διάμεσο αν όμως είναι περιττός αριθμός κάτι τέτοιο δεν ισχύει. Συγκεκριμένα για τη διάμεσο έχουμε:

Αν το μέγεθος του δείγματος n είναι άρτιο. Χωρίς να στερούμαστε γενικότητα ας υποθέσουμε πως οι παρατηρήσεις είναι σε αύξουσα σειρά. Τότε αφαιρώντας μια παρατήρηση, αν η παρατήρηση είναι μικρότερη από τη διάμεσο θα ισχύει $\hat{\theta}_{(i)} = X_{\frac{n}{2}+1}$ ενώ αν είναι μεγαλύτερη από τη διάμεσο θα είναι $\hat{\theta}_{(i)} = X_{n/2}$. Επομένως θα έχουμε

$$\hat{\theta}_{(i)} = \begin{cases} X_{n/2} & i \geq \frac{n}{2}+1 \\ X_{\frac{n}{2}+1} & i < \frac{n}{2}+1 \end{cases}$$

Επομένως $\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\frac{n}{2}\left(X_{n/2} + X_{\frac{n}{2}+1}\right) = \frac{1}{2}\left(X_{n/2} + X_{\frac{n}{2}+1}\right)$ το οποίο ταυτίζεται με τη δειγματική διάμεσο.

Αν τώρα το μέγεθος του δείγματος n είναι περιττό. Και πάλι υποθέτουμε πως οι παρατηρήσεις είναι σε αύξουσα σειρά. Τότε όταν αφαιρέσουμε μια παρατήρηση μικρότερη από τη διάμεσο του δείγματος θα έχουμε

$$\hat{\theta}_{(i)} = \frac{1}{2}\left(X_{\frac{n+1}{2}} + X_{\frac{n+1}{2}+1}\right)$$

ενώ αν αφαιρέσουμε από το δεύτερο μισό θα έχουμε

$$\hat{\theta}_{(i)} = \frac{1}{2}\left(X_{\frac{n+1}{2}} + X_{\frac{n+1}{2}-1}\right).$$

Αν τώρα αφαιρέσουμε την ίδια τη δειγματική διάμεσο θα πάρουμε

$$\hat{\theta}_{\left(\frac{n+1}{2}\right)} = \frac{1}{2}\left(X_{\frac{n+1}{2}+1} + X_{\frac{n+1}{2}-1}\right),$$

δηλαδή θα έχουμε

$$2\hat{\theta}_{(i)} = \begin{cases} X_{\frac{n+1}{2}} + X_{\frac{n+1}{2}+1} & i < \frac{n+1}{2} \\ X_{\frac{n+1}{2}+1} + X_{\frac{n+1}{2}-1} & i = \frac{n+1}{2} \\ X_{\frac{n+1}{2}} + X_{\frac{n+1}{2}-1} & i > \frac{n+1}{2} \end{cases}$$

και συνεπώς

$$\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{2n}\left(\frac{n-1}{2}+1\right)\left(X_{\frac{n+1}{2}+1} + X_{\frac{n+1}{2}-1}\right) + \frac{1}{2n}(n-1)X_{\frac{n+1}{2}}$$

και άρα η jackknife διάμεσος θα είναι

$$\hat{\theta}_J = \left[n - \frac{1}{2n}(n-1)^2\right]X_{\frac{n+1}{2}} - (n-1)\frac{1}{2n}\left(\frac{n-1}{2}+1\right)\left(X_{\frac{n+1}{2}+1} + X_{\frac{n+1}{2}-1}\right) \quad .$$

Επομένως διαφέρει από τη δειγματική διάμεσο και κάνει χρήση των 3 κεντρικών παρατηρήσεων και όχι μόνο της διαμέσου.

Για παράδειγμα αν έχουμε 15 παρατηρήσεις η jackknife διάμεσος θα είναι

$$\hat{\theta}_J = \frac{218X_8 - 56\left(X_9 + X_7\right)}{15},$$

όπου X_j είναι η j διατεταγμένη παρατήρηση.

4.3.3 Διακύμανση

Για τη διακύμανση μπορεί κανείς να δει το εξής ενδιαφέρον. Ξέρουμε πως αν ο πληθυσμός είναι κανονικός, η εκτιμήτρια μεγίστης πιθανοφάνειας της παραμέτρου σ^2 είναι η δειγματική διακύμανση $\hat{\theta} = s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ για την οποία ξέρουμε πως είναι μεροληπτική για κανονικούς πληθυσμούς. Μπορεί να αποδειχτεί πως :

Η jackknife εκτιμήτρια της διακύμανσης είναι

$$\hat{\theta}_J = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

που είναι αμερόληπτη.

Απόδειξη: Γενικά η δειγματική διακύμανση είναι γνωστή μέσω του τύπου

$$s^2 = \frac{\sum\limits_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n} = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$$

παρόλα αυτά αναπτύσσοντας το τετράγωνο της μέσης τιμής μπορεί κανείς να δείξει πως

$$s^2 = \frac{(n-1)\sum\limits_{i=1}^{n} x_i^2 - \sum\limits_{i=1}^{n}\sum\limits_{\substack{j=1\\ j\neq i}}^{n} x_i x_j}{n^2}$$

και επομένως

$$ns^2 = n\hat{\theta} = \frac{(n-1)\sum\limits_{i=1}^{n} x_i^2 - \sum\limits_{i=1}^{n}\sum\limits_{\substack{j=1\\ j\neq i}}^{n} x_i x_j}{n}$$

Αφαιρώντας την i παρατήρηση με παρόμοιο τρόπο μπορεί να δειχτεί πως

$$(n-1)\hat{\theta}_{(i)} = \frac{(n-2)\sum\limits_{\substack{j=1\\ j\neq i}}^{n} x_j^2 - \sum\limits_{\substack{k=1\\ k\neq i}}^{n}\sum\limits_{\substack{j=1\\ j\neq i}}^{n} x_k x_j}{(n-1)}$$

και άρα

$$\begin{aligned}
(n-1)\bar{\hat{\theta}}_{(\cdot)} &= \frac{(n-1)}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \\
&= \frac{(n-1)}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \\
&= \frac{(n-1)}{n}\sum_{i=1}^{n}\left[\frac{(n-2)\sum\limits_{\substack{j=1\\ j\neq i}}^{n} x_j^2 - \sum\limits_{\substack{k=1\\ k\neq i}}^{n}\sum\limits_{\substack{j=1\\ j\neq i}}^{n} x_k x_j}{(n-1)}\right]
\end{aligned}$$

'Ομως στο άθροισμα τετραγώνων κάθε παρατήρηση θα εμφανίζεται ακριβώς $(n-1)$ φορές ενώ στο γινόμενο κάθε όρος θα εμφανίζεται ακριβώς $(n-2)$ φορές και επομένως

$$\begin{aligned}(n-1)\bar{\hat{\theta}}_{(\cdot)} &= \frac{(n-1)}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \\ &= \frac{1}{n}\left[(n-2)\sum_{i=1}^{n}x_i^2 - \frac{n-2}{n-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}x_i x_j\right]\end{aligned}$$

Επομένως η jackknife εκτιμήτρια $\hat{\theta}_{jack}$ θα είναι

$$\begin{aligned}\hat{\theta}_{jack} &= n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\cdot)} \\ &= \frac{n-1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}x_i x_j \\ &\quad -\frac{1}{n}\left[(n-2)\sum_{i=1}^{n}x_i^2 - \frac{n-2}{n-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}x_i x_j\right] \\ &= \frac{(n-1)\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}x_i x_j}{n(n-1)} \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\end{aligned}$$

δηλαδή η αμερόληπτη εκτιμήτρια της διακύμανσης.

4.3.4 Ποσοστά

Ας δούμε λεπτομερώς ένα ακόμα παράδειγμα όπου η μέθοδος βελτιώνει τη μεροληψία.

'Εστω ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ δοκιμών Bernoulli. Δηλαδή κάθε X_i παίρνει την τιμή 1 με πιθανότητα p και 0 με πιθανότητα $1-p$. 'Εστω ότι θέλουμε να εκτιμήσουμε την ποσότητα p^2. Δείξτε ότι η εκτιμήτρια jackknife βελτιώνει τη μεροληψία της εκτιμήτριας μεγίστης πιθανοφάνειας.

Αν ορίσουμε ως $R = \sum_{i=1}^{n} X_i$ αυτό μετρά τον αριθμό των επιτυχιών (των μονάδων δηλαδή) στις n δοκιμές. Είναι γνωστό πως η εκτιμήτρια μεγίστης πιθανοφάνειας για το p είναι R/n. Επομένως από τις ιδιότητες της μεθόδου μεγίστης πιθανοφάνειας η εκτιμήτρια μεγίστης πιθανοφάνειας $\hat{\theta}$ της παραμέτρου $\theta = p^2$ θα είναι

$$\hat{\theta} = \left(\frac{R}{n}\right)^2.$$

Οπότε θα ισχύει

$$E(\hat{\theta}) = E\left[\left(\frac{R}{n}\right)^2\right] = \frac{1}{n^2}E(R^2)$$

και επειδή η τυχαία μεταβλητή R ακολουθεί τη διωνυμική κατανομή (ως άθροισμα Bernoulli) θα ισχύει

$$E(R) = np,\ Var(R) = np(1-p).$$

Επομένως

$$E(R^2) = Var(R) + [E(R)]^2 = np(1-p) + n^2p^2$$

και άρα

$$E(\hat{\theta}) = \frac{p(1-p)}{n} + p^2$$

κάτι το οποίο υποδηλώνει ότι η εκτιμήτρια είναι μεροληπτική.

Ας βρούμε τώρα την jackknife εκτιμήτρια. Θα ισχύει πως

$$\begin{aligned}\hat{\theta}_{(i)} &= \left(\frac{R - X_i}{n-1}\right)^2 \\ &= \frac{R^2 - 2RX_i + X_i^2}{(n-1)^2}\end{aligned}$$

και άρα

$$\begin{aligned}\bar{\hat{\theta}}_{(\bullet)} &= \frac{1}{n}\sum_{i=1}^{n}\frac{R^2 - 2RX_i + X_i^2}{(n-1)^2} \\ &= \frac{nR^2 - 2R\sum\limits_{i=1}^{n} X_i + \sum\limits_{i=1}^{n} X_i^2}{n(n-1)^2}.\end{aligned}$$

Επειδή τα X_i παίρνουν μόνο τιμές 0 και 1 θα ισχύει $R = \sum\limits_{i=1}^{n} X_i = \sum\limits_{i=1}^{n} X_i^2$ οπότε

$$\bar{\hat{\theta}}_{(\bullet)} = \frac{nR^2 - 2R^2 + R}{n(n-1)^2}.$$

Μπορούμε τώρα να υπολογίσουμε την jackknife εκτιμήτρια ως

$$\begin{aligned}\hat{\theta}_J &= n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\bullet)} \\ &= n\frac{R^2}{n^2} - (n-1)\frac{nR^2 - 2R^2 + R}{n(n-1)^2} \\ &= \frac{R^2}{n} - \frac{nR^2 - 2R^2 + R}{n(n-1)} \\ &= \frac{R^2(n-1) - nR^2 + 2R^2 - R}{n(n-1)} \\ &= \frac{R^2 - R}{n(n-1)} = \frac{R(R-1)}{n(n-1)}\end{aligned}$$

Τώρα η αναμενόμενη τιμή αυτής της εκτιμήτριας είναι

$$\begin{aligned} E(\hat{\theta}_J) &= E\left[\frac{R(R-1)}{n(n-1)}\right] \\ &= \frac{1}{n(n-1)} E\left[R^2 - R\right] \\ &= \frac{1}{n(n-1)}\left[E(R^2) - E(R)\right] \\ &= \frac{np(1-p) + p^2n^2 - np}{n(n-1)} \\ &= \frac{n(n-1)p^2}{n(n-1)} = p^2 \end{aligned}$$

επομένως η jackknife εκτιμήτρια είναι αμερόληπτη.

4.4 Εκτίμηση τυπικών σφαλμάτων και μεροληψίας

Από τον ορισμό της εκτιμήτριας jackknife με τη χρήση των ψευδοτιμών προκύπτει πως

$$Var(\hat{\theta}_J) = \frac{\sum_{i=1}^{n} Var(p_i)}{n^2} = \frac{Var(p_1)}{n},$$

αλλά παρατηρείστε πως τα p_i έχουν όλα την ίδια διακύμανση για αυτό μπορούμε να χρησιμοποιούμε, έστω, την τιμή p_1.

Μια αμερόληπτη εκτίμηση για τη διακύμανση των ψευδοτιμών είναι η

$$s_p^2 = \frac{1}{n-1}\sum_{i=1}^{n}(p_i - \bar{p})^2$$

και επομένως μπορούμε να εκτιμήσουμε τη διακύμανση της jackknife εκτιμήτριας ως

$$\begin{aligned} s_{\hat{\theta}_j}^2 &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(p_i - \bar{p})^2 \\ &= \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_{(\bullet)})^2 \end{aligned}$$

δηλαδή είναι η διακύμανση των τιμών της εκτιμήτριας όταν αφαιρέσουμε μια παρατήρηση πολλαπλασιασμένη με $n-1$. Αυτός ο πολλαπλασιαστής στην πραγματικότητα 'μεγαλώνει' τη διακύμανση επειδή τα δείγματα που παίρνουμε όταν αφαιρούμε μόνο μια παρατήρηση είναι πολύ όμοια και άρα για να φέρουμε την εκτίμηση της διακύμανσης σε σωστό μέγεθος πρέπει να μεγαλώσουμε την ποσότητα αυτή. Επομένως είδαμε πως μπορούμε πολύ εύκολα να υπολογίσουμε το τυπικό σφάλμα της εκτιμήτριας jackknife.

Είδαμε προηγουμένως πως για μια μεγάλη ομάδα εκτιμητριών η εκτιμήτρια jackknife ταυτίζεται με την απλή εκτιμήτρια και επομένως μπορούμε να εκτιμήσουμε το τυπικό σφάλμα της απλής εκτιμήτριας χρησιμοποιώντας αυτό της jackknife που είναι πιο εύκολο.

Επίσης μπορούμε να εκτιμήσουμε τη μεροληψία της εκτιμήτριας $\hat{\theta}$ ως

$$bias(\hat{\theta}) = (n-1)(\bar{\hat{\theta}}_{(\bullet)} - \hat{\theta})$$

'Εχοντας υπολογίσει την τυπική απόκλιση της εκτιμήτριας μπορούμε να κατασκευάσουμε προσεγγιστικά διαστήματα εμπιστοσύνης για την εκτιμήτρια χρησιμοποιώντας ασυμπτωτική κανονικότητα (η οποία είναι μάλλον μια ρεαλιστική υπόθεση αν η εκτιμήτρια έχει τη μορφή μιας μέσης τιμής όπως αυτής που είδαμε πριν, ή η εκτιμήτρια είναι εκτιμήτρια μεγίστης πιθανοφάνειας και σε πολλές άλλες περιπτώσεις).

Επίσης είναι ενδιαφέρον να δούμε πως η διακύμανση της jackknife εκτιμήτριας θα είναι περίπου ίση με τη διακύμανση της αρχικής εκτιμήτριας. Αυτό μπορεί να δειχτεί καθώς από τον ορισμό

$$Var(\hat{\theta}_J) = n^2 Var(\hat{\theta}) + (n-1)^2 Var(\bar{\hat{\theta}}_{(\bullet)}) - 2n(n-1)Cov(\hat{\theta}, \bar{\hat{\theta}}_{(\bullet)})$$

'Ομως περιμένουμε πως η διακύμανση των $\hat{\theta}$ και $\bar{\hat{\theta}}_{(\bullet)}$ να είναι περίπου ίδια, ενώ η συνδιακύμανση τους θα είναι περίπου ίδια με την κοινή τους διακύμανση. Επομένως προκύπτει πως

$$Var(\hat{\theta}_J) \approx Var(\hat{\theta})$$

Συνεπώς η εκτιμήτρια της διακύμανσης της jackknife εκτιμήτριας μπορεί να χρησιμοποιηθεί και ως εκτιμήτρια της διακύμανσης της αρχικής εκτιμήτριας .

Παράδειγμα 4.1 (συνέχεια).

Για το παράδειγμα μας μπορούμε να υπολογίσουμε με τη χρήση των αποτελεσμάτων του πίνακα πως $s^2_{\bar{x}} = 0.815$ και $s^2_M = 1.25$. Βλέπουμε πως η διακύμανση της εκτιμήτριας της διαμέσου είναι αρκετά μεγαλύτερη.

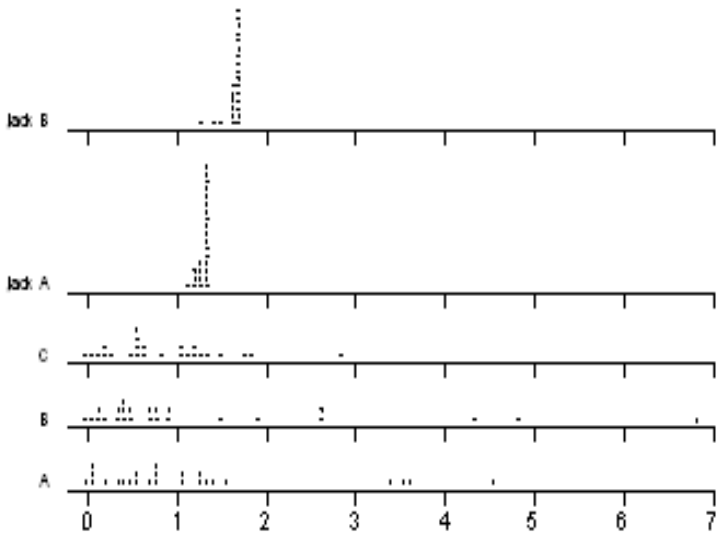
Παράδειγμα 4.2

Ας δούμε ένα παράδειγμα που η χρήση ασυμπτωτικών σφαλμάτων είναι ουσιαστικά αδύνατη. Ο συντελεστής Gini, χρησιμοποιείται από τους οικονομολόγους για να μετρήσουν την ανισοκατανομή του εισοδήματος σε έναν πληθυσμό. Ο συντελεστής δίνεται από τον τύπο

$$G = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$$

'Οπως μπορεί να δει κανείς η μορφή του συντελεστή είναι αρκετά πολύπλοκη για να μπορέσει κανείς να υπολογίσει εύκολα, με τη χρήση τύπων, την τυπική απόκλιση του δειγματικού συντελεστή. Για αυτό το λόγο η μέθοδος jackknife μοιάζει να είναι κάποια λύση. Τα δεδομένα που ακολουθούν στον πίνακα 4.2 αφορούν τυχαία δείγματα από τρία διαφορετικά χωριά και τα εισοδήματα τους και ενδιαφερόμαστε να δούμε αν έχουν διαφορετικό συντελεστή Gini και άρα διαφορετική κατανομή εισοδήματος.

Στον πίνακα 4.2 μπορεί κανείς να δει τις τιμές για κάθε δείγμα καθώς επίσης και τις ψευδοτιμές που προκύπτουν αν αφαιρέσουμε τη συγκεκριμένη παρατήρηση. Ενδιαφέρον είναι να δει κανείς και το γράφημα 4.1 όπου υπάρχουν τα διαγράμματα σημείων για τα 3 δείγματα αλλά και διάγραμμα σημείων για τις ψευδοτιμές των δύο πρώτων δειγμάτων. Παρατηρείστε πως οι ψευδοτιμές είναι σχετικά παρόμοιες και επομένως απλά η διακύμανση τους θα μας έδινε μια υποεκτίμηση της πραγματικής διακύμανσης. Αυτό διορθώνει ο όρος $(n-1)$. Στον πίνακα 4.3 μπορεί κανείς να δει την εκτίμηση του συντελεστή από το δείγμα, καθώς και το μέσο των ψευδοτιμών. Καθώς αυτές είναι ίσες συμπεραίνουμε πως η jackknife εκτιμήτρια της μεροληψίας είναι 0, δηλαδή ο δειγματικός συντελεστής Gini είναι αμερόληπτος. Αυτό ήταν κάτι αναμενόμενο καθώς ο συντελεστής Gini είναι μια γραμμική συνάρτηση των δεδομένων. Τέλος, χρησιμοποιώντας την ασυμπτωτική κανονικότητα (που εδώ δεν είναι και τόσο λογική λόγω του μικρού δείγματος) κατασκευάσαμε 95% διαστήματα εμπιστοσύνης για το συντελεστή Gini του πληθυσμού. Παρατηρείστε πως καθώς τα διαστήματα επικαλύπτονται μοιάζει λογικό να υποστηρίξει κανείς πως δεν διαφέρουν οι συντελεστές των 3 πληθυσμών. Βέβαια το δείγμα είναι αρκετά μικρό για πιο αξιόπιστα αποτελέσματα.



Γράφημα 4.1: Διαγράμματα σημείων για τα 3 χωριά καθώς και οι ψευδοτιμές για τα χωριά Α και Β.

				jackknife values		
Χωριό	A	B	C	A	B	C
1	0.38	0.10	1.28	1.309	1.651	0.739
2	0.05	0.88	1.86	1.290	1.667	0.700
3	0.68	0.42	0.83	1.317	1.668	0.749
4	1.06	0.18	0.17	1.315	1.656	0.731
5	0.55	2.64	1.34	1.315	1.582	0.735
6	0.03	0.31	0.44	1.288	1.663	0.745
7	1.23	6.82	0.52	1.311	1.262	0.748
8	1.43	0.70	1.76	1.302	1.670	0.707
9	0.52	0.40	0.26	1.314	1.668	0.737
10	0.77	0.49	0.55	1.318	1.670	0.749
11	3.42	1.86	1.45	1.181	1.625	0.728
12	0.01	0.77	1.15	1.287	1.669	0.744
13	1.56	0.01	0.57	1.295	1.644	0.750
14	1.07	1.49	0.07	1.315	1.642	0.724
15	3.60	0.01	2.80	1.167	1.645	0.622
16	1.30	0.13	1.06	1.308	1.654	0.747
17	1.28	2.58	1.18	1.309	1.586	0.744
18	0.00	0.43	0.59	1.286	1.668	0.750
19	3.52	0.35	1.04	1.174	1.666	0.747
20	0.18	4.83	0.14	1.298	1.428	0.729
21	4.53	0.50	0.00	1.090	1.670	0.718
22	0.78	4.36	0.53	1.318	1.464	0.749
23	0.32	0.78	1.18	1.306	1.669	0.743
24	0.76	0.67	0.22	1.318	1.670	0.734
25	0.51	0.92	0.63	1.314	1.665	0.750

Πίνακας 4.2: Εισοδήματα για ένα τυχαίο δείγμα 25 νοικοκυριών σε 3 διαφορετικά χωριά και οι ψευδοτιμές για κάθε χωριό

	Χωριά		
	A	B	Γ
Δειγματική εκτίμηση	1.281	1.620	0.732
Μέσος ψευδοτιμών	1.281	1.620	0.732
Εκτίμηση τυπικού σφάλματος	0.2888	0.4664	0.1288
Κάτω όριο 95% διαστήματος εμπιστοσύνης	0.715	0.706	0.479
’Ανω όριο 95% διαστήματος εμπιστοσύνης	1.847	2.534	0.984

Πίνακας 4.3: Αποτελέσματα χρησιμοποίησης της μεθόδου Jackknife στα δεδομένα των εισοδημάτων

4.5 Εφαρμογές

Η μέθοδος jackknife βρίσκει πολλές εφαρμογές στη δειγματοληψία ως μια μέθοδος εκτίμησης τυπικών σφαλμάτων σε πολύπλοκα δειγματοληπτικά σχέδια αλλά και σε πολύπλοκες ποσότητες. Δύο τέτοια παραδείγματα είναι τα εξής

- Σε πολλές δειγματοληπτικές έρευνες το ενδιαφέρον εστιάζεται στην εκτίμηση του λόγου 2 μέσων τιμών. Αν $\bar{X}$ και $\bar{Y}$ οι δύο μέσες τιμές που μας ενδιαφέρουν τότε μας ενδιαφέρει η εκτιμήτρια λόγου $Z = \bar{X}/\bar{Y}$ για να εκτιμήσουμε το λόγο των μέσων τιμών στον πληθυσμό. Είναι γνωστό πως από την ανισότητα Jensen η εκτιμήτρια αυτή είναι μεροληπτική. Χρησιμοποιώντας jackknife εκτιμήτριες καταφέρνουμε όχι μόνο να μειώσουμε τη μεροληψία αλλά να πάρουμε και μια εκτίμηση της διακύμανσης της ποσότητας που μας ενδιαφέρει κάτι που δεν είναι εύκολο με τη χρήση ασυμπτωτικών αποτελεσμάτων.

- Σε πολύπλοκα δειγματοληπτικά σχέδια που αφορούν συνήθως πολυσταδιακή δειγματοληψία δεν είναι καθόλου εύκολο να εκτιμήσουμε τα τυπικά σφάλματα των εκτιμητριών που μας ενδιαφέρουν. Η μέθοδος jackknife είναι πολύ χρήσιμη σε αυτές τις περιπτώσεις καθώς μπορεί να λάβει υπόψη της και στοιχεία από το δειγματοληπτικό σχέδιο. Για παράδειγμα αν το σχέδιο αφορά δειγματοληψία σε ομάδες (cluster sampling) μπορούμε να κάνουμε jackknife αφαιρώντας κάθε φορά μια ομάδα. Έτσι ακόμα και σε ιδιαίτερα πολύπλοκα δειγματοληπτικά σχέδια (τα οποία όμως είναι αρκετά διαδεδομένα σε μεγάλες έρευνες στην πράξη) η jackknife αποτελεί πολύτιμο εργαλείο.

4.6 Συμπεράσματα

Ολοκληρώνοντας λοιπόν αυτή τη σύντομη παρουσίαση θα πρέπει να τονίσουμε τα εξής:

- Οι εκτιμήτριες jackknife μειώνουν την μεροληψία σε σχέση με τις απλές εκτιμήτριες και μπορούμε σχετικά εύκολα να υπολογίσουμε τα τυπικά τους σφάλματα.

- Αν η απλή εκτιμήτρια έχει τη μορφή μιας γραμμικής συνάρτησης των δεδομένων, τότε η jackknife εκτιμήτρια ταυτίζεται με αυτή και άρα μπορούμε να βρούμε μια εκτίμηση του τυπικού της σφάλματος απλά υπολογίζοντας το τυπικό σφάλμα της jackknife εκτιμήτριας

- Αν η μορφή της εκτιμήτριας όμως είναι διαφορετική (π.χ. διάμεσος, μέγιστη τιμή) τότε η μέθοδος jackknife δεν είναι ικανοποιητική και πρέπει να χρησιμοποιείται με προσοχή. Προσέξτε στο παράδειγμα μας για τη διάμεσο ότι επειδή αφαιρούμε μόνο μια παρατήρηση κάθε φορά οι ψευδοτιμές που πήραμε μπορούσαν να πάρουν μόλις 2 διαφορετικές τιμές. Επομένως η εκτίμηση του τυπικού σφάλματος είναι μάλλον απλοϊκή και υπερεκτιμημένη.

  Σε αυτές τις περιπτώσεις μπορούμε να χρησιμοποιήσουμε μια γενίκευση της μεθόδου όπου αφαιρούμε όχι μόνο μια παρατήρηση τη φορά αλλά περισσότερες. Έχει αποδειχθεί θεωρητικά πως αυτό μπορεί να διορθώσει το πρόβλημα σε αρκετές περιπτώσεις.

- Η μέθοδος jackknife είναι ξεκάθαρα μια μη παραμετρική μέθοδος που δεν βασίζεται σε καμιά παραμετρική υπόθεση σχετικά με τον πληθυσμό. Επομένως τα αποτελέσματα της δεν εξαρτώνται από καμιά υπόθεση. Για παράδειγμα, στο παράδειγμα με το συντελεστή Gini και τα εισοδήματα, δεν χρειαστήκαμε καμιά υπόθεση για την κατανομή του πληθυσμού και άρα τα αποτελέσματα είναι ανεξάρτητα της κατανομής του πληθυσμού (πχ αν αυτή είναι Pareto όπως συνήθως υποθέτουμε για εισοδήματα ή εκθετική)

Από τα παραπάνω προκύπτει πως η μέθοδος jackknife αποτελεί έναν τρόπο να εκτιμήσουμε τη μεροληψία και την τυπική απόκλιση μιας εκτιμήτριας για την οποία κλασσικές μέθοδοι βασισμένες σε ασυμπτωτικά αποτελέσματα δεν είναι εύκολο να χρησιμοποιηθούν. Στο επόμενο κεφάλαιο θα δούμε μια άλλη μέθοδο, που ουσιαστικά επεκτείνει τη μέθοδο jackknife, τη μέθοδο bootstrap με την οποία μπορούμε και πάλι να εκτιμήσουμε τυπικά σφάλματα και τη μεροληψία. Η μέθοδος bootstrap μπορεί να μας δώσει τις απαιτούμενες εκτιμήσεις σε πολλά προβλήματα όπου μας ενδιαφέρει αλλά έχει κάποια άλλα μειονεκτήματα. Μια πιο λεπτομερής σύγκριση θα γίνει λοιπόν αργότερα

4.7 Cross-Validation

4.7.1 Εισαγωγή

Η μέθοδος Cross-validation [1] είναι μια μέθοδος για να εξετάσουμε την καλή προσαρμογή ενός στατιστικού μοντέλου στα δεδομένα μας. Ως στατιστικό μοντέλο εννοούμε μια ποικιλία από διαφορετικές στατιστικές μεθόδους. 'Ετσι, η μέθοδος cross-validation έχει χρησιμοποιηθεί σε μια πληθώρα εφαρμογών που μπορούν να ειδωθούν σε αυτό το πρίσμα. Επεκτείνοντας λίγο την ιδέα, αφού μπορούμε να εξετάσουμε την καλή προσαρμογή ενός μοντέλου στα δεδομένα μπορούμε με την ίδια λογική να συγκρίνουμε και διαφορετικά μοντέλα, επομένως η μέθοδος cross-validation είναι μια τεχνική που μας επιτρέπει να διαλέξουμε ανάμεσα σε μοντέλα. 'Οπως θα δούμε σε λίγο αυτή η ιδιότητά της είναι ιδιαίτερα χρήσιμη σε πολλές περιπτώσεις.

Πρέπει να αναφερθεί πως η μέθοδος Cross-validation δεν είναι μέθοδος για να εξετάσουμε την ποιότητα εκτιμητριών όπως η μέθοδος jackknife που μόλις αναφέρθηκε και η μέθοδος bootstrap που θα συζητήσουμε σε λίγο. Είδαμε προηγούμενα μια εφαρμογή της μεθόδου στην εκτίμηση πυκνότητας πιθανότητας με kernels. Στο παρόν κεφάλαιο θα δούμε πιο αναλυτικά τη μεθοδολογία της.

4.7.2 Περιγραφή της μεθόδου

Μια πρώτη προσέγγιση για να δούμε την καλή προσαρμογή ενός μοντέλου είναι να προσαρμόσουμε το μοντέλο στα δεδομένα μας και μετά χρησιμοποιώντας κάποιο κατάλληλο μέτρο να εξετάσουμε την καλή προσαρμογή τους. Κλασσικό τέτοιο παράδειγμα είναι ο συντελεστής προσδιορισμού στη γραμμική παλινδρόμηση. Η προσέγγιση αυτή έχει ένα μεγάλο μειονέκτημα. Χρησιμοποιούμε τα ίδια δεδομένα

[1] Μια ελληνική μετάφραση της λέξης cross-validation είναι διεπικύρωση. Στις σημειώσεις αυτές θα χρησιμοποιήσουμε μόνο την αγλική ορολογία

δύο φορές: για να προσαρμόσουμε το μοντέλο αλλά και για να δούμε πόσο καλό είναι. Επομένως, είναι αναμενόμενο πως κάποια ακραία φαινόμενα, δηλαδή παρατηρήσεις ολότελα ασύμφωνες με το μοντέλο, θα έχουν εξαλειφθεί καθώς όλα τα δεδομένα συμμετέχουν στην εκτίμηση του μοντέλου. Για παράδειγμα, στην απλή γραμμική παλινδρόμηση, αν στα δεδομένα μας υπάρχει μια ακραία παρατήρηση τότε επειδή αυτή λαμβάνεται υπόψη για την εκτίμηση του μοντέλου, το μοντέλο που προσαρμόσαμε θα έχει προσανατολιστεί και προς αυτή την παρατήρηση και επομένως δεν θα μας δείξει την ποιότητα του μοντέλου πλήρως. Το φαινόμενο αυτό ονομάζεται overfitting καθώς το μέτρο που χρησιμοποιούμε για να μετρήσουμε την ποιότητα του μοντέλου στηρίζεται σε όλες τις παρατηρήσεις και επομένως έχει "βελτιστοποιηθεί" ως προς τα δεδομένα.

Για να το αποφύγουμε αυτό θα μπορούσαμε να αφήσουμε κάποιες παρατηρήσεις έξω κατά τη διάρκεια της εκτίμησης και να τις χρησιμοποιήσουμε αργότερα μόνο για να δούμε αν το μοντέλο είναι καλό. Αυτή η προσέγγιση έχει το μειονέκτημα πως δεν χρησιμοποιεί όλα τα δεδομένα και επίσης μπορεί το αποτέλεσμα σχετικά με την ποιότητα του μοντέλου να εξαρτάται από το ποιες παρατηρήσεις θα χρησιμοποιήσουμε για ποιόν σκοπό, δηλαδή ποιες θα κρατήσουμε μόνο για την εκτίμηση του μοντέλου και ποιες για τον έλεγχο της προσαρμοστικότητας του μοντέλου.

Μια εναλλακτική προσέγγιση προσφέρει η μέθοδος Cross-validation. Με αυτή τη μέθοδο αφήνουμε έξω μια παρατήρηση κάθε φορά και στη συνέχεια κάνουμε πρόβλεψη με βάση το μοντέλο που προσαρμόσαμε για αυτή την παρατήρηση που αφήσαμε έξω. Στη συνέχεια επαναλαμβάνουμε αυτή τη διαδικασία αφήνοντας κάθε φορά μια παρατήρηση έξω μέχρι να τις έχουμε αφήσει όλες ανά μία έξω από το δείγμα. Με αυτό τον τρόπο έχουμε πρόβλεψη για όλες τις παρατηρήσεις από μοντέλα που δεν τις χρησιμοποίησαν και άρα μπορούμε να έχουμε ένα σκορ που να μας δείχνει την καλή προσαρμογή του μοντέλου. Στην πράξη η μέθοδος cross-validation μοιάζει πολύ με τη μέθοδο jackknife, αλλά διαφέρει στο ότι η παρατήρηση που δεν χρησιμοποιήθηκε για την εκτίμηση του μοντέλου, χρησιμοποιείται τελικά για τον έλεγχο του μοντέλου.

Η εφαρμογή της μεθόδου cross-validation στη γραμμική παλινδρόμηση ίσως βοηθήσει στην κατανόηση της χρησιμότητας της.

4.7.3 Cross Validation στη γραμμική παλινδρόμηση

Έστω ότι έχουμε παρατηρήσεις (X_i, Y_i), $i = 1, 2, ..., n$ και θέλουμε να προσαρμόσουμε ένα μοντέλο με τη μεταβλητή Y ως εξαρτημένη και κάποια συνάρτηση της μεταβλητής X ως ανεξάρτητη (η ιδέα μπορεί να γενικευτεί στην περίπτωση πολλών διαφορετικών επεξηγηματικών μεταβλητών). Ας υποθέσουμε επίσης πως θέλουμε να εξετάσουμε ποιο από δύο μοντέλα έχει καλύτερη προσαρμογή. Τα δύο μοντέλα είναι

$$\begin{aligned} Y_i &= a + \beta X_i + \varepsilon_i &&= h(x_i) + \varepsilon_i \\ Y_i &= a' + \beta' X_i + \gamma X_i^2 + \varepsilon_i' &&= g(x_i) + \varepsilon_i' \end{aligned}$$

δηλαδή ένα απλό γραμμικό μοντέλο και ένα πολυώνυμο δευτέρου βαθμού.

Η ιδέα είναι να αφήσουμε μια παρατήρηση έξω κάθε φορά. να εκτιμήσουμε τις παραμέτρους κάθε μοντέλου χρησιμοποιώντας μόνο τις υπόλοιπες παρατηρήσεις,

και στη συνέχεια, με βάση τις εκτιμήσεις που έχουμε πάρει για τις παραμέτρους του μοντέλου, να προβλέψουμε την τιμή που είχαμε αφήσει έξω. Αν συμβολίσουμε με $\hat{g}_{-i}(x_i)$και $\hat{h}_{-i}(x_i)$ την πρόβλεψη για την τιμή x_i για κάθε μοντέλο όταν το μοντέλο έχει εκτιμηθεί χωρίς αυτή την παρατήρηση τότε οι τιμές $y_i - \hat{g}_{-i}(x_i)$ και $y_i - \hat{h}_{-i}(x_i)$ είναι τα σφάλματα της πρόβλεψης για κάθε μοντέλο (error predictions). Στη γραμμική παλινδρόμηση τα κατάλοιπα αυτά ονομάζονται studentized ή deleted κατάλοιπα και χρησιμοποιούνται γιατί μπορούν να δώσουν μια πιο καλή εικόνα για το ποιες παρατηρήσεις έχουν μεγάλα κατάλοιπα από ότι τα συνηθισμένα κατάλοιπα.

Για να βρούμε αν το μοντέλο είναι καλό μπορούμε να χρησιμοποιήσουμε τη συνάρτηση

$$CV(g) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{g}_{-i}(x_i)\right)^2$$

και αντίστοιχα για το μοντέλο h. Πρέπει να αναφερθεί πως η ποσότητα $nCV(g)$ αναφέρεται ως άθροισμα τετραγώνων του προβλεπτικού σφάλματος (prediction error sum of squares, PRESS) και προσφέρεται σε πολλά από τα υπάρχοντα στατιστικά πακέτα. Το παραπάνω κριτήριο είναι ένα άθροισμα τετραγώνων των studentized καταλοίπων.

Συγκρίνοντας τα $CV(g)$ και $CV(h)$ μπορεί κανείς να δει ποιο μοντέλο είναι καλύτερο. 'Ετσι αν $CV(g) < CV(h)$ τότε το μοντέλο g είναι καλύτερο από το h καθώς έχει μικρότερο PRESS.

Αν θέλουμε να συγκρίνουμε το PRESS με το γνωστό και κλασσικό άθροισμα τετραγωνικών καταλοίπων $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$πρέπει να πούμε τα εξής:

- Ξεκάθαρα η μέθοδος Cross-validation απαιτεί πολλούς υπολογισμούς, αν και υπάρχουν σε πολλές περιπτώσεις ικανοποιητικοί αλγόριθμοι για να μειώσουν τον όγκο των υπολογισμών που απαιτούνται. Θα δούμε πως στην περίπτωση της γραμμικής παλινδρόμησης ο υπολογισμός του κριτηρίου μπορεί να γίνει εξαιρετικά απλά.

- Η μέθοδος Cross-validation δεν επηρεάζεται από ακραίες τιμές με μεγάλη επίδραση (influential values) που ουσιαστικά δεν οφείλονται στο μοντέλο και οι οποίες όμως μπορούν να καθορίζουν τις εκτιμήτριες του μοντέλου. Συνεπώς το κριτήριο είναι απαλλαγμένο από τέτοιες επιδράσεις και άρα μπορεί να 'δει' καλύτερα την καλή προσαρμογή ή όχι του μοντέλου.

- Αν τα μοντέλα είναι φωλιασμένα (nested) τότε γνωρίζουμε πολύ καλά πως το άθροισμα τετραγωνικών καταλοίπων θα μειώνεται κάθε φορά που προσθέτουμε μια νέα μεταβλητή άσχετα πόσο καλή είναι αυτή. Αυτό δεν συμβαίνει με το PRESS και επομένως το κριτήριο μπορεί να χρησιμοποιηθεί για την σύγκριση φωλιασμένων μοντέλων.

- Είναι επίσης ενδιαφέρον να παρατηρήσει κανείς πως μπορούμε να χρησιμοποιήσουμε διάφορες άλλες προσεγγίσεις για να μετρήσουμε την καλή προσαρμογή του μοντέλου όπως κριτήρια της μορφής $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{g}_{-i}(x_i)|$, αντί

δηλαδή να υψώσουμε στο τετράγωνο χρησιμοποιούμε απόλυτες τιμές. Γενικά η μέθοδος Cross-validation μας προσφέρει την δυνατότητα να δημιουργήσουμε διάφορα κριτήρια που μας επιτρέπουν να εξετάσουμε την καλή προσαρμογή του μοντέλου μας

- Τέλος παρατηρείστε πως το κριτήριο PRESS στην ουσία εκτός από την καλή προσαρμογή του μοντέλου μας εξετάζει και την προβλεπτική του ικανότητα, μια άλλη παράμετρος ενός γραμμικού μοντέλου που αν και πολλές φορές είναι το κλειδί για την επιτυχία ενός μοντέλου, συνήθως το παραβλέπουμε λόγω έλλειψης εργαλείων για να δούμε και αυτή την πλευρά του γραμμικού μοντέλου. Το απλό άθροισμα τετραγωνικών καταλοίπων δεν είναι σε θέση να κάνει κάτι τέτοιο καθώς η μέθοδος ελάχιστων τετραγώνων είναι εξ ορισμού τέτοια ώστε να κάνει το άθροισμα αυτό ελάχιστο και επομένως ενδιαφέρεται μόνο για την καλή προσαρμογή του μοντέλου και όχι για την προβλεπτική του ικανότητα

Αξίζει να αναφερθεί πως στην περίπτωση του γραμμικού μοντέλου που εξετάσαμε, η μέθοδος μπορεί να γίνει σχετικά εύκολα και οι υπολογισμοί να μειωθούν στο ελάχιστο ως εξής: Γνωρίζουμε πως γενικά το γραμμικό μοντέλο με τη χρήση πινάκων μπορεί να γραφτεί ως εξής

$$\mathbf{Y} = \mathbf{X}\beta$$

όπου $\mathbf{Y}$ είναι ένα $n \times 1$ διάνυσμα με τις τιμές της εξαρτημένης μεταβλητής, $\mathbf{X}$ είναι ένας $n \times (p+1)$ πίνακας σχεδιασμού με τις p ανεξάρτητες μεταβλητές (και συνήθως την πρώτη στήλη όλο μονάδες για να αναπαριστούν τη σταθερά) και β είναι ένα $(p+1) \times 1$ διάνυσμα με τους συντελεστές.

'Οπως είναι γνωστό η εκτιμήτρια ελαχίστων τετραγώνων είναι η

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

και επομένως το διάνυσμα με τις προβλέψεις $\hat{\mathbf{Y}}$ δίνεται από το

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}\mathbf{Y}$$

όπου $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ είναι ένας πίνακας που ονομάζεται hat matrix και έχει μερικές σημαντικές ιδιότητες. Αν συμβολίσουμε με a_{ij} το ij στοιχείο αυτού του πίνακα (ο οποίος εύκολα μπορεί κανείς να δει πως έχει διαστάσεις $n \times n$) τότε αποδεικνύεται εύκολα η σχέση

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1 - a_{ii}}$$

και επομένως δεν χρειάζεται κανείς να προσαρμόσει παρά μόνο μια φορά το μοντέλο και να αποθηκεύσει κάπου τον hat matrix. Οι τιμές a_{ii} συνήθως αναφέρονται στη βιβλιογραφία ως leverages και χρησιμοποιούνται για να μετρήσουμε αν μια παρατήρηση μπορεί να θεωρηθεί πως έχει μεγάλη επίδραση στην εκτίμηση του μοντέλου. Πρέπει να αναφερθεί πως ο hat matrix εμφανίζεται με όμοιο τρόπο και στην περίπτωση μη παραμετρικής παλινδρόμησης είτε με τη μέθοδο των kernels

που συζητήσαμε είτε με τη χρήση splines και επομένως και σε αυτές τις περιπτώσεις μπορεί κανείς να απλοποιήσει τους απαιτούμενους υπολογισμούς για cross-validation κριτήρια.

Παράδειγμα 4.3: Ας δούμε ένα παράδειγμα χρήσης της μεθόδου. Τα δεδομένα που υπάρχουν στον πίνακα 4.4 αφορούν 18 παρατηρήσεις όπου για μια συγκεκριμένη σοδιά καλαμποκιού χρησιμοποιήθηκαν λιπάσματα με φώσφορο τόσο σε οργανική όσο και σε ανόργανη μορφή. Σκοπός του πειράματος ήταν να μελετηθεί κατά πόσο ο φώσφορος πέρασε στην παραγωγή του καλαμποκιού. 'Ετσι τα δεδομένα αποτελούνται από 3 μεταβλητές

Y = Ποσότητα φωσφόρου στην παραγωγή καλαμποκιού

X_1 = Ποσότητα ανόργανου φωσφόρου στο λίπασμα

X_2= Ποσότητα οργανικού φωσφόρου στο λίπασμα

Y	0.4	0.4	3.1	0.6	4.7	1.7	9.4	10.1	11.6
X_1	64	60	71	61	54	77	81	93	93
X_2	53	23	19	34	24	65	44	31	29
Y	12.6	10.9	23.1	23.1	21.6	23.1	1.9	26.8	29.9
X_1	51	76	96	77	93	95	54	168	99
X_2	58	37	46	50	44	56	36	58	51

Πίνακας 4.4: Δεδομένα για το παράδειγμα 4.3

Σκοπός της μελέτης ήταν να μελετηθεί η σχέση ανάμεσα στο φώσφορο του λιπάσματος και της παραγωγής προσαρμόζοντας κατάλληλα γραμμικά μοντέλα. Στο Γράφημα 4.2 με τα διαγράμματα σημείων μπορεί κανείς να παρατηρήσει πως υπάρχει μια παρατήρηση αρκετά έξω από τις άλλες για την οποία υπάρχουν βάσιμες υποψίες πως έχει μεγάλη επίδραση στην εκτίμηση των παραμέτρων του μοντέλου. Στον πίνακα 4.5 έχουμε χρησιμοποιήσει το απλό γραμμικό μοντέλο $Y = a + \beta X_1$ και μπορεί κανείς να δει τις προσαρμοσθείσες τιμές για το πλήρες μοντέλο (fitted values) καθώς και τις προσαρμοσθείσες τιμές με τη μέθοδο cross-validation. Επίσης στον πίνακα υπάρχουν οι τιμές των leverages καθώς και οι εκτιμήσεις των συντελεστών του μοντέλου όταν αφαιρέσουμε την i παρατήρηση.

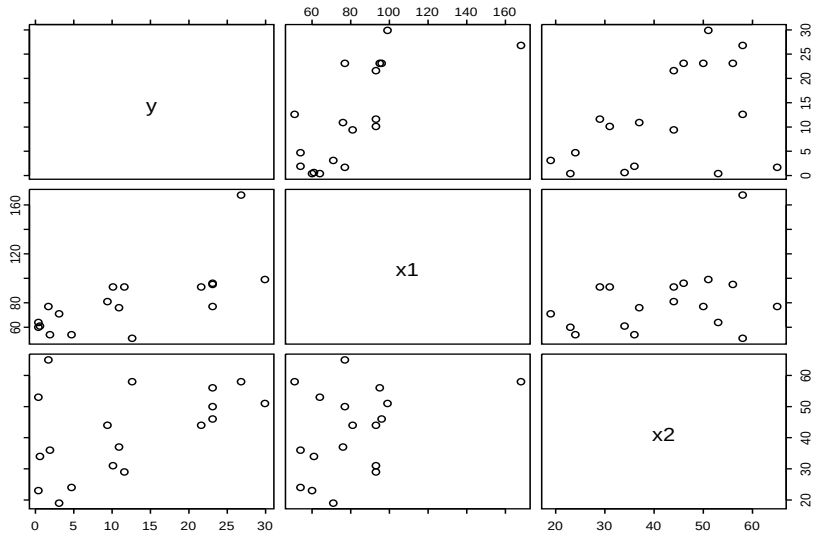
Αυτό που είναι πολύ ενδιαφέρον είναι πως η παρατήρηση 17 έχει μεγάλη επίδραση στην εκτίμηση των παραμέτρων, όπως μπορεί να δει κανείς και στο γράφημα 4.3, όπου απεικονίζονται οι 18 ευθείες που προσαρμόσαμε αφαιρώντας κάθε φορά μια παρατήρηση.

Μέχρι τώρα όμως βλέπουμε μόνο τις παρενέργειες της μεθόδου cross-validation. Ας δούμε πως θα διαλέξουμε το κατάλληλο μοντέλο με το κριτήριο PRESS

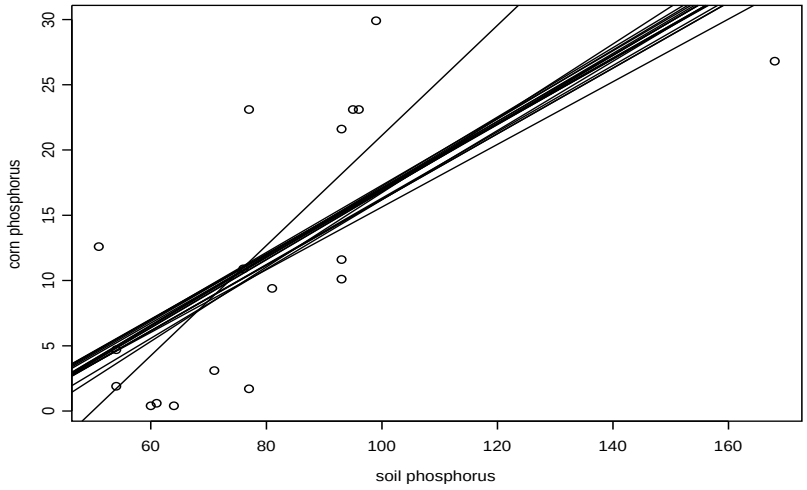
'Οπως μπορεί κανείς να δει στον πίνακα 4.6 το κριτήριο PRESS δεν έχει τη μονότονη συμπεριφορά του συντελεστή προσδιορισμού και επομένως εισάγοντας

Y	X_1	X_2	Fitted values	Fitted with CV	Leverages	$\hat{a}_{(i)}$	$\hat{\beta}_{(i)}$
0.4	64	53	7.4380	8.0471	0.07965	-7.9630	0.250157
0.4	60	23	6.3947	7.0028	0.092098	-7.9661	0.249482
3.1	71	19	9.2638	9.6858	0.064081	-8.4446	0.255358
0.6	61	34	6.6556	7.2453	0.088744	-8.0014	0.249945
4.7	54	24	4.8298	4.8468	0.115612	-9.2201	0.260498
1.7	77	65	10.8287	11.3808	0.057033	-8.4451	0.257479
9.4	81	44	11.8720	12.0174	0.055562	-9.1044	0.260763
10.1	93	31	15.0019	15.3519	0.066646	-9.3666	0.265791
11.6	93	29	15.0019	15.2448	0.066646	-9.3323	0.264270
12.6	51	58	4.0473	2.7745	0.129549	-11.7520	0.284833
10.9	76	37	10.5679	10.5475	0.057804	-9.2863	0.260972
23.1	96	46	15.7843	15.2078	0.07305	-8.9308	0.251443
23.1	77	50	10.8287	10.0865	0.057033	-10.3427	0.265315
21.6	93	44	15.0019	14.5307	0.066646	-9.1037	0.254133
23.1	95	56	15.5235	14.9466	0.070754	-8.9736	0.251791
1.9	54	36	4.8298	5.2128	0.115612	-8.4777	0.253528
26.8	168	58	34.5635	49.8080	0.662576	-21.0659	0.421868
29.9	99	51	16.5668	15.3931	0.080906	-8.3739	0.240071

Πίνακας 4.5: Χαρακτηριστικά των δεδομένων σχετικά με το φώσφορο για τις 18 παρατηρήσεις



Γράφημα 4.2: Διαγράμματα νέφους σημείων για τις 3 μεταβλητές του παραδείγματος



Γράφημα 4.3: Οι 18 ευθείες που προκύπτουν αν κάθε φορά αφαιρέσουμε μια παρατήρηση

μια καινούρια επεξηγηματική μεταβλητή στις ήδη υπάρχουσες δεν βελτιώνεται αναγκαστικά. Επίσης το κριτήριο μπορεί να χρησιμοποιηθεί ανάμεσα και σε φωλιασμένα μοντέλα. Με βάση λοιπόν το κριτήριο το απλό μοντέλο με μόνο τη X_1 ως επεξηγηματική είναι προτιμότερο. Παρατηρείστε πως έχει κατά πολύ μικρότερο συντελεστή προσδιορισμού και αυτό οφείλεται στην παρατήρηση 17 η οποία έχει πολύ μεγάλα κατάλοιπα και επομένως μεγαλύτερα μοντέλα προσπαθούν απλά να βελτιώσουν την προσαρμογή του μοντέλου ως προς την παρατήρηση αυτή.

Επεξηγηματικές Μεταβλητές	PRESS	Συντελεστής Προσδιορισμού
X_1	1526,6	48%
X_2	1809,10	21%
$X_1 + X_2$	1601,55	53%
$X_1 + X_1^2$	26248,5	55%
$X_1 + X_1^2 + X_1^3$	19676,2	68%

Πίνακας 4.6: Κριτήρια επιλογής μοντέλου για τα δεδομένα του φωσφόρου.

Τελειώνοντας με το παράδειγμα θα πρέπει να αναφέρουμε πως η εκτίμηση των παραμέτρων αφήνοντας έξω μια παρατήρηση κάθε φορά ξεκάθαρα οδηγεί σε μια jackknife μέθοδο εκτίμησης. Στην συγκεκριμένη περίπτωση θα μπορούσε να χρησιμοποιηθούν οι μέσοι των $\hat{a}_{(i)}$,$\hat{\beta}_{(i)}$ ως εκτιμήσεις των συντελεστών του μοντέλου βασισμένοι στην ιδέα του jackknife. Μια τέτοια προσέγγιση θα διόρθωνε το πρόβλημα με την παρατήρηση 17 που έχει μεγάλη επίδραση στο αποτέλεσμα.

4.7.4 'Αλλες Εφαρμογές

Το πιο σημαντικό ίσως είναι πως η μέθοδος cross-validation χρησιμοποιείται σε μια πλειάδα εφαρμογών πέραν αυτής για την καλή προσαρμογή ενός γραμμικού μοντέλου. Σε πολλές εφαρμογές, όπου δεν υπάρχει τεράστια θεωρητική ανάπτυξη του προβλήματος όπως στην περίπτωση της γραμμικής παλινδρόμησης, η μέθοδος cross-validation είναι το κύριο εργαλείο για να μελετήσουμε. Αυτές είναι:

- Είδαμε σε προηγούμενο κεφάλαιο πως μπορούμε να τη χρησιμοποιήσουμε για να βρούμε την τιμή του παράθυρου στην εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας με τη χρήση των kernels.

- Παρομοίως μπορούμε να τη χρησιμοποιήσουμε σε παρόμοια προβλήματα smoothing, όπως σε προβλήματα με splines, εξομάλυνσης χρονολογικών σειρών κλπ.

- Το κριτήριο PRESS που αναφέραμε μπορεί να χρησιμοποιηθεί σε πολλές άλλες περιπτώσεις επιλογής μοντέλων, όπως μη γραμμικά μοντέλα, γενικευμένα γραμμικά μοντέλα κλπ.

- Πολλές εφαρμογές έχει η μέθοδος και σε προβλήματα τεχνητής νοημοσύνης. Συγκεκριμένα πολλοί αλγόριθμοι μάθησης στην τεχνητή νοημοσύνη έχουν σκοπό να δημιουργήσουν κανόνες οι οποίοι να μπορούν να προβλέψουν σωστά την τιμή κάποιου χαρακτηριστικού με βάση διάφορα άλλα χαρακτηριστικά. Για παράδειγμα χρησιμοποιώντας πολλά συμπτώματα και την παρουσία ή απουσία τους να προβλέψουν την ασθένεια ενός αρρώστου. Το κύριο εργαλείο για να μετρηθεί η καλή ποιότητα τέτοιων κανόνων είναι η μέθοδος Cross-validation.

- Μια άλλη πολύ γνωστή εφαρμογή της μεθόδου είναι και στη διακριτική ανάλυση προκειμένου να αξιολογηθεί το μοντέλο που προσαρμόσαμε.

Η μέθοδος είναι μη παραμετρική καθώς δεν χρειαζόμαστε κάποιες υποθέσεις για να τη χρησιμοποιήσουμε ακόμα και αν το μοντέλο έχει συγκεκριμένες παραμετρικές υποθέσεις.

Θα πρέπει να σημειωθεί πως η μέθοδος cross-validation που μόλις περιγράψαμε αποτελεί στην ουσία την περίπτωση leave-one out cross-validation, δηλαδή αφήνουμε μόνο μια παρατήρηση έξω κάθε φορά. Σε μερικά προβλήματα αυτό δεν είναι αρκετό και μπορεί κανείς να γενικεύσει στην περίπτωση του leave-k out cross-validation όπου πια αφήνουμε k παρατηρήσεις έξω κάθε φορά μειώνοντας ακόμα περισσότερο την επίδραση ομάδων παρατηρήσεων. Βέβαια σε αυτή την περίπτωση τα πράγματα γίνονται αρκετά πιο περίπλοκα.

Για παράδειγμα, αν χρησιμοποιήσουμε k-fold cross-validation, δηλαδή αφήσουμε k παρατηρήσεις κάθε φορά έξω, προκύπτει το πρόβλημα ποιες k-άδες θα αφήσουμε κάθε φορά έξω. Για μέγεθος δείγματος 100 υπάρχουν $\begin{pmatrix} 100 \\ k \end{pmatrix}$ ομάδες για να δημιουργήσουμε και επομένως θα ήταν επικίνδυνο να χρησιμοποιήσουμε κάποιες αυθαίρετα. Σε αυτές τις περιπτώσεις έχουν αναπτυχθεί ειδικές τεχνικές ώστε να

κατασκευάζονται οι απαιτούμενες ομάδες και να μπορεί να εφαρμοσθεί απρόσκοπτα η μέθοδος cross-validation, στις οποίες όμως δεν θα επεκταθούμε.

4.8 Περισσότερα αποτελέσματα για τη μέθοδο jackknife

Ας δούμε εν συντομία κάποια ακόμα στοιχεία για τη μέθοδο jackknife

΄Εστω ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές κι έστω $\hat{\theta}$ μια εκτιμήτρια της άγνωστης παραμέτρου θ του πληθυσμού. Ας χωρίσουμε το δείγμα σε g ομάδες από h παρατηρήσεις η κάθε μια και έστω πως $n = gh$. Τότε μια γενίκευση της εκτιμήτριας jackknife που είδαμε είναι η εξής: ΄Εστω $\hat{\theta}_{-i}$ η αντίστοιχη εκτιμήτρια όταν έχουμε απαλείψει την i ομάδα, επομένως το μέγεθος του δείγματος είναι $(g-1)h$. Τότε, υπολόγισε την ποσότητα

$$\hat{\theta}_i = g\hat{\theta} - (g-1)\hat{\theta}_{-i}$$

Η jackknife εκτιμήτρια προκύπτει ως

$$\hat{\theta}_J = \frac{1}{g}\sum_{i=1}^{g}\hat{\theta}_i = g\hat{\theta} - (g-1)\frac{1}{g}\sum_{i=1}^{g}\hat{\theta}_{-i}$$

Οι ομοιότητες με την απλή εκτιμήτρια που είδαμε προηγουμένως είναι σαφείς.

Ας δούμε εν συντομία γιατί η μέθοδος jackknife μειώνει τη μεροληψία. Γενικά για μια εκτιμήτρια $\hat{\theta}$ και χρησιμοποιώντας σειρές Taylor για να εκφράσουμε την μερληψία προκύπτει πως

$$E(\hat{\theta}) = \theta + \frac{a_1(\theta)}{n} + \frac{a_2(\theta)}{n^2} + \ldots$$

όπου οι συναρτήσεις $a_i(\theta)$ δεν εξαρτώνται από το μέγεθος του δείγματος αλλά μπορούν να εξαρτώνται από το θ. Για ευκολία γράφουμε την αναμενόμενη τιμή ως

$$E(\hat{\theta}) = \theta + \frac{a_1(\theta)}{n} + R(n^2)$$

όπου $R(n^2)$ υπονοεί όλους του υπόλοιπους όρους.

Μπορεί κανείς τότε να δει πως

$$\begin{aligned}
E(\hat{\theta}_J) &= E\left[g\hat{\theta} - (g-1)\frac{1}{g}\sum_{i=1} g\hat{\theta}_{-i}\right] \\
&= gE(\hat{\theta}) - (g-1)\frac{1}{g}\sum_{i=1} gE(\hat{\theta}_{-i}) \\
&= gE(\hat{\theta}) - (g-1)\frac{1}{g}\sum_{i=1} gE(\hat{\theta}) \\
&= g\left[\theta + \frac{a_1(\theta)}{n} + R(n^2)\right] - (g-1)\left[\theta + \frac{a_1(\theta)}{n-1} + R((n-1)^2)\right] \\
&= g\theta + a_1(\theta) - (g-1)\theta - a_1(\theta) + R(n^2) - R((n-1)^2) \\
&= \theta + R(n^2) - R((n-1)^2)
\end{aligned}$$

δηλαδή ο όρος που αναφερόταν στο n εξαφανίστηκε και η ποσότητα $R(n^2)-R((n-1)^2)$ είναι προφανώς μικρότερη από το $R(n^2)$ που είχαμε. Συνεπώς αποδείξαμε πως η μεροληψία μειώθηκε.

Αξίζει να σημειωθεί πως μπορεί κανείς να εξαφανίσει και όρους της μεροληψίας που οφείλονται στο n^2 με παρόμοιο τρόπο ορίζοντας εκτιμήτριες jackknife δεύτερης τάξης ως εξής

$$\begin{aligned}\theta_J^{(2)} &= \frac{1}{n-1}\left[n^2\hat{\theta} - (2n^2-2n+1)(n-1)\frac{1}{n}\sum_{i=1} n\hat{\theta}_{-i}\right. \\ &\left. +(n-1)^2(n-2)\left(\frac{2}{n(n-1)}\sum_{i<j} n\hat{\theta}_{-ij}\right)\right]\end{aligned}$$

όπου $\hat{\theta}_{-ij}$ είναι η εκτίμηση του θ όταν αφαιρέσουμε τις παρατηρήσεις X_i, X_j. Η ιδέα μπορεί να γενικευτεί περισσότερο για να απαλείψουμε μεροληψία που οφείλεται σε μεγαλύτερης τάξεις δυνάμεις του μεγέθους του δείγματος.

4.9 Ασκήσεις

1. 'Εστω δείγμα μεγέθους 3 και έστω οι διατεταγμένες παρατηρήσεις $X_1 < X_2 < X_3$. Δείξτε ότι ο jackknife εκτιμητής της δειγματικής διαμέσου δίνεται από τον τύπο
$$\hat{m}_J = \frac{7X_2 - 2(X_1+X_3)}{3}$$

2. 'Εστω ένα τυχαίο δείγμα από 7 παρατηρήσεις $X_1, X_2, ..., X_7$ που δεν είναι απαραίτητα σε αύξουσα ή φθίνουσα σειρά. Να βρείτε την jackknife εκτιμήτρια του εύρους $R = X_{\max} - X_{\min}$ όπου $X_{\max}, X_{\min}$ η μεγαλύτερη και η μικρότερη παρατήρηση αντίστοιχα. Τι παρατηρείτε;

3. 'Εστω ένα δείγμα μεγέθους 8, με τιμές 10,11,14,16,18,20,10,12. Εκτιμήστε με τη μέθοδο jackknife τη διακύμανση του λόγου μέση τιμή/διάμεσος.

4. 'Εστω ένα τυχαίο δείγμα από πληθυσμό με κατανομή $g(x \mid \theta)$ όπου θ είναι η άγνωστη παράμετρος που θέλουμε να εκτιμήσουμε. Πιστεύετε ότι η μέθοδος cross-validation μπορεί να χρησιμοποιηθεί για την επιλογή της καλύτερης εκτιμήτριας $\hat{\theta}$;

5. Ποιά είναι η εκτιμήτρια jackknife δευτέρας τάξης για τη δειγματική διακύμανση; Είναι αμεροληπτη;

6. Μια στατιστική συνάρτηση λεμε πως ειναι τετραγωνικής μορφής αν μπορεί να γραφτεί ως
$$\hat{\theta} = \mu + \sum_{i=1}^{n}\alpha(x_i) + \sum_{1\le i<j\le n}\beta(x_i, x_j)$$

όπου μ είναι μια σταθερά και $\alpha(x), \beta(x, y)$ είναι συναρτήσεις μόνο των x και των x και y αντίστοιχα. Για τη μέση τιμή και τη διακύμανση βρεστε τις συναρτήσεις αυτές.

7. Για τα δεδομένα του πίνακα 2.4 βρείτε τη μορφή της σχέσης ανάμεσα στη συμμετοχή και τη διαφορά κάνοντας cross-validation.

8. Ένας ερευνητής έχει στα χέρια του ένα τυχαίο δείγμα μεγέθους $n = 53$. Ποιά θα είναι η jackknife εκτιμήτρια της διαμέσου;

9. Έστω ένα δείγμα $X_1, X_2, \ldots, X_n$. Να βρείτε την jackknife εκτιμήτρια της ποσότητας $\hat{\theta} = n^{-1} \sum_{i=1}^{n} X_i^2$. Τι παρατηρείτε;

10. Ένας ερευνητής έχει στα χέρια του ένα τυχαίο δείγμα μεγέθους $n = 147$. Η ποσότητα η οποία τον ενδιαφέρει είναι η $\hat{\theta} = max(X_i)$, δηλαδή η μεγαλύτερη τιμή του δείγματος. Ποιά θα είναι η jackknife εκτιμήτρια της ποσότητας αυτής; Πως θα εκτιμούσατε το τυπικό σφάλμα; Θα εμπιστευόσαστε αυτή την εκτίμηση του τυπικού σφάλσματος;

11. Έχουμε 50 μετρήσεις για καθένα από δυο πληθυσμούς. Θέλουμε να κατασκευάσουμε έναν κανόνα κατάταξης που να κατατάσσει μελλοντικές παρατηρήσεις σε έναν από τους 2 πληθυσμούς. Για το σκοπό αυτό εκτιμάμε για κάθε πληθυσμό την συνάρτηση πυκνότητας πιθανότητας και ο κανόνας μας είναι πως κατατάσσουμε τη νέα παρατήρηση στον πληθυσμό που έχει τη μεγαλύτερη εκτιμώμενη πυκνότητα. Ένας επιστήμονας υποστηρίζει πως η μέθοδος αυτή δεν είναι καλή και πως αρκεί να κατατάξουμε τη νέα παρατήρηση στον πληθυσμό με μέση τιμή πιό κοντά στην νέα τιμή. Περιγράψτε πως θα χρησιμοποιήσουμε τη μέθοδο cross-validation για να δούμε ποιά μέθοδος είναι η καλύτερη.

Βιβλιογραφία

Efron B., Tibshirani (1993) *An Introduction to the Bootstrap*. Marcel and Decker

Αν και το βιβλίο πραγματεύεται τη μέθοδο bootstrap περιέχει μερικά ενδαφέροντα αποτελέσματα για τη μέθοδο jackknife κυρίως σε πρακτικό επίπεδο

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer

Σε αυτό το βιβλίο περιέχεται μια θεωρητική προσέγγιση της μεθόδου και αποδεικνύονται πολλά θεωρήματα που δείχνουν πως και γιατί δουλεύει η μέθοδος. Είναι αρκετά τεχνικά και χρειάζονται αρκετή γνώση μαθηματικών και στατιστικής

Wolter, K.M (1985) *Variance Estimation* Springer

Αν και το βιβλίο όπως μαρτυρά και ο τίτλος ασχολείται γενικά με εκτίμηση διακύμανσης σε δειγματοληπτικές έρευνες περιέχει κάποια κεφάλαια με πολύ ενδιαφέρουσες εφαρμογές της μεθόδου jackknife στη δειγματοληψία.

Miller, R.G. (1974). The jackknife - a review. *Biometrics*, 61, 1-15

Το άρθρο αυτό παρουσιάζει πολλά θεωρητικά αποτελέσματα μέχρι και τη δεκαετία του 1970. Αν και από τότε υπάρχουν πολλά καινούρια αποτελέσματα το άρθρο έχει αρκετό ενδιαφέρον.

Quenouille, M.H. (1956) Notes on bias in estimation *Biometrika*, 43, 353-360

Στο άρθρο αυτό πρωτοπαρουσιάστηκε η ιδέα του jackknife.

Κεφάλαιο 5

Η Μέθοδος Bootstrap

5.1 Εισαγωγή

Στα προηγούμενα κεφάλαια παρουσιάσαμε διάφορες μεθόδους που στηρίχτηκαν στην ιδέα να παίρνουμε δείγματα από τα ήδη υπάρχοντα δείγματα και να τα χρησιμοποιούμε για λόγους στατιστικής συμπερασματολογίας. Παρόλα αυτά, σε πολλές περιπτώσεις οι μέθοδοι που είδαμε είχαν σημαντικά μειονεκτήματα. Για παράδειγμα οι έλεγχοι τυχαιοποίησης μπορούσαν να ελέγξουν μόνο συγκεκριμένες μηδενικές υποθέσεις. Η μεθοδολογία Monte Carlo βασίστηκε σε μια πολύ σημαντική υπόθεση: ότι ξέρουμε την κατανομή του πληθυσμού. Δηλαδή τα δείγματα που χρειαζόμασταν τα προσομοιώναμε από κάποια κατανομή που γνωρίζαμε ποια είναι. Τι μπορούμε να κάνουμε όμως όταν η κατανομή δεν είναι γνωστή; Η μέθοδος jackknife είχε προβλήματα καθώς τα δείγματα που παίρναμε έμοιαζαν πολύ μεταξύ τους και άρα δεν μπορούσε να δουλέψει ικανοποιητικά για στατιστικές συναρτήσεις που δεν είναι εξ ορισμού τους ιδιαίτερα λείες (smooth) όπως η διάμεσος. Στο κεφάλαιο αυτό θα δούμε τη μέθοδο bootstrap η οποία μοιάζει να ξεπερνά κάποια από τα προβλήματα που μόλις περιγράψαμε και κυρίως δεν χρειάζεται να γνωρίζουμε τίποτα για την κατανομή του πληθυσμού. Η μέθοδος bootstrap εμφανίστηκε στο τέλος της δεκαετίας του 70 και σε σχέση με τη ραγδαία πρόοδο των υπολογιστών αποτελεί σήμερα ένα ισχυρό εργαλείο σε πληθώρα εφαρμογών.

Η ιδέα είναι σχετικά απλή:

Δεν ξέρουμε την κατανομή του πληθυσμού αλλά στην πραγματικότητα έχουμε ένα δείγμα από αυτήν (τα δεδομένα μας). Επομένως γιατί να μην χρησιμοποιήσουμε την εμπειρική κατανομή των δεδομένων μας ως μια εκτίμηση της πραγματικής, αλλά άγνωστης, κατανομής του πληθυσμού;

Επομένως η μέθοδος bootstrap χρησιμοποιεί την εμπειρική κατανομή για να προσομοιώσει από αυτή δείγματα. Ποια είναι όμως η εμπειρική κατανομή; Είναι η κατανομή που δίνει πιθανότητα $1/n$ σε κάθε μια από τις n παρατηρήσεις του δείγματος μας και 0 σε οποιαδήποτε άλλη τιμή.

Δηλαδή έχουμε:

Τιμή	X_1	X_2	X_3	...	X_n	Οποιαδήποτε άλλη τιμή
Πιθανότητα	$1/n$	$1/n$	$1/n$	...	$1/n$	0

Αυτή είναι και η βασική διαφορά της μεθόδου bootstrap από τη μέθοδο Monte Carlo. 'Οτι χρησιμοποιεί την εμπειρική κατανομή αντί για κάποια γνωστή κατανομή. Από εκεί και πέρα πρέπει να πάρουμε δείγματα από αυτή την κατανομή και να προχωρήσουμε σύμφωνα με όσα είπαμε και για τις μεθόδους Monte Carlo, με μικρές σε κάποιες περιπτώσεις διαφοροποιήσεις. Πριν όμως δούμε με λεπτομέρεια τη χρήση της μεθόδου σε διάφορες εφαρμογές της στατιστικής ας δούμε πως μπορούμε να πάρουμε ένα δείγμα από την εμπειρική κατανομή.

Για να πάρουμε ένα δείγμα από την εμπειρική κατανομή πραγματοποιούμε δειγματοληψία με επανάθεση. (Παρατήρηση: σε ένα δείγμα μεγέθους n είναι πιθανό κάποια τιμή να εμφανίζεται περισσότερες από μια φορές, έστω για παράδειγμα 2 φορές. Σε αυτή την περίπτωση είναι κατανοητό ότι έχουμε $n-1$ διαφορετικές τιμές και πως αυτή η τιμή έχει πια πιθανότητα $2/n$. Αυτό δεν είναι πρόβλημα καθώς είναι ισοδύναμο με το να διαλέγουμε κάθε μια από τις n παρατηρήσεις με πιθανότητα $1/n$).

Μια πλήρης περιγραφή της δειγματοληψίας είναι η εξής: Διαλέγουμε δηλαδή μια πρώτη τιμή από τις παρατηρήσεις μας και μετά την επιστρέφουμε και διαλέγουμε ξανά από το σύνολο των παρατηρήσεων. Επομένως ένα δείγμα bootstrap μπορεί να περιέχει κάποια τιμή περισσότερες από μια φορές αλλά μπορεί να μην περιέχει κάποια άλλη τιμή. Φανταστείτε τις 10 τιμές 1, 3, 6, 8, 9, 11, 14, 16 19, 18. Κάνοντας δειγματοληψία με επανάθεση μπορεί το δείγμα που θα προκύψει να είναι 1, 1, 3, 3, 3, 3, 8, 14, 16 19 , δηλαδή η τιμή 3 εμφανίζεται 4 φορές, η τιμή 1 εμφανίζεται 2 φορές ενώ οι τιμές 6, 9, 11, 18 δεν εμφανίστηκαν σε αυτό το δείγμα. Είναι ευνόητο ότι καθώς η μεθοδολογία υποθέτει πως παίρνουμε αρκετά τέτοια δείγματα, τελικά όλες οι παρατηρήσεις θα εμφανίζονται με τη συχνότητα που υποθέτει η εμπειρική κατανομή και άρα δεν υπάρχει πρόβλημα. Αυτό από την άλλη σημαίνει πως πρέπει να πάρουμε αρκετά δείγματα για να έχουμε μεγαλύτερη σιγουριά πως η προσέγγιση μας είναι ικανοποιητική.

Συνεπώς η βασική ιδέα της μεθόδου bootstrap είναι πως κάνουμε δειγματοληψία με επανάθεση από το υπάρχον δείγμα και άρα θεωρούμε πως η εμπειρική κατανομή είναι μια καλή προσέγγιση της κατανομής του πληθυσμού. Αυτή η τελευταία υπόθεση είναι θεμελιώδης. Μπορεί να παρατηρήσει αμέσως κανείς πως όταν αυτό δεν ισχύει (πχ μικρό μέγεθος δείγματος, πολυμεταβλητά προβλήματα κλπ) η μέθοδος bootstrap είναι καταδικασμένη να μην δουλεύει καλά.

Η μέθοδος bootstrap χρησιμοποιείται κυρίως για στατιστική συμπερασματολογία. Μπορούμε να εκτιμήσουμε τυπικά σφάλματα και τη μεροληψία εκτιμητριών, να κάνουμε ελέγχους υποθέσεων ακόμα και σε ιδιαίτερα πολύπλοκες μηδενικές υποθέσεις καθώς και να προσεγγίσουμε την κατανομή πολύπλοκων συναρτήσεων (κάτι που δεν μπορούσαμε να το κάνουμε με τη μέθοδο jackknife για παράδειγμα). Στη συνέχεια θα προσπαθήσουμε να περιγράψουμε κάποιες από τις εφαρμογές της μεθόδου. Πριν προχωρήσουμε πρέπει να αναφέρουμε πως αν και η βασική ιδέα της μεθόδου είναι απλή, σε πολλά στατιστικά προβλήματα χρειαζόμαστε ειδική

προσέγγιση που μειώνει ή δυσκολεύει τη χρήση της μεθόδου.

5.2 Bootstrap 'Ελεγχοι Υποθέσεων

Ας ξεκινήσουμε την περιγραφή των εφαρμογών της μεθόδου με ελέγχους υποθέσεων, αφού με αυτούς έχουμε ασχοληθεί αρκετά μέχρι τώρα. Ας ξεκινήσουμε με ένα παράδειγμα

Παράδειγμα 5.1: 'Εστω οι 10 παρατηρήσεις:
-0.89, -0.47, 0.05, 0.155, 0.279, 0.775, 1.0016, 1.23, 1.89, 1.96.
Να ελεγχθεί η υπόθεση ότι ο μέσος του πληθυσμού είναι 1 έναντι της εναλλακτικής ότι διαφέρει.

Κατά αρχάς δεν έχουμε καμιά πληροφορία σχετικά με την κατανομή του πληθυσμού και άρα δεν μπορούμε να εφαρμόσουμε ελέγχους Monte Carlo. Η μέθοδος bootstrap μοιάζει μια καλή εναλλακτική μέθοδος. Δεδομένου λοιπόν πως δεν ξέρουμε τίποτα για την κατανομή του πληθυσμού θα προσομοιώσουμε τα δείγματα που χρειαζόμαστε από την εμπειρική κατανομή.

Ας δούμε όμως τις υποθέσεις που έχουμε. 'Ετσι έχουμε
$H_0 : \mu = 1$ έναντι της
$H_1 : \mu \neq 1$

Δεδομένου ότι η εναλλακτική είναι δίπλευρη μια καλή ελεγχοσυνάρτηση είναι η $T = |\bar{x} - 1|$ η οποία θα είναι 0 αν ισχύει η μηδενική υπόθεση ενώ μεγάλες τιμές υποδηλώνουν απόκλιση από τη μηδενική υπόθεση. Και πάλι δεν χρειάζεται να ξέρουμε την κατανομή της ελεγχοσυνάρτησης αφού έτσι κι αλλιώς θα την προσεγγίσουμε με προσομοίωση. 'Αρα μπορούμε να διαλέξουμε σχετικά απλές ελεγχοσυναρτήσεις. Για τα δεδομένα μας υπολογίζουμε $\bar{x} = 0.598$ και $T = 0.402$.

Ας σταματήσουμε εδώ να δούμε κάποιο πρόβλημα που υπάρχει. Είπαμε και πριν στους ελέγχους Monte Carlo πως θέλουμε να προσομοιώσουμε από την κατανομή που η μηδενική υπόθεση υποδεικνύει. Η μηδενική υπόθεση υποθέτει μέση τιμή 1. Αν πάρουμε bootstrap δείγματα από τα δεδομένα επειδή αυτά έχουν μέση τιμή 0.598 στην πραγματικότητα δεν προσομοιώνουμε από τη μηδενική υπόθεση.

Επομένως μια λύση είναι να μετακινήσουμε τα δεδομένα μας ώστε να ικανοποιούν τη μηδενική υπόθεση. Δηλαδή σε κάθε παρατήρηση προσθέτουμε 0.402 οπότε τώρα η μέση τιμή είναι 1 και άρα είναι λογικό να προσομοιώσουμε από αυτή. Παρατηρείστε όμως πως αν και διορθώνουμε τη μέση τιμή άλλα χαρακτηριστικά όπως η διακύμανση, η ασυμμετρία κλπ παραμένουν σταθερά.

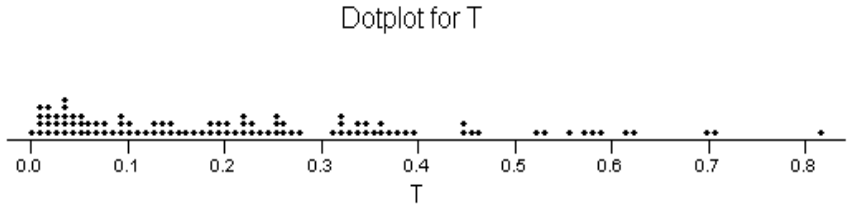
Μπορούμε τώρα να περιγράψουμε τον έλεγχο ως :

- Προσομοιώνουμε k, έστω 99, δείγματα από την εμπειρική κατανομή μετακινημένη κατά 0.402 ώστε να ικανοποιεί τη μηδενική υπόθεση. Κάθε τέτοιο δείγμα έχει 10 παρατηρήσεις και η δειγματοληψία γίνεται διαλέγοντας τυχαία παρατηρήσεις με επανάθεση.

- Για κάθε δείγμα υπολογίζουμε την ελεγχοσυνάρτηση

- Στη συνέχεια βρίσκουμε πόσες από τις 99 τιμές της ελεγχοσυνάρτησης είναι πιο μεγάλες από αυτή του δείγματος μας.

Από τις 99 επαναλήψεις βρήκαμε πως m =18, δηλαδή 18 παρατηρήσεις είναι μεγαλύτερες από 0.402. Επομένως εκτιμούμε το p-value ως $\tilde{p} = \frac{m+1}{k+1} = \frac{18+1}{99+1} = 0.18$. Γενικά για το p-value και την επιλογή του αριθμού των επαναλήψεων k ισχύουν όσα έχουμε πει προηγουμένως για τους ελέγχους Monte Carlo. 'Αρα σε επίπεδο στατιστικής σημαντικότητας 5% δεν απορρίπτουμε τη μηδενική υπόθεση. Οι τιμές υπάρχουν στο γράφημα 5.1. Παρατηρείστε πως η κατανομή είναι ιδιαίτερα ασύμμετρη.



Γράφημα 5.1: Διάγραμμα σημείων για την ελεγχοσυνάρτηση

Αν αναλογιστεί κανείς την ομοιότητα στην κατασκευή ενός διαστήματος εμπιστοσύνης και σε έναν δίπλευρο έλεγχο, η αντιστοιχία στην περίπτωση μας θα ήταν να κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για τη μέση τιμή και να δούμε αν περιέχει την τιμή 1. Το πως κατασκευάζουμε διαστήματα εμπιστοσύνης με τη χρήση bootstrap θα το δούμε αργότερα.

Και πάλι θα μπορούσε να πει κανείς ότι έλεγχος για τη μέση τιμή είναι μάλλον τετριμμένος. 'Ομως αν μας ενδιέφερε να ελέγξουμε την υπόθεση ότι η διάμεσος του πληθυσμού είναι 1 έναντι της εναλλακτικής ότι δεν είναι, ένας τέτοιος έλεγχος είναι ιδιαίτερα δύσκολος να γίνει με τις συμβατικές μεθόδους.

Συνεπώς μπορούμε να περιγράψουμε έναν έλεγχο υποθέσεων με τη χρήση bootstrap ως εξής:

Bootstrap 'Ελεγχος Υποθέσεων

- Βήμα 1ο: Θέσε τη μηδενική υπόθεση .
- Βήμα 2ο: Διάλεξε την ελεγχοσυνάρτηση, υπολόγισε την τιμή της t_{obs} για τα δεδομένα που έχεις.
- Βήμα 3ο: Προσομοίωσε αρκετά δείγματα από την εμπειρική κατανομή, προσαρμοσμένη για να ικανοποιεί τη μηδενική υπόθεση αν χρειάζεται (η δειγματοληψία θα γίνει με επανάθεση από τα δεδομένα μετά την τροποποίηση). Για κάθε δείγμα υπολόγισε την τιμή της ελεγχοσυνάρτησης

- Βήμα 4ο: Υπολόγισε το p-value ελέγχοντας πόσες ακραίες τιμές πήρες από τα δείγματα bootstrap σε σχέση με την τιμή που είχες στα δεδομένα σου

Από τα παραπάνω γίνεται κατανοητό ότι οι έλεγχοι υποθέσεων με bootstrap είναι συμπληρωματικοί των ελέγχων Monte Carlo στις περιπτώσεις που δεν ξέρουμε την κατανομή του πληθυσμού. Δηλαδή δεν χρειάζονται σχεδόν καθόλου υποθέσεις για να εφαρμοστούν. Σε πολλές περιπτώσεις όπως για παράδειγμα σε ελέγχους καλής προσαρμογής, η μηδενική υπόθεση μας καθορίζει και την κατανομή του πληθυσμού οπότε σε αυτή την περίπτωση χρησιμοποιούμε έλεγχο Monte Carlo. Θα πρέπει εδώ να παρατηρήσουμε πως σε πολλά βιβλία η μέθοδος Monte Carlo αναφέρεται ως παραμετρική Bootstrap μέθοδος (parametric bootstrap).

Ας δούμε ένα ακόμα παράδειγμα.

Παράδειγμα 5.2

16 ποντίκια συμμετείχαν σε ένα πείραμα, σε 7 από αυτά δόθηκε ένα καινούριο φάρμακο ενώ στα υπόλοιπα 9 δόθηκε ένα άλλο ήδη υπάρχον φάρμακο. Σκοπός ήταν να μελετηθεί η επιβίωση σε μέρες και αν το καινούριο φάρμακο μεγαλώνει την επιβίωση των ποντικιών. Οι τιμές εμφανίζονται στον πίνακα 5.1:

Φάρμακο	Παρατηρήσεις	Μέση τιμή	Τυπική απόκλιση της μέσης τιμής
Νέο	94, 197, 16, 38, 99, 141, 23	86.86	25.24
Παλιό	52, 104, 146, 10, 50, 31, 40, 27, 46	56.22	14.14

Πίνακας 5.1: Δεδομένα του παραδείγματος 5.2

Από τα δεδομένα μπορούμε να δούμε πως η μέση διαφορά είναι $d = \bar{x} - \bar{y}$=30.63 δηλαδή το νέο φάρμακο έχει κατά 30 περίπου μέρες μεγαλύτερη επιβίωση. Είναι αυτή η διαφορά στατιστικά σημαντική;

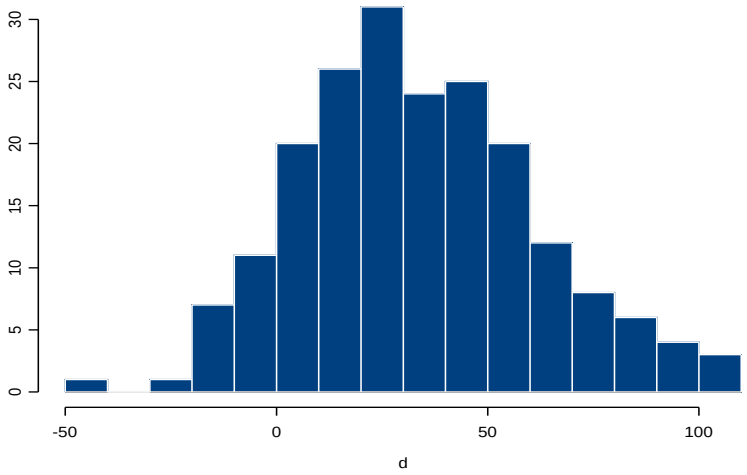
Αν δουλέψουμε με τη μέση τιμή μπορεί πάλι κάποιος να επικαλεστεί το κεντρικό οριακό θεώρημα. Με βάση λοιπόν αυτή τη προσέγγιση η τυπική απόκλιση του d είναι $se(d) = \sqrt{25.24^2 + 14.14^2} = 28.93$ και επομένως κάποιος μπορεί να δει πως επειδή η τυπική απόκλιση είναι μεγάλη τελικά η διαφορά δεν είναι στατιστικά σημαντική.

Για να συμπερασματολογήσουμε χρησιμοποιώντας bootstrap θα πρέπει να φτιάξουμε την κατανομή της στατιστικής συνάρτησης d. Για να γίνει αυτό θα πρέπει να φτιάξουμε δείγματα από τις δύο ομάδες (δειγματοληψία με επανάθεση μέσα σε κάθε ομάδα και όχι ανακατεύοντας παρατηρήσεις της κάθε ομάδας) και να υπολογίσουμε την τιμή του d για κάθε τέτοιο δείγμα. Χρησιμοποιώντας λοιπόν αυτή την τεχνική και για 199 επαναλήψεις βρήκαμε πως η τυπική απόκλιση του d για αυτές τις 199 τιμές είναι 27.32. Η διαφορά οφείλεται όπως μπορεί να φανταστεί κανείς στο ότι πήραμε 199 τιμές. Αν αυξήσουμε τον αριθμό των επαναλήψεων περιμένουμε η τυπική απόκλιση να πλησιάζει αυτή που θεωρητικά περιμένουμε. Στο γράφημα 5.2 βλέπουμε το ιστόγραμμα των 199 τιμών της d που πήραμε από το bootstrap. Παρατηρείστε ότι η τιμή 0 είναι κοντά στο κέντρο της κατανομής, δηλαδή μια τέτοια τιμή δεν είναι καθόλου απίθανη.

Μπορεί να παρατηρήσει κάποιος πως το πρόβλημα είναι όμοιο με αυτό που περιγράψαμε στον έλεγχο τυχαιοποίησης. Εκεί βασισμένοι στο γεγονός ότι η μηδενική

υπόθεση δηλώνει ότι οι μέσοι είναι ίσοι, αντί να πάρουμε δείγμα από κάθε ομάδα απλά ανακατεύαμε όλες τις παρατηρήσεις χωρίς επανάθεση.

Βέβαια οι bootstrap έλεγχοι υπόθεσης έχουν έναν μεγάλο περιορισμό. Πρέπει να προσομοιώσουμε από την εμπειρική κατανομή κατάλληλα διαμορφωμένη ώστε να αντικατοπτρίζει τη μηδενική υπόθεση. Αν και ο μετασχηματισμός είναι εύκολος για κάποιους ελέγχους σε μερικές περιπτώσεις είναι πολύ δύσκολο να το επιτύχουμε και συνεπώς δεν μπορούμε να χρησιμοποιήσουμε τη μέθοδο bootstrap. Σε μερικές περιπτώσεις μπορεί να χρησιμοποιηθεί η ιδέα των διαστημάτων εμπιστοσύνης, λόγω της ισοδυναμίας τους με δίπλευρους ελέγχους αλλά και αυτό δεν είναι πάντα εφικτό. 'Ενα τέτοιο παράδειγμα είναι όταν έχουμε πολυδιάστατα δεδομένα και θέλουμε να ελέγξουμε τη μηδενική υπόθεση πως ο πίνακας συσχετίσεων είναι ο μοναδιαίος (και άρα τα δεδομένα είναι ασυσχέτιστα). Για να κάνουμε τον οποιοδήποτε έλεγχο θα πρέπει να μετασχηματίσουμε τα δεδομένα έτσι ώστε να είναι ασυσχέτιστα. Αν και μαθηματικά αυτό μπορεί να γίνει πιθανότατα ο μετασχηματισμός θα μας χαλάσει διάφορα άλλα χαρακτηριστικά, οπότε τα πράγματα περιπλέκονται. Θα δούμε αργότερα το πρόβλημα αυτό.



Γράφημα 5.2: Ιστόγραμμα των 199 τιμών bootstrap για τη διαφορά d.

5.3 Pivotal Ελεγχοσυναρτήσεις

Στους ελέγχους τυχαιοποίησης η επιλογή της ελεγχοσυνάρτησης γινόταν με βάση την ικανότητα της ελεγχοσυνάρτησης να διακρίνει ανάμεσα στη μηδενική και την εναλλακτική υπόθεση. Το ίδιο και στους ελέγχους Monte Carlo. Δεδομένου πως

και στη μια περίπτωση αλλά και στην άλλη τα δείγματα προέρχονταν είτε από τυχαιοποίηση είτε από προσομοίωση από μια γνωστή κατανομή δεν ενδιαφερόμασταν ιδιαίτερα για τις ιδιότητες της ελεγχοσυνάρτησης. Δυστυχώς δεν ισχύει το ίδιο με τους ελέγχους bootstrap. Ο λόγος είναι πως προσομοιώνουμε τα δείγματα από την εμπειρική κατανομή και συνεπώς οποιοδήποτε λάθος δειγματοληψίας υπάρχει στα δεδομένα θα περάσει και στα δείγματα που θα πάρουμε και άρα στα αποτελέσματα. Για παράδειγμα αν εφαρμόσουμε έλεγχο Bootstrap για έναν μέσο, μπορεί να διορθώνουμε πριν κάνουμε τον έλεγχο το δείγμα μας μετατοπίζοντας τις τιμές, αλλά στην ουσία η διακύμανση δεν διορθώνεται με αποτέλεσμα να παίρνουμε δείγματα από κατανομή με διαφορετική διακύμανση.

Για να αποφύγουμε λοιπόν τέτοια λάθη, θα πρέπει η ελεγχοσυνάρτηση μας να δίνει όσο γίνεται μικρότερη σημασία στην κατανομή της πραγματικής συνάρτησης του πληθυσμού και αν είναι δυνατόν η κατανομή της να μην εξαρτάται από παραμέτρους. Σημειώνουμε πως με τη μέθοδο Bootstrap ουσιαστικά προσεγγίζουμε αυτή την κατανομή που μας είναι άγνωστη και επομένως δεν είναι εύκολο να δούμε αν η ελεγχοσυνάρτηση έχει αυτή την ιδιότητα ή όχι. Μια ελεγχοσυνάρτηση (και γενικότερα μια στατιστική συνάρτηση) με την ιδιότητα ότι η κατανομή της δεν εξαρτάται από τις παραμέτρους του πληθυσμού ονομάζεται pivotal. Για παράδειγμα, ξέρουμε πως ασυμπτωτικά η κατανομή της μέσης τιμής είναι κανονική με κάποια μέση τιμή μ και διακύμανση σ^2/n. Αντίθετα η ελεγχοσυνάρτηση $\sqrt{n}\left(\bar{x}-\mu\right)/\sigma$ ακολουθεί ασυμπτωτικά την τυποποιημένη κανονική κατανομή. Επομένως η ελεγχοσυνάρτηση αυτή είναι απαλλαγμένη από όποια προβλήματα έχουν να κάνουν με το ότι προσομοιώνουμε αντί για την πραγματική κατανομή του πληθυσμού από την εμπειρική κατανομή για τους σκοπούς της μεθόδου Bootstrap.

Για να γίνει πιο κατανοητή πόσο σημασία έχει η ελεγχοσυνάρτηση να είναι pivotal ας κοιτάξουμε τον πίνακα 5.2, όπου παρουσιάζονται τα αποτελέσματα ενός πειράματος.

Σε αυτόν τον πίνακα κάνουμε έλεγχο για μια μέση τιμή. Τα δεδομένα είναι προσομοιωμένα από εκθετική κατανομή με μέση τιμή θ και στη συνέχεια κάνουμε έλεγχο της μηδενικής υπόθεσης ότι η μέση τιμή είναι θ έναντι της εναλλακτικής ότι είναι διάφορη του θ. Χρησιμοποιώντας δύο ελεγχοσυναρτήσεις αλλά και το ασυμπτωτικό αποτέλεσμα του κεντρικού οριακού θεωρήματος (δηλαδή το t-test) μετρήσαμε πόσες φορές στις 10000 επαναλήψεις απορρίψαμε τη μηδενική υπόθεση σε επίπεδο στατιστικής σημαντικότητας 5%. Οι δύο ελεγχοσυναρτήσεις ήταν οι $T_1 = |\bar{x} - \theta|$ και $T_2 = \sqrt{n}\frac{\bar{x}-\theta}{s}$. Από τον ορισμό και τον τρόπο που κατασκευάσαμε τους ελέγχους bootstrap περιμένουμε πως το ποσοστό θα είναι κοντά στο 5%. Βέβαια αυτό το ποσοστό θα επιτευχθεί όταν ο αριθμός των επαναλήψεων bootstrap είναι σχετικά μεγάλος. Στην πράξη είναι και άλλοι παράγοντες που θα οδηγήσουν το μέγεθος του ελέγχου λίγο πιο μακριά από το 5%. Τέλος στον πίνακα έχουμε και την περίπτωση που χρησιμοποιείται το κλασικό t-test με βάση το κεντρικό οριακό θεώρημα περί κανονικότητας της ελεγχοσυνάρτησης, οπότε απορρίπτουμε τη μηδενική υπόθεση όταν $\left|\sqrt{n}\frac{\bar{x}-\theta}{s}\right| \geq 1.964$.

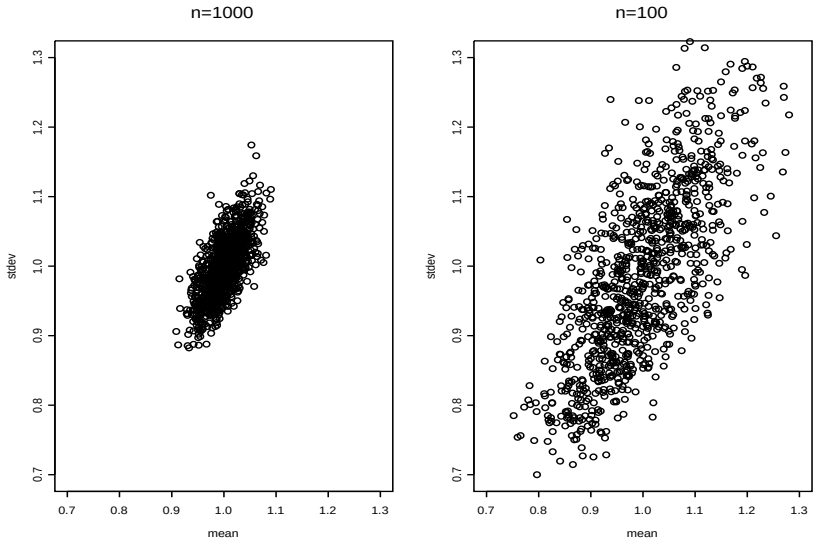
Πρώτα από όλα να παρατηρήσουμε πως το κλασικό t-test αποτυγχάνει ιδιαίτερα για μικρά μεγέθη δείγματος. 'Ετσι και αλλιώς αυτή η προσέγγιση δεν εξαρτάται από τον αριθμό των bootstrap δειγμάτων που θα πάρουμε για αυτό εμφανίζουμε μόνο μια τιμή για κάθε μέγεθος του δείγματος. 'Οσο αυξάνει το μέγεθος του δείγματος

		$B = 100$		$B = 500$		
	n	T_1	T_2	T_1	T_2	t-test
	10	0.14740	0.07100	0.14250	0.06890	0.12880
	20	0.10550	0.06190	0.10680	0.06180	0.09750
$\theta = 1$	50	0.06900	0.05480	0.07430	0.05380	0.07120
	100	0.05830	0.05130	0.06100	0.05040	0.06060
	250	0.05250	0.05070	0.05460	0.05020	0.05320
	n	$B = 100$		$B = 500$		
	10	0.14320	0.06810	0.14380	0.06550	0.12640
	20	0.10790	0.06490	0.10400	0.05880	0.09440
$\theta = 3$	50	0.07750	0.05560	0.07540	0.05670	0.07320
	100	0.06530	0.05440	0.05790	0.05280	0.05650
	250	0.05540	0.05150	0.05080	0.05140	0.05190

Πίνακας 5.2: Το ποσοστό των φορών που απορρίψαμε τη μηδενική υπόθεση ότι η μέση τιμή του πληθυσμού είναι θ έναντι δίπλευρης εναλλακτικής υπόθεσης. Τα δεδομένα είχαν προσομοιωθεί από εκθετική κατανομή με μέση τιμή θ. Χρησιμοποιήθηκαν οι 3 ελεγχοσυναρτήσεις που αναφέραμε και 2 διαφορετικά πλήθη Bootstrap επαναλήψεων $B = 100$ και 500.

ο έλεγχος πλησιάζει το 5%. Ανάμεσα στις 2 ελεγχοσυναρτήσεις για τους ελέγχους bootstrap αυτή που είναι pivotal συμπεριφέρεται πολύ καλύτερα, δηλαδή το μέγεθος του ελέγχου είναι κοντά στο 5% που θέλουμε, πολύ πιο γρήγορα. Η αιτία είναι πως η ελεγχοσυνάρτηση δεν εξαρτάται τόσο πολύ από το γεγονός πως η κατανομή κάτω από την οποία προσομοιώνουμε από τη μηδενική υπόθεση μπορεί να περιέχει σφάλμα. Το γράφημα 5.3 δείχνει την κατάσταση. Προσομειωθήκαν 1000 δείγματα από την εκθετική με μέση τιμή 1. Το διάγραμμα δείχνει τη μέση τιμή και την τυπική απόκλιση κάθε δείγματος. Δεδομένου πως τα δείγματα προέρχονται από τη μηδενική υπόθεση προκύπτει πως το σημείο (1,1) είναι το σημείο όπου όλα τα δείγματα έπρεπε να είναι, στην ιδανική περίπτωση. Παρόλα αυτά υπάρχει μεταβλητότητα γύρω από αυτό το σημείο. Οι έλεγχοι bootstrap διορθώνουν τη μέση τιμή, δηλαδή προσομοιώνουμε από μια κατανομή με μέση τιμή 1, αλλά η διακύμανση δεν είναι αυτή που θα έπρεπε. Στην περίπτωση της ελεγχοσυνάρτησης T_1 ουσιαστικά η κατανομή της ελεγχοσυνάρτησης εξαρτάται από την διακύμανση του δείγματος που σύμφωνα με τη μεθοδολογία Bootstrap γίνεται και διακύμανση του πληθυσμού. Επομένως ουσιαστικά η κατανομή της ελεγχοσυνάρτησης δεν προσεγγίζεται καλά αφού έχουμε κάποιο σφάλμα δειγματοληψίας και από επανάληψη σε επανάληψη διαφέρει. Από την άλλη μεριά η ελεγχοσυνάρτηση T_2 περιμένουμε να μην εξαρτάται από τη διακύμανση του δείγματος και άρα του πληθυσμού και επομένως μπορεί καλύτερα να ελέγξει την υπόθεση που μελετάμε.

Το γράφημα 5.3 είναι αποκαλυπτικό για το ότι με τη μέθοδο bootstrap στην πραγματικότητα έχουμε πάντα μια απόκλιση από τον πραγματικό πληθυσμό, ιδιαίτερα για μικρά μεγέθη δείγματος.



Γράφημα 5.3: Μέση τιμή και τυπική απόκλιση για δείγματα διαφορετικού μεγέθους προσομοιωμένα από εκθετική κατανομή με μέση τιμή 1. Παρατηρείστε την πολύ μεγαλύτερη διασπορά των σημείων όταν το μέγεθος του δείγματος είναι πιο μικρό.

Βέβαια είναι πολύ δύσκολο να βρει κανείς μια ελεγχοσυνάρτηση που να είναι pivotal για κάθε περίπτωση ελέγχου και τις περισσότερες φορές για να γίνει αυτό χρειάζεται πολύ καλή γνώση του προβλήματος. Υπάρχουν έλεγχοι όπου δεν υπάρχουν pivotal ελεγχοσυναρτήσεις. Συνεπώς μια ιδιότητα που θέλουμε για τις ελεγχοσυναρτήσεις μας, ιδιαίτερα σε ελέγχους bootstrap, είναι να εξαρτώνται ως προς την κατανομή τους όσο γίνεται λιγότερο από την κατανομή του πληθυσμού. Από τον πίνακα 5.2 μπορούμε να παρατηρήσουμε ότι όσο αυξάνει ο αριθμός των επαναλήψεων τόσο ο έλεγχος προσεγγίζει το πραγματικό επίπεδο στατιστικής σημαντικότητας και άρα η συμβουλή είναι να χρησιμοποιούμε αρκετές επαναλήψεις.

5.4 Τυπικά Σφάλματα

Η μέθοδος bootstrap είναι πολύ χρήσιμη για τον υπολογισμό τυπικών σφαλμάτων διαφόρων εκτιμητριών και γενικά στατιστικών συναρτήσεων. Για να εκτιμήσουμε λοιπόν ένα τυπικό σφάλμα μιας ποσότητας $\hat{\theta} = T(X_1, X_2, ..., X_n)$ του δείγματος ακολουθούμε την εξής διαδικασία:

- Δημιούργησε k bootstrap δείγματα, με δειγματοληψία με επανάθεση
- Για κάθε bootstrap δείγμα $(X_1^*, X_2^*, ..., X_n^*)$ (χρησιμοποιούμε το * για να μην το μπερδεύουμε με το αρχικό δείγμα) υπολόγισε την τιμή $\hat{\theta}_i^* = T(X_1^*, X_2^*, ..., X_n^*)$ η οποία είναι η τιμή της συνάρτησής μας στο i bootstrap δείγμα.

- Εκτίμησε το τυπικό σφάλμα της $\hat{\theta}$ ως

$$se(\hat{\theta}) = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{\theta}_i^* - \bar{\hat{\theta}}^*\right)^2} \quad ,$$

όπου $\bar{\hat{\theta}}^* = \frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_i^*$, δηλαδή χρησιμοποιούμε την τυπική απόκλιση των τιμών $\hat{\theta}_i^*$.

Πρέπει να παρατηρήσουμε εδώ πως γενικά η $\bar{\hat{\theta}}^*$ διαφέρει από την $\hat{\theta}$ και αυτή τη διαφορά θα τη χρησιμοποιήσουμε για να εκτιμήσουμε τη μεροληψία της εκτιμήτριας $\hat{\theta}$.

Η κεντρική ιδέα πίσω από τον προηγούμενο τύπο είναι πως οι εκτιμήτριες $\bar{\hat{\theta}}^*$ και $\hat{\theta}$ έχουν διακύμανση περίπου ίδια (κάτω από κάποιες υποθέσεις που συνήθως πληρούνται). Ας δούμε ένα παράδειγμα.

Παράδειγμα 5.2 (συνέχεια): Ας χρησιμοποιήσουμε τις 9 παρατηρήσεις με τον χρόνο επιβίωσης των ποντικιών και ας υποθέσουμε πως θέλουμε να εκτιμήσουμε το τυπικό σφάλμα της δειγματικής διαμέσου. Από τα δεδομένα βρίσκουμε εύκολα πως η διάμεσος είναι 40. Στην πραγματικότητα αυτή την τιμή δεν την χρειαζόμαστε κάπου.

Στη συνέχεια κατασκευάζουμε 19 δείγματα bootstrap τα οποία μπορείτε να δείτε στον πίνακα 5.3. Παρατηρείστε πως στο πρώτο δείγμα η τιμή 40 εμφανίζεται 4 φορές. Στην τελευταία στήλη του πίνακα 5.3 μπορεί κάποιος να δει τη διάμεσο για τα δεδομένα του συγκεκριμένου δείγματος

Μόλις τελειώσει λοιπόν η παραγωγή των δειγμάτων και ο υπολογισμός των διαμέσων τους, μπορούμε πια να χρησιμοποιήσουμε αυτές τις 19 τιμές για να εκτιμήσουμε το τυπικό σφάλμα της δειγματικής διαμέσου με την τυπική απόκλιση αυτών των 19 παρατηρήσεων. Βρίσκουμε πως $\bar{\hat{\theta}}^* = 35.95$ με τυπικό σφάλμα $se(\hat{\theta}) = 7.72$

Αξίζει σε αυτό το σημείο να αναφερθεί πως για την περίπτωση της εκτίμησης ενός τυπικού σφάλματος με τη μέθοδο bootstrap, ακόμα και μικρός αριθμός δειγμάτων κοντά στο 20 είναι αρκετός για να μας δώσει μια πληροφορία για την τάξη μεγέθους του σφάλματος. Αν το ενδιαφέρον περιορίζεται στην εκτίμηση του τυπικού σφάλματος 200 επαναλήψεις είναι, εκτός από εξαιρέσεις, ικανοποιητικές. Αν όμως τα τυπικά σφάλματα χρησιμοποιηθούν και για άλλους σκοπούς περισσότερες επαναλήψεις θα δώσουν καλύτερη εκτίμηση (π.χ. για τη δημιουργία διαστημάτων εμπιστοσύνης χρειαζόμαστε περισσότερες επαναλήψεις).

Τέλος χρησιμοποιώντας τη μέθοδο bootstrap μπορούμε να εκτιμήσουμε τη μεροληψία της εκτιμήτριας ως

$$bias(\hat{\theta}) = \bar{\hat{\theta}}^* - \hat{\theta}$$

Θα πρέπει εδώ να συγκρίνουμε τη μέθοδο bootstrap με τη μέθοδο jackknife ως προς τις εκτιμήσεις τυπικών σφαλμάτων και της μεροληψίας. 'Οπως είδαμε, και οι 2 μέθοδοι μας δίνουν εκτιμήσεις για το τυπικό σφάλμα και τη μεροληψία μιας εκτιμήτριας. Στην πράξη η μέθοδος jackknife χρειάζεται να επαναληφθεί ακριβώς

Δείγμα	Παρατηρήσεις Δείγματος									$\hat{\theta}_i^*$
1	40	27	40	46	40	40	27	10	27	40
2	27	40	27	27	27	10	27	31	52	27
3	40	27	40	46	31	27	40	31	31	31
4	27	146	10	40	46	10	10	46	10	27
5	146	46	52	52	27	146	52	31	52	52
6	46	146	46	31	46	31	40	10	40	40
7	27	27	27	27	40	27	10	31	46	27
8	146	40	31	52	27	40	10	31	27	31
9	52	146	40	146	52	46	27	10	27	46
10	10	31	10	27	146	31	10	27	46	27
11	10	146	52	146	27	46	27	40	40	40
12	40	146	146	40	31	27	27	27	27	31
13	27	146	40	52	40	27	27	52	40	40
14	40	10	27	27	27	40	52	31	40	31
15	10	10	31	52	27	146	46	27	27	27
16	46	27	27	46	31	146	46	46	52	46
17	27	40	31	52	27	40	146	52	10	40
18	146	10	52	40	52	27	27	31	52	40
19	52	27	146	40	27	27	146	27	40	40

Πίνακας 5.3: 19 δείγματα bootstrap και η εκτίμηση της διαμέσου για καθένα από αυτά.

n φορές, ενώ η μέθοδο bootstrap χρειάζεται έναν αριθμό δειγμάτων που εμείς καθορίζουμε αλλά που συνήθως είναι μεγαλύτερος του 100. Αν επομένως το μέγεθος του δείγματος είναι μικρό χρειαζόμαστε μικρότερο υπολογιστικό φόρτο από ότι με τη μέθοδο bootstrap. Από την άλλη η μέθοδο bootstrap δουλεύει σε μερικές περιπτώσεις που η jackknife αποτυγχάνει.

Επίσης είναι σημαντικό να παρατηρήσουμε πως η μέθοδος bootstrap στην ουσία προσεγγίζει την κατανομή της στατιστικής συνάρτησης που μας ενδιαφέρει, και συνεπώς οι τιμές που έχουμε μας δίνουν μια καλή εικόνα για τη μορφή της κατανομής, πχ αν είναι συμμετρική ή όχι, αν έχει μεγάλες ουρές κλπ. Επομένως με τη μέθοδο bootstrap μπορούμε να εκτιμήσουμε πολύ περισσότερα πράγματα από την τυπική απόκλιση και τη μεροληψία. Στην ουσία μπορούμε να εκτιμήσουμε την ίδια την κατανομή.

5.5 Διαστήματα Εμπιστοσύνης

Με τη μέθοδο bootstrap μπορούμε να κατασκευάσουμε και διαστήματα εμπιστοσύνης για τις στατιστικές συναρτήσεις που μας ενδιαφέρουν. Στη βιβλιογραφία έχουν προταθεί αρκετές μέθοδοι οι οποίες βασισμένες στο bootstrap κατασκευάζουν διαστήματα εμπιστοσύνης. Θα αναφέρουμε εδώ μερικές από αυτές αφού πρώτα υπενθυμίσουμε κάποια λίγα πράγματα σχετικά με τα διαστήματα εμπιστοσύνης.

'Οταν μιλάμε για ένα $(1-\alpha)$ διάστημα εμπιστοσύνης εννοούμε πως η πιθανότητα το διάστημα αυτό να περιέχει την πραγματική αλλά άγνωστη τιμή είναι ακριβώς $(1-\alpha)$. Είναι επίσης συνηθισμένο στη στατιστική να κατασκευάζουμε διαστήματα εμπιστοσύνης τα οποία να αφήνουν και στις δύο ουρές την ίδια πιθανότητα $\alpha/2$. Για παράδειγμα αν $\theta^{(a)}$ είναι το α ποσοστιαίο σημείο δηλαδή $P(\Theta \leq \theta^{(a)}) = a$ τότε συνήθως τα διαστήματα εμπιστοσύνης κατασκευάζονται ως $\theta^{(a/2)}, \theta^{(1-a/2)}$ οπότε υπάρχει πιθανότητα $\alpha/2$ σε κάθε ουρά. Προφανώς αυτά δεν είναι τα μοναδικά διαστήματα που μπορεί να κατασκευάσει κανείς και μόνο αν η κατανομή είναι συμμετρική μπορεί να το δικαιολογήσει πως είναι τα διαστήματα με το μικρότερο μήκος.

Ας δούμε τώρα πως κατασκευάζουμε διαστήματα εμπιστοσύνης με τη μέθοδο bootstrap.

5.5.1 Κλασικά bootstrap διαστήματα εμπιστοσύνης

'Ενας τρόπος για να κατασκευάσουμε τα διαστήματα είναι ο εξής:

'Ενα $(1-\alpha)-$διάστημα εμπιστοσύνης για την στατιστική συνάρτηση θ είναι το

$$\hat{\theta} \pm \Phi^{-1}(1-a/2)se(\hat{\theta})$$

όπου $\hat{\theta}$ είναι η εκτιμήτρια από το δείγμα

$\Phi^{-1}(1-a/2)$ είναι το $(1-\alpha/2)$ ποσοστιαίο σημείο από την τυποποιημένη κανονική κατανομή,

$se(\hat{\theta})$είναι το τυπικό σφάλμα της $\hat{\theta}$ υπολογισμένο με τη μέθοδο bootstrap όπως είδαμε στην προηγούμενη ενότητα.

Τα παραπάνω διαστήματα εμπιστοσύνης είναι συμμετρικά ως προς $\hat{\theta}$ και στηρίζονται στην παραδοχή ότι η κατανομή της εκτιμήτριας $\hat{\theta}$ είναι η κανονική. Στην

πράξη αυτή η υπόθεση δεν είναι κακή αλλά μπορεί να είναι προβληματική αν το μέγεθος του δείγματος είναι μικρό ή η μορφή της $\hat{\theta}$ δεν είναι απλή. Θα δούμε αργότερα ένα τέτοιο παράδειγμα. Μοιάζουν δε με τα κλασικά διαστήματα εμπιστοσύνης για μια μέση τιμή.

Για το παράδειγμα με τη διάμεσο που είδαμε πριν έχουμε ότι $\hat{\theta} = 40$, $\Phi^{-1}(1 - a/2) = 1.965$ (για $\alpha = 5\%$), $se(\hat{\theta}) = 7.72$, επομένως ένα 95% διάστημα εμπιστοσύνης είναι το διάστημα $(24.84, 55.16)$.

5.5.2 Bootstrap t-διαστήματα εμπιστοσύνης

Το προηγούμενο διάστημα εμπιστοσύνης στηρίχτηκε σε μια αυθαίρετη παραδοχή πως η κατανομή της στατιστικής συνάρτησης που μας ενδιαφέρει είναι η κανονική και για αυτό χρησιμοποιήσαμε τα ποσοστιαία σημεία της κανονικής κατανομής. 'Ομως αφού η μέθοδος bootstrap μας δίνει τη δυνατότητα να εκτιμήσουμε την κατανομή της ελεγχοσυνάρτησης γιατί να πρέπει να καταφύγουμε σε αμφίβολης ποιότητας ασυμπτωτικά αποτελέσματα τα οποία είναι καταδικασμένα να μην δουλέψουν ειδικά για μικρά δείγματα;

Επομένως μπορούμε να κατασκευάσουμε καλύτερα διαστήματα εμπιστοσύνης αν αντί να χρησιμοποιήσουμε το ποσοστιαίο σημείο της κανονικής κατανομής χρησιμοποιήσουμε τα ποσοστιαία σημεία της κατανομής της ελεγχοσυνάρτησης τα οποία θα πρέπει να εκτιμήσουμε από την κατανομή που έχουμε σχηματίσει με τη μέθοδο bootstrap. Αν συμβολίσουμε με $D(a)$ το a-ποσοστιαίο σημείο της κατανομής τότε μπορούμε να το εκτιμήσουμε ως

$$\{\#\ \text{τιμών}\ Z(\theta_i^*) \leq D(a)\}/\mathrm{B}{=}\alpha$$

όπου B είναι ο αριθμός των bootstrap επαναλήψεων και $Z(\theta_i^*) = \frac{\theta_i^* - \hat{\theta}}{s_i}$, όπου s_i είναι η τυπική απόκλιση του i bootstrap δείγματος, δηλαδή $Z(\theta_i^*)$ είναι μια μορφή τυποποιημένων τιμών της θ_i^*. Δηλαδή απλά εκτιμούμε το ζητούμενο ποσοστιαίο σημείο με το ποσοστιαίο σημείο των τυποποιημένων τιμών της θ_i^*. 'Ετσι το διάστημα εμπιστοσύνης γίνεται

$$\hat{\theta} + D(a/2)se(\hat{\theta}),\ \hat{\theta} + D(1 - a/2)se(\hat{\theta})$$

(παρατηρείστε πως περιμένουμε πως $D(a/2)$ θα είναι αρνητικό). Αυτό το διάστημα εμπιστοσύνης δεν είναι συμμετρικό κατά ανάγκη.

Για το παράδειγμα με τη διάμεσο που χρησιμοποιήσαμε και στο προηγούμενο διάστημα εμπιστοσύνης βρίσκουμε, από 1000 επαναλήψεις πως τα ποσοστιαία σημεία $D(0.025) = -0.6453$ και $D(0.975) = 1.5518$ οπότε το διάστημα εμπιστοσύνης γίνεται

$$40 - (0.6453)(7.72), 40 + (1.5518)(7.72) = (35.0183, 51.9799)$$

Θα δούμε αργότερα, αλλά το επισημαίνουμε από τώρα, πως για να εκτιμήσει κανείς με ακρίβεια ποσοστιαία σημεία χρειάζεται αρκετά μεγάλο αριθμό επαναλήψεων. Τα διαστήματα εμπιστοσύνης που μόλις είδαμε δεν είναι ακριβή αλλά διορθώνουν αρκετά την κατάσταση σε σχέση με τα διαστήματα εμπιστοσύνης που βασίζονται σε μια αυθαίρετη αποδοχή της κανονικότητας που είδαμε στην προηγούμενη παράγραφο.

Θα πρέπει επίσης να παρατηρήσουμε πως για το συγκεκριμένο παράδειγμα που χρησιμοποιήσαμε η μέθοδος είναι καταδικασμένη να μην δουλεύει καλά επειδή το μέγεθος του δείγματος είναι πολύ μικρό (μόλις 9) και η συνάρτηση που μελετάμε (διάμεσος) δεν έχει καθόλου λεία μορφή. Επίσης, σε αυτή την περίπτωση τα διαστήματα εμπιστοσύνης που μόλις περιγράψαμε, είναι επιρρεπή σε ακραίες τιμές.

5.5.3 Bootstrap διαστήματα εμπιστοσύνης βασισμένα σε ποσοστιαία σημεία

Αν θυμηθούμε τη διαδικασία με την οποία εκτιμήσαμε το τυπικό σφάλμα της $\hat{\theta}$, είχαμε πάρει πολλές τιμές $\hat{\theta}_i^*$ από τα bootstrap δείγματα. Αυτές οι τιμές είναι λογικό να υποθέσουμε ότι αποτελούν μια καλή προσέγγιση της κατανομής της $\hat{\theta}$, και επομένως μπορούμε από αυτή την προσέγγιση να κατασκευάσουμε διαστήματα εμπιστοσύνης. Η διαδικασία για να κατασκευάσουμε ένα $(1-a)$ διάστημα εμπιστοσύνης είναι:

- Βάζουμε σε αύξουσα σειρά τις τιμές $\hat{\theta}_i^*$
- Βρίσκουμε τα $a/2$ και $1-a/2$ ποσοστιαία σημεία της κατανομής αυτής

Το διάστημα εμπιστοσύνης που κατασκευάζουμε με αυτό τον τρόπο δεν είναι απαραίτητα συμμετρικό. Συνήθως τα διαστήματα εμπιστοσύνης με αυτή τη μέθοδο είναι σωστά στην πράξη (σωστά σημαίνει πως πραγματικά η πιθανότητα να περιέχουν την τιμή του πληθυσμού είναι κοντά στο $(1-a)$). Πρόβλημα μπορεί να παρουσιαστεί αν η κατανομή των $\hat{\theta}_i^*$ δεν είναι καλή προσέγγιση της πραγματικής κατανομής, όπως σε ένα παράδειγμα που θα δούμε σε λίγο.

Για το παράδειγμα μας με τη διάμεσο, παίρνοντας 99 δείγματα βρίσκουμε πως το 95% διάστημα εμπιστοσύνης είναι από την 3η μεγαλύτερη παρατήρηση μέχρι την 97η. 'Αρα το διάστημα που προκύπτει είναι το (31, 104). Παρατηρούμε μεγάλη διαφορά ανάμεσα στα 2 διαφορετικά διαστήματα που οφείλεται στη διακριτότητα της διαμέσου (δηλαδή παίρνει μόνο κάποιες συγκεκριμένες τιμές). Τα διαστήματα εμπιστοσύνης που βασίζονται σε ποσοστιαία σημεία αποτυγχάνουν όταν η προσεγγιστική κατανομή που χρησιμοποιούμε (αυτή δηλαδή των bootstrap τιμών $\hat{\theta}_i^*$) είναι διακριτή.

Αν επιστρέψουμε πίσω στο παράδειγμα 5.2 με τα ποντίκια και την κατανομή της διαφοράς d μπορούμε να βρούμε πως ένα 95% διάστημα εμπιστοσύνης είναι το $(-16.5, 88.5)$ το οποίο σαφώς και περιλαμβάνει το 0. Επομένως αν θυμηθούμε την αναλογία διαστημάτων εμπιστοσύνης και δίπλευρου ελέγχου η μηδενική υπόθεση ότι η διαφορά είναι 0 δεν μπορεί να απορριφθεί.

Τα διαστήματα εμπιστοσύνης βασισμένα στα ποσοστιαία σημεία μπορεί να αποδειχτεί πως είναι ακριβή αν η κατανομή της στατιστικής συνάρτησης είναι συμμετρική, σε αντίθετη περίπτωση δεν είναι ακριβή και επομένως υποεκτιμούν ή υπερεκτιμούν την πραγματική πιθανότητα να περιέχουν την πραγματική τιμή.

5.5.4 BC$_\alpha$ διαστήματα εμπιστοσύνης

Αν προσπαθήσουμε να ανακεφαλαιώσουμε τα προβλήματα των διαστημάτων εμπιστοσύνης που μέχρι τώρα συζητήσαμε θα δούμε πως αυτά είναι αφενός η μη

κανονικότητα της εκτιμήτριας, η τυχόν μεροληψία και η διαφορετική μορφή (συμμετρία και κύρτωση πχ). Μια ιδέα λοιπόν θα ήταν να επιφέρουμε αλλαγές στα άκρα του διαστήματος ώστε να λάβουν υπόψη τους αυτά τα προβλήματα. Έτσι τα άκρα ενός BC_α διαστήματος εμπιστοσύνης δίνονται από τα

$$\theta^{*(a_1)}, \theta^{*(a_2)},$$

όπου $\theta^{*(a)}$ είναι το a-ποσοστιαίο σημείο της κατανομής των bootstrap τιμών που έχουμε, και τα a_1 και a_2 υπολογίζονται ως

$$a_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(a)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(a)}\right)}\right)$$

και

$$a_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-a)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(1-a)}\right)}\right)$$

όπου $z^{(a)}$είναι το a ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής, πχ $z^{(0.95)} = 1.645$, και $\Phi(a)$ είναι η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής, πχ $\Phi(1.645) = 0.95$, δηλαδή ισχύει πως $\Phi^{-1}(a) = z^{(a)}$ Στην πραγματικότητα υπάρχουν δύο άγνωστες ποσότητες $\hat{z}_0, \hat{a}$ που πρέπει να εκτιμηθούν. Η πρώτη διορθώνει ως προς τη μεροληψία (bias correction) ενώ η δεύτερη διορθώνει ως προς την απόκλιση από την κανονική κατανομή. Παρατηρείστε πως αν $\hat{z}_0 = \hat{a} = 0$ τότε προκύπτει εύκολα πως $a_1 = a$ και $a_2 = 1 - a$ και επομένως τα διαστήματα εμπιστοσύνης ταυτίζονται με αυτά της μεθόδου των ποσοστιαίων σημείων.

Πως όμως υπολογίζουμε τις ποσότητες $\hat{z}_0$ και $\hat{a}$; Δύο απλοί τύποι για τον υπολογισμό τους είναι

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\hat{\theta}^*_i < \hat{\theta}}{B}\right),$$

δηλαδή είναι μια ποσότητα που βασίζεται στη μεροληψία της εκτιμήτριας. Αν η εκτιμήτρια $\hat{\theta}$ είναι η διάμεσος των bootstrap τιμών τότε $\hat{z}_0 = 0$.

Για να υπολογίσουμε το $\hat{a}$ που συχνά αναφέρεται ως επιταχυντής (acceleration parameter), χρησιμοποιούμε τον τύπο

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^2\right]^{3/2}},$$

όπου $\hat{\theta}_{(i)}$ είναι η i jackknife τιμή της παραμέτρου όταν αφαιρέσουμε την i παρατήρηση. Παρατηρείστε πως η ποσότητα αυτή μετρά την ασυμμετρία των jackknife ψευδοτιμών. Θα πρέπει να τονιστεί ότι πολλές φορές χρησιμοποιούμε τα παραπάνω διαστήματα χωρίς τη χρήση του επιταχυντή. Συνήθως αυτά τα απλούστερα διαστήματα εμπιστοσύνης ονομάζονται απλά BC και χρησιμοποιούνται συχνότερα από τα BC_α διαστήματα εμπιστοσύνης.

5.5.5 'Αλλα διαστήματα εμπιστοσύνης

Εκτός από αυτές τις μεθόδους για δημιουργία διαστημάτων εμπιστοσύνης, υπάρχουν και μερικές παραλλαγές οι οποίες βασισμένες σε αυτές προσπαθούν να διορθώσουν κάποια προβλήματά τους. Για παράδειγμα στα διαστήματα εμπιστοσύνης από τα ποσοστιαία σημεία, υπάρχει πρόβλημα στα άκρα του διαστήματος ιδίως όταν το μέγεθος του δείγματος είναι μικρό. 'Η τα συμμετρικά διαστήματα της πρώτης μεθόδου ίσως χρειάζονται βελτίωση. Μια ιδέα είναι να μετασχηματίσουμε τις τιμές σε κάποια συνάρτηση που να τις κάνει να προσεγγίζουν καλύτερα την κανονική κατανομή και να φτιάξουμε διαστήματα εμπιστοσύνης για τις μετασχηματισμένες τιμές χρησιμοποιώντας την κανονική προσέγγιση. Στη συνέχεια με τον αντίστροφο μετασχηματισμό επανερχόμαστε στις αρχικές μας τιμές.

Δεν θα περιγράψουμε τέτοιες παραλλαγές. Καλό είναι όταν σκοπός μας είναι η κατασκευή διαστημάτων εμπιστοσύνης να αυξάνουμε τον αριθμό των δειγμάτων που παίρνουμε ώστε να μεγαλώσουμε την αξιοπιστία του διαστήματος. Ειδικά η μέθοδος με τα ποσοστά μπορεί να δώσει πολύ λάθος αποτελέσματα αν πάρουμε μικρό αριθμό επαναλήψεων.

5.6 Σύγκριση της μεθόδου bootstrap με τη μέθοδο jackknife

Σε προηγούμενο κεφάλαιο είδαμε τη μέθοδο jackknife ως μια μέθοδο εκτίμησης της μεροληψίας και του τυπικού σφάλματος μιας εκτιμήτριας. Η μέθοδος Bootstrap χρησιμοποιείται για τους ίδιους σκοπούς αλλά επιπλέον μας επιτρέπει να εκτιμήσουμε την ίδια την κατανομή της συνάρτησης που χρησιμοποιούμε. Μάλιστα η μέθοδος bootstrap μπορεί να μας δώσει αποτελέσματα σε περιπτώσεις που η μέθοδος jackknife αποτυγχάνει (πχ διάμεσος). Από άποψη υπολογισμών η μέθοδος jackknife είναι προτιμότερη για μικρά μεγέθη δείγματος.

Μια άλλη διαφορά είναι πως η μέθοδος jackknife οδηγεί σε έναν αριθμό τον οποίο ο οποιοσδήποτε ερευνητής μπορεί να επαληθεύσει. Δεν συμβαίνει το ίδιο με τη μέθοδο bootstrap, όπου λόγοι τυχαιότητας μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα (αν και σίγουρα τα αποτελέσματα θα είναι πολύ κοντά το ένα στο άλλο). Σε γενικές γραμμές όμως οι δύο μέθοδοι αναμένονται να δώσουν τυπικά σφάλματα της ίδιας τάξη μεγέθους.

Παράδειγμα 5.3

Ας θυμηθούμε λίγο το παράδειγμα 4.2 με τα τρία χωριά και την εκτίμηση του τυπικού σφάλματος του συντελεστή Gini με τη χρήση της μεθόδου jackknife. Θα προσπαθήσουμε να χρησιμοποιήσουμε τη μέθοδο bootstrap για να εκτιμήσουμε τη διακύμανση του δειγματικού συντελεστή αλλά και για να κατασκευάσουμε διαστήματα εμπιστοσύνης. Συγκεκριμένα χρησιμοποιήσαμε 500 Bootstrap δείγματα καθώς και τη μέθοδο παραμετρικού bootstrap υποθέτοντας πως η κατανομή του πληθυσμού ήταν η εκθετική. Στην περίπτωση αυτή τα δείγματα προσομοιώθηκαν όχι από την εμπειρική κατανομή αλλά από εκθετική κατανομή με παράμετρο εκτιμημένη από το κάθε δείγμα. Υπενθυμίζουμε πως η παράμετρος της εκθετικής κατανομής μπορεί να εκτιμηθεί από το δειγματικό μέσο (ή τον αντίστροφο του ανάλογα με την παραμετροποίηση που έχουμε χρησιμοποιήσει). Τα αποτελέσματα μαζί με αυτά της

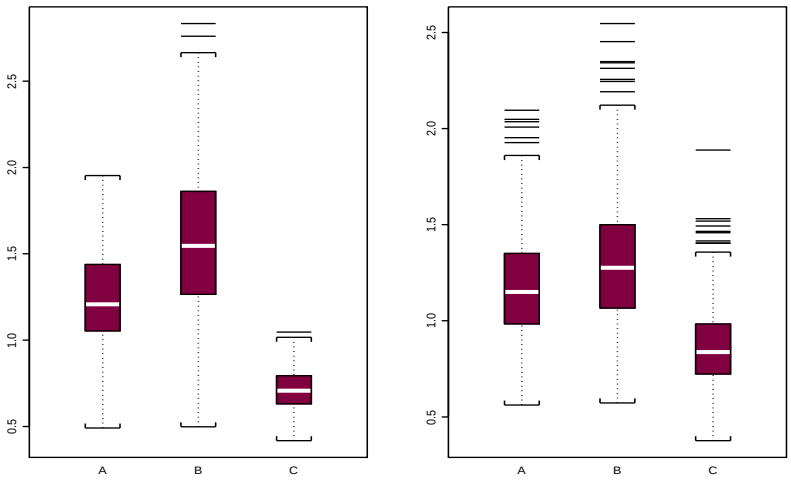
μεθόδου jackknife μπορεί κανείς να τα δει στον πίνακα 5.4

	Χωριά		
	Α	Β	Γ
Δειγματική εκτίμηση	1.281	1.620	0.732
Jackknife			
Μέσος ψευδοτιμών	1.281	1.620	0.732
Εκτίμηση τυπικού σφάλματος	0.2888	0.4664	0.1288
Κάτω όριο 95% διαστήματος εμπιστοσύνης	0.715	0.706	0.480
'Ανω όριο 95% διαστήματος εμπιστοσύνης	1.847	2.534	0.984
Bootstrap			
Μέσος bootstrap	1.242	1.534	0.705
Εκτίμηση τυπικού σφάλματος	0.260	0.446	0.117
Κάτω όριο 95% διαστήματος εμπιστοσύνης	0.686	0.663	0.498
'Ανω όριο 95% διαστήματος εμπιστοσύνης	1.712	2.383	0.945
Παραμετρικό bootstrap (εκθετική κατανομή)			
Μέσος bootstrap	1.185	1.339	0.862
Εκτίμηση τυπικού σφάλματος	0.271	0.321	0.200
Κάτω όριο 95% διαστήματος εμπιστοσύνης	0.696	0.791	0.515
'Ανω όριο 95% διαστήματος εμπιστοσύνης	1.722	2.064	1.346

Πίνακας 5.4: Αποτελέσματα χρησιμοποίησης των μεθόδων Jackknife και bootstrap στα δεδομένα των εισοδημάτων

Μπορεί κανείς να παρατηρήσει πως και οι δύο μέθοδοι δίνουν εκτιμήσεις του τυπικού σφάλματος αρκετά όμοιες, κάτι που τουλάχιστον μας κάνει σίγουρους για την τάξη μεγέθους του τυπικού σφάλματος. Μην ξεχνάμε πως πρόκειται για εκτίμηση του τυπικού σφάλματος και άρα περιμένουμε να υπάρχουν διαφορές. Ως προς τα διαστήματα εμπιστοσύνης αυτά διαφέρουν περισσότερο καθώς η υπόθεση της κανονικότητας των Bootstrap τιμών δεν φαίνεται να ισχύει. Για παράδειγμα από το γράφημα 5.4 μπορεί κανείς να δει πως η κατανομή ιδιαίτερα των χωριών Α και Γ έχει αρκετά μεγάλες ουρές. Από το γράφημα αυτό καθώς και από τον πίνακα 5.4 μπορεί να δει κανείς πως υπάρχουν διαφορές ανάμεσα στα τυπικά σφάλματα που βρήκαμε με απλό και παραμετρικό bootstrap. Η διαφορά είναι μεγαλύτερη για το χωριό Β. Παρατηρείστε επίσης πως η κατανομή του συντελεστή και για τα τρία χωριά έχει σχετικά διαφορετική όψη (όπως μπορούμε να δούμε από το boxplot). Η διαφορά οφείλεται πως επειδή το δείγμα είναι σχετικά μικρό, η απλή μέθοδος bootstrap είναι πιο επιρρεπής σε ακραίες τιμές, αν το δείγμα ήταν μεγαλύτερο θα είχαμε μεγαλύτερη συμφωνία. Βέβαια αν κανείς γνωρίζει την κατανομή του πληθυσμού είναι καλύτερα να χρησιμοποιήσει αυτή την πληροφορία και να κάνει παραμετρικό bootstrap. Επίσης αν αντί για την εκθετική κατανομή χρησιμοποιούσαμε κάποια άλλη κατανομή τότε τα αποτελέσματα μπορεί να ήταν ολότελα διαφορετικά ειδικά αν η κατανομή που υποθέταμε ήταν πολύ διαφορετική από την εμπειρική (πχ για το

παράδειγμα μας η κανονική κατανομή μοιάζει να είναι εντελώς διαφορετική από ότι τα δεδομένα δείχνουν).

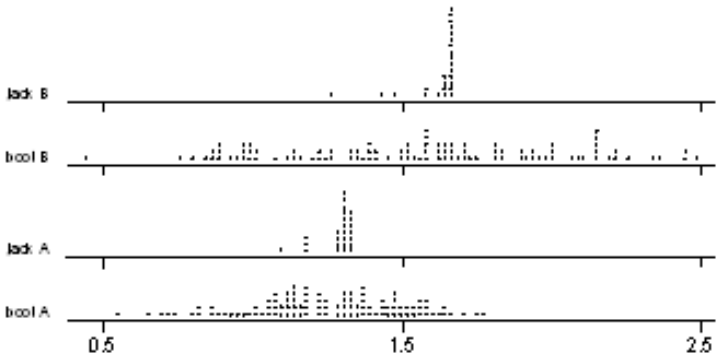


Γράφημα 5.4: Διάγραμμα πλαισίου και απολήξεων για τις 500 bootstrap τιμές. Το γράφημα a αναφέρεται στην περίπτωση του απλού bootstrap, ενώ το b στην περίπτωση παραμετρικού bootstrap

Επίσης ενδιαφέρον παρουσιάζει και το γράφημα 5.5 όπου κανείς μπορεί να δει τις jackknife ψευδοτιμές μαζί με τις αντίστοιχες τιμές των bootstrap δειγμάτων για τα χωριά A και B. Αν και η διασπορά των bootstrap τιμών αντικατοπτρίζει τη διασπορά της εκτιμήτριας κάτι τέτοιο δεν συμβαίνει με τις jackknife ψευδοτιμές. Παρατηρείστε όμως πως για το χωριό B και οι δύο μέθοδοι φανερώνουν τη μεγαλύτερη αβεβαιότητα που υπάρχει.

5.7 Τροποποιήσεις της μεθόδου Bootstrap

Στη βιβλιογραφία έχουν προταθεί αρκετοί τρόποι για να βελτιώσει κανείς μερικά αδύνατα σημεία της μεθόδου αυτής. Έτσι για παράδειγμα αν στην πραγματικότητα έχουμε κάποια πληροφορία ή κάποιο λόγο να πιστεύουμε πως η κατανομή του πληθυσμού έχει κάποια συγκεκριμένη μορφή τότε μπορούμε να χρησιμοποιήσουμε παραμετρικό bootstrap. Αν για παράδειγμα πιστεύουμε πως ο πληθυσμός μας είναι κανονικός αντί να χρησιμοποιήσουμε την εμπειρική κατανομή μπορούμε να εκτιμήσουμε από τα δεδομένα τις παραμέτρους της κανονικής κατανομής και να



Γράφημα 5.5: Διάγραμμα σημείων για τα χωριά A και B. Εμφανίζονται τόσο οι bootstrap τιμές όσο και οι jackknife ψευδοτιμές.

χρησιμοποιήσουμε την κανονική κατανομή. Στην ουσία δηλαδή χρησιμοποιούμε Monte Carlo.

5.7.1 Smoothed Bootstrap

'Ενας άλλος τρόπος να βελτιώσουμε τη μέθοδο bootstrap, και ιδίως το γεγονός ότι η εμπειρική κατανομή είναι διακριτή και μπορεί να πάρει μόνο n διαφορετικές τιμές, είναι να χρησιμοποιήσουμε μια άλλη πιο λεία εκτίμηση της συνάρτησης πυκνότητας πιθανότητας, όπως αυτή που προκύπτει από τη μέθοδο των kernels. Θυμηθείτε πως η εκτίμηση με τη μέθοδο των kernels ήταν απλά η κατανομή του αθροίσματος των τυχαίων μεταβλητών X και Y, όπου X ακολουθούσε την εμπειρική κατανομή και Y την κατανομή του kernel. Δηλαδή η τυχαία μεταβλητή Y κάνει την εμπειρική κατανομή από διακριτή συνεχής. Αντί δηλαδή να πάρουμε δείγμα από την εμπειρική κατανομή (που είναι διακριτή) παίρνουμε δείγμα από μια συνεχή κατανομή που την εκτιμά.

5.7.2 Επαναληπτική μέθοδος Bootstrap

Τέλος μια άλλη μέθοδος είναι η λεγόμενη επαναληπτική μέθοδος bootstrap. Με αυτή τη μέθοδο, έχοντας πάρει τις τιμές bootstrap $\hat{\theta}_i^*$, τις χρησιμοποιούμε για να ξανακάνουμε bootstrap, κάνοντας πάλι τυχαία δειγματοληψία με επανάθεση από αυτές. Με αυτό τον τρόπο κάνουμε ακόμα πιο λεία την εκτίμηση και βελτιώνουμε την ακρίβεια του τυπικού σφάλματος. Η μέθοδος μπορεί να βελτιώσει την εκτίμηση στην περίπτωση μη λείων συναρτήσεων, όπως για παράδειγμα είναι η διάμεσος.

5.7.3 Bayesian bootstrap

Μια παραλλαγή της μεθόδου είναι και η χρήση της Μπευζιανής bootstrap προσέγγισης. Συγκεκριμένα αντί να δίνουμε σε κάθε παρατήρηση πιθανότητα $1/n$ αντιστοιχούμε σε κάθε παρατήρηση μια πιθανότητα, έστω g_i η οποία έχει διάμεσο την τιμή $1/n$ αλλά μεταβάλλεται από επανάληψη σε επανάληψη. Η μέθοδος ονομάζεται Μπευζιανή καθώς έχει την ερμηνεία της Μπευζιανής προσέγγισης ως εκ των υστέρων κατανομής σε αντίθεση με την κλασική bootstrap προσέγγιση.

Συγκεκριμένα η μέθοδος δημιουργεί τα δείγματα bootstrap ως εξής:

'Εστω διάνυσμα παρατηρήσεων $x' = (x_1, x_2, ..., x_n)$. Η κλασική μέθοδος bootstrap δίνει σε κάθε σημείο πιθανότητα $1/n$, άρα αν συμβολίσουμε με p_i την πιθανότητα της i παρατήρησης έχουμε $p' = (p_1, p_2, ..., p_n) = (1/n, 1/n, ..., 1/n)$. Η Μπευζιανή bootstrap αντίθετα προχωράει ως εξής:

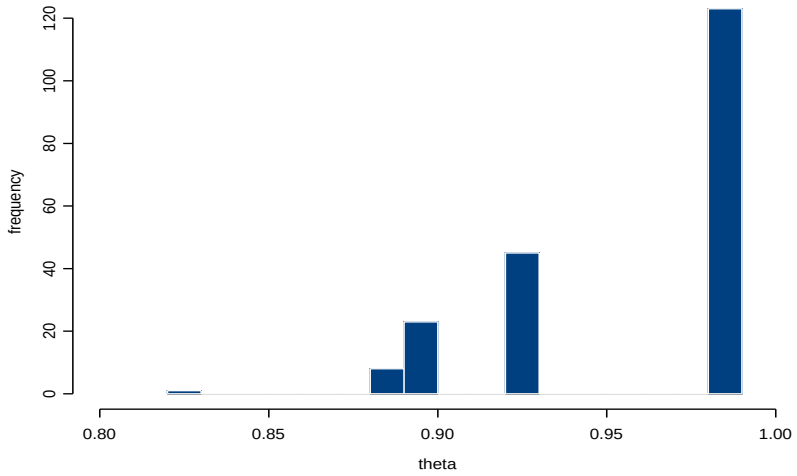
Προσομοιώνουμε $n-1$ τυχαίες μεταβλητές από την ομοιόμορφη κατανομή στο διάστημα (0,1) και τις διατάσσουμε. 'Εστω πως $u_{(1)}, u_{(2)}, ..., u_{(n-1)}$ οι διατεταγμένες αυτές τιμές. 'Εστω $u_{(0)} = 0, u_{(n)} = 1$. Υπολόγισε τις ποσότητες $g_i = u_{(i)} - u_{(i-1)}$. Το Bootstrap δείγμα θα δημιουργηθεί με δειγματοληψία με επανάθεση αλλά οι πιθανότητες κάθε παρατήρησης δεν θα είναι το p_i αλλά τα g_i που υπολογίσαμε. Μάλιστα για το επόμενο bootstrap δείγμα θα πρέπει να ξαναδημιουργήσουμε το διάνυσμα με τα g_i και αυτό θα επαναλαμβάνεται σε κάθε δείγμα. Μπορεί κανείς εύκολα να δει πως η αναμενόμενη τιμή της πιθανότητας κάθε παρατήρησης είναι ίση με $1/n$.

Η μέθοδος χρησιμοποιείται σε Μπευζιανές εφαρμογές καθώς η κατανομή της παραμέτρου που δημιουργούμε με τις bootstrap επαναλήψεις μπορεί να θεωρηθεί ως εκ των υστέρων κατανομή κατά τη Μπευζιανή προσέγγιση.

5.8 Περιπτώσεις που η μέθοδος bootstrap αποτυγχάνει

'Εστω ότι έχουμε δεδομένα $X_1, X_2, ..., X_n$ από μια ομοιόμορφη κατανομή στο διάστημα $(0, \theta)$. Γνωρίζουμε πως η εκτιμήτρια μεγίστης πιθανοφάνειας του θ είναι η μεγαλύτερη τιμή του δείγματος. 'Εστω ότι $\theta = 1$ και πως το μέγεθος του δείγματος είναι 50. Παράγοντας ένα δείγμα από την ομοιόμορφη κατανομή στο (0,1), βρήκαμε πως η μεγαλύτερη τιμή είναι 0.9832. Στη συνέχεια πήραμε 200 bootstrap δείγματα από το αρχικό δείγμα, και οι τιμές $\hat{\theta}_i^*$, i=1,..,200 που πήραμε φαίνονται στο γράφημα 5.6.

Χρησιμοποιώντας τη μέθοδο bootstrap η εικόνα που παίρνουμε είναι μάλλον άσχημη. Ο λόγος είναι πως στην πραγματικότητα προσπαθούμε να προσεγγίσουμε μια συνεχή κατανομή (ομοιόμορφη) με μια διακριτή (την εμπειρική). Δεδομένου και πως $\theta = \max X_i$ είναι μια συνάρτηση που χρησιμοποιεί πληροφορία μόνο από τη μεγαλύτερη παρατήρηση το αποτέλεσμα δεν είναι καθόλου ικανοποιητικό. Η πιθανότητα η τιμή 0.9832 να μην περιληφθεί σε ένα δείγμα bootstrap είναι $\left(1 - \frac{1}{n}\right)^n$ για ένα δείγμα μεγέθους n, οπότε γενικά περιμένουμε πως θα υπάρχει σε $1 - \left(1 - \frac{1}{n}\right)^n \to e^{-1} \approx 0.632$ των δειγμάτων. Στην πραγματικότητα την είχαμε στα 124 από τα 200 δείγματα. Γενικά η μέθοδος bootstrap αποτυγχάνει όταν έχουμε να εκτιμήσουμε ακραίες τιμές.



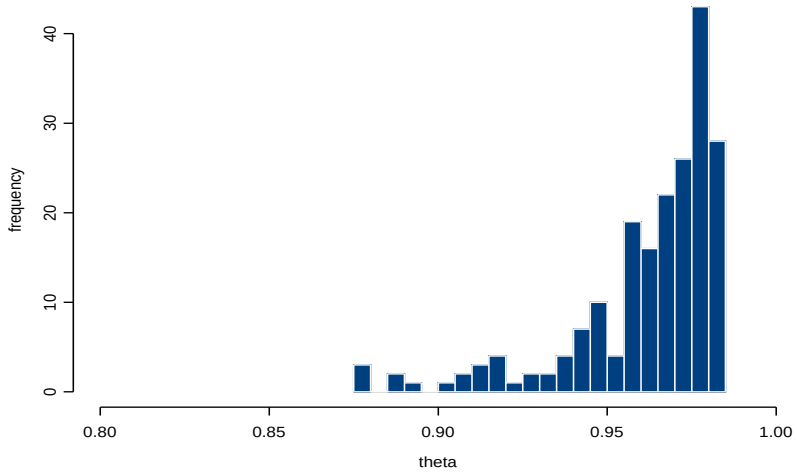
Γράφημα 5.6: Ιστόγραμμα 200 τιμών για το μέγιστο βασισμένο στη μέθοδο bootstrap

Μια λύση σε αυτές τις περιπτώσεις είναι να χρησιμοποιήσουμε parametric bootstrap. Δηλαδή στο παράδειγμα μας, αντί να γεννήσουμε τα δείγματα από την εμπειρική κατανομή θα μπορούσαμε να χρησιμοποιήσουμε μια ομοιόμορφη στο διάστημα $(0, \hat{\theta})$. Αυτό θα διόρθωνε αρκετά το πρόβλημα. Δείτε τώρα το ιστόγραμμα στο γράφημα 5.7 όταν χρησιμοποιήσαμε parametric bootstrap για την ίδια περίπτωση. Η κατανομή είναι αρκετά καλύτερη.

Επίσης πρόβλημα υπάρχει και αν προσπαθήσουμε να εκτιμήσουμε ακραία ποσοστιαία σημεία. Στην περίπτωση αυτή θα πρέπει να κάνουμε μεγάλο αριθμό επαναλήψεων για να μπορούμε να είμαστε σίγουροι για το αποτέλεσμά μας.

'Αλλες περιπτώσεις που η μέθοδος μπορεί να αποτύχει είναι όταν:

- Το μέγεθος του δείγματος είναι πολύ μικρό. Είπαμε και σε προηγούμενη ευκαιρία πως επειδή η μέθοδος bootstrap στην ουσία θεωρεί πως η εμπειρική κατανομή είναι μια καλή εκτίμηση της άγνωστης κατανομής του πληθυσμού. Για μικρά δείγματα κάτι τέτοιο δεν είναι αλήθεια καθώς η προσέγγιση της εμπειρικής κατανομής είναι μάλλον άσχημη. Συνεπώς τα bootstrap δείγματα που παίρνουμε θα περιέχουν συνέχεια το δειγματοληπτικό σφάλμα το οποίο εξαιτίας του μικρού μεγέθους του δείγματος θα είναι μεγάλο. Επίσης σε μικρά μεγέθη δείγματος, ο αριθμός διαφορετικών bootstrap δειγμάτων που μπορούμε να πάρουμε είναι σχετικά μικρός και άρα αν πάρουμε πολλές επαναλήψεις είναι πολύ πιθανό να πάρουμε περισσότερες από μια φορές το ίδιο δείγμα.



Γράφημα 5.7: Ιστόγραμμα 200 τιμών για το μέγιστο βασισμένο σε parametric bootstrap. Τα δείγματα προέρχονται από ομοιόμορφη κατανομή στο διάστημα $(0, \hat{\theta})$.

- Προσπαθήσουμε να εκτιμήσουμε ποσότητες οι οποίες δεν έχουν καλά ορισμένες ροπές. Για παράδειγμα αν τα δεδομένα προέρχονται από την κατανομή Cauchy ξέρουμε πως η αναμενόμενη τιμή της κατανομής δεν υπάρχει. Επομένως αν προσπαθήσουμε να χρησιμοποιήσουμε bootstrap για να εκτιμήσουμε το τυπικό σφάλμα της δειγματικής μέσης τιμής είμαστε καταδικασμένοι να αποτύχουμε.

- Το πείραμα που θα περιγράψουμε σκοπεύει στο να δείξει το γεγονός πως η μέθοδος Bootstrap μπορεί να αποτύχει να δώσει καλά αποτελέσματα σε μερικές περιπτώσεις. Η ιδέα είναι να χρησιμοποιηθεί parametric bootstrap για να εκτιμηθεί ένα ποσοστιαίο σημείο. Προκειμένου να έχουμε πλήρη έλεγχο στο πείραμα, τα δείγματα έχουν προσομοιωθεί από γνωστές κατανομές, και συγκεκριμένα από τυποποιημένη κανονική κατανομή και από την κατανομή t με 5 βαθμούς ελευθερίας. Για να δούμε καλύτερα το πρόβλημα έχουμε επίσης εκτιμήσει διάφορες άλλες ποσότητες για να συγκρίνουμε την ποιότητα της εκτίμησης με αυτή της μεθόδου bootstrap. 'Ετσι η ιδέα ήταν η εξής:

  Προσομοίωσαμε ένα δείγμα μεγέθους n από τη γνωστή κατανομή. Από το δείγμα αυτό πήραμε B bootstrap δείγματα και για κάθε bootstrap δείγμα βρήκαμε τις ποσότητες που ενδιαφέρουν (μέση τιμή, τυπική απόκλιση, διάμεσος, 99^o ποσοστιαίο σημείο). Στον πίνακα 5.5 υπάρχουν οι μέσες τιμές και οι αντίστοιχες τυπικές αποκλίσεις, δηλαδή οι bootstrap εκτιμήσεις των τυπικών αποκλίσεων των δειγματικών ποσοτήτων.

  'Οπως μπορεί κανείς να διακρίνει εύκολα η μέθοδος bootstrap καταφέρνει

Τυποποιημένη Κανονική κατανομή

		$\mu = 0$		M=0		$z_{0.99} = 2.3263$		$\sigma = 1$	
n	B	$\bar{x}$	$se_B(\bar{x})$	$\hat{M}$	$se_B(\hat{M})$	$\hat{z}_{0.99}$	$se_B(\hat{z}_{0.99})$	s	$se_B(s)$
10	100	0.023	0.274	0.040	0.353	1.512	0.565	0.966	0.203
10	500	0.000	0.320	0.012	0.387	1.459	0.565	0.971	0.236
100	100	0.007	0.107	-0.001	0.135	2.177	0.317	0.999	0.076
100	500	-0.002	0.103	-0.006	0.131	2.133	0.306	0.996	0.071
500	100	0.002	0.050	0.002	0.060	2.281	0.156	1.000	0.035
500	500	0.001	0.046	0.001	0.058	2.290	0.155	0.998	0.032

t με 5 βαθμούς ελευθερίας

		$\mu = 0$		$M = 0$		$z_{0.99} = 3.36$		$\sigma = 1.2909$	
n	B	$\bar{x}$	$se_B(\bar{x})$	$\hat{M}$	$se_B(\hat{M})$	$\hat{z}_{0.99}$	$se_B(\hat{z}_{0.99})$	s	$se_B(s)$
10	100	0.016	0.403	-0.002	0.394	1.952	1.013	1.226	0.421
10	500	0.057	0.446	0.055	0.411	2.021	1.023	1.210	0.390
100	100	0.000	0.126	-0.002	0.129	3.031	0.710	1.279	0.151
100	500	-0.004	0.128	-0.015	0.128	3.063	0.785	1.306	0.153
500	100	-0.001	0.058	0.000	0.058	3.291	0.411	1.290	0.090
500	500	-0.008	0.059	-0.012	0.059	3.321	0.357	1.295	0.078

Πίνακας 5.5: Εκτιμηθείσες ποσότητες βασισμένες σε Bδείγματα bootstrap

ακόμα και με μικρά δείγματα να εκτιμά σωστά τη μέση τιμή και την τυπική απόκλιση. Αποτυγχάνει όμως και μάλιστα παταγωδώς για το 99 ποσοστιαίο σημείο, ενώ για τη διάμεσο που είναι πιο κοντά στον κύριο όγκο των δεδομένων δεν τα πηγαίνει άσχημα. Ο πίνακας είναι μια καλή ένδειξη πως για να εκτιμήσουμε με τη μέθοδο bootstrap παραμέτρους που βασίζονται σε ακραίες τιμές χρειαζόμαστε πολύ μεγάλους αριθμούς επαναλήψεων. Παρατηρείστε πως το τυπικό σφάλμα για το ποσοστιαίο σημείο μικραίνει όσο μεγαλώνει το μέγεθος του δείγματος αρκετά πιο έντονα από ότι με τις άλλες ποσότητες.

Βέβαια και σε αυτές τις περιπτώσεις υπάρχουν στη βιβλιογραφία κάποιες παραλλαγές της μεθόδου bootstrap οι οποίες μπορούν να βελτιώσουν την ποιότητα της μεθόδου και για τις οποίες δεν θα αναφερθούμε περισσότερο.

5.9 Bootstrap για τη γραμμική παλινδρόμηση

Μέχρι τώρα συζητήσαμε την απλή περίπτωση εκτίμησης μιας παραμέτρου και συγκεκριμένα είχαμε μια μεταβλητή και προσπαθούσαμε να εκτιμήσουμε κάποια παράμετρο. Ας δούμε μια πιο σύνθετη περίπτωση, που αφορά τη γραμμική παλινδρόμηση. 'Εστω ότι έχουμε 2 μεταβλητές Y και X και θέλουμε να προσαρμόσουμε ένα γραμμικό μοντέλο της μορφής $Y = \alpha + \beta X$. Θα πρέπει να τονιστεί ότι η παρουσίαση μπορεί να γενικευτεί και στην περίπτωση περισσοτέρων από μια μεταβλητών.

Το ερώτημα είναι πως μπορούμε να πάρουμε δείγματα bootstrap σε μια τέτοια περίπτωση; Αν πάρουμε τα δείγματα μετά τα υπόλοιπα μπορούν εύκολα να γίνουν

με βάση τα όσα περιγράψαμε μέχρι τώρα.

Μια πρώτη ιδέα είναι να κάνουμε δειγματοληψία με επανάθεση από τα ζεύγη $(Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n)$. Για κάθε δείγμα εκτιμάμε τις ποσότητες που μας ενδιαφέρουν (πχ τα α και β ή το συντελεστή προσδιορισμού) και στη συνέχεια χρησιμοποιούμε αυτές τις τιμές για συμπερασματολογία σχετικά με τις παραμέτρους.

Εναλλακτικά θα μπορούσε κανείς να προσαρμόσει ένα μοντέλο στα πραγματικά δεδομένα, να βρει τα $\hat{a}, \hat{\beta}$ και στη συνέχεια να υπολογίσει τα κατάλοιπα $e_i = Y_i - \hat{a} - \hat{\beta} X_i$. Στη συνέχεια τα bootstrap δείγματα τα κατασκευάζουμε χρησιμοποιώντας τα πραγματικά X_i αλλά υπολογίζοντας τα Y_i^* ως $Y_i^* = \hat{a} + \hat{\beta} X_i + e_i^*$ όπου e_i^* είναι δείγματα bootstrap από τις τιμές e_i που πήραμε πριν.

Είναι άμεσα αντιληπτό πως με αυτό τον τρόπο μπορούμε να κάνουμε συμπερασματολογία για τις παραμέτρους του γραμμικού μοντέλου χωρίς την υπόθεση της κανονικότητας. Οι υπόλοιπες υποθέσεις χρειάζονται κυρίως γιατί έχουν να κάνουν με τη λογική της μεθόδου των ελαχίστων τετραγώνων που χρησιμοποιούμε για την εκτίμηση των παραμέτρων.

Στην πράξη και οι δύο μέθοδοι δίνουν παρόμοια αποτελέσματα. Για καθαρά φιλοσοφικούς λόγους που έχουν να κάνουν κυρίως με την υπόθεση της γραμμικής παλινδρόμησης πως ο πίνακας σχεδιασμού είναι από πριν γνωστός η μέθοδος με τα κατάλοιπα είναι στατιστικά πιο σωστή.

Παράδειγμα 5.4

'Ενας ορνιθολόγος κατέγραψε για 12 σπουργίτια την ηλικία τους σε μέρες και το μήκος των φτερών τους σε εκατοστά με σκοπό να ελέγξει αν υπάρχει μια γραμμική σχέση του μήκος των φτερών με την ηλικία. Τα δεδομένα μπορεί να τα δει κάποιος στον πίνακα 5.6.

1. Να εκτιμήσετε τα τυπικά σφάλματα των συντελεστών α και β της παλινδρόμησης και τη συνδιακύμανση τους.

2. Είναι η σταθερά διαφορετική της μονάδας;

3. Υπάρχει σχέση ανάμεσα στις 2 μεταβλητές;

4. Από προηγούμενη έρευνα είχε εκτιμηθεί πως η διακύμανση σ^2 είναι 0.05. Ισχύει κάτι τέτοιο στα δεδομένα μας;

Τα δεδομένα φαίνονται στο γράφημα 5.8 μαζί με την προσαρμοσθεισα με τη μέθοδο ελαχίστων τετραγώνων ευθεία. Παρατηρείστε πως πραγματικά εμφανίζεται μια γραμμική σχέση ανάμεσα στις 2 μεταβλητές.

Θα χρησιμοποιήσουμε τη μέθοδο με τα κατάλοιπα. Χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων βρίσκουμε πως $\hat{a} = 0.779$ και $\hat{\beta} = 0.266$ και άρα το μοντέλο είναι της μορφής

$$Y = 0.779 + 0.266X,$$

όπου X είναι η ηλικία (σε μέρες) και Y είναι το μήκος των φτερών (σε cm). Επομένως τα κατάλοιπα υπολογίζονται ως

$$\hat{e}_i = Y_i - 0.779 - 0.266X_i$$

Μήκος φτερών (σε cm)	Ηλικία (σε μέρες)	Κατάλοιπα
1.40	3	-0.176623
1.50	3	-0.076623
2.20	5	0.091415
2.40	6	0.025435
3.10	8	0.193473
3.20	9	0.027492
3.20	10	-0.238489
3.90	11	0.195530
4.10	12	0.129549
4.70	14	0.197588
4.50	15	-0.268393
5.20	17	-0.100355

Πίνακας 5.6: Τα δεδομένα για το παράδειγμα μαζί με τα κατάλοιπα του μοντέλου

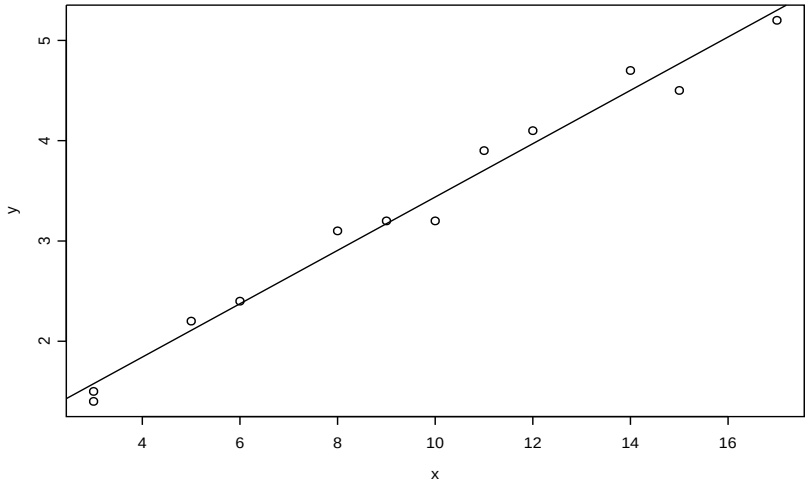
και μπορεί κανείς να τα δει στην τρίτη στήλη του πίνακα 5.6. Θα πρέπει εδώ να υπενθυμίσουμε πως η υπόθεση της κανονικότητας είναι αδιάφορη για την εφαρμογή της μεθόδου των ελαχίστων τετραγώνων. Στην περίπτωση μας όλη η συμπερασματολογία θα βασιστεί στο bootstrap και επομένως καμιά υπόθεση περί κανονικότητας των καταλοίπων δεν χρειάζεται.

Τα βήματα για το bootstrap στην περίπτωση αυτή είναι τα εξής:

- Πάρε ένα bootstrap δείγμα από τα κατάλοιπα χρησιμοποιώντας δειγματοληψία με επανάθεση. Το δείγμα θα έχει 12 τιμές έστω $e_1^*, e_2^*, ..., e_{12}^*$ και στη συνέχεια υπολόγισε $Y_i^* = \hat{a} + \hat{\beta}X_i + e_i^* = 0.779 + \beta 0.266X_i + e_i^*$ για κάθε τιμή. Παρατηρείστε πως χρησιμοποιούμε τα γνωστά X_i και τις τιμές των συντελεστών που εκτιμήσαμε πριν.
- Χρησιμοποίησε τα ζεύγη τιμών (Y_i^*, X_i) και εκτίμησε με τη μέθοδο ελαχίστων τετραγώνων τους συντελεστές της παλινδρόμησης και την εκτίμηση $\hat{\sigma}^2 = MSE$ της παλινδρόμησης αυτής
- Επανέλαβε αυτή τη διαδικασία πολλές φορές (πχ 1000).

Το αποτέλεσμα που θα πάρει κανείς είναι 1000 τιμές για τα $\hat{\alpha}, \hat{\beta}$ και $\hat{s}^2$ της παλινδρόμησης και σύμφωνα με τα όσα είπαμε πριν αυτές οι 1000 τιμές αποτελούν μια καλή προσέγγιση της πραγματικής κατανομής αυτών των ποσοτήτων και άρα μπορούν να χρησιμοποιηθούν για την εκτίμηση τυπικών σφαλμάτων, διαστημάτων εμπιστοσύνης κλπ.

Αυτή ακριβώς τη διαδικασία ακολουθήσαμε. Η επιλογή 1000 επαναλήψεων έγινε γιατί επιθυμούμε την κατασκευή διαστημάτων εμπιστοσύνης και άρα αρκετές επαναλήψεις είναι χρήσιμες για την κατασκευή των διαστημάτων αυτών. Ο πίνακας



Γράφημα 5.8: Διάγραμμα σημείων για τις 2 μεταβλητές του παραδείγματος

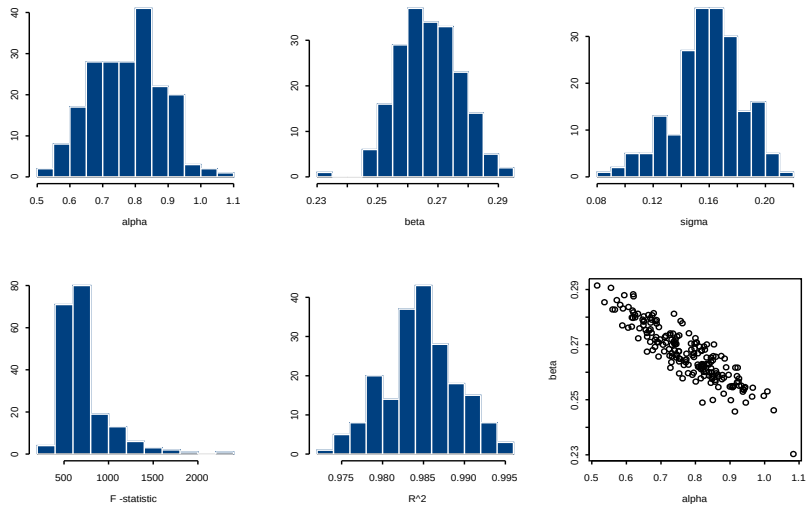
5.7 περιέχει αρκετές ποσότητες βασισμένες στις bootstrap αυτές επαναλήψεις. Εκτός από τις παραμέτρους του γραμμικού μοντέλου για τις οποίες υπάρχουν κάποια γνωστά αποτελέσματα, στον πίνακα εκτιμούμε τα τυπικά σφάλματα αλλά και διαστήματα εμπιστοσύνης και για άλλες πιο πολύπλοκες ποσότητες όπως το συντελεστή προσδιορισμού R^2 και το στατιστικό F που χρησιμοποιούμε για να ελέγξουμε τη σημαντικότητα του μοντέλου. Θα πρέπει να τονιστεί ότι για παράδειγμα η κατανομή του δειγματικού συντελεστή προσδιορισμού δεν είναι γενικά γνωστή ακόμα και στην περίπτωση που η υπόθεση της κανονικότητας ισχύει. Συνεπώς χρησιμοποιώντας bootstrap βρισκόμαστε στην πλεονεκτική θέση να μπορούμε να συμπερασματολογήσουμε και για τέτοιες ποσότητες για τις οποίες δεν υπάρχουν θεωρητικά αποτελέσματα.

			95% δ.ε.		95% δ.ε.	
	Μέση τιμή	Τυπική απόκλιση	κλασσική μέθοδος		ποσοστιαία σημεία	
$\hat{s}^2$	0.026	0.00776	0.011	0.041	0.012	0.042
$\hat{a}$	0.777	0.10961	0.563	0.991	0.554	0.986
$\hat{\beta}$	0.266	0.01052	0.245	0.286	0.246	0.286
$\hat{F}$	717.536	288.291	406.493	1566.891	417.432	1499.78
$\hat{R}^2$	0.984	0.0044	0.975	0.993	0.972	0.991

Πίνακας 5.7: Τυπικά σφάλματα και διαστήματα εμπιστοσύνης βασισμένα σε bootstrap

Η συνδιακύμανση των συντελεστών μπορεί να εκτιμηθεί απλά σαν η συνδιακύμανση των bootstrap τιμών τους και είναι -0.001 ενώ ο συντελεστής συσχέτισης μπορεί να εκτιμηθεί ως $Corr(\hat{a}, \hat{\beta}) = -.902$. Μπορεί επίσης κάποιος να δει ιστογράμματα και για όλες τις εκτιμήτριες, από τα οποία μπορεί να αποκτήσει μια εικόνα για την κατανομή των εκτιμητριών.

Από τον πίνακα 5.7 βλέπουμε πως τα 2 διαφορετικά διαστήματα εμπιστοσύνης σχεδόν συμπίπτουν, κάτι το οποίο οφείλεται στην περίπου κανονικότητα των τιμών (όπως φαίνεται και από το ιστόγραμμα (γράφημα 5.9)). Παρατηρούμε πως οι μέσοι των bootstrap δειγμάτων είναι πολύ κοντά στις τιμές που πήραμε από τη γραμμική παλινδρόμηση και άρα η μεροληψία που εκτιμάμε είναι πολύ μικρή.



Γράφημα 5.9: Ιστογράμματα των bootstrap τιμών για όλες τις ποσότητες που υπολογίσαμε. Επίσης ένα διάγραμμα σημείων για τα $\hat{\alpha}, \hat{\beta}$ που φανερώνει τη μεγάλη εξάρτηση των εκτιμητριών.

Με τις ποσότητες που υπολογίσαμε μπορούμε τώρα να απαντήσουμε τις ερωτήσεις. Βλέπουμε λοιπόν πως τα τυπικά σφάλματα του $\hat{a}$και του $\hat{\beta}$ είναι 0.1096 και 0.01052 αντίστοιχα ενώ η συνδιακύμανσή τους είναι -0.012 όπως είπαμε. Η συσχέτιση όμως είναι -0.89, δηλαδή έντονα αρνητική όπως μπορεί να φανεί και στο διάγραμμα σημειών (γράφημα 5.9.

Επειδή το 95% διάστημα εμπιστοσύνης για το a δεν περιέχει το 1, συνεπώς σε επίπεδο στατιστικής σημαντικότητας 5% απορρίπτουμε τη μηδενική υπόθεση πως $\alpha = 1$ έναντι της εναλλακτικής ότι διαφέρει. Θα μπορούσε κανείς να κάνει απευθείας bootstrap έλεγχο υπόθεσης αλλά κάτι τέτοιο είναι περιττό στην περίπτωση μας.

Η ύπαρξη σχέσης ανάμεσα στις 2 μεταβλητές αντιστοιχεί σε έλεγχο για το β. Το γεγονός ότι το διάστημα εμπιστοσύνης δεν περιέχει το 0, ισοδυναμεί με απόρριψη της υπόθεσης ότι $\beta = 0$ και άρα υπάρχει σχέση ανάμεσα στις μεταβλητές.

Τέλος για το s^2 έχουμε πως διαφέρει στατιστικά σημαντικά από το 0.05 καθώς το διάστημα εμπιστοσύνης δεν το περιέχει.

Είναι πολύ βασικό να παρατηρήσει κανείς ότι η συμπερασματολογία δεν βασίστηκε σε καμιά υπόθεση για τα κατάλοιπα. Επίσης παρατηρείστε πως γενικά όταν κάνουμε την υπόθεση της κανονικότητας, η συμπερασματολογία για το s^2 δεν είναι εύκολη καθώς η κατανομής της εκτιμήτριας δεν είναι γνωστή παρά μόνο αν ισχύει πως $\beta = 0$. Από τον πίνακα όμως 5.7 έχουμε μια καλή πληροοφορία για την άγνωστη διακύμανση του πληθυσμού και μπορούμε να προβούμε σε στατιστική συμπερασματολογία.

Επομένως η χρήση της μεθόδου bootstrap στη γραμμική παλινδρόμηση μας επιτρέπει να απαλλαχτούμε από την υπόθεση της κανονικότητας και να κάνουμε πλήρη στατιστική συμπερασματολογία για τις παραμέτρους χωρίς αυτή την υπόθεση.

Τελειώνοντας θα πρέπει να αναφέρουμε την περίπτωση παραμετρικού Bootstrap, όπου τα σφάλματα μπορούν να ακολουθούν οποιαδήποτε κατανομή. Βέβαια για να είναι λογικό το μοντέλο πρέπει η αναμενόμενη τιμή των σφαλμάτων να είναι μηδέν. Για παράδειγμα ας υποθέσουμε πως η κατανομή των σφαλμάτων είναι η $t - Student$ με άγνωστους βαθμούς ελευθερίας. Η υπόθεση της ομοσκεδαστικότητας ισχύει και πάλι. Μπορεί κανείς να δουλέψει ως εξής:

- *Βήμα 1ο:* Προσάρμοσε το μοντέλο με τη χρήση της μεθόδου ελαχίστων τετραγώνων. Θυμηθείτε πως η μέθοδος είναι ανεξάρτητη από την κατανομή των σφαλμάτων (αν και χρειάζεται την υπόθεση της ομοσκεδαστικότητας)

- *Βήμα 2ο:* Εκτίμησε την παράμετρο της κατανομής t από τα κατάλοιπα του μοντέλου, έστω λοιπόν $\hat{\nu}$ η εκτίμηση των βαθμών ελευθερίας.

  Από εδώ και στο εξής προσομοιώνουμε τα Bootstrap δείγματα που θέλουμε ως εξής:

- *Βήμα 3α:* Προσομοίωσε το διάνυσμα των τυχαίων σφαλμάτων από την κατανομή t με $\hat{\nu}$ βαθμούς ελευθερίας.

- *Βήμα 3β:*Δημιούργησε τις παρατηρήσεις με τη χρήση του τύπου $Y_i^* = \hat{a} + \hat{\beta}X_i + e_i^*$

- *Βήμα 4ο:* Προσάρμοσε το γραμμικό μοντέλο στα δεδομένα (Y_i^*, X_i) και υπολόγισε όποιες ποσότητες σε ενδιαφέρουν

- *Βήμα 5ο:* Επανέλαβε τη διαδικασία πολλές φορές .

Με την προσέγγιση που μόλις περιγράψαμε μπορεί κανείς να προσαρμόσει το γραμμικό μοντέλο χωρίς να χρειάζεται την υπόθεση της κανονικότητας, επομένως παρουσιάσαμε μια μέθοδο γραμμικής παλινδρόμησης απαλλαγμένη από την υπόθεση της κανονικότητας. Στατιστική συμπερασματολογία μπορεί να στηριχθεί στη μέθοδο bootstrap και στις κατανομές των ζητούμενων στατιστικών ποσοτήτων, όπως αυτές έχουν εκτιμηθεί με τη μέθοδο bootstrap.

5.10 Bootstrap στην ανάλυση κυρίων συνιστωσών

Ας δούμε τώρα ένα άλλο παράδειγμα χρήσης της μεθόδου bootstrap που αφορά πολυμεταβλητά δεδομένα και μια προχωρημένη στατιστική μέθοδο για να πάρουμε μια ιδέα για τις εφαρμογές της μεθόδου σε πολύπλοκα προβλήματα.

5.10.1 Η μέθοδος της ανάλυσης σε κύριες συνιστώσες

Η μέθοδος των κυρίων συνιστωσών (Principal Components Analysis) είναι μια μέθοδος η οποία έχει σκοπό να δημιουργήσει γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί να είναι ασυσχέτιστοι μεταξύ τους αλλά να περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης. Το κέρδος από μια τέτοια διαδικασία είναι πως ξεκινώντας από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών που για κάποιες στατιστικές μεθόδους είναι περισσότερο χρήσιμες. Επίσης, αν οι κύριες συνιστώσες που θα προκύψουν μπορούν να ερμηνεύσουν ένα μεγάλο ποσοστό της διακύμανσης τότε αυτό σημαίνει πως αντί να έχουμε p μεταβλητές, όπως είχαμε αρχικά, έχουμε λιγότερες, μειώνουμε δηλαδή τις διαστάσεις του προβλήματος. 'Ενα άλλο μεγάλο πλεονέκτημα (το οποίο από την άλλη ίσως είναι και μειονέκτημα για πολλούς) είναι πως με τη μέθοδο των κυριών συνιστωσών μπορούμε να εξετάσουμε τις συσχετίσεις ανάμεσα στις μεταβλητές και να διαπιστώσουμε πόσο αυτές μοιάζουν ή όχι. Επίσης η μέθοδος μας επιτρέπει να δώσουμε ονόματα στις καινούριες μεταβλητές (τις συνιστώσες) παρατηρώντας ποιες από τις αρχικές μεταβλητές έχουν μεγάλη επίδραση σε αυτές. Με αυτό τον τρόπο επιτυγχάνουμε να ποσοτικοποιήσουμε μη μετρήσιμες (και μερικές φορές αφηρημένες) έννοιες.

'Εστω λοιπόν πως έχουμε ένα σύνολο από k μεταβλητές $(X_1, X_2, ..., X_k)$ και θέλουμε να δημιουργήσουμε τις κύριες συνιστώσες $(Y_1, Y_2, ..., Y_k)$ οι οποίες να είναι γραμμικός συνδυασμός των αρχικών μεταβλητών , δηλαδή

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1k}X_k \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ... + a_{2k}X_k \\ &\quad ..... \\ Y_k &= a_{k1}X_1 + a_{k2}X_2 + ... + a_{kk}X_k \end{aligned}$$

Υπό μορφή πινάκων μπορεί να γραφτεί ως $\mathbf{Y} = \mathbf{X}'\mathbf{A}$ όπου $\mathbf{Y}, \mathbf{X}$ είναι διανύσματα $k \times 1$ και $\mathbf{A}$ είναι $k \times k$ πίνακας με στοιχεία

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1k} \\ a_{21} & a_{22} & ... & a_{2k} \\ ... & & ... & \\ a_{k1} & a_{k2} & ... & a_{kk} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & ... & a_k \end{bmatrix}$$

όπου a_j είναι το διάνυσμα στήλη με στοιχεία

$$a'_j = [a_{1j}, a_{2j}, \ldots, a_{kj}], \quad j = 1, ..., k$$

και για να μην έχουμε προβλήματα ταυτοποίησης θέτουμε πως

$$\sum_{i=1}^{k} a_{ij}^2 = a_j' a_j = 1.$$

Επομένως το πρόβλημα εύρεσης των κυρίων συνιστωσών είναι το πρόβλημα της εύρεσης των στοιχείων του πίνακα $\mathbf{A}$. Έχουμε όμως έναν επιπλέον περιορισμό, ότι δηλαδή οι κύριες συνιστώσες πρέπει να είναι σε φθίνουσα σειρά ως προς τη διακύμανση τους, δηλαδή η πρώτη να έχει τη μεγαλύτερη διακύμανση, η δεύτερη τη δεύτερη μεγαλύτερη και ούτω καθεξής. Μπορεί να δειχτεί πως ο πίνακας $\mathbf{A}$ είναι ο πίνακας που περιέχει τα ιδιοδιανύσματα του πίνακα συσχετίσεων (ή του πίνακα διακύμανσης ανάλογα με τον ποιόν έχουμε χρησιμοποιήσει). Επίσης η κάθε ιδιοτιμή, σε φθίνουσα σειρά είναι η διακύμανση κάθε κύριας συνιστώσας και επομένως ο λόγος της τιμής προς το άθροισμα όλων των ιδιοτιμών είναι το ποσοστό της διακύμανσης που κάθε συνιστώσα ερμηνεύει.

Επομένως, μερικά ενδιαφέροντα αποτελέσματα είναι τα εξής:

- Για να κατασκευάσουμε τις κύριες συνιστώσες χρειάζεται να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα ($\mathbf{S}$ ή $\mathbf{R}$) που χρησιμοποιούμε. Πρέπει να τονίσουμε πως μπορεί να χρησιμοποιήσει κανείς τόσο τον πίνακα δειγματικών συσχετίσεων $\mathbf{R}$ ή τον πίνακα δειγματικών διακυμάνσεων $\mathbf{S}$. Δεν θα ασχοληθούμε εδώ με το πρόβλημα επιλογής πίνακα αν και στο παράδειγμα θα χρησιμοποιήσουμε τον $\mathbf{R}$.

- Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμά της αντιστοιχούν στην πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη ιδιοτιμή στη δεύτερη κύρια συνιστώσα κλπ.

- Η διακύμανση της κάθε κύριας συνιστώσας είναι ίση με την ιδιοτιμή που της αντιστοιχεί. Έτσι αν συμβολίσουμε με λ_j την j μεγαλύτερη ιδιοτιμή τότε έχουμε πως $Var(Y_j) = \lambda_j$.

- Η συνολική διακύμανση των κύριων συνιστωσών θα είναι η ίδια με τη συνολική διακύμανση των αρχικών μεταβλητών εξαιτίας των ιδιοτήτων του ίχνους συμμετρικού και τετραγωνικού πίνακα. Δηλαδή θα ισχύει $tr(\mathbf{\Sigma}) = tr(\mathbf{\Lambda})$ και άρα η συνολική διακύμανση διατηρείται. Ο πίνακας $\mathbf{\Lambda}$ είναι ο διαγώνιος πίνακας που περιέχει τις ιδιοτιμές του $\mathbf{\Sigma}$ από τη φασματική ανάλυση του $\mathbf{\Sigma}$.

- Επίσης η γενικευμένη διακύμανση των κυριών συνιστωσών είναι η ίδια με τη γενικευμένη διακύμανση των αρχικών μεταβλητών. Αυτό προκύπτει εύκολα καθώς η ορίζουσα ενός τετραγωνικού πίνακα είναι το γινόμενο των ιδιοτιμών της και άρα ισχύει

$$|\Sigma| = \prod_{i=1}^{p} \lambda_i = |\Lambda|$$

.

- Η ποσότητα $\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ μας δείχνει το ποσοστό της συνολικής διακύμανσης που εξηγεί η i συνιστώσα. Είναι ευνόητο πως αν κάποιος πάρει όλες τις συνιστώσες τότε θα διατηρήσει όλη τη διακύμανση, ενώ αν τελικά διώξει κάποιες συνιστώσες κάποιο ποσοστό θα χαθεί. Προφανώς συμφέρει να διατηρούμε τις πρώτες συνιστώσες που εξηγούν μεγαλύτερο κομμάτι.

Καταλαβαίνει κανείς πως επειδή στην ανάλυση σε κύριες συνιστώσες δουλεύουμε με ιδιοτιμές και ιδιοδιανύσματα είναι πολύ δύσκολο να έχουμε αναλυτικά αποτελέσματα σχετικά με την κατανομή των ιδιοτιμών και των ιδιοδιανυσμάτων. Παρόλο που υπάρχουν ασυμπτωτικά αποτελέσματα βασισμένα στην υπόθεση πολυμεταβλητού κανονικού πληθυσμού, είναι πολλές φορές άχρηστα.

'Ενα επίσης ενδιαφέρον θέμα που προκύπτει έχει να κάνει με την επιλογή του αριθμού των κυρίων συνιστωσών που θα κρατήσει ο ερευνητής για παραπέρα ανάλυση. 'Ενα από τα κριτήρια που συνήθως χρησιμοποιούνται είναι το κριτήριο του Kaiser το οποίο προτείνει να κρατήσουμε τόσες συνιστώσες όσες έχουν ιδιοτιμή μεγαλύτερη από το μέσο όρο των ιδιοτιμών. Αν δουλεύουμε με τον πίνακα συσχέτισης αυτό ισοδυναμεί με ιδιοτιμές μεγαλύτερες της μονάδας. Η λογική είναι πως ιδιοτιμές πάνω από το μέσο όρο δεν έχουν προκύψει από ασυσχέτιστα δεδομένα και άρα έχει νόημα να κρατήσουμε αυτή τη συνιστώσα.

Από την άλλη ακόμα και αν τα δεδομένα είναι στην πραγματικότητα ασυσχέτιστα, τουλάχιστον μια ιδιοτιμή θα είναι μεγαλύτερη από τη μονάδα. Συνεπώς είναι ενδιαφέρον να δει κανείς αν η τιμή που βρήκαμε είναι πραγματικά μεγαλύτερη της μονάδας και δεν οφείλεται απλά στην τυχαία δειγματοληψία. Επομένως θα ήταν ενδιαφέρον να ξέρουμε την τυπική απόκλιση κάθε ιδιοτιμής ώστε να μπορούμε να βρούμε αν είναι όντως μεγαλύτερη από το μέσο όρο και όχι απλά για λόγους τυχαιότητας.

5.10.2 Τα δεδομένα: 'Επταθλο Ολυμπιακών αγώνων 2000

Θα δούμε λοιπόν τη χρήση της μεθόδου bootstrap σε αυτό το πρόβλημα καθώς και άλλα σχετικά με την ανάλυση σε κύριες συνιστώσες εξετάζοντας ένα σετ από πραγματικά δεδομένα. Τα δεδομένα που θα αναλύσουμε αφορούν το αγώνισμα του επτάθλου στους Ολυμπιακούς αγώνες του Σίδνευ το 2000. Συγκεκριμένα μόνο οι 26 αθλήτριες που βαθμολογήθηκαν και στα 7 αγωνίσματα θα χρησιμοποιηθούν για την ανάλυση. Σκοπός της ανάλυσης είναι να δούμε αν, και κατά πόσο, τα 7 αγωνίσματα μπορούν να αντικατασταθούν από συνιστώσες που μετράνε κάποιες μη παρατηρήσιμες ποσότητες, όπως η δύναμη ή η ταχύτητα. Εμείς θα παρουσιάσουμε το παράδειγμα επικεντρώνοντας περισσότερο στο τι μπορεί να μας προσφέρει η μέθοδος bootstrap. Τα δεδομένα υπάρχουν στον πίνακα 5.8. Για τα αγωνίσματα των δρόμων (100μ με εμπόδια, 200 μέτρα και 800 μέτρα) τα δεδομένα είναι σε δευτερόλεπτα ενώ για τα υπόλοιπα (ρίψεις και άλματα) σε μέτρα.

Κατά αρχάς θα πρέπει να τονίσουμε πως θα δουλέψουμε με τον πίνακα συσχετίσεων, λόγω των διαφορετικών μονάδων μέτρησης σε κάθε μεταβλητή (αγώνισμα). Επίσης επειδή για τους δρόμους καλές επιδόσεις είναι οι μικρές τιμές (μικροί χρόνοι) θα αλλάξουμε τα πρόσημα των παρατηρήσεων ώστε σε όλα τα αγωνίσματα οι καλές επιδόσεις να είναι οι μεγάλες τιμές. Αυτό διευκολύνει πολύ την όποια

ερμηνεία των συνιστωσών.

Αθλήτρια	Χώρα εμπόδια	100μ.	Ύψος βολία	Σφαιρο-200 μ.	200 μ.	Μήκος σμός	Ακοντι-	800 μ.
Azzizi	ALG	13.64	1.6	14.17	24.59	5.88	46.28	141.82
Bacher	ITA	13.82	1.75	12.75	24.96	5.84	41.14	129.08
Biswas	IND	14.11	1.63	11.69	24.73	5.64	39.59	142.17
Braun	GER	13.49	1.81	14.33	24.74	6.22	48.56	139.14
Ganapathy	IND	14.22	1.69	11.14	24.69	5.96	36.02	140.86
Garcva	CUB	13.46	1.66	13.29	24.58	5.92	50.31	139.64
Hautala	FIN	13.62	1.78	13.31	25.00	6.12	45.4	134.9
Jamieson	AUS	14.09	1.81	13.59	25.27	6.09	45.32	136.57
Kabanova	UZB	14.89	1.72	11.56	27.27	5.22	36.61	140.11
Kazanina	KZK	14.71	1.75	12.97	25.04	5.84	43.53	130.45
Koritskaya	RUS	13.88	1.72	13.53	24.08	5.56	40.67	129.77
Kovalenkoa	UKR	15.12	1.72	13.57	26.36	5.57	42.5	133.52
Lewis	GBR	13.23	1.75	15.55	24.34	6.48	50.19	136.83
Mark	TRI	13.72	1.66	11.44	25.35	5.9	48.99	152.36
Nathan Le	USA	13.74	1.78	14.22	24.84	6.06	43.48	136.67
Naumenko	KZK	14.26	1.84	11.26	25.19	5.88	32.53	138.49
Prokhorova	RUS	13.63	1.81	13.21	23.72	6.59	45.05	130.32
Rajamaki	FIN	13.6	1.66	13.87	24.03	6.36	37	138.47
Roshchupkina	RUS	13.7	1.84	14.03	23.53	5.47	43.87	132.24
Sazanovich	BLR	13.45	1.84	14.79	24.12	6.5	43.97	136.41
Skujyte	LIT	14.37	1.78	15.09	25.35	5.97	45.43	140.25
Specht-Ertl	GER	13.43	1.78	13.55	24.64	6.22	42.7	136.25
Teppe	FRA	14.02	1.72	13.44	26.39	5.94	46.98	138.56
Teteryuk	UKR	14.53	1.72	13.56	25.76	5.89	44.57	139.94
Tigau	ROM	13.39	1.72	11.53	24.8	6.01	43.38	139.65
Wlodarczyk	POL	13.33	1.78	14.45	24.29	6.31	46.16	132.15

Πίνακας 5.8: Οι επιδόσεις των 26 επταθλητριών στην ολυμπιάδα του Σίδνευ, 2000

Οι συναρτήσεις για τις οποίες ενδιαφερόμαστε είναι οι εξής:

- Η ορίζουσα του πίνακα συσχέτισης είναι ένα μέτρο συμμεταβλητότητας των δεδομένων. Αν τα δεδομένα ήταν ασυσχέτιστα η τιμή της ορίζουσας θα έπρεπε να είναι 1 ενώ αν ήταν πλήρως συσχετισμένα θα έπρεπε να είναι 0. Σε πραγματικά δεδομένα περιμένουμε μικρές τιμές κοντά στο 0. Η κατανομή της ορίζουσας είναι άγνωστη και επομένως θα την εκτιμήσουμε με τη χρήση της μεθόδου Bootstrap.

- Ποια είναι η κατανομή των ιδιοτιμών του πίνακα συσχετίσεων; Πόσες ιδιοτιμές είναι μεγαλύτερες της μονάδας αν κανείς λάβει υπόψη την τυχαιότητα λόγω του δείγματος (οι επταθλήτριες είναι ένα δείγμα από τον πληθυσμό των επταθλητριών, δεν θα ασχοληθούμε με το θέμα αν και κατά πόσο είναι τυχαίο δείγμα ή άλλα τέτοια προβλήματα).

- Η πρώτη συνιστώσα που προκύπτει ως ένας σταθμικός μέσος των αγωνισμάτων δίνει διαφορετικά βάρη σε κάθε αγώνισμα ή όχι;

Θα πρέπει να προειδοποιήσουμε τον αναγνώστη πως κάποια από τα θέματα αυτά απαιτούν καλές γνώσεις πολυμεταβλητής στατιστικής. Σκοπός της εφαρμογής

είναι να αναδείξει τη χρησιμότητα της μεθόδου bootstrap και όχι τα υπέρ και τα κατά της ανάλυσης σε κύριες συνιστώσες. Σε κάθε περίπτωση όμως θα δούμε πως λειτουργεί η μέθοδος bootstrap σε ένα πολυμεταβλητό παράδειγμα όπου κλασσικές μέθοδοι βασισμένες σε υποθέσεις και ασυμπτωτικά αποτελέσματα είναι μάλλον δύσκολο να χρησιμοποιηθούν.

Για να πάρουμε ένα bootstrap δείγμα από τα δεδομένα μας αρκεί να πάρουμε με επανάθεση παρατηρήσεις, δηλαδή αθλήτριες. Ως παρατήρηση εννοούμε όλο το διάνυσμα με τα 7 αγωνίσματα. Συνεπώς για το παράδειγμα μας το bootstrap δείγμα μπορεί να περιέχει περισσότερες από μια φορές κάποια αθλήτρια. Στη συνέχεια για κάθε bootstrap δείγμα που δημιουργήσαμε θα προχωρήσουμε σε ανάλυση σε κύριες συνιστώσες με τη χρήση του πίνακα συσχετίσεων τις ποσότητες που μας ενδιαφέρουν: τις ιδιοτιμές, τα ιδιοδιανύσματα και την ορίζουσα του πίνακα συσχετίσεων.

Πριν προχωρήσουμε στην ανάλυση θα πρέπει να τονίσουμε πως πολλές φορές (αν όχι τις περισσότερες) η ανάλυση σε κύριες συνιστώσες έχει καθαρά περιγραφικό σκοπό και δεν ενδιαφερόμαστε για στατιστική συμπερασματολογία. Επίσης πολλές φορές τα δεδομένα αφορούν τον πληθυσμό. Θα παρουσιάσουμε τη μεθοδολογία αφήνοντας στην άκρη τέτοια ζητήματα.

5.10.3 Η κατανομή των ιδιοτιμών

Ας ξεκινήσουμε από το πρόβλημα που σχετίζεται με τον αριθμό των κυρίων συνιστωσών που πρέπει να χρησιμοποιήσουμε. 'Οπως είπαμε, ένα από τα κριτήρια που συνήθως χρησιμοποιούμε, αυτό του Kaiser, προτρέπει να κρατήσουμε τόσες συνιστώσες όσες και ιδιοτιμές μεγαλύτερες από το 1 έχουμε. Στην περίπτωση μας αυτές είναι 3, όπως μπορούμε να δούμε στον πίνακα 5.9. Η τρίτη ιδιοτιμή είναι 1.05 και είναι αρκετά λογικό να έχει αυτή την τιμή απλά για λόγους τυχαιότητας και όχι γιατί απαραίτητα αντιστοιχεί σε κάποια συνιστώσα με σημαντική συνεισφορά.

Α/Α	Ιδιοτιμή	% της συνολικής διακύμανσης	Αθροιστικό Ποσοστό
1	2.96009	42.29%	42.287%
2	1.51991	21.71%	64.000%
3	1.05053	15.01%	79.008%
4	0.64641	9.23%	88.242%
5	0.38602	5.51%	93.757%
6	0.27779	3.97%	97.725%
7	0.15924	2.27%	100%

Πίνακας 5.9: Ιδιοτιμές του πίνακα συσχετίσεων

Θέλουμε επομένως να εκτιμήσουμε τη διακύμανση κάθε ιδιοτιμής με σκοπό να δούμε αν αυτή είναι πράγματι μεγαλύτερη από τη μονάδα ή αν απλά έτυχε για λόγους τυχαίας δειγματοληψίας. Η μέθοδος bootstrap μπορεί να βοηθήσει σε αυτή την κατεύθυνση. Παρατηρείστε πως η κατανομή μιας ιδιοτιμής είναι ιδιαίτερα δύσκολο να μελετηθεί θεωρητικά αν και υπάρχουν θεωρητικά αποτελέσματα τα οποία στηρίζονται είτε σε κανονικότητα του πληθυσμού είτε σε ασυμπτωτικά

θεωρήματα. Προφανώς η μέθοδος bootstrap δεν χρειάζεται τέτοιες υποθέσεις.

Α/Α	Ιδιοτιμή	Μέση τιμή	Τυπική απόκλιση	95% δ.ε. (percentile μέθοδος)
1	2.96009	3.057	0.450	(2.229,3.896)
2	1.51991	1.627	0.206	(1.227,2.043)
3	1.05053	1.049	0.191	(0.685,1.416)
4	0.64641	0.624	0.145	(0.375,0.932)
5	0.38602	0.350	0.097	(0.183,0.552)
6	0.27779	0.198	0.067	(0.080,0.333)
7	0.15924	0.095	0.041	(0.025,0.184)

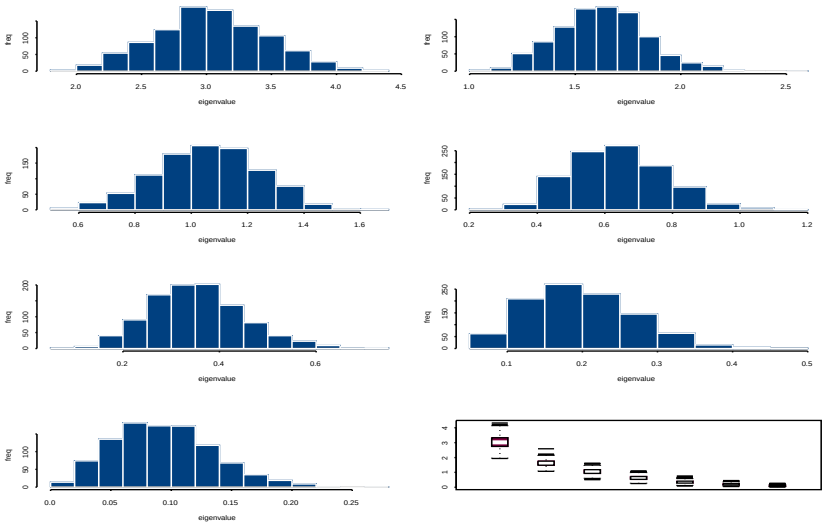
Πίνακας 5.10: Εκτιμηθείσες τυπικές αποκλίσεις και 95% διαστήματα εμπιστοσύνης για τις ιδιοτιμές από 1000 bootstrap επαναλήψεις

Στον πίνακα 5.10 μπορεί κανείς να δει τη μέση τιμή για κάθε ιδιοτιμή καθώς και την εκτίμηση της τυπικής της απόκλισης βασισμένες σε 1000 bootstrap δείγματα. Επίσης έχουμε κατασκευάσει ένα 95% διάστημα εμπιστοσύνης για κάθε ιδιοτιμή, βασισμένο στη μέθοδο των ποσοστιαίων σημείων. Αν κοιτάξει κανείς στο γράφημα 5.10 μπορεί να δει τα ιστογράμματα των ιδιοτιμών καθώς και ένα διάγραμμα πλαισίου και απολήξεων για όλες τις ιδιοτιμές. Είναι πολύ ενδιαφέρον να παρατηρήσει κανείς πως το διάστημα εμπιστοσύνης για την τρίτη ιδιοτιμή περιέχει ξεκάθαρα την τιμή 1 και συνεπώς η τρίτη ιδιοτιμή δεν μπορεί να ισχυριστεί κανείς πως είναι στατιστικά σημαντικά μεγαλύτερη από το 1, δηλαδή δεν υπάρχει κάποιος λόγος να κρατήσουμε την τρίτη συνιστώσα. Επίσης παρατηρείστε πως η κατανομή των ιδιοτιμών δεν μοιάζει με την κανονική για όλες τις ιδιοτιμές. Συγκεκριμένα οι ουρές της κατανομής είναι συνήθως πιο παχιές από αυτές της κανονικής κατανομής, ιδιαίτερα για τις ιδιοτιμές μεγάλης τάξης. Κάνοντας έλεγχο κανονικότητας Anderson Darling απορρίπτουμε την υπόθεση της κανονικότητας για όλες τις ιδιοτιμές εκτός από τη δεύτερη.

Επομένως η μέθοδος bootstrap μπορεί να χρησιμοποιηθεί για να μελετηθεί η κατανομή των ιδιοτιμών και άρα να επιλεγεί ο αριθμός των συνιστωσών που θα κρατηθούν για περαιτέρω ανάλυση.

5.10.4 Η κατανομή της ορίζουσας

Μια άλλη ποσότητα που μας ενδιαφέρει είναι η ορίζουσα του πίνακα συσχετίσεων. Για το παράδειγμα μας η τιμή της είναι 0.05217. Μικρές τιμές αντιστοιχούν σε έντονα συσχετισμένα δεδομένα αλλά σίγουρα η τιμή δεν λέει και πολλά πράγματα καθώς το απόλυτο μέγεθος της εξαρτάται από το πλήθος των μεταβλητών που χρησιμοποιούμε. Γενικά είναι πολύ δύσκολο να μελετήσει κανείς την κατανομή της ορίζουσας αλλά με τη μέθοδο bootstrap μπορεί κανείς να την προσεγγίσει ικανοποιητικά. Στο γράφημα 5.11 μπορεί κανείς να δει το ιστόγραμμα βασισμένο σε 1000 επαναλήψεις. Η κατανομή είναι ιδιαίτερα ασύμμετρη. Η μέση τιμή από τις 1000 επαναλήψεις είναι 0.02467 και συνεπώς η δειγματική ορίζουσα ως εκτίμηση



Γράφημα 5.10: Ιστογράμματα και διάγραμμα Boxplot για τις ιδιοτιμές.

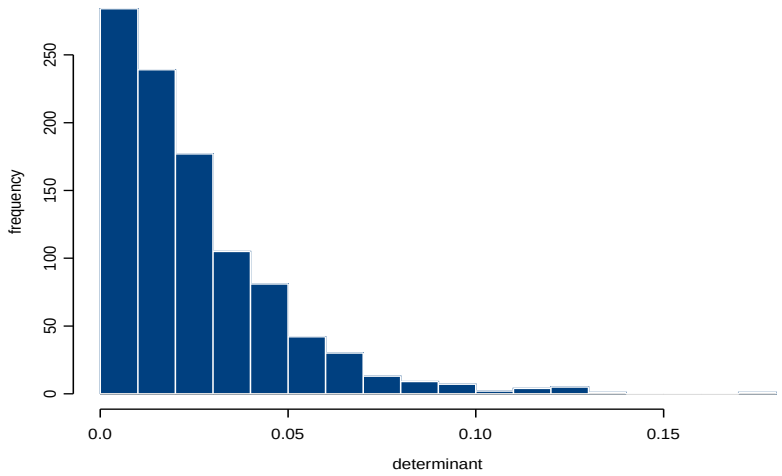
της ορίζουσας του πίνακα συσχέτισης του πληθυσμού είναι μεροληπτική. Η τυπική της απόκλιση εκτιμήθηκε με τη μέθοδο bootstrap και είναι 0.02417 ενώ ένα 95% διάστημα εμπιστοσύνης βασισμένο στη μέθοδο των ποσοστιαίων σημείων είναι το (0.00168879, 0.0793803) .

Αν η ορίζουσα από μόνη της δεν μας δίνει πολλές πληροφορίες είναι ενδιαφέρον να παρατηρήσουμε πως στην πολυμεταβλητή ανάλυση υπάρχουν κάποιες ελεγχοσυναρτήσεις που τη χρησιμοποιούν. Για παράδειγμα ο έλεγχος σφαιρικότητας του Bartlett χρησιμοποιείται για να ελέγξει την υπόθεση ότι οι μεταβλητές είναι ασυσχέτιστες. Ο έλεγχος χρησιμοποιεί την ελεγχοσυνάρτηση

$$L = -\left[n - \frac{1}{6(2p+5)}\right] \ln |R|$$

όπου p είναι το πλήθος των μεταβλητών και n το μέγεθος του δείγματος. Η ελεγχοσυνάρτηση χρησιμοποιείται για να ελέγξει αν ο πίνακας συσχέτισης είναι μοναδιαίος, και άρα δεν υπάρχουν συσχετίσεις, έναντι της εναλλακτικής πως δεν είναι ο μοναδιαίος και άρα υπάρχουν συσχετίσεις.

Είναι ευνόητο πως μπορεί κανείς να προχωρήσει σε έλεγχο bootstrap. Με βάση αυτά που έχουμε πει θα πρέπει κανείς να μετασχηματίσει τα δεδομένα έτσι ώστε να έχουν πίνακα συσχέτισης το μοναδιαίο και στη συνέχεια να πάρει δείγματα Bootstrap από αυτά τα δεδομένα. Ποιος όμως μετασχηματισμός μπορεί να μας οδηγήσει σε ασυσχέτιστα δεδομένα; Μια λύση είναι να χρησιμοποιηθούν οι κύριες



Γράφημα 5.11: Ιστόγραμμα βασισμένο σε 1000 επαναλήψεις για τη δειγματική ορίζουσα

συνιστώσες οι οποίες είναι ασυσχέτιστες εξ ορισμού και να προχωρήσει κανείς στον έλεγχο. Επίσης είναι ευνόητο πως μπορούμε να σχηματίσουμε την κατανομή της ελεγχοσυνάρτησης του Bartlett η οποία κάτω από την υπόθεση πως ο πληθυσμός ακολουθεί την πολυμεταβλητή κανονική κατανομή έχει κατανομή τη χ^2 κατανομή με $p(p-1)/2$ βαθμούς ελευθερίας. Συνεπώς αν κάποιος ακολουθούσε τον έλεγχο του Bartlett, θα απέρριπτε τη μηδενική υπόθεση αν η ελεγχοσυνάρτηση είχε τιμή μεγαλύτερη από 32.67 (το 95% ποσοστιαίο σημείο της κατανομής χ^2 με 21 βαθμούς ελευθερίας). Για τα δεδομένα μας η τιμή της ελεγχοσυνάρτησης είναι 76.758 κατά πολύ μεγαλύτερη από την τιμή 32.67 κι επομένως οδηγούμαστε σε απόρριψη της μηδενικής υπόθεσης. Βέβαια η κριτική τιμή έχει στηριχτεί στην υπόθεση της κανονικότητας, παρόλα αυτά η απόκλιση των δεδομένων μας από αυτή την τιμή είναι πολύ μεγάλη και μας προσφέρει αρκετή βεβαιότητα για την απόφαση μας. Επομένως αν αναλογιστούμε τη σχέση διαστημάτων εμπιστοσύνης και δίπλευρων ελέγχων ξεκάθαρα θα απορρίπταμε τη μηδενική υπόθεση περί ανεξαρτησίας των δεδομένων.

5.10.5 Ερμηνεία των ιδιοδιανυσμάτων - συνιστωσών

Μια άλλη ενδιαφέρουσα χρησιμότητα της μεθόδου bootstrap στην ανάλυση σε κύριες συνιστώσες είναι πως μπορούμε να κάνουμε συμπερασματολογία σχετικά με τις ίδιες τις συνιστώσες και τους συντελεστές τους. Για παράδειγμα από τα

δεδομένα μας προκύπτει πως η πρώτη κύρια συνιστώσα είναι η

$$\begin{aligned} Y_1 &= 0.466 \times (100\ \mu) + 0.246 \times (\text{ύψος}\ ) + 0.420 \times (\text{σφαιροβολία}) \\ &+ 0.433 \times (200\ \mu) + 0.463 \times (\text{άλμα εις μήκος}) \\ &+ 0.323 \times (\text{ακοντισμός}) + 0.202 \times (800\mu) \end{aligned}$$

Μια ερμηνεία αυτής της συνιστώσας θα μπορούσε να είναι ότι είναι ένας σταθμικός μέσος των αγωνισμάτων (θυμηθείτε πως έχουμε αλλάξει πρόσημα στους δρόμους ώστε μεγάλες τιμές δείχνουν μεγάλες επιδόσεις). Και το ερώτημα που προκύπτει είναι:

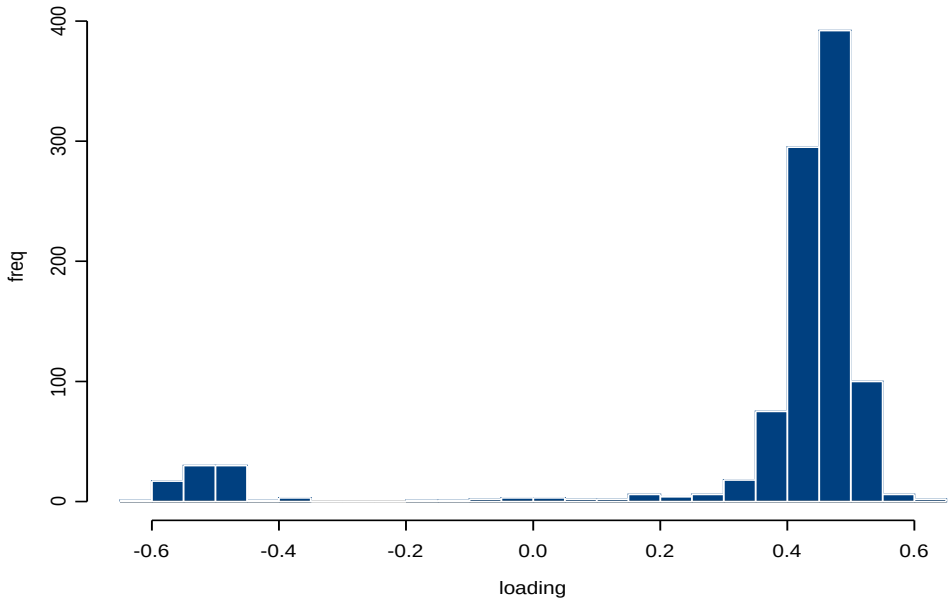
Μπορούμε να κάνουμε συμπερασματολογία για κάθε συντελεστή ξεχωριστά; 'Ενας μηδενικός συντελεστής σημαίνει πως η μεταβλητή δεν είναι σημαντική για τη συνιστώσα (κατ' αναλογία με το ότι συμβαίνει και στη γραμμική παλινδρόμηση). Μπορούμε επίσης να πούμε πως όλες οι μεταβλητές έχουν την ίδια στάθμιση και πως απλά έτυχε για λόγους τυχαιότητας να βρούμε διαφορετικούς συντελεστές;

Θα εξετάσουμε και τα δύο ερωτήματα. Κατά αρχάς ας θυμηθούμε πως το άθροισμα τετραγώνων των συντελεστών αθροίζει αναγκαστικά στη μονάδα, επομένως αν κάποιος αλλάξει τα πρόσημα μπορεί να πάρει πάλι μια λύση. Αν δούμε το ιστόγραμμα που ακολουθεί (γράφημα 5.12) και αφορά το συντελεστή των 100 μέτρων με εμπόδια παρατηρείστε πως έχει ξεκάθαρα μια δίκορφη μορφή. Οι δύο κορυφές εξηγούνται λόγω της δυνατότητας να αλλάξει κανείς αυθαίρετα πρόσημο στη συνιστώσα. Παρατηρείστε μια κάποια συμμετρία ως προς το 0. Τα υπάρχοντα στατιστικά πακέτα, αποφασίζουν αυθαίρετα αν θα κρατήσουν τη λύση με το θετικό ή το αρνητικό πρόσημο για τον πρώτο συντελεστή. Κάποιος που θα χρησιμοποιήσει τη μέθοδο bootstrap στην ανάλυση σε κύριες συνιστώσες πρέπει να έχει το νου του σε αυτό το πρόβλημα.

Για να αποφύγουμε αυτό το πρόβλημα αποφασίσαμε πως για κάθε λύση το πρόσημο του πρώτου συντελεστή θα είναι θετικό. Με αυτό τον περιορισμό πήραμε τα ιστογράμματα και τα διαγράμματα boxplot που βλέπετε στο γράφημα 5.13.

Θα πρέπει να τονίσουμε πως στην πράξη εμφανίζονται μερικά ακόμα προβλήματα, που σχετίζονται με τη μέθοδο. Είδαμε ένα από αυτά το οποίο έχει να κάνει με το πρόσημο των συντελεστών. Επίσης υπάρχει πρόβλημα με την περιστροφή των αξόνων καθώς οι κύριες συνιστώσες που προκύπτουν δεν είναι μοναδικές αλλά αν πολλαπλασιάσουμε με έναν ορθογώνιο πίνακα προκύπτει μια ισοδύναμη λύση. Κάτι τέτοιο δημιουργεί πρόβλημα κατά τη διάρκεια του bootstrap. Επίσης συνιστώσες που αντιστοιχούν σε μικρές ιδιοτιμές μπορεί να αλλάξουν σειρά (order) κατά τη διάρκεια του bootstrap με αποτέλεσμα οι συντελεστές να εμφανίζουν δίκορφες κατανομές. Τέτοια προβλήματα δεν αντιμετωπίστηκαν σε αυτή την εισαγωγική περιγραφή

Παρατηρείστε τη μεγάλη αριστερή ουρά των περισσοτέρων συντελεστών. Ο συντελεστής του ύψους, του ακοντισμού και των 800 μέτρων είναι συστηματικά μικρότερος από τα άλλα αγωνίσματα κάτι που μάλλον δείχνει πως αυτά τα αγωνίσματα έχουν μικρότερο βάρος.



Γράφημα 5.12: Ιστόγραμμα για το συντελεστή των 100 μέτρων στην πρώτη συνιστώσα. Δεν έχει τεθεί περιορισμός σχετικά με το πρόσημο.

Ενδιαφέρον παρουσιάζει και ο πίνακας 5.11 με τις μέσες τιμές και τις τυπικές αποκλίσεις των συντελεστών καθώς και 95% διαστήματα εμπιστοσύνης για τον καθένα.

Α/Α	Τιμή συντελεστή	Μέση τιμή	Τυπική απόκλιση	95% Δ.ε (percentile μέθοδος)
1	0.466	0.448	0.067	(0.267,0.561)
2	0.246	0.238	0.150	(-0.137,0.478)
3	0.420	0.404	0.079	(0.230,0.509)
4	0.433	0.415	0.081	(0.228,0.527)
5	0.463	0.437	0.092	(0.206,0.546)
6	0.323	0.300	0.131	(-0.084,0.471)
7	0.202	0.200	0.143	(-0.133,0.464)

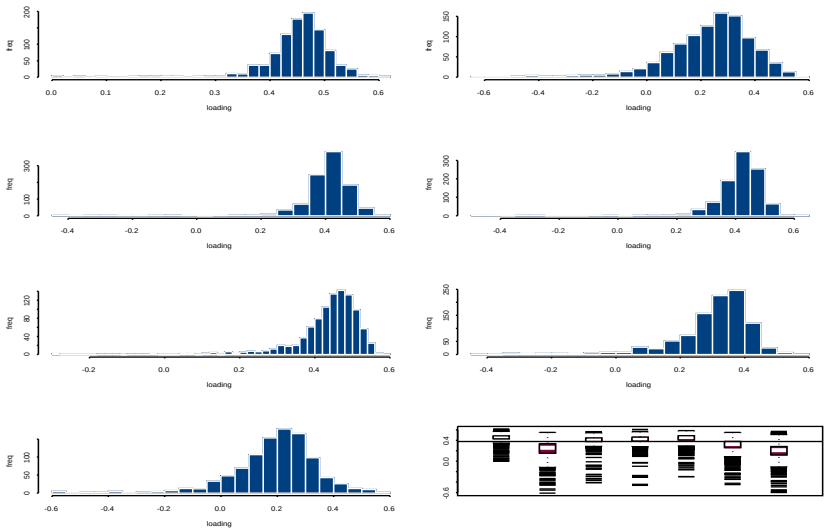
Πίνακας 5.11: Οι τιμές των συντελεστών και εκτιμήσεις των τυπικών τους σφαλμάτων με τη μέθοδο Bootstrap βασισμένες σε 1000 επαναλήψεις. Μπορείτε επίσης να δείτε και ένα 95% διάστημα εμπιστοσύνης για τους συντελεστές

Μπορεί κανείς να παρατηρήσει πως το 95% διάστημα εμπιστοσύνης του συντελεστή για το ύψους, τον ακοντισμό και τα 800 μέτρα περιέχει την τιμή 0 και επομένως αυτοί οι συντελεστές δεν διαφέρουν στατιστικά σημαντικά από το 0, δηλαδή οι μεταβλητές δεν είναι συσχετισμένες με τη συνιστώσα.

'Ενα άλλο ενδιαφέρον θέμα είναι κατά πόσο οι συντελεστές είναι όλοι ίδιοι. Αν σκεφτούμε πως το άθροισμα τετραγώνων είναι 1 τότε κάθε συντελεστής θα έπρεπε να είναι 0.377. Επομένως θέλω να ελέγξω αν το διάνυσμα των συντελεστών είναι το (0.377, 0.377, 0.377, 0.377, 0.377, 0.377, 0.377). Στην πολυμεταβλητή στατιστική ένας γνωστός έλεγχος που μπορεί κανείς να χρησιμοποιήσει είναι ο έλεγχος του Hotelling για να ελέγξει τη μηδενική υπόθεση ότι ο μέσος είναι ίσος με ένα συγκεκριμένο διάνυσμα. μ_0. Ισοδύναμα θα μπορούσε κανείς να κατασκευάσει διαστήματα εμπιστοσύνης για το διάνυσμα των συντελεστών, αυτό θα είναι ένα από κοινού διάστημα εμπιστοσύνης που θα λαμβάνει υπόψη του τις συσχετίσεις ανάμεσα στους συντελεστές. Δηλαδή αν δει τα απλά διαστήματα εμπιστοσύνης για κάθε συντελεστή ξεχωριστά παρατηρεί πως η τιμή 0.377 ανήκει μέσα σε όλα τα διαστήματα για τους συντελεστές. Αυτά όμως τα διαστήματα εμπιστοσύνης δεν λαμβάνουν υπόψη τους τις συσχετίσεις ανάμεσα στους συντελεστές οι οποίες λόγω του περιορισμού (το άθροισμα τετραγώνων των συντελεστών αθροίζει αναγκαστικά στη μονάδα) θα πρέπει να είναι μεγάλες.

Κάτω από την υπόθεση της πολυμεταβλητής κανονικότητας για το διάνυσμα των συντελεστών θα μπορούσε κανείς να κατασκευάσει ένα από κοινού 95% διάστημα εμπιστοσύνης το οποίο θα περιείχε το 95% των παρατηρήσεων. Συγκεκριμένα αν με $\bar{X}$ συμβολίσουμε το διάνυσμα με τις μέσες τιμές από το δείγμα και με $\mathbf{S}$ το δειγματικό πίνακα διακύμανσης συνδιακύμανσης, τότε το ελλειψοειδές που περικλείεται μέσα στα σημεία που ικανοποιούν τη σχέση

$$(x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \leq C$$

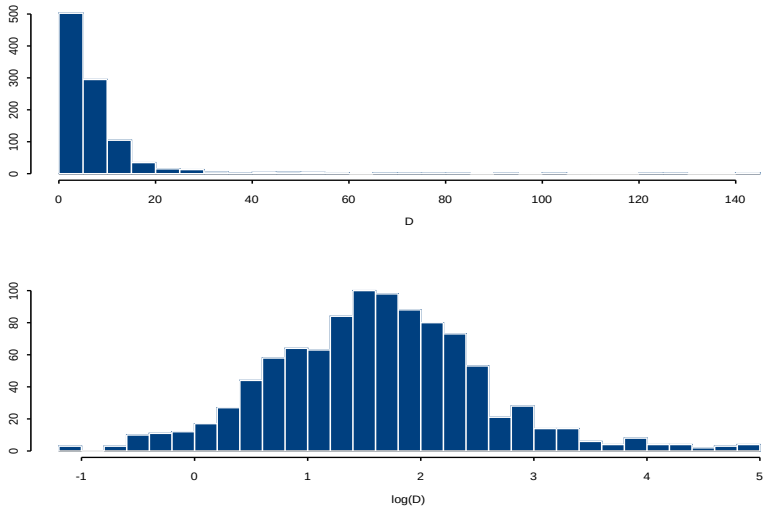


Γράφημα 5.13: Ιστογράμματα και διάγραμμα Boxplot για τους συντελεστές της πρώτης συνιστώσας.

είναι ένα διάστημα εμπιστοσύνης. Η τιμή Cείναι μια κατάλληλα επιλεγμένη κριτική τιμή από μια F κατανομή (για την ακρίβεια είναι συνάρτηση μιας F κριτικής τιμής). Στην περίπτωση μας, όπου δεν μπορούμε και δεν θέλουμε να υποθέσουμε κανονικότητα για τα δεδομένα μας, θα πρέπει να κατασκευάσουμε από τις bootstrap τιμές την κατανομή της συνάρτησης $D_i = (a_i - \bar{a})' S^{-1} (a_i - \bar{a})$, όπου a_i είναι το διάνυσμα των συντελεστών από το i bootstrap δείγμα, $\bar{a}$ είναι το διάνυσμα των μέσων από τις 1000 επαναλήψεις, δηλαδή $\bar{a}$= (0.448, 0.238, 0.404, 0.415, 0.437, 0.300, 0.200) και **S** είναι ο πίνακας διακύμανσης - συνδιακύμανσης των συντελεστών, τον οποίο έχουμε εκτιμήσεις από τις 1000 bootstrap τιμές που πήραμε. Θα πρέπει να σημειωθεί πως υπάρχουν σημαντικές συσχετίσεις ανάμεσα στους συντελεστές. Ο πίνακας συσχετίσεων , όπως εκτιμήθηκε με τη μέθοδο Bootstrap είναι ο

$$\mathbf{R} = \begin{bmatrix} 1.00000 & & & & & & \\ -0.59102 & 1 & & & & & \\ -0.02199 & 0.20992 & 1 & & & & \\ 0.29812 & -0.08829 & -0.20557 & 1 & & & \\ 0.36047 & -0.17295 & -0.09661 & 0.12633 & 1 & & \\ 0.40529 & -0.35193 & 0.26834 & -0.39413 & 0.15731 & 1 & \\ -0.54235 & 0.64511 & 0.18617 & 0.15738 & -0.34500 & -0.43607 & 1 \end{bmatrix}$$

και μπορεί κανείς να δει καθαρά ότι κάποιες συσχετίσεις είναι αρκετά μεγάλες.



Γράφημα 5.14: Ιστόγραμμα συχνοτήτων για τη συνάρτηση D και το λογάριθμο της.

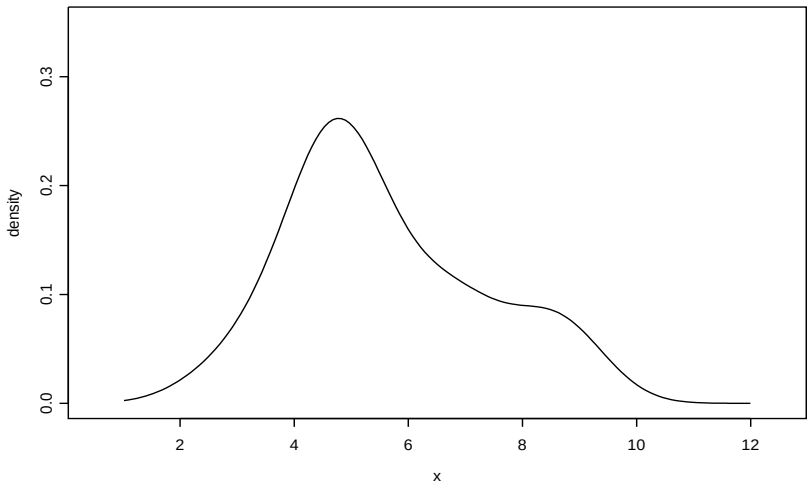
Στο γράφημα 5.14 μπορεί κανείς να δει την κατανομή της συνάρτησης D. Επείδή η κατανομή έχει πολύ μεγάλη ουρά στο ίδιο γράφημα μπορεί κανείς να δει και το ιστόγραμα της $\log D$. Παρατηρείστε πως υπάρχει μια πολύ μεγάλη τιμή κοντά στο 150, δηλαδή για ένα bootstrap δείγμα οι εκτιμηθέντες συντελεστές απείχαν πάρα πολύ από ότι συνέβηκε στα υπόλοιπα bootstrap δείγματα. Η μέση τιμή είναι 8.34 αλλά λόγω της ύπαρξης πολλών ακραίων τιμών δεν είναι αξιόπιστο μέτρο θέσης. Η διάμεσος είναι 4.98 και ο 5% περικομμένος μέσος 6.20. Η τυπική απόκλιση της είναι 13.67. Η κατανομή είναι ιδιαίτερα ασύμμετρη και έχει μεγάλη δεξιά ουρά. Αν κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης με τη μέθοδο των ποσοστιαίων σημείων, αυτό είναι το (0.816, 35.157). Βέβαια θα χρειαζόντουσαν περισσότερες επαναλήψεις για να είναι το διάστημα εμπιστοσύνης πιο αξιόπιστο, θυμηθείτε την ύπαρξη κάποιων πολύ ακραίων τιμών. Παρόλα αυτά το διάνυσμα (0.377, 0.377, 0.377, 0.377, 0.377, 0.377, 0.377) που μας ενδιαφέρει έχει τιμή $D = 5.142$ η οποία σε καμιά περίπτωση δεν είναι ακραία. Μάλιστα από τις 1000 τιμές που έχουμε η τιμή αυτή είναι μεγαλύτερη μόλις από τις 489 και επομένως δεν μπορούμε να ισχυριστούμε πως το διάνυσμα αυτό έχει πολύ μικρή πιθανότητα να παρατηρηθεί. Δηλαδή η υπόθεση πως η πρώτη συνιστώσα είναι ένας σταθμικός μέσος των αγωνισμάτων δείχνει να ευσταθεί.

5.11 Εφαρμογή του bootstrap στην εκτίμηση με kernels

'Οπως είδαμε στο κεφάλαιο 1, παράγραφος 1.7.7. η μέθοδος Bootstrap είναι μια εύχρηστη μέθοδος για την κατασκευή διαστημάτων εμπιστοσύνης σε διάφορες εφαρμογές. Εκεί είδαμε, χωρίς λεπτομέρειες και πως μπορούμε να κατασκευά-

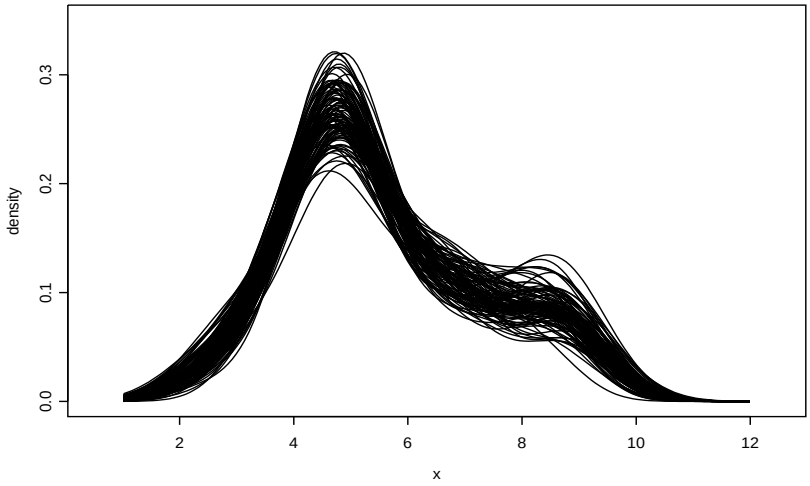
σουμε μια ζώνη εμπιστοσύνης γύρω από την εκτιμήτρια $\hat{f}(x)$. Ας δούμε τώρα πιο αναλυτικά την προσέγγιση σε πρακτικό επίπεδο. Ας υποθέσουμε πως έχουμε ένα δείγμα από n παρατηρήσεις. Διαλέγουμε bootstrap δείγματα από αυτές με επανάθεση και εκτιμούμε την συνάρτηση πυκνότητας πιθανότητας. Παρατηρείστε πως για να πετύχουμε την εκτίμηση σε συγκεκριμένα σημεία για όλες τα bootstrap δείγματα θα πρέπει να έχουμε επιλέξει εκ των προτέρων τα σημεία στα οποία ενδιαφερόμαστε να εκτιμήσουμε. Το παράθυρο το εκτιμούμε από κάθε δείγμα με τον ίδιο τρόπο καθώς είναι συνάρτηση του δείγματος.

Στο γράφημα 5.15 μπορούμε να δούμε την εκτίμηση για 130 προσομοιωμένες παρατηρήσεις



Γράφημα 5.15: Εκτίμηση της συνάρτησης πυκνότητας πιθανότητας

'Εχοντας επαναλάβει τη διαδικασία 100 φορές βρίσκουμε για κάθε σημείο όπου έχουμε υπολογίσει την συνάρτηση πυκνότητας πιθανότητας το 95% διάστημα εμπιστοσύνης βασισμένο στα ποσοστιαία σημεία. Το διάστημα είναι ένα 95% Δ.Ε. για κάθε σημείο και συνεπώς όχι για ολόκληρη την καμπύλη. Αν και υπάρχουν μεθοδολογίες για να κατασκευάσουμε ένα 95% Δ.Ε. για ολόκληρη την καμπύλη κάτι τέτοιο είναι δύσκολο τόσο να υπολογιστεί όσο και να ερμηνευθεί. Στο γράφημα 5.16 μπορεί κανείς να δει τις 100 εκτιμήσεις που αντιστοιχούν στα 100 bootstrap δείγματα. Παρατηρείστε πόσο μεγαλύτερη μεταβλητότητα έχουμε κοντά στις κορυφές, όπως επίσης και πόσο διαφορετική είναι η μεταβλητότητα σε διάφορα σημεία.



Γράφημα 5.16: Οι εκτιμήσεις της συνάρτησης πυκνότητας πιθανότητας για τα 100 bootstrap δείγματα

Τέλος στο γράφημα 5.17 έχουμε ενώσει τα σημεία που ορίζουν το πάνω και κάτω όριο των Δ.Ε. για κάθε σημείο, ώστε να δημιουργήσουμε μια ζώνη και από τις δύο μεριές της αρχικής μας εκτιμήτριας από τα δεδομένα.

Με τον ίδιο τρόπο μπορούμε να μελετήσουμε τη μεταβλητότητα διαφόρων μοντέλων που αντιστοιχούν στην προσαρμογή μιας καμπύλης. Τέτοια παραδείγματα είναι η μη παραμετρική παλινδρόμηση που είδαμε στο κεφάλαιο 1, μη γραμμικά μοντέλα και άλλα. Ακόμα και σε ένα γραμμικό μοντέλο μπορεί κάποιος να χρησιμοποιήσει την τεχνική αυτή αλλά εκεί είναι πολύ πιο εύκολο να μελετήσει τη μεταβλητότητα της γραμμής με βάση τη μεταβλητότητα των παραμέτρων που ορίζουν τη γραμμή.

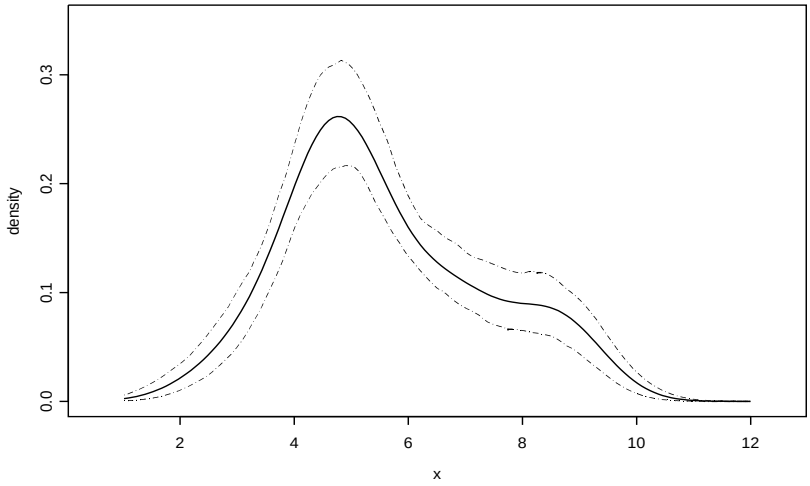
5.12 Χρονολογικές Σειρές

'Εστω το αυτοπαλίνδρομο (autoregressive) μοντέλο p βαθμού $AR(p)$,

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + ... + \beta_p y_{t-p} + \varepsilon_t, \;\; t = 1, ..., T$$

Συνήθως υποθέτουμε κανονικότητα για τα κατάλοιπα και προχωράμε σε εκτίμηση και συμπερασματολογία βασισμένα στην υπόθεση της κανονικότητας. 'Οπως είδαμε και στην περίπτωση της γραμμικής παλινδρόμησης μπορούμε να αποφύγουμε την υπόθεση της κανονικότητας χρησιμοποιώντας bootstrap.

Και στην περίπτωση αυτή θα χρησιμοποιήσουμε τα κατάλοιπα για να δημιουργήσουμε τα bootstrap δείγματα.



Γράφημα 5.17: Η εκτιμήτρια μας και μια ζώνη εμπιστοσύνης γύρω από αυτήν, που κατασκευάσαμε ενώνοντας τα πάνω και κάτω όρια των διαστημάτων εμπιστοσύνης για τα σημεία που υπολογίσαμε την συνάρτηση πυκνότητας πιθανότητας.

Αλγοριθμικά ο τρόπος που θα δουλέψουμε είναι ο εξής

- *Βήμα 1:* Προσάρμοσε το μοντέλο $AR(p)$ στα δεδομένα. Βρες τις εκτιμήσεις $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$
- *Βήμα 2:* Βρες τα κατάλοιπα του εκτιμηθέντος μοντέλου. Έστω ότι αυτά είναι $(\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_t)$.
- *Βήμα 3:* Ξεκίνησε το bootstrap: Πάρε δείγμα από τα εκτιμηθέντα κατάλοιπα με επανάθεση. Έστω ότι το bootstrap δείγμα είναι το $(\hat{\varepsilon}^*_1, \hat{\varepsilon}^*_2, ..., \hat{\varepsilon}^*_t)$.

  Κατασκεύασε την σειρά ως εξής:

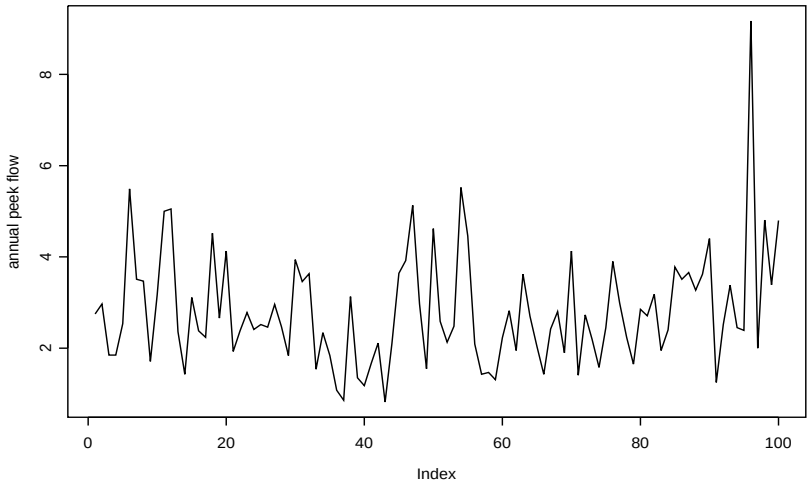
  1. Θέσε $y^*_i = y_i, i = 1, ..., p$
  2. Θέσε $y^*_t = \hat{\beta}_1 y^*_{t-1} + \hat{\beta}_2 y^*_{t-2} + ... + \hat{\beta}_p y^*_{t-p} + \varepsilon^*_t$, για $t = p+1, ..., T$
  3. Προσάρμοσε το μοντέλο AR(p) στη νέα σειρά $y^*_t, t = 1, ..., T$ και εκτίμησε τις παραμέτρους
  4. Επανέλαβε τη διαδικασία B φορές.

- *Βήμα 4:* Χρησιμοποίησε τις B τιμές των παραμέτρων από τα Bootstrap δείγματα για συμπερασματολογία (πχ ελέγχους υποθέσεων, διαστήματα εμπιστοσύνης κλπ)

'Οπως μπορεί κανείς εύκολα να δει, η υπόθεση της κανονικότητας δεν χρειάζεται πουθενά. Αν για κάποιους λόγους, χρειάζεται να κάνουμε μια παραμετρική υπόθεση, τότε αυτό που αλλάζει είναι πως στο βήμα 3.1 αντί να γεννήσουμε τα κατάλοιπα από την εμπειρική τους κατανομή τα προσομοιώνουμε από την παραμετρική υπόθεση που έχουμε κάνει. Πχ αν έχουμε υποθέσει ότι τα κατάλοιπα προέρχονται από μια κατανομή t τότε απλά στο βήμα 3.1 προσομοιώνουμε από αυτή την κατανομή.

Παράδειγμα 5.5

Στο γράφημα 5.18 μπορεί κανείς να δει μια χρονολογική σειρά που αφορά την ετήσια μέγιστη ροή σε ένα συγκεκριμένο σημείο του ποταμού Missouri για τα έτη 1898-1997. Οι παρατηρήσεις αφορούν κυβικά μέτρα ανά δευτερόλεπτο. Στα δεδομένα αυτά προσαρμόσαμε ένα AR(2) μοντέλο.



Γράφημα 5.18: Ετήσιες μέγιστες ροές σε ένα σημείο του ποταμού Missouri για τα έτη 1898-1997.

Αποφεύγοντας κάποια παραμετρική υπόθεση, προχωρήσαμε σε εκτίμηση του μοντέλο. Οι εκτιμήτριες υπάρχουν στην τελευταία στήλη του πίνακα 5.12. Στη συνέχεια εφαρμόσαμε bootstrap για να εκτιμήσουμε τα τυπικά σφάλματα των εκτιμητριών. Τα αποτελέσματα φαίνονται στον πίνακα 5.12.

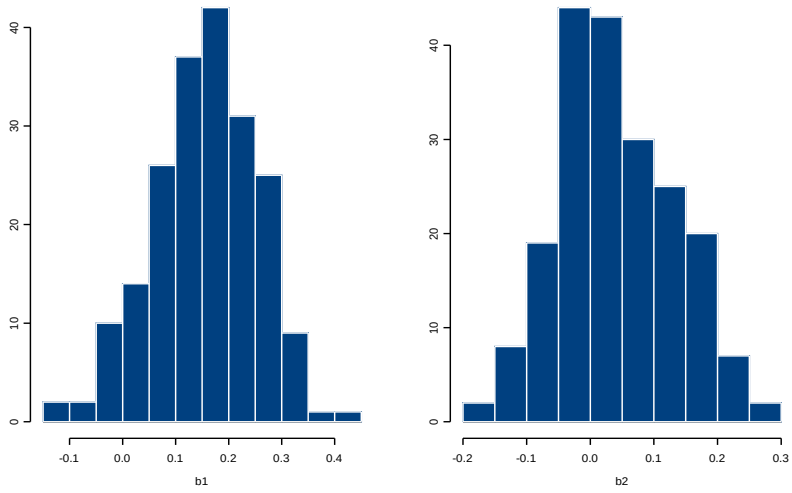
Στο γράφημα 5.19 μπορεί κανείς να δει ιστογράμματα για τις παραμέτρους $\hat{\beta}_1$ και $\hat{\beta}_2$. Παρατηρείστε μια έντονη ασυμμετρία και στις δύο εκτιμήτριες. Από τον πίνακα 5.12 μπορεί επίσης κάποιος να δει τη μεροληψία των εκτιμητριών. Κοιτώντας το γράφημα 5.20 που παρουσιάζει ένα ιστόγραμμα των καταλοίπων διαπιστώνει κανείς ότι η κατανομή του είναι εντόνως ασύμμετρη. Αυτό αποτελεί

Παράμετρος	Μέση τιμή	Bootstrap Τυπικά σφάλματα	95% δ.ε. (ποσοστιαία σημεία)	$\hat{\theta}$
β_1	0.155	0.0964	(-0.032, 0.333)	0.112
β_2	0.042	0.0923	(-0.118, 0.226)	0.063
σ_2	1.669	0.415	(1.087, 2.533)	1.579

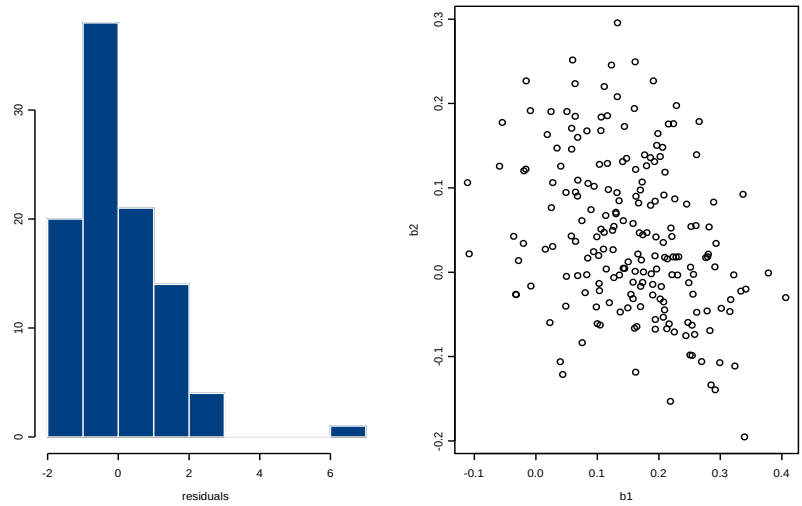
Πίνακας 5.12: Εκτιμήσεις με bootstrap $(B = 1000)$. Η εκτιμηθείσα συσχέτιση ανάμεσα στα $\hat{\beta}_1$ και $\hat{\beta}_2$ είναι -0.16

μια ένδειξη πως αν είχαμε υποθέσει κανονικότητα των καταλοίπων τότε η υπόθεση αυτή δεν θα ήταν σωστή. Η ασυμμετρία μπορεί να εξηγηθεί να σκεφτούμε πως τα δεδομένα αφορούν μέγιστες παρατηρήσεις.

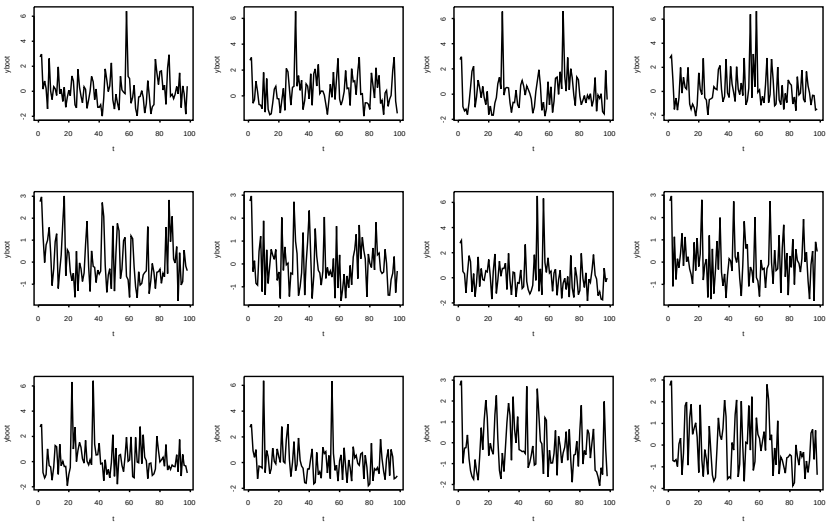
Τέλος στο γράφημα 5.21 μπορεί κανείς να δει κάποιες από τις bootstrap χρονολογικές σειρές που κατασκευάσαμε με τον αλγόριθμος που είδαμε. Αν τις συγκρίνει με τα πραγματικά δεδομένα παρατηρεί πως γενικά οι μέγιστες τιμές δεν εμφανίζονται απαραίτητα στα ίδια σημεία. Επίσης από τον πίνακα 5.12 μπορεί κάποιος να δει ότι και οι δύο παράμετροι δεν είναι στατιστικά σημαντικοί, κάτι που υποδηλώνει ότι το μοντέλο δεν είναι ικανοποιητικό.



Γράφημα 5.19: Ιστόγραμμα των Bootstrap τιμών για τις παραμέτρους $\hat{\beta}_1$ και $\hat{\beta}_2$



Γράφημα 5.20: Ιστόγραμμα καταλοίπων (αριστερά) και διάγραμμα σημείων για τις Bootstrap τιμές των παραμέτρων $\hat{\beta}_1$ και $\hat{\beta}_2$.



Γράφημα 5.21: Διάφορες σειρές που δημιουργήθηκαν με Bootstrap

5.13 Μια άλλη ματιά στο Bootstrap

Ας υποθέσουμε πως γνωρίζουμε την άγνωστη κατανομή F του πληθυσμού και πως σκοπός μας είναι να υπολογίσουμε μια ποσότητα του πληθυσμού η οποία μπορεί να εκφραστεί ως αναμενόμενη τιμή, τότε

$$\mu(F) = \int \phi(y) dF(y)$$

Χρησιμοποιούμε τον συμβολισμό $\mu(F)$ για να δηλώσουμε πως εξαρτάται από την κατανομή F του πληθυσμού. Επίσης πρέπει να αναφερθεί πως αν και εκφράζουμε την $\mu(F)$ ως ολοκλήρωμα τα αποτελέσματα ισχύουν και για άλλης μορφής ποσότητες.

Είναι γνωστό από τη θεωρία πως για να εκτιμήσουμε την παραπάνω ποσότητα αρκεί να έχουμε ένα τυχαίο δείγμα $(y_1, y_2, ..., y_M)$ από την κατανομή F τότε μπορούμε να εκτιμήσουμε την $\mu(F)$ με την

$$\hat{\mu}(F) = \frac{1}{M} \sum_{i=1}^{M} \phi(y_i)$$

Αυτή την ιδέα χρησιμοποιούμε συνέχεια όταν έχουμε ένα δείγμα στα χέρια μας και θέλουμε να εκτιμήσουμε κάποια ποσότητα του πληθυσμού. Επεκτείνοντας την ιδέα, μη έχοντας δείγμα, αν είμαστε σε θέση να προσομοιώσουμε από την F τότε μπορούμε να χρησιμοποιήσουμε την αντίστοιχη ποσότητα από το δείγμα για να εκτιμήσουμε την άγνωστη τιμή του πληθυσμού. Η αρχή αυτή, που σε κάποια βιβλία αναφέρεται ως plug-in principle, μπορεί να χρησιμοποιηθεί για οποιεσδήποτε συναρτήσεις, όχι μόνο για ολοκληρώματα. (Στην ουσία προσεγγίζουμε το ολοκλήρωμα με ένα άθροισμα στον πιο πάνω τύπο). 'Ενα τέτοιο παράδειγμα είναι η δειγματική διάμεσος που εκτιμά τη διάμεσο του πληθυσμού. Επομένως αν γνωρίζουμε την κατανομή του πληθυσμού αρκεί να πάρουμε ένα δείγμα από αυτήν (πχ με προσομοίωση) για να εκτιμήσουμε την ποσότητα που μας ενδιαφέρει.

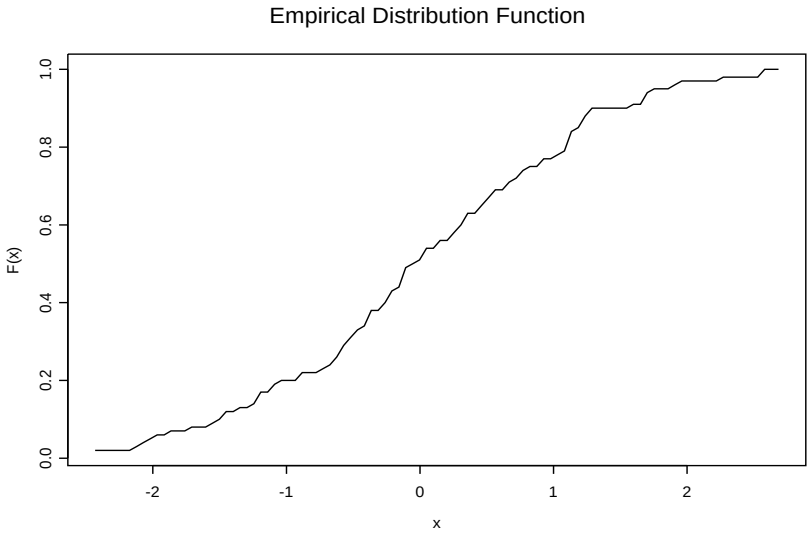
Τι γίνεται όμως αν δεν γνωρίζουμε την κατανομή του πληθυσμού;

Η ιδέα είναι να εκτιμήσουμε την άγνωστη κατανομή του πληθυσμού και να χρησιμοποιήσουμε αυτή την εκτιμήτρια $\hat{F}$ για να δημιουργήσουμε τα δείγματα μας. Επομένως χρειαζόμαστε μια εκτιμήτρια που να έχει καλές ιδιότητες. Μια τέτοια εκτιμήτρια, που είναι συνεπής, δηλαδή ασυμπτωτικά, όταν το $n \to \infty$, είναι η εμπειρική συνάρτηση κατανομής, που ορίζεται ως

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

όπου $I(\cdot)$ είναι η δείκτρια συνάρτηση. Αποδεικνύεται ότι η $\hat{F}_n \to F$ όταν $n \to \infty$ και επομένως η $\hat{F}_n$ είναι μια καλή εκτιμήτρια της άγνωστης κατανομής.

'Ενα παράδειγμα εμπειρικής συνάρτησης κατανομής μπορείτε να δείτε στο γράφημα 5.22.



Γράφημα 5.22: Παράδειγμα εμπειρικής συνάρτησης κατανομής (βασισμένη σε προσομοιωμένο δείγμα μεγέθους 100).

Η εμπειρική κατανομή χρησιμοποιείται και για άλλες στατιστικές εφαρμογές όπως για παράδειγμα στον έλεγχο Kolmogorov-Smirnov ή στην κατασκευή p-p γραφημάτων.

Συνεπώς η βασική ιδέα της μεθόδου bootstrap είναι να χρησιμοποιήσουμε την $\hat{F}_n$ στη θέση της F και συνεπώς η bootstrap εκτιμήτρια θα είναι η

$$\mu(\hat{F}) = \int \phi(y) d\hat{F}_n$$

Σε πολλές περιπτώσεις η εκτιμήτρια αυτή μπορεί να βρεθεί σε κλειστή μορφή, δηλαδή για πολλές εκτιμήτριες μπορούμε να έχουμε με bootstrap την εκτιμήτρια σε κλειστή μορφή. 'Ενα τέτοιο παράδειγμα είναι η μέση τιμή. Μπορεί να αποδειχτεί πως ο δειγματικός μέσος είναι η bootstrap εκτιμήτρια της αναμενόμενης τιμής του πληθυσμού.

Σε αρκετές όμως περιπτώσεις που κάτι τέτοιο δεν είναι εφικτό ή, ακόμα και αν η bootstrap εκτιμήτρια μπορεί να βρεθεί σε κλειστή μορφή δεν ισχύει το ίδιο για το τυπικό της σφάλμα το οποίο μας ενδιαφέρει περισσότερο από την ίδια την εκτιμήτρια στις εφαρμογές.

Στις περιπτώσεις που δεν είναι εφικτό να βρούμε την Bootstrap εκτιμήτρια σε κλειστή μορφή τότε χρησιμοποιούμε την εκτιμήτρια της βασισμένη σε προσομοίωση, δηλαδή παίρνουμε δείγμα από την $\hat{F}_n$ και χρησιμοποιούμε την

$$\hat{\mu}(\hat{F}_n) = \frac{1}{M} \sum_{i=1}^{M} \phi(y_i^*),$$

όπου $(y_1^*, y_2^*, ..., y_M^*)$ είναι ένα δείγμα από την $\hat{F}_n$. Συνεπώς το πρόβλημα έχει μεταφερθεί στο να προσομοιώσουμε από την $\hat{F}_n$, κάτι το οποίο είναι πολύ εύκολο όπως είδαμε στην εισαγωγή.

Αν συμβολίσουμε με $\hat{\theta}$ την εκτιμήτρια από το δείγμα της παραμέτρου θ και με $\hat{\theta}_B$ την bootstrap εκτιμήτρια, τότε αν $\hat{F}_n \to F$ τότε και $\hat{\theta}_B \to \hat{\theta}$. Το πιο σημαντικό όμως αποτέλεσμα είναι πως η εκτίμηση του τυπικού σφάλματος της $\hat{\theta}_B$ είναι μια καλή εκτίμηση και του τυπικού σφάλματος της $\hat{\theta}$ και καθώς είναι σχετικά εύκολο να βρούμε την τυπική απόκλιση της $\hat{\theta}_B$ αφού πολύ εύκολα παίρνουμε ένα τυχαίο δείγμα από αυτή (το μόνο που χρειάζεται είναι να προσομοιώσουμε από την $\hat{F}_n$, κάτι που όπως είπαμε είναι απλό) συνεπώς καταλήγουμε να έχουμε μια εκτίμηση και για τη μεταβλητότητα της $\hat{\theta}$. Επίσης οι κατανομές των $\hat{\theta}$ και $\hat{\theta}_B$ μοιάζουν πολύ καθώς η εμπειρική κατανομή είναι μια καλή εκτίμηση της άγνωστης κατανομής του πληθυσμού και συνεπώς το τυχαίο δείγμα από την κατανομή της $\hat{\theta}_B$ είναι μια καλή προσέγγιση ενός τυχαίου δείγματος από την $\hat{\theta}$.

Ολοκληρώνοντας αυτή την ενότητα, καταλήγουμε πως ξεκινάμε από την εμπειρική κατανομή που είναι μια καλή εκτιμήτρια της άγνωστης κατανομής, και βρίσκοντας εκτιμήτριες βασισμένες σε αυτή ουσιαστικά μεταφέρουμε τις καλές ιδιότητες.

5.14 Bootstrap για εξαρτημένα δεδομένα

'Οπως είπαμε η βασική αρχή πάνω στην οποία στηρίζεται το bootstrap είναι ότι παίρνουμε ανεξάρτητα δείγματα. Επίσης, στηριζόμαστε στο γεγονός ότι η εμπειρική κατανομή είναι μια καλή εκτίμηση της άγνωστης κατανομής του πληθυσμού. Και τα δύο δεν ισχύουν στην περίπτωση που έχουμε δεδομένα που δεν είναι ανεξάρτητα (πχ χρονολογικές σειρές). Προκειμένου να χρησιμοποιήσουμε λοιπόν bootstrap θα πρέπει να καταφύγουμε σε διαφορετικές προσεγγίσεις.

Το μεγαλύτερο πρόβλημα είναι πως μπορεί να έχουμε διαφορετικούς τρόπους εξάρτησης, οπότε δεν υπάρχει γενικά μια μέθοδος που να δουλεύει σε κάθε περίπτωση. Θα δούμε σύντομα κάποιες από τις μεθόδους αυτές

5.14.1 Block Bootstrap

Η ιδέα είναι πως προκειμένου να διατηρήσουμε στα δεδομένα μας κάποια από την εξάρτηση που έχουν, αντί να κάνουμε δειγματοληψία από τις παρατηρήσεις, κάνουμε δειγματοληψία από blocks παρατηρήσεων. Τα blocks αν είναι καλά διαλεγμένα διατηρούν αρκετή από την πληροφορία που μας ενδιαφέρει και συνεπώς μπορούμε να έχουμε αρκετή από την πληροφορία στα δείγματα μας. Η ιδέα είναι λοιπόν πως από το σύνολο των n παρατηρήσεων κατασκευάζουμε b blocks από ℓ διαδοχικές παρατηρήσεις στο καθένα. Σε περίπτωση που δεν ισχύει ότι $b\ell = n$, τότε κάποιο block μπορεί να έχει λιγότερες παρατηρήσεις. Η ιδέα είναι λοιπόν αντί να κάνουμε δειγματοληψία με επανάθεση από τις παρατηρήσεις, κάνουμε δειγματοληψία με επανάθεση από τα blocks. Επειδή μέσα σε κάθε block έχουμε διαδοχικές παρατηρήσεις, ουσιαστικά κρατάμε την πληροφορία για εξάρτηση μέχρι τάξεως ℓ. Στα σημεία που ενώνονται τα blocks χάνουμε κάθε πληροφορία. Συνεπώς με αυτόν τον τρόπο μπορούμε να αναπαράγουμε μέρος της πληροφορίας αλλά σίγουρα προσθέτουμε θόρυβο στα δεδομένα και η αναπαράσταση της εξάρτησης δεν είναι

πλήρης. Παρόλα αυτά έχει μελετηθεί αρκετά το πόσο πιστά μια τέτοια προσέγγιση μπορεί να αναπαραστήσει την εξάρτηση και τα αποτελέσματα είναι ενθαρρυντικά.

Παράδειγμα:Ας υποθέσουμε πως έχουμε πως έχουμε 12 παρατηρήσεις $(x_1, ..., x_{12})$. Τότε δημιουργούμε 4 blocks από 3 παρατηρήσεις, δηλαδή ορίζουμε $y_1 = (x_1, x_2, x_3)$, $y_2 = (x_4, x_5, x_6)$, $y_3 = (x_7, x_8, x_9)$ και $y_4 = (x_{10}, x_{11}, x_{12})$. Στη συνέχεια κάνουμε δειγματοληψία με επανάθεση από τα $y_i, i = 1, 2, 3, 4$. Δεδομένου πως το μήκος των blocks είναι 3, δεν είμαστε σε θέση να αναπαραστήσουμε την αυτοσυσχέτιση βαθμού μεγαλύτερου από 2 στα δεδομένα μας.

5.14.2 Moving blocks

'Αμεσα σχετισμένη με την παραπάνω μέθοδος είναι η μέθοδος moving blocks. Σε αυτή την περίπτωση, δημιουργούμε τα Block με επικάλυψη, δηλαδή διαφορετικά Blocks περιέχουν την ίδια παρατήρηση. Σε σχέση με το παράδειγμα που μόλις είδαμε δημιουργούμε τα Blocks ως

$y_1 = (x_1, x_2, x_3), \quad y_2 = (x_2, x_3, x_4), \quad y_3 = (x_3, x_4, x_5) \quad ... \quad y_{11} = (x_{11}, x_{12}, x_1)$

και $y_{12} = (x_{12}, x_1, x_2)$ και στη συνέχεια κάνουμε δειγματοληψία με επανάθεση από τα $y_i, i = 1, ..., 12$.

Με αυτή την προσέγγιση ο θόρυβος που προσθέτουμε είναι λιγότερος και επομένως διατηρούμε περισσότερη πληροφορία.

5.14.3 Παραμετρικές μέθοδοι

Τέλος μια άλλη προσέγγιση σε αρκετά προβλήματα εξάρτησης και κυρίως χρονολογικές σειρές, προσαρμόζουμε ένα παραμετρικό μοντέλο στα δεδομένα μας και στη συνέχεια χρησιμοποιούμε parametric bootstrap. Για παράδειγμα, χρονολογικές σειρές με γραμμική συσχέτιση μέχρι κάποια τάξη μπορούν να προσεγγιστούν πολύ καλά από αυτοπαλίνδρομα μοντέλα AR. Συνεπώς προσαρμόζουμε στα δεδομένα μας ένα τέτοιο μοντέλο το οποίο στη συνέχεια χρησιμοποιούμε για να δημιουργήσουμε τα bootstrap δείγματα.

5.15 Subsampling

Η μέθοδος subsampling βασίζεται στην ιδέα του jackknife αλλά θα δούμε σε λίγο ότι μοιάζει πολύ με τη μέθοδο bootstrap. Συγκεκριμένα, η μέθοδος jackknife αφαιρεί κάθε φορά μια παρατήρηση και στη συνέχεια χρησιμοποιούμε το δείγμα που προκύπτει με τις $n-1$ παρατηρήσεις. Η ιδέα μπορεί να γενικευτεί αφαιρώντας περισσότερες από μια παρατηρήσεις κάθε φορά. Ας υποθέσουμε ότι αφαιρούμε s παρατηρήσεις κάθε φορά. Σύμφωνα με τη μέθοδο πρέπει να κατασκευάσουμε όλα τα δυνατά δείγματα μεγέθους $n - s$ τα οποία είναι $\begin{pmatrix} n \\ s \end{pmatrix}$. Στην πραγματικότητα ο αριθμός αυτός είναι απαγορευτικά μεγάλος. Εναλλακτικά αντί να πάρουμε όλα τα δυνατά δείγματα αρκεί να πάρουμε τυχαία ένα δείγμα από αυτά. Αυτή είναι η ιδέα του subsampling.

Σε αυτή την προσέγγιση υπάρχει μια σημαντική διαφορά με το bootstrap.Αυτή είναι πως τα δείγματα που παίρνουμε είναι ουσιαστικά, δείγματα από την κατανομή του πληθυσμού απευθείας και όχι από την εμπειρική συνάρτηση κατανομής. Επομένως σε σχέση με το bootstrap υπάρχει ένα ξεκάθαρο πλεονέκτημα. Το μειονέκτημα

όμως είναι πως τα δείγματα δεν είναι μεγέθους n όπως το αρχικό δείγμα αλλά μικρότερα, κάτι που σημαίνει πως δεν έχουμε πληροφορία για την εκτιμήτρια από δείγμα μεγέθους n με συνέπεια να πρέπει να πολλαπλασιάσουμε το τυπικό σφάλμα με μια κατάλληλη ποσότητα για να το επαναφέρουμε στη σωστή κλίμακα.

Επίσης πρέπει να τονιστεί πως στο subsampling τα δείγματα λαμβάνονται με δειγματοληψία χωρίς επανάθεση σε αντίθεση με το bootstrap. Επίσης η αποτελεσματικότητα του subsampling δηλαδή πόσο καλό είναι, εξαρτάται από την τιμή του s σε σχέση με το n.

5.16 Ασκήσεις

1. Περιγράψτε με ακρίβεια τα βήματα για την κατασκευή ενός 99% διαστήματος εμπιστοσύνης για το συντελεστή συσχέτισης του Pearson δύο μεταβλητών που ακολουθούν από κοινού μια διμεταβλητή εκθετική κατανομή

2. 'Ενας ερευνητής από ένα δείγμα 100 παρατηρήσεων ενδιαφέρεται να μελετήσει τη μέση τιμή του πληθυσμού. Από το δείγμα βρήκε πως $\bar{x} = 160$ και $s^2 = 25$. Παίρνοντας 10 bootstrap δείγματα βρήκε τα εξής

Bootstrap δείγμα	Μέση τιμή	Τυπική απόκλιση
1	160.199	5.17
2	160.093	4.92
3	160.790	4.63
4	159.707	4.75
5	160.713	5.27
6	160.077	4.98
7	161.015	5.09
8	159.034	5.26
9	160.018	5.06
10	159.871	5.27

   Να κατασκευαστεί ένα 80% bootstrap διάστημα εμπιστοσύνης για τη μέση τιμή χρησιμοποιώντας την bootstrap-t μέθοδο και τη μέθοδο των ποσοστιαίων σημείων. Ποια μέθοδο προτιμάτε για το συγκεκριμένο πρόβλημα;

3. Οι 10 μεγαλύτερες συγκεντρώσεις όζοντος πάνω από την πόλη της Φρουτοπίας ήταν για το 2004

   85 78 105 82 84 123 111 134 96 102

   σε μικρογραμμάρια ανά κυβικό χιλιόμετρο. 'Ενας ερευνητής υποστηρίζει πως η μέση τιμή των μέγιστων συγκεντρώσεων είναι 110. Ελέγξτε την υπόθεση του ερευνητή αυτού χρησιμοποιώντας τη μέθοδο bootstrap. Από 99 bootstrap δείγματα που πήραμε κατασκευάσαμε το διάγραμμα μίσχου φύλλου που ακολουθεί. Περιγράψτε τα βήματα του ελέγχου. Ποιό είναι το p-value του ελέγχου αν μας ενδιαφέρει να δείξουμε πως η μέση μέγιστη συγκέντρωση

είναι μεγαλύτερη από 110;

1	8	5
2	8	7
3	8	8
6	9	111
11	9	22333
24	9	4444555555555
36	9	666667777777
48	9	888899999999
(16)	10	0000000011111111
36	10	222222333333
24	10	4444555
17	10	66667
12	10	888999
6	11	001
3	11	2
2	11	44

4. 'Εστω 30 παρατηρήσεις από την κατανομή Γάμμα. Βρήκαμε πως η μέση απόλυτη απόκλιση του δείγματος είναι 12. Περιγράψτε πως θα ελέγξετε την υπόθεση πως η μέση απόλυτη απόκλιση στον πληθυσμό είναι 15, χρησιμοποιώντας τη μέθοδο Monte Carlo. Μπορεί να χρησιμοποιηθεί η μέθοδος Bootstrap; Ποια από τις δύο μεθόδους προτιμάτε και γιατί;

5. Οι χρόνοι σε λεπτά 10 μαραθωνοδρόμων σε μια κούρσα μαραθωνίου δρόμου ήταν οι εξής

   132 135 140 137 162 135 140 137 137 131

   Εκτιμήστε χρησιμοποιώντας τη μέθοδο bootstrap με 10 δείγματα (B=10) το τυπικό σφάλμα και τη μεροληψία της εκτιμήτριας $\hat{\theta} = \frac{X_{\max} - X_{\min}}{2}$ όπου X_{min} και X_{max} η μικρότερη και μεγαλύτερη τιμή του δείγματος αντίστοιχα. Περιγράψτε πλήρως τη διαδικασία. (Περιγράψτε πλήρως τα δείγματα που σχηματίσατε, την τιμή της ελεγχοσυνάρτησης κλπ) και φτιάξτε ένα 90% διάστημα εμπιστοσύνης για την αντίστοιχη ποσότητα του πληθυσμού. Σχολιάστε.

6. Σε μια κλινική για να εξεταστεί η επίδραση ενός καινούριου φαρμάκου για τον πονοκέφαλο κατέγραψαν τον χρόνο ανακούφισης για το καινούριο φάρμακο αλλά και για το υπάρχον σε 10 ασθενείς. Τα δεδομένα δίνονται στον πίνακα που ακολουθεί. Εκτιμήστε με τη μέθοδο bootstrap τη μεροληψία και το τυπικό σφάλμα της εκτιμήτριας $\hat{\theta} = \frac{\bar{x}}{\bar{y}}$. Πιστεύετε πως η καινούρια θεραπεία είναι πιο αποτελεσματική;

Ασθενής	Καινούρια Θεραπεία	Υπάρχουσα θεραπεία
	x_i	y_i
1	9	10
2	8	8
3	4	5
4	5	8
5	3	10
6	7	9
7	8	7
8	3	4
9	4	4
10	9	12

7. Δίνονται 120 παρατηρήσεις που αφορούν τους χρόνους επιβίωσης σε κάποιο φάρμακο σε ώρες. Σε αυτές τις κατανομές προσαρμόστε την κατανομή με συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{p^2}{p+1}(x+1)\exp(-xp), \quad x, p > 0$$

χρησιμοποιώντας τη μέθοδο μεγίστης πιθανοφάνειας.

Βρείτε το τυπικό σφάλμα της εκτιμήτριας για το p και κάντε έναν bootstrap έλεγχο υπόθεσης για το πόσο καλά προσαρμόζει η κατανομή στα δεδομένα

Οι παρατηρήσεις είναι

0.97128	3.42496	1.01747	1.08612	0.39861	2.12319	5.54058	4.48896	3.26076
0.84119	0.66519	2.83693	1.58362	1.77516	2.68224	1.15717	1.23641	2.71796
2.23851	0.03431	3.80094	3.21728	0.03547	0.91004	0.98571	2.27045	2.94202
3.58498	3.62628	0.31190	0.05294	1.66590	5.41478	0.55897	0.72802	2.17938
2.43169	1.23713	2.13827	0.85678	1.76452	1.00452	2.94100	2.82921	3.57832
4.21120	0.16659	1.33669	1.86040	1.31140	3.45106	4.32229	1.65638	1.23359
2.83094	0.00857	0.48440	0.53558	4.62487	3.98649	0.65555	0.51840	6.40268
0.74812	0.35051	2.46949	0.10932	2.30631	1.48821	7.14091	1.24166	1.03077
5.30861	4.99441	1.56971	4.30958	1.21858	1.39507	0.53234	5.25738	2.01314
1.13676	1.06746	0.28033	1.92605	0.03841	4.22988	2.75366	0.50909	1.27649
1.67831	1.46088	5.60007	3.38162	3.84487	2.34105	2.85175	3.17934	2.56831
2.77271	4.75760	4.63306	0.21283	6.15438	0.03796	4.62189	0.61090	0.25604
2.66532	4.33870	0.46040	0.87105	2.43106	0.21129	0.05779	1.47573	0.74043
4.62278	2.23271	1.08709						

8. Δίνονται οι παρακάτω 15 παρατηρήσεις που αφορούν το μέσο όρο τάξεων στις κατατακτήριες εξετάσεις (LSAT) και στο βαθμό πτυχίου (GPA). Χρησιμοποιώντας αυτά τα δεδομένα να κατασκευάσετε όσα Bootstrap διαστήματα εμπιστοσύνης μπορείτε για τις εξής ποσότητες

   - Μέση τιμή για κάθε μεταβλητή
   - Συντελεστή μεταβλητότητας για κάθε μεταβλητή

- Συντελεστή συσχέτισης
- Και να εκτιμήσετε τη διαφορά και τη συνδιακύμανση των μέσων τιμών των δύο μεταβλητών

LSAT	576	635	558	578	666	580	555	661
GPA	3.39	3.30	2.81	3.03	3.44	3.07	3	3.43
LSAT	651	605	653	575	545	572	594	
GPA	3.36	3.13	3.12	2.74	2.76	2.88	2.96	

9. Οι 100 παρατηρήσεις που ακολουθούν αφορούν το ετήσιο μέγιστο της ροής ενός ποταμού σε εκατοντάδες χιλιάδες κυβικά μέτρα ανά δευτερόλεπτο. Σκοπός της εργασίας είναι να χρησιμοποιήσετε παραμετρικό και μη παραμετρικό Bootstrap για να ελέγξετε κατά πόσο το μοντέλο είναι πράγματι AR(1). Στην περίπτωση παραμετρικού Bootstrap υποθέστε ότι τα κατάλοιπα σας όταν διαιρέσετε με την τυπική απόκλιση τους ακολουθούν την κατανομή t με 5 βαθμούς ελευθερίας .

Έτος	Μέγιστη ροή	Έτος	Μέγιστη ροή	Έτος	Μέγιστη ροή	Έτος	Μέγιστη ροή
1898	2.63	1923	3.08	1948	3.79	1973	6.03
1899	2.63	1924	3.45	1949	2.54	1974	3.33
1900	1.88	1925	2.36	1950	2.91	1975	3.02
1901	1.73	1926	2.96	1951	6.32	1976	2.00
1902	2.35	1927	5.37	1952	4.20	1977	2.64
1903	6.69	1928	4.10	1953	2.08	1978	3.62
1904	4.06	1929	4.28	1954	1.51	1979	3.80
1905	4.69	1930	1.67	1955	1.82	1980	2.10
1906	1.97	1931	2.81	1956	1.33	1981	2.90
1907	2.84	1932	1.77	1957	2.69	1982	3.64
1908	4.57	1933	1.82	1958	3.69	1983	5.34
1909	5.10	1934	1.02	1959	2.00	1984	3.48
1910	2.96	1935	4.64	1960	4.62	1985	4.65
1911	1.41	1936	1.43	1961	4.05	1986	3.51
1912	3.61	1937	1.90	1962	3.16	1987	8.20
1913	2.64	1938	2.40	1963	1.74	1988	1.90
1914	2.12	1939	2.93	1964	2.89	1989	2.53
1915	4.96	1940	1.16	1965	3.54	1990	5.02
1916	3.22	1941	2.91	1966	1.90	1991	2.42
1917	4.49	1942	4.76	1967	5.74	1992	3.07
1918	1.86	1943	5.77	1968	1.96	1993	9.70
1919	2.45	1944	5.91	1969	3.71	1994	4.45
1920	3.01	1945	4.19	1970	3.74	1995	8.50
1921	3.09	1946	2.14	1971	2.09	1996	3.77
1922	4.51	1947	5.01	1972	2.51	1997	5.41

10. Για την εφαρμογή που είδαμε στην ανάλυσης σε κύριες συνιστώσες προσπαθήστε να δουλέψετε με ους συντελεστές της πρώτης κύριας συνιστώσας, δηλαδή να εκτιμήσετε με Bootstrap την κατανομή τους. Στη συνέχεια δημιουργήστε τα γραφήματα των ζευγών των συντελεστών αυτών. Παρατηρείτε κάτι περίεργο; Πως αυτό που βλέπετε συνδέεται με τη μοναδικότητα ή όχι των κυρίων συνιστωσών και με την περιστροφή τους;

11. Για το παράδειγμα της παλινδρόμησης που χρησιμοποιήσαμε προχωρήστε με parametric bootstrap υποθέτοντας πως η κατανομή των σφαλμάτων είναι κατανομή $t(5)$ δηλαδή t-student με 5 βαθμούς ελευθερίας και ένα μείγμα από 2 κανονικές κατανομές της μορφής $0.5N(0,1) + 0.5N(0,10)$

    Στη συνέχεια υποθέστε πάλι το ίδιο αλλά τώρα οι παράμεροι των κατανομών δεν είναι γνωστές και πρέπει να εκτιμηθούν από τα δεδομένα, δηλαδή οι βαθμοί ελευθερίας της κατανομής t-student και οι διακυμάνσεις στη μείξη των δυο κανονικών. Συγκρίνετε τα αποτελέσματα

12. Στο προηγούμενο κεφάλαιο είδαμε ότι η εκτιμήτρια μεγίστης πιθανοφάνειας του p^2 από δοκιμές Bernoulli είναι μεροληπτική και πως η jackknife έκδοση της εξαφανίζει τη μεροληψία. Χρησιμοποιήστε παραμετρικό bootstrap για να εκτιμήσετε τη μεροληψία των δύο εκτιμητριών και να επαληθεύσετε πως η μέθοδος jackknife μειώσε τη μεροληψία.

13. Έστω μια τμ Y και πολλές υποψήφιες ερμηνευτικές μεταβλητές $X_1, X_2, \ldots, X_k$. Το κριτήριο για το μοντέλο M ορίζεται ως (προσοχή στη βιβλιογραφία υπάρχουν πολλοί ισοδύναμοι ορισμοί)

$$AIC_M = 2L(M) - 2D_M$$

    όπου $L(M)$ είναι η λογαριθμική πιθανοφάνεια του μοντέλου M και D_M ο αριθμός των παραμέτρων του αντίστοιχα. Το κριτήριο του Akaike (AIC) επιλέγει ανάμεσα από πολλά υποψήφια μοντέλα το μοντέλο με τη μεγαλύτερη τιμή.

    Για τα δεδομένα που μπορείτε να βρείτε στην ιστοσελίδα

    `http://www.stat-athens.aueb.gr/~karlis/dataformscproject.txt`

    έχουμε μια τμ Y και τρεις υποψήφιες επεξηγηματικές μεταβλητές X_1, X_2, X_3. Τα υποψήφια μοντέλα είναι

$$\begin{aligned} Y &= \alpha_0 + \alpha X_1 + \beta X_2 \\ Y &= \alpha_0 + \alpha X_1 + \gamma X_3 \\ Y &= \alpha_0 + \beta X_2 + \gamma X_3 \\ Y &= \alpha_0 + \alpha X_1 \end{aligned}$$

    (a) Υποθέτοντας κανονικά σφάλματα και χρησιμοποιώντας bootstrap, χρησιμοποιήστε το κριτήριο AIC για να επιλέξετε ανάμεσα στα μοντέλα.

    (b) Για το μοντέλο που επιλέξατε χρησιμοποιώντας παραμετρικό Bootstrap κατασκευάστε με όσες μεθόδους μπορείτε 90% διαστήματα εμπιστοσύνης για το R^2 και το adjusted R^2.

    (c) Στη συνέχεια θεωρήστε το μοντέλο που επιλέξατε ως το σωστό μοντέλο και χρησιμοποιώντας bootstrap κατασκευάστε ένα 87% διάστημα εμπιστοσύνης για το AIC κάθε μοντέλου. Εκτιμήστε τον πίνακα συσχέτισης των διαφορετικών AIC.

(d) Εκτιμήστε χρησιμοποιώντας bootstrap τις πιθανότητες με τις οποίες θα επιλέγατε κάθε ένα από τα υποψήφια μοντέλα.

(e) Η μέθοδος model averaging αντί να χρησιμοποιεί ένα μόνο μοντέλο ουσιαστικά χρησιμοποιεί όλα τα υποψήφια μοντέλα για να κατασκευάσει καλύτερη πρόβλεψη. Έτσι η τιμή πρόβλεψης $\hat{Y}_i$ για την Y_i δίνεται ως

$$\hat{Y}_i = \frac{\sum_{M=1}^{p} w_M \hat{y}_i^{(M)}}{\sum_{M=1}^{p} w_M},$$

όπου $\hat{y}_i^{(M)}$ είναι η πρόβλεψη για την i τιμή όταν χρησιμοποιήσουμε το μοντέλο M, p είναι το πλήθος των διαφορετικών μοντέλων και w_M είναι το βάρος που δίνουμε σε κάθε μοντέλο και που μια λογική εκτίμηση του είναι οι πιθανότητες που υπολογίσατε στο προηγούμενο ερώτημα. Κατασκευάστε ένα 95% διάστημα εμπιστοσύνης για κάθε μια από τις τιμές $\hat{Y}_1, \ldots, \hat{Y}_{10}$.

14. Ένας ερευνητής θέλει να κατασκευάσει ένα 80% διάστημα εμπιστοσύνης για την παράμετρο θ. Το αρχικό του δείγμα είχε μέγεθος 30. Χρησιμοποιώντας bootstrap με 20 επαναλήψεις προέκυψαν από τα 20 αυτά δείγματα οι παρακάτω bootstrap τιμές $\hat{\theta}_i^*$

-1.34	-1.29	-0.59	-0.44	-0.28
-0.10	0.00	0.00	0.01	0.02
0.05	0.40	0.57	1.03	1.12
1.35	1.47	1.60	1.98	2.02

(Δίνονται $\hat{\theta} = 3.8$, $\sum \hat{\theta}_i^* = 7.557$ και $\sum (\hat{\theta}_i^*)^2 = 21.36$). Να κατασκευάσετε ένα μη συμμετρικό διάστημα εμπιστοσύνης βασισμένοι στις τιμές αυτές και να εκτιμήσετε τη μεροληψία της εκτιμήτριας $\hat{\theta}$. Πόσο καλό είναι το διάστημα που φτιάξατε;

15. Ένας ερευνητής, δουλεύοντας με εκθετικό πληθυσμό θέλει να κατασκευάσει ένα 88% διάστημα εμπιστοσύνης για την ποσότητα $T = exp(-2\hat{\lambda})$ όπου $\hat{\lambda}$ είναι η εκτιμήτρια του για να εκτιμήσει την πιθανότητα $P(X \geq 2)$ όταν τα δεδομένα προέρχονται από εκθετικό πληθυσμό. Κάνοντας 100 επαναλήψεις παραμετρικού bootstrap βρήκε τις παρακάτω 100 τιμές για το $\hat{\lambda}$:

2.11159	3.08912	3.16649	3.47683	3.56999	3.63735	3.72116	3.74121
3.79615	3.96288	3.98867	4.04300	4.05999	4.06005	4.12463	4.13644
4.17099	4.21123	4.21150	4.21242	4.21837	4.23021	4.23250	4.24699
4.27903	4.27910	4.33348	4.36302	4.37487	4.45416	4.55862	4.60402
4.60782	4.61151	4.61800	4.62797	4.70439	4.72964	4.74251	4.76682
4.78447	4.85957	4.90635	4.90764	4.91145	4.95217	4.96730	5.00858
5.01341	5.03636	5.03841	5.05075	5.07505	5.08309	5.08763	5.16117
5.18034	5.24989	5.26450	5.26561	5.26860	5.28735	5.29099	5.35492
5.38282	5.44814	5.49966	5.53346	5.54038	5.55005	5.55993	5.57750
5.58174	5.59428	5.68202	5.68815	5.86001	5.86112	5.93445	5.94656
5.94815	5.99680	6.03802	6.06910	6.08651	6.09264	6.10453	6.22306
6.33531	6.39227	6.53478	6.58061	6.64255	6.71888	6.75187	6.81866
6.92348	6.92936	7.48469	7.56646				

Να κατασκευάσετε ένα 88% δε για το T με όποια μέθοδο θέλετε.

16. Ενας ερευνητής για να περιορίσει τον υπολογιστικό φόρτο της κατασκευής των bootstrap-t διαστημάτων προτεινε το εξής διάστημα:

$$\hat{\theta} + D(a/2)se(\hat{\theta}),\ \hat{\theta} + D(1-a/2)se(\hat{\theta})$$

όπου ως συνήθως $\hat{\theta}$ είναι η εκτιμήτρια από το δείγμα, $se(\hat{\theta})$ είναι η τυπική απόκλιση της εκτιμήτριας (υπολογισμένη με bootstrap) και $D(a)$ είναι το a ποσοστιαιο σημείο της ποσότητας

$$Z(\theta_i^*) = \frac{\theta_i^* - \hat{\theta}}{s_B},$$

όπου s_B είναι η bootstrap τυπική απόκλιση της εκτιμήτριας $\hat{\theta}$ και θ_i^* είναι η i bootstrap τιμή.

Ποιές οι διαφορές του με το γνωστό bootstrap-t διάστημα; Μειώνεται ο υπολογιστικός φόρτος; Συμφωνειτε η όχι με ένα τέτοιο διάστημα εμπιστοσύνης; Εξηγήστε διεξοδικά.

Βιβλιογραφία

Efron B., Tibshirani (1993) *An Introduction to the Bootstrap*. Marcel and Decker

Το καλύτερο βιβλίο για να ξεκινήσει κανείς. Εισάγει την ιδέα του bootstrap με πολύ απλό τρόπο και περιέχει πάρα πολλά παραδείγματα.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge.

Επίσης πολύ καλό βιβλίο για εισαγωγή αν και περιέχει περισσότερες μαθηματικές αποδείξεις και θεωρητικά αποτελέσματα.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer

Σε αυτά τα δύο βιβλία περιέχεται μια θεωρητική προσέγγιση της μεθόδου και αποδεικνύονται πολλά θεωρήματα που δείχνουν πως και γιατί δουλεύει η μέθοδος. Είναι αρκετά τεχνικά και χρειάζονται αρκετή γνώση μαθηματικών και στατιστικής

Chernick, M. R. (1999) *Bootstrap Methods: A Practitioner's Guide*. Wiley

Εξαιρετικό βιβλίο για αυτούς που ενδιαφέρονται για την πρακτική πλευρά της μεθόδου. Περιέχει λίγα μαθηματικά, πολύ λίγες αποδείξεις αλλά πλήθος εφαρμογών της μεθόδου σε διάφορες επιστήμες και περιπτώσεις.

Κεφάλαιο 6

Ο Αλγόριθμος ΕΜ

6.1 Μέθοδος Μεγίστης Πιθανοφάνειας

6.1.1 Εισαγωγή

Η Μέθοδος Μεγίστης Πιθανοφάνειας (ΜΜΠ) αποτελεί την πιο γνωστή και διαδεδομένη μέθοδο εκτίμησης στη στατιστική και για αυτό το λόγο χρησιμοποιείται σε πλήθος από ποικίλες και διαφορετικές εφαρμογές. Βασικό συστατικό της είναι η πιθανοφάνεια (likelihood) του δείγματος που μετράει πόσο πιθανό είναι τα δεδομένα να έχουν προέλθει από το συγκεκριμένο μοντέλο. Όταν λέμε μοντέλο αυτό προϋποθέτει την ύπαρξη ενός σαφώς ορισμένου πιθανοθεωρητικού μοντέλου όπου πιθανότατα υπάρχουν και κάποιες άγνωστες παράμετροι που θα πρέπει να εκτιμηθούν. Η ΜΜΠ εκτιμά αυτές τις παραμέτρους χρησιμοποιώντας τις τιμές που μεγιστοποιούν την πιθανοφάνεια των δεδομένων.

Ας υποθέσουμε λοιπόν πως έχουμε ένα τέτοιο μοντέλο με συνάρτηση πυκνότητας πιθανότητας $f(x \mid \theta)$, όπου θ είναι το διάνυσμα με τις άγνωστες παραμέτρους που θέλουμε να εκτιμήσουμε. Συνήθως δουλεύουμε με ένα τυχαίο δείγμα μεγέθους n το οποίο και συμβολίζουμε ως $(x_1, x_2, \ldots, x_n)$ όπου κάθε τυχαία μεταβλητή του δείγματος προέρχεται από το μοντέλο $f(x \mid \theta)$ που έχουμε υποθέσει.

Η συνάρτηση πιθανοφάνειας του δείγματός μας δίνεται από την

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

και είναι συνάρτηση του θ. Σκοπός μας σύμφωνα με τη ΜΜΠ είναι να μεγιστοποιήσουμε τη συνάρτηση αυτή ως προς θ. Στην πράξη δεν είναι εύκολο να μεγιστοποιήσουμε τη συνάρτηση πιθανοφάνειας για αυτό και καταφεύγουμε στη μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας (loglikelihood). Επειδή ο λογάριθμος μιας συνάρτησης είναι μονότονη συνάρτηση προκύπτει πως το σημείο στο οποίο μεγιστοποιείται η λογαριθμική συνάρτηση πιθανοφάνειας είναι και το σημείο στο οποίο μεγιστοποιείται και η συνάρτηση πιθανοφάνειας. Έτσι η λογαριθμική συνά-

ρτηση πιθανοφάνειας ορίζεται ως

$$\ell(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

Επομένως το πρόβλημα είναι να μεγιστοποιήσουμε την $\ell(\theta)$ ως προς θ. Βλέποντας το πρόβλημα αυστηρά μαθηματικά, αυτό μπορεί να επιλυθεί εξισώνοντας τις πρώτες παραγώγους ως προς το 0, λύνοντας το σύστημα που προκύπτει και στη συνέχεια επιβεβαιώνοντας από τις δεύτερες παραγώγους ότι αυτό που βρήκαμε είναι μέγιστο. Θα πρέπει βέβαια να διευκρινιστεί πως πολλές φορές λανθασμένα δεν προχωράμε στις δεύτερες παραγώγους, αυτό οφείλεται σε κάποια σύγχυση καθώς για κάποιες απλές κατανομές η μελέτη της δεύτερης παραγώγου είναι απλοϊκή. Συνήθως αυτή η προσέγγιση αποτυγχάνει εκτός (ευτυχώς και δυστυχώς) από κάποιες απλές κατανομές. Όσο όμως τα μοντέλα που θέλουμε να προσαρμόσουμε γίνονται πιο πολύπλοκα τόσο και η μεγιστοποίηση γίεται ολοένα και δυσκολότερη. Τα προβλήματα που παρουσιάζονται σε πραγματικά προβλήματα είναι τα εξής:

1. Το σύστημα εξισώσεων που προκύπτει δεν είναι γραμμικό κι επομένως δεν είναι εύκολο να λυθεί.

2. Το σύστημα εξισώσεων μπορεί να μην έχει λύση μέσα στο επιτρεπτό πεδίο ορισμού των παραμέτρων, δηλαδή να βρεθεί κάποια λύση η οποία όμως να μην είναι αποδεκτή.

3. Οι παράγωγοι δεν είναι εύκολο να βρεθούν και επομένως και το σύστημα που προκύπτει δεν λύνεται εύκολα.

Ας δούμε κάποια παραδείγματα για να γίνει σαφές το επίπεδο δυσκολίας.

Παράδειγμα 6.1.Κατανομή Poisson: ΄Ενα απλό παράδειγμα είναι η περίπτωση της κατανομής Poisson όπου

$$f(x_i \mid \lambda) = \frac{exp(-\lambda)\lambda^{x_i}}{x_i!}.$$

Η λογαριθμική συνάρτηση πιθανοφάνειας δίνεται από

$$\ell(\lambda) = -n\lambda + \log\lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log x_i!$$

και επομένως για να μεγιστοποιήσουμε χρειαζόμαστε την παράγωγό της ως προς λ. Λύνοντας βρίσκουμε εύκολα πως η εκτιμήτρια μεγίστης πιθανοφάνειας είναι ο δειγματικός μέσος, δηλαδή $\hat{\lambda} = \bar{x}$.

Παράδειγμα 6.2. Ομοιόμορφη κατανομή: Ας υποθέσουμε τώρα πως $f(x_i \mid \theta) = 1/\theta$, $x_i \leq \theta$, δηλαδή το δείγμα προέρχεται από μια ομοιόμορφη κατανομή στο διάστημα $(0, \theta)$. Η λογαριθμική πιθανοφάνειας γίνεται

$$\ell(\theta) = -n \log\theta$$

και η παραγωγός της είναι $-n/\theta$, που ποτέ δεν παίρνει την τιμή 0. Συνεπώς η προσέγγιση με την παράγωγο αποτυγχάνει. Ένα απλό όμως γράφημα της λογαριθμικής συνάρτησης πιθανοφάνειας αποκαλύπτει πως αυτή παίρνει τη μέγιστη της τιμή όταν $\hat{\theta} = \max(x_i)$. Αυτή είναι λοιπόν η εκτιμήτρια μεγίστης πιθανοφάνειας (ΕΜΠ). Παρατηρείστε πως επειδή από τον ορισμό της συνάρτησης πυκνότητας πιθανότητας πρέπει όλες οι παρατηρήσεις να είναι μεγαλύτερες ή ίσες από θ οι αποδεκτές τιμές για το θ είναι μεγαλύτερες ή ίσες από τη μεγαλύτερη παρατήρηση.

Παράδειγμα 6.3 Κατανομή Γάμμα: Η περίπτωση τη κατανομής Γάμμα είναι πιο πολύπλοκη. Ας υποθέσουμε πως η συνάρτηση πυκνότητας πιθανότητας είναι η

$$f(x_i \mid \alpha, \beta) = \frac{x_i^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)} exp(-\beta x_i), \tag{6.1}$$

όπου $x, \alpha, \beta > 0$ και $\Gamma(x)$ είναι η συνάρτηση Γάμμα. Η λογαριθμική πιθανοφάνεια έχει τη μορφή

$$\ell(\alpha, \beta) = (\alpha - 1)\sum_{i=1}^{n} \log(x_i) + n\alpha \log(\beta) - n\log(\Gamma(\alpha)) - \beta \sum_{i=1}^{n} x_i$$

Οι παράγωγοί της εξισώνοντας με το 0 γίνονται

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= \sum_{i=1}^{n} \log(x_i) + n\log(\beta) - n\Psi(\alpha) = 0 \\ \frac{\partial \ell}{\partial \beta} &= \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0, \end{aligned}$$

όπου $\Psi(x)$ είναι η Δίγαμμα συνάρτηση που ορίζεται ως $\Gamma'(x)/\Gamma(x)$. Το παραπάνω σύστηνα αποτελείται από μη γραμμικές εξισώσεις και επομένως δεν υπάρχει λύση σε κλειστή μορφή. Για την επίλυση τους απαιτούνται αριθμητικές μέθοδοι.

Συνεπώς, με βάση τη σύντομη συζήτηση που προηγήθηκε γίνεται σαφές πως προκειμένου να βρεθούν οι εκτιμήτριες μέγιστης πιθανοφάνειας τα πράγματα συνήθως δεν είναι απλά και χρειάζονται αριθμητικές μέθοδοι είτε για να επιλυθούν τα συστήματα εξισώσεων που προκύπτουν είτε για την απευθείας μεγιστοποίηση των συναρτήσεων. Ο αλγόριθμος ΕΜ που θα μας απασχολήσει στη συνέχεια είναι μια τέτοια τεχνική.

Στην επόμενη ενότητα θα δούμε κάποια στοιχεία από τη θεωρία των εκτιμητριών μεγίστης πιθανοφάνειας και θα μιλήσουμε για μια απλή μέθοδο λύσης μη γραμμικών συστημάτων όπως αυτά που προέκυψαν στα παραδείγματα, τη μέθοδο Newton-Rahson. Η μέθοδος θα χρησιμοποιηθεί απλά σαν μέτρο σύγκρισης με τον αλγόριθμο ΕΜ για τον οποίο θα μιλήσουμε εκτενώς αμέσως μετά. Προφανώς στη βιβλιογραφία υπάρχουν ποικίλες άλλες μέθοδοι για τις οποίες όμως δεν θα αναφερθούμε στις σημειώσεις αυτές.

6.1.2 Βασικά στοιχεία της θεωρίας για τη Μέθοδο Μεγίστης Πιθανοφάνειας

Ας θυμηθούμε λίγα πράγματα σχετικά με τη θεωρία των εκτιμητριών μεγίστης πιθανοφάνειας που θα μας χρειαστούν στη συνέχεια. Ας υποθέσουμε ότι μας ενδιαφέρει η εκτίμηση του διανύσματος $\theta = (\theta_1, \ldots, \theta_p)$, δηλαδή θέλουμε να εκτιμήσουμε p παραμέτρους.

Το διάνυσμα των σκορ $U(\theta)$ είναι το διάνυσμα που περιέχει τις πρώτες παραγώγους της λογαριθμικής συνάρτησης πιθανοφάνειας για κάθε παράμετρο, δηλαδή

$$U(\theta) = \left(\frac{\partial \ell(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_p} \right)'$$

Είναι προφανές πως για την ΕΜΠ $\hat{\theta}$ ισχύει ότι

$$U(\hat{\theta}) = 0$$

Ασυμπτωτικά (όταν το μέγεθος του δείγματος τείνει στο ∞) ισχύει ότι (υπάρχουν και κάποιες συνθήκες που πρέπει επίσης να πληρούνται και οι οποίες συνήθως ισχύουν)

$$U(\theta) \sim N_p(0, J(\theta)),$$

δηλαδή ασυμπτωτικά ακολουθούν κανονική κατανομή, όπου $J(\theta)$ είναι ένας $p \times p$ πίνακας με στοιχεία J_{ij} που ορίζονται ως

$$J_{ij}(\theta) = -E\left[\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} \right]$$

Όταν υπολογίσουμε τον πίνακα J στην τιμή της ΕΜΠ $\hat{\theta}$, δηλαδή πάρουμε τον $J(\hat{\theta})$ τότε αυτός είναι ο πίνακας αναμενόμενης πληροφορίας του Fisher (expected Fisher Information marix) . Ισχύει επίσης πως

$$\hat{\theta} \sim N_p(\theta, J^{-1}(\theta))$$

Συνήθως στην πράξη ο πίνακας αναμενόμενης πληροφορίας του Fisher προσεγγίζεται από τον παρατηρούμενο πίνακα πληροφορίας (observed Fisher information matrix) στην τιμή $\hat{\theta}$ ο οποίος ορίζεται ως $I(\hat{\theta})$ με στοιχεία $I_{ij}(\theta)$ που υπολογίζονται ως

$$I_{ij}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j}$$

υπολογισμένα πάντα στην τιμή $\hat{\theta}$. Επομένως στην πράξη χρησιμοποιούμε το αποτέλεσμα πως ασυμπτωτικά για την ΕΜΠ έχουμε πως

$$\hat{\theta} \sim N_p(\hat{\theta}, I^{-1}(\hat{\theta}))$$

Μπορεί επίσης να δειχθεί πως ισχύει ότι

$$E\left[-\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} \right] = E\left[\frac{\partial \ell(\theta)}{\partial \theta_i} \frac{\partial \ell(\theta)}{\partial \theta_j} \right]$$

κι επομένως οι απαιτούμενοι υπολογισμοί μπορούν να περιοριστούν. Τα παραπάνω αποτελέσματα είναι πολύ χρήσιμα για τη στατιστική συμπερασματολογία σχετικά με ΕΜΠ.

6.2 Αριθμητικές μέθοδοι

Όπως είδαμε και στα προηγούμενα παραδείγματα πολλές φορές οι ΕΜΠ δεν υπάρχουν σε κλειστή μορφή και προκειμένου να βρεθούν θα πρέπει να λυθεί ένα πρόβλημα μεγιστοποίησης. Συνήθως το πρόβλημα ανάγεται στην επίλυση ενός συστήματος μη γραμμικών εξισώσεων. Είναι προφανές πως παρόμοια προβλήματα επίλυσης μη γραμμικών συστημάτων καθώς και προβλήματα μεγιστοποίησης πολύπλοκων συναρτήσεων εμφανίζονται σε πολλές άλλες επιστήμες.

Υπάρχουν ποικίλες αριθμητικές μέθοδοι για την επίλυση τέτοιων προβλημάτων. Σκοπός των σημειώσεων αυτών δεν είναι να παρουσιάσουν τις μεθόδους αυτές αλλά να περιγράψουν μια αμιγώς στατιστική προσέγγιση μέσω του αλγορίθμου ΕΜ. Παρόλα αυτά για λόγους σύγκρισης στη συνέχεια θα περιγράψουμε μια τέτοια μέθοδο και συγκεκριμένα τη μέθοδο Newton-Raphson. Θα πρέπει να σημειωθεί πως σε πολλά παρόμοια προβλήματα στη στατιστική μπορούν να χρησιμοποιηθούν και άλλες πιο ad-hoc μέθοδοι.

6.2.1 Η μέθοδος Newton-Raphson

Ας ξεκινήσουμε με την απλούστερη περίπτωση αυτής που υπάρχει μόνο μια μεταβλητή. Το πρόβλημα είναι να επιλυθεί η εξίσωση

$$F(\theta) = 0$$

Σε πολλά προβλήματα η εξίσωση που θέλουμε να λύσουμε δεν έχει στο δεξί μέλος το 0. Η μέθοδος απαιτεί να υπάρχει το 0 στο δεξί μέλος και αυτό μπορεί εύκολα να επιτευχθεί με απλή αφαίρεση.

Η μέθοδος Newton-Raphson είναι μια επαναληπτική μέθοδος επίλυσης εξισώσεων, δηλαδή σε κάθε επαναληψη βελτιώνουμε την ήδη υπάρχουσα λύση. Η μέθοδος ξεκινά με μια αρχική λύση $\theta^{(0)}$ (η οποία μπορεί να επιλεγεί είτε τυχαία είτε με βάση κάποια λογική επιλογή) και σε κάθε επανάληψη ανανεώνουμε τη λύση μας. Αν είμαστε στην r επανάληψη τότε η καινούρια λύση προκύπτει χρησιμοποιώντας τον παρακάτω τύπο

$$\theta^{(r+1)} = \theta^{(r)} - \frac{F(\theta^{(r)})}{F'(\theta^{(r)})}$$

Οι επαναλήψεις συνεχίζονται μέχρι κάποιο κριτήριο τερματισμού να ικανοποιηθεί, για παράδειγμα σταματάμε αν η διαφορά ανάμεσα σε δύο διαδοχικές λύσεις είναι μικρότερη από 10^{-4}, δηλαδή αν $|\theta^{(r+1)} - \theta^{(r)}| \leq 10^{-4}$.

Η μέθοδος Newton-Raphson είναι εξαιρετικά απλή και επιτρέπει πολύ απλή γεωμετρική ερμηνεία. Χρειάζεται μόνο ο υπολογισμός της πρώτης παραγώγου. Πλεονεκτήματα της μεθόδου είναι η ευκολία της στον προγραμματισμό σχεδόν σε οποιοδήποτε πακέτο και η μεγάλη ταχύτητα σύγκλισης, δηλαδή η λύση προκύπτει μετά από λίγα μόνο βήματα.

Μερικά από τα μειονεκτήματα της μεθόδου είναι πως:

- μπορεί η λύση να είναι έξω από τα επιτρεπτά για το πρόβλημά μας όρια,
- μπορεί να σταματήσει σε μια λύση η οποία όμως δεν είναι μοναδική και
- η λύση που θα βρεθεί αν υπάρχουν περισσότερες από μια λύσεις εξαρτάται από την αρχική τιμή.

Θα πρέπει να παρατηρήσει κανείς πως αυτά είναι μειονεκτήματα σχεδόν των περισσότερων επαναληπτικών αριθμητικών μεθόδων και όχι μόνο της μεθόδου Newton-Raphson.

Ας δούμε όμως ένα απλό παράδειγμα εφαρμογής της μεθόδου για την εύρεση μιας ΕΜΠ.

6.2.2 Περικομμένη κατανομή Poisson

Η περικομμένη κατανομή Poisson (truncated Poisson distribution) προκύπτει από την απλή κατανομή Poisson όταν η τιμή 0 δεν μπορεί να παρατηρηθεί. Κάτι τέτοιο συμβαίνει σε πολλές επιστημονικές περιοχές. Για παράδειγμα στην οικολογία όταν θέλουμε να μελετήσουμε τον αριθμό των ζώων που απαρτίζουν ένα κοπάδι η τιμή 0 δεν μπορεί να παρατηρηθεί. Το ίδιο ισχύει στη λογομετρία όπου μελετάμε τον αριθμό των λέξεων που απαρτίζουν μια πρόταση. Κάτι τέτοιο είναι πολύ χρήσιμο για έξυπνα συστήματα αναγνώρισης κειμένων. Στην επιδημιολογία αν μας ενδιαφέρει ο αριθμός των ατόμων που αρρώστησαν από έναν ιό (επιδημία) η τιμή 0 δεν μπορεί να παρατηρηθεί. Υπάρχουν πολλά άλλα παραδείγματα χρήσης παρόμοιων μοντέλων.

Η συνάρτηση πιθανότητας της κατανομής δίνεται από τον τύπο

$$P(X=x)=\frac{exp(-\lambda)\lambda^x}{x!(1-exp(-\lambda))}=\frac{\lambda^x}{x!(e^\lambda-1)},\ \ \lambda>0$$

Από ένα τυχαίο δείγμα $(x_1,x_2,\ldots,x_n)$ η λογαριθμική συνάρτηση πιθανοφάνειας δίνεται ως

$$\ell(\lambda)=\log\lambda\sum_{i=1}^{n}x_i-n\log\left(e^\lambda-1\right)-\sum_{i=1}^{n}(\log x_i!)$$

Η παράγωγος ως προς λ είναι

$$\frac{d\ell(\lambda)}{d\lambda}=\frac{\sum\limits_{i=1}^{n}x_i}{\lambda}-\frac{ne^\lambda}{e^\lambda-1}=0$$

η οποία δεν μπορεί να λυθεί σε κλειστή μορφή και είναι προφανώς μη γραμμική. Για να χρησιμοποιήσουμε τη μέθοδο Newton-Raphson μπορούμε να γράψουμε την εξίσωση στη μορφή

$$F(\lambda)=\frac{\lambda}{\bar{x}}-1+e^{-\lambda}=0$$

και με βάση αυτά που είπαμε προηγουμένως η επαναληπτική διαδικασία θα έχει τη μορφή (λ^{old} είναι η τιμή μέχρι τώρα)

$$\lambda^{(new)}=\lambda^{old}-\frac{\frac{\lambda^{old}}{\bar{x}}-1+e^{-\lambda^{old}}}{\frac{1}{\bar{x}}-e^{-\lambda^{old}}}$$

x	1	2	3	4	5	6	7
συχνότητα	797	301	77	17	6	1	1

Πίνακας 6.1: Συχνότητα εμφάνισης λέξεων

Το επαναληπτικό σχήμα σταματά όταν η αλλαγή στην τιμή του λ είναι μικρότερη από κάποια μικρή τιμή που εμείς επιλέγουμε (πχ 10^{-4}).

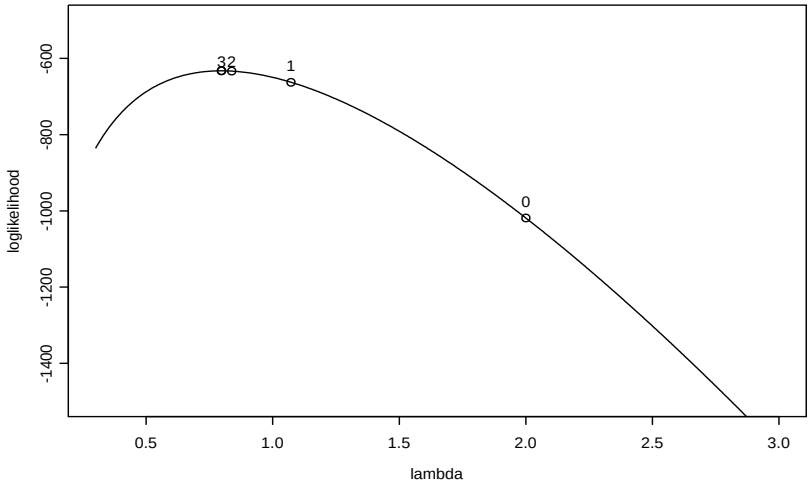
6.2.3 Εφαρμογή

Τα δεδομένα του πίνακα 6.1 αναφέρονται στην συχνότητα εμφάνισης των λέξεων σε ένα κείμενο. Έτσι 797 λέξεις εμφανίζονται μόνο μια φορά στο κείμενο, 301 λέξεις εμφανίζονται 2 φορές κλπ. Το χαρακτηριστικό αυτό (συχνότητα εμφάνισης λέξεων) χρησιμοποιείται πολύ για να αναγνωρίσει την ταυτότητα του συγγραφέα ενός κειμένου.

Θέλουμε να προσαρμόσουμε την περικομμένη κατανομή Poisson σε αυτά τα δεδομένα βρίσκοντας την ΕΜΠ της παραμέτρου λ. Όπως είδαμε η ΕΜΠ δεν μπορεί να βρεθεί σε κλειστή μορφή και επομένως θα χρησιμοποιήσουμε τη μέθοδο Newton-Raphson για να βρούμε την ΕΜΠ. Ξεκινώντας με αρχική τιμή $\lambda^{(0)} = 2$, βρήκαμε πως $\hat{\lambda} = 0.7969$ μετά από 4 επαναλήψεις. Στην ίδια λύση καταλήγουμε για κάθε αρχική τιμή $\lambda^{(0)} \geq 0.373138$. Αν η αρχική τιμή είναι $\lambda^{(0)} < 0.373138$ τότε ο αλγόριθμος δίνει ως λύση την τιμή 0 η οποία δεν είναι αποδεκτή καθώς έχουμε θέσει πως $\lambda > 0$. Το κριτήριο που χρησιμοποιήσαμε για να σταματήσουμε ήταν $|\lambda^{(r+1)} - \lambda^{(r)}| < 10^{-6}$.

Στο γράφημα 6.1 μπορεί να δει κανείς τη λογαριθμική συνάρτηση πιθανοφάνειας. Δεν έχει αφαιρεθεί η σταθερή ποσότητα $\sum_{i=1}^{n}(\log x_i!)$. Με τους αριθμούς συμβολίζουμε τις διαδοχικές λύσεις που βρίσκουμε από τον αλγόριθμό μας. Γιατί όμως η μέθοδος αποτυγχάνει να μας δώσει τη λύση αν ξεκινήσουμε από κακή αρχική τιμή; Η απάντηση μπορεί να δοθεί με τη μελέτη του Γραφηματος 6.2.

Στο Γράφημα 6.2 βλέπουμε τη συνάρτηση $F(\lambda) = \frac{\lambda}{\bar{x}} - 1 + e^{-\lambda}$. Η παράλληλη γραμμή αντιστοιχεί στην τιμή 0 κι επομένως οι λύσεις της εξίσωσης $F(\lambda) = 0$ είναι οι τιμές που η συνάρτηση τέμνεται με την παράλληλη γραμμή. Επομένως υπάρχουν 2 διαφορετικές λύσεις εκ των οποίων η μια δεν είναι συμβατή με το μοντέλο που χρησιμοποιούμε κι άρα απορρίπτεται (επειδή $\lambda < 0$). Η μέθοδος Newton-Raphson είναι μια αριθμητική μέθοδος επίλυσης εξισώσεων και δεν λαμβάνει υπόψη της τυχόν στατιστικές υποθέσεις. Αυτό είναι και το μειονέκτημα της σε πολλά προβλήματα στατιστικού περιεχομένου. Τέλος θα πρέπει να παρατηρήσουμε ότι η τιμή 0.373138 η οποία καθορίζει ποια λύση θα βρεθεί δεν είναι παρά το ελάχιστο σημείο της συνάρτησης $F(\lambda)$. Αυτό αποδεικνύεται εύκολα καθώς η μέθοδος Newton-Raphson χρησιμοποιεί τις εφαπτομένες για να καθορίσει το επόμενο σημείο με αποτέλεσμα αν βρεθεί σε τιμή μικρότερη του 0.373138 θα κινηθεί προς τα αριστερά και θα βρει την άλλη (μη αποδεκτή) λύση.



Γράφημα 6.1: Η λογαριθμική συνάρτηση πιθανοφάνειας για τα δεδομένα μας. Τα σημεία υποδεικνύουν τις λύσεις σε κάθε επανάληψη

6.2.4 Η μέθοδος Newton-Raphson στις περισσότερες διαστάσεις

Η μέθοδος Newton-Raphson μπορεί εύκολα να γενικευτεί σε περισσότερες από μια διαστάσεις. Θα δούμε τη μέθοδο απευθείας για τη μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας αλλά είναι εύκολο να δει κανείς τον τρόπο με τον οποίο αποτελεί απλή γενίκευση της μονοδιάστατης περίπτωσης. Έστω πως θέλουμε να μεγιστοποίησουμε τη λογαριθμική πιθανοφάνεια $\ell(\theta)$ ως προς κάποιο διάνυσμα $\theta = (\theta_1, \ldots, \theta_p)$. Ο αλγόριθμος έχει την εξής μορφή:

Ξεκινάμε με μια αρχική τιμή $\theta^{(0)}$ και στη συνέχεια χρησιμοποιούμε τον επαναληπτικό τύπο

$$\theta^{(r+1)} = \theta^{(r)} - \mathbf{A}^{-1}\theta^{(r)} g(\theta^{(r)})$$

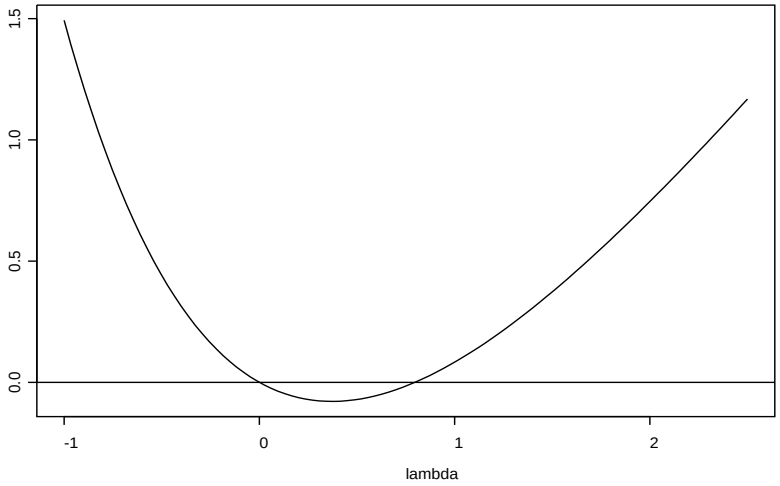
όπου

$$g(\theta) = \left(\frac{\partial \ell(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_p} \right)'$$

(ουσιαστικά το διάνυσμα των σκορ που είδαμε προηγουμένως) και $\mathbf{A}$ είναι ένας $p \times p$ πίνακας με στοιχεία a_{ij} που ορίζονται ως

$$a_{ij} = \frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j}$$

Οι επαναλήψεις σταματάνε όταν κάποιο κριτήριο ικανοποιηθεί.



Γράφημα 6.2: Η συνάρτηση $F(\lambda) = \frac{\lambda}{\bar{x}} - 1 + e^{-\lambda}$. Οι λύσεις είναι τα σημεία που η συνάρτηση τέμνει την παράλληλη γραμμή που αντιστοιχεί στο 0.

Θα πρέπει να τονιστεί ότι στην περίπτωση που $p = 1$ η μέθοδος είναι ίδια με αυτή που είδαμε πριν. Οι όποιες διαφορές οφείλονται στο ότι τώρα μεγιστοποιούμε άμεσα την πιθανοφάνεια που αντιστοιχεί στο να εξισώσουμε τις πρώτες παραγώγους με 0.

6.2.5 Παράδειγμα: Η κατανομή Γάμμα

Είδαμε σε ένα προηγούμενο παράδειγμα πως για την κατανομή Γάμμα οι ΕΜΠ δεν υπάρχουν σε κλειστή μορφή και πως το σύστημα εξισώσεων που προκύπτει είναι μη γραμμικό. Θα χρησιμοποιήσουμε τη μέθοδο Newton-Raphson για να βρούμε τις ΕΜΠ.

Η λογαριθμική πιθανοφάνεια είναι

$$\ell(\alpha, \beta) = (\alpha - 1) \sum_{i=1}^{n} \log(x_i) + n\alpha \log(\beta) - n \log(\Gamma(\alpha)) - \beta \sum_{i=1}^{n} x_i \qquad (6.2)$$

κι επομένως

$$g(\theta) = \left(\sum_{i=1}^{n} \log(x_i) + n \log(\beta) - n\Psi(\alpha), \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i \right)'$$

ενώ

$$\mathbf{A} = \begin{bmatrix} -n\Psi_3(\alpha) & \frac{n}{\beta} \\ \frac{n}{\beta} & -\frac{n\alpha}{\beta^2} \end{bmatrix}$$

όπου $\Psi(x)$ είναι η Δίγαμμα συνάρτηση και $\Psi_3(x)$ η τρίγαμμα συνάρτηση που ορίζεται απλά ως η παράγωγος της δίγαμμα συνάρτησης. Με βάση αυτά το επαναληπτικό σχήμα γίνεται ως εξής: Από τις μέχρι τώρα τιμές α και β, βρες τις καινούριες τιμές με τη χρήση του παρακάτω τύπου

$$\begin{bmatrix} \alpha^{new} \\ \beta^{new} \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - \begin{bmatrix} -n\Psi_3(\alpha) & \frac{n}{\beta} \\ \frac{n}{\beta} & -\frac{n\alpha}{\beta^2} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} \log(x_i) + n\log(\beta) - n\Psi(\alpha) \\ \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i \end{bmatrix}$$

Τα περισσότερα στατιστικά πακέτα παρέχουν συναρτήσεις για τον υπολογισμό των συναρτήσεων $\Psi(x)$ και $\Psi_3(x)$.

6.3 Η μέθοδος Scoring

Η μέθοδος scoring είναι επίσης πολύ χρήσιμη για στατιστικές εφαρμογές και αποτελεί παραλλαγή της μεθόδου Newton-Raphson. Η βασική ιδέα είναι πως στον επαναληπτικό τύπο (6.2.4) αντί να χρησιμοποιήσουμε τον πίνακα $\mathbf{A}$, χρησιμοποιούμε την αναμενόμενη του τιμή $E(\mathbf{A})$. Σε πολλές περιπτώσεις οι δύο μέθοδοι συμπίπτουν. Για παράδειγμα στην κατανομή Γάμμα που είδαμε πριν από λίγο ο πίνακας $\mathbf{A}$ έχει μόνο σταθερές κι επομένως $E(\mathbf{A}) = \mathbf{A}$, δηλαδή οι δύο μέθοδοι συμπίπτουν.

Μερικά σημαντικά στοιχεία για τη σύγκριση των δύο μεθόδων είναι τα εξής

- Η μέθοδος scoring οδηγεί στην αποφυγή του υπολογισμού των δευτέρων παραγώγων κι έτσι μπορεί να είναι πιο οικονομική από ότι η μέθοδος Newton-Raphson.
- Από την άλλη πλευρά οι δεύτερες παράγωγοι είναι χρήσιμες για τον υπολογισμό των τυπικών σφαλμάτων των ΕΜΠ και επομένως αυτό δίνει ένα πλεονέκτημα στη μέθοδο Newton-Raphson
- Και στις δύο περιπτώσεις απαιτείται ο αντίστροφος ενός πίνακα διαστάσεων $p \times p$ σε κάθε επανάληψη. Αυτό, ειδικά αν το p είναι μεγάλο μπορεί να είναι πολύ χρονοβόρο

6.3.1 Μερικές χρήσιμες ιδέες

Σε κάποιες περιπτώσεις μπορούμε να βελτιώσουμε την αποδοτικότητα των αριθμητικών μεθόδων παρατηρώντας πως κάποιες από τις εξισώσεις που προκύπτουν μπορούν να λυθούν και επομένως μπορούμε να μειώσουμε τη διάσταση του συστήματος.

Για παράδειγμα έστω η κατανομή Weibull με συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \alpha\beta(\beta x)^{\alpha-1} exp\left((-\beta x)^{\alpha}\right)$$

Από ένα τυχαίο δείγμα $(x_1, x_2, \ldots, x_n)$ η λογαριθμική συνάρτηση πιθανοφάνειας γίνεται

$$\ell(\alpha, \beta) = n\log(\alpha) + n\alpha\log(\beta) + (\alpha - 1)\sum_{i=1}^{n} \log(x_i) - \beta^{\alpha}\sum_{i=1}^{n} x_i^{\alpha}$$

Εξισώνοντας τις πρώτες παραγώγους με 0 παίρνουμε

$$\frac{\partial \ell}{\alpha} = \frac{n}{\alpha} + n\log(\beta) - \beta^{\alpha}\sum_{i=1}^{n} x_i^{\alpha}\log(\beta x_i) + \sum_{i=1}^{n}\log x_i \tag{6.3}$$

$$\frac{\partial \ell}{\beta} = \frac{n\alpha}{\beta} - \alpha\beta^{\alpha-1}\sum_{i=1}^{n} x_i^{\alpha} \tag{6.4}$$

όμως η δεύτερη εξίσωση μπορεί να λυθεί εύκολα δίνοντας

$$\hat{\beta} = \left(\frac{n}{\sum\limits_{i=1}^{n} x_i^{\alpha}}\right)^{1/\alpha}$$

Αντικαθιστώντας στην πρώτη εξίσωση έχουμε πια μόνο μια εξίσωση να λύσουμε και χρειαζόμαστε τη μέθοδο Newton-Raphson στη μια διάσταση αντί για τις 2 όπως πριν.

Εκτός από τις μεθόδους που επιλεκτικά αναφέραμε θα πρέπει να τονίσουμε την ύπαρξη πολλών άλλων αριθμητικών μεθόδων μεγιστοποίησης και επίλυσης μη γραμμικών συστημάτων. Εκτός από τις κλασικές αριθμητικές μεθόδους τα τελευταία χρόνια έχουν αναπτυχθεί και αρκετές άλλες μέθοδοι που βασίζονται στην προσομοίωση και οι οποίες έχουν πολλές ενδιαφέρουσες εφαρμογές στη στατιστική. Αυτοί οι αλγόριθμοι ονομάζονται και στοχαστικοί αλγόριθμοι και έχουν την πολύ χρησιμη ιδιότητα ότι δεν χρειαζεται ο υπολογισμός παραγώγων και αρα μπορεί να χρησιμοποιηθουν και για προβλήματα βελτιστοποιήσης που ειναι από τη φύση τους διακριτά. Θα δούμε στη συνέχεια, και πριν μιλήσουμε για τον ΕΜ, μερικούς τετοιους αλγορίθμους.

6.4 Στοχαστικοί αλγόριθμοι

Στις ενότητες που ακολουθούν θα χρησιμοποιήσουμε την προσομοίωση για τη βελτιστοποίηση συναρτήσεων. Σε πολλά προβλήματα σκοπός μας είναι να βρούμε τις τιμές των παραμέτρων που βελτιστοποιούν μια αντικειμενική συνάρτηση. Για παράδειγμα στη στατιστική ενδιαφερόμαστε να μεγιστοποιήσουμε την πιθανοφάνειας ενός δείγματος. Στο πρόβλημα του πωλητή ενδιαφερόμαστε να βρούμε τη βέλτιστη διαδρομή. Μέχρι τώρα σε πολλά προβλήματα αυτής της μορφής χρησιμοποιούσαμε αριθμητικές μεθόδους βελτιστοποίησης. Σε αυτό το κεφάλαιο θα δούμε στοχαστικούς αλγόριθμους, δηλαδή αλγόριθμους που χρησιμοποιούν ιδέες από την προσομοίωση για την βελτιστοποίηση συναρτήσεων.

Αυτή η κατηγορία των αλγόριθμων μεγιστοποίησης αναπτύσσεται αρκετά ραγδαία τα τελευταία χρόνια εξαιτίας της χρήσης ισχυρών υπολογιστών που μπορούν να κάνουν πολλές πράξεις σε μικρό χρονικό διάστημα. Η βασική τους ιδέα είναι ότι ψάχνουν με τυχαίο τρόπο τον παραμετρικό χώρο για να βρουν που μεγιστοποιείτε [1]

[1] θα χρησιμοποιούμε τη λέξη μεγιστοποίηση αλλά σε πολλά προβλήματα μας ενδιαφέρει η ελαχιστοποίηση. Στην πράξη αρκεί να αλλάξουμε το πρόσημο της αντικειμενικής συνάρτησης και να μετατρέψουμε ένα πρόβλημα ελαχιστοποίησης σε πρόβλημα μεγιστοποίησης

μια συνάρτηση. Το μεγάλο τους πλεονέκτημα είναι πως δεν απαιτούν τον υπολογισμό παραγώγων οι οποίες σε πολλά προβλήματα είναι ιδιαίτερα δύσκολες και χρονοβόρες. Επομένως αρκεί η δυνατότητα υπολογισμού της αντικειμενικής συνάρτησης (δηλαδή της συνάρτησης που θέλουμε να βελτιστοποιήσουμε) για να δουλέψουν. Από αυτή την άποψη οι μέθοδοι αυτές μπορούν να χρησιμοποιηθούν και να προγραμματιστούν σχετικά εύκολα σε μια μεγάλη ποικιλία προβλημάτων.

Στη συνέχεια θα εξετάσουμε εν συντομία κάποιες από αυτές τις μεθόδους.

6.5 Τυχαίο Ψάξιμο (Random Search)

Η βασική ιδέα του πιο απλού στοχαστικού αλγόριθμου είναι ότι απλά διαλέγουμε τυχαία σημεία μέσα στον παραμετρικό χώρο, υπολογίζουμε στα σημεία αυτά την αντικειμενική συνάρτηση και κρατάμε το σημείο που έδωσε μεγαλύτερη τιμή. Η ιδέα είναι λοιπόν απλή αρκεί να μπορούμε να πάρουμε εύκολα σημεία στον παραμετρικό χώρο κάτι που συνήθως δεν είναι δύσκολο.

Μπορεί εύκολα να δει κανείς ότι αν το πλήθος των σημείων τείνει στο ∞ ο αλγόριθμος θα βρει το μέγιστο. Στην πραγματικότητα αν το πλήθος των σημείων τείνει στο ∞ ουσιαστικά κάνουμε πλήρες ψάξιμο σε όλο τον παραμετρικό χώρο. Στην πράξη βέβαια δεν μπορούμε να πάρουμε παρά έναν πεπερασμένο αριθμό σημείων και σε αυτή την περίπτωση είναι πιθανό να μην βρούμε το μέγιστο αλλά κάποιο σημείο πολύ κοντά σε αυτό

Από την άλλη αν το πλήθος των παραμέτρων της συνάρτησης που θέλουμε να μεγιστοποιήσουμε είναι μεγάλο αυτό σημαίνει πως αφενός χρειαζόμαστε ολοένα και περισσότερα σημεία για να βρεθεί το μέγιστο (ή καλύτερα για να βρούμε μια τιμή όσο γίνεται πιο κοντά σε αυτό) ενώ από την άλλη και η επιλογή τυχαία των σημείων γίνεται ολοένα και πιο χρονοβόρα. Για παράδειγμα φανταστείτε ότι θέλουμε να μεγιστοποιήσουμε την πιθανοφάνεια ενός σετ δεδομένων από την κανονική κατανομή. Στη μια διάσταση έχουμε μόλις 2 παραμέτρους και η επιλογή σημείων είναι σχετικά απλή. Αλλά αν πάμε στις 10 διαστάσεις η επιλογή σημείων που αφορούν τον πίνακα διακύμανσης δεν είναι απλή αφού αυτός πρέπει να είναι θετικά ημιορισμένος και επομένως ακόμα και η επιλογή σημείων ίσως είναι χρονοβόρα αλλά σε κάθε περίπτωση χρειαζόμαστε περισσότερα σημεία για να έχουμε ερευνήσει μεγαλύτερο κομμάτι του παραμετρικού χώρου.

Παράδειγμα: Ευρεση εκτιμητριών μεγίστης πιθανοφάνειας

Η Μέθοδος Μεγίστης Πιθανοφάνειας (ΜΜΠ) αποτελεί την πιο γνωστή και διαδεδομένη μέθοδο εκτίμησης στη στατιστική και για αυτό το λόγο χρησιμοποιείται σε ποικίλες και διαφορετικές εφαρμογές. Βασικό συστατικό της είναι η πιθανοφάνεια (likelihood) του δείγματος που μετράει πόσο πιθανό είναι τα δεδομένα να έχουν προέλθει από το συγκεκριμένο μοντέλο. Η ΜΜΠ εκτιμά αυτές τις παραμέτρους χρησιμοποιώντας τις τιμές που μεγιστοποιούν την πιθανοφάνεια των δεδομένων.

Ας υποθέσουμε λοιπόν πως έχουμε ένα τέτοιο μοντέλο με συνάρτηση πυκνότητας πιθανότητας $f(x \mid \theta)$, όπου θ είναι το διάνυσμα με τις άγνωστες παραμέτρους που θέλουμε να εκτιμήσουμε. Έστω ένα τυχαίο δείγμα $(x_1, x_2, \ldots, x_n)$ μεγέθους n το από την $f(x \mid \theta)$ που έχουμε υποθέσει.

Η συνάρτηση πιθανοφάνειας του δείγματός μας δίνεται από την

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

και είναι συνάρτηση του θ. Σκοπός μας σύμφωνα με τη ΜΜΠ είναι να μεγιστοποιήσουμε τη συνάρτηση αυτή ως προς θ. Στην πράξη δεν είναι εύκολο να μεγιστοποιήσουμε τη συνάρτηση πιθανοφάνειας για αυτό και καταφεύγουμε στη μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας (loglikelihood). Έτσι η λογαριθμική συνάρτηση πιθανοφάνειας ορίζεται ως

$$\ell(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

Επομένως το πρόβλημα είναι να μεγιστοποιήσουμε την $\ell(\theta)$ ως προς θ.

Ας υποθέσουμε ότι τα δεδομένα προέρχονται από την κατανομή Γάμμα. Η συνάρτηση πυκνότητας πιθανότητας είναι η

$$f(x_i \mid \alpha, \beta) = \frac{x_i^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)} exp(-\beta x_i),$$

όπου $x, \alpha, \beta > 0$ και $\Gamma(x)$ είναι η συνάρτηση Γάμμα. Η λογαριθμική πιθανοφάνεια έχει τη μορφή

$$\ell(\alpha, \beta) = (\alpha - 1)\sum_{i=1}^{n} \log(x_i) + n\alpha \log(\beta) - n \log(\Gamma(\alpha)) - \beta \sum_{i=1}^{n} x_i \qquad (6.5)$$

Οι παράγωγοί της εξισώνοντας με το 0 γίνονται

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= \sum_{i=1}^{n} \log(x_i) + n \log(\beta) - n\Psi(\alpha) = 0 \\ \frac{\partial \ell}{\partial \beta} &= \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0, \end{aligned}$$

όπου $\Psi(x)$ είναι η Δίγαμμα συνάρτηση που ορίζεται ως $\Gamma'(x)/\Gamma(x)$. Το παραπάνω σύστημα αποτελείται από μη γραμμικές εξισώσεις και επομένως δεν υπάρχει λύση σε κλειστή μορφή. Για την επίλυση τους απαιτούνται αριθμητικές μέθοδοι.

Ένας από τους λόγους για τη χρήση στοχαστικών αλγόριθμων είναι πως άλλες πιο απλές μέθοδοι μεγιστοποίησης απαιτούν τη χρήση παραγώγων της λογαριθμικής συνάρτησης πιθανοφάνειας, κάτι που οδηγεί στην ανάγκη υπολογισμού παραγώγων της συνάρτησης Γάμμα, κάτι όχι τόσο απλό, και εύκολο χωρίς τη χρήση ειδικού λογισμικού.

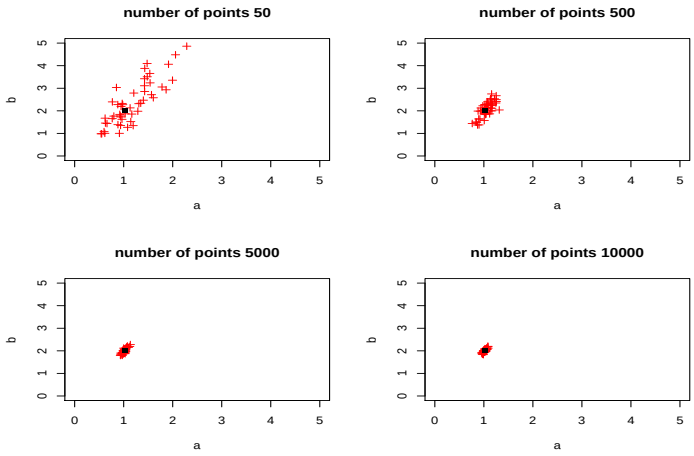
Ας υποθέσουμε λοιπόν ότι θέλουμε να μεγιστοποιήσουμε την πιθανοφάνεια ενός σετ δεδομένων με 1000 παρατηρήσεις από μια κατανομή Γάμμα. Η λογαριθμική συνάρτηση πιθανοφάνειας δίνεται από την (6.5), επομένως αυτή είναι η συνάρτηση

που θέλουμε να μεγιστοποιήσουμε. Για το σκοπό αυτό θα επιλέξουμε τυχαία m σημεία για τις παραμέτρους (α, β) από μια ομοιόμορφη κατανομή στο διάστημα $(0, 10)$. Είναι κατανοητό ότι όσο μεγαλύτερο είναι αυτό το διάστημα τόσο μεγαλύτερα προβλήματα θα αντιμετωπίσουμε για την εύρεση του μέγιστου.

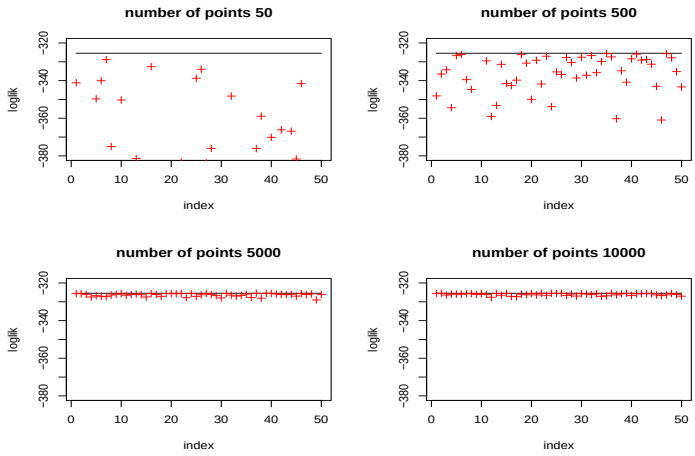
Στα γραφήματα 6.3 και 6.4 μπορεί κανείς να δει τα αποτελέσματα από την επιλογή 50 σετ από m σημεία. Έχουμε επιλέξει $m = 50, 500, 5000, 10000$ για να δείξουμε τη σημασία του πλήθους των σημείων. Το πείραμα επαναλήφθηκε 50 φορές για να γίνει κατανοητό ότι με τους στοχαστικούς αλγόριθμους η λύση που θα βρεθεί δεν είναι πάντα ή ίδια, κι επομένως οι 50 επαναλήψεις μας δίνουν μια εικόνα από το μέγεθος του λάθους που μπορούμε να κάνουμε.

Στο γράφημα 6.3 μπορεί κανείς αν δει το πραγματικό μέγιστο για τα α και β το οποίο αντιστοιχεί στο σημείο ■ ενώ γύρω από αυτό βλέπουμε τις εκτιμήσεις από τις 50 επαναλήψεις. Όπως θα περίμενε κανείς όταν το m είναι μικρό υπάρχει ο κίνδυνος λάθους, δηλαδή οι εκτιμήτριες μας να είναι μακριά από τις πραγματικές, αλλά όσο μεγαλώνει το m κάτι τέτοιο γίνεται ολοένα και πιο απίθανο. Παρατηρείστε πόσο μειώνεται η μεταβλητότητα όταν μεγαλώνει το m.

Στο γράφημα 6.4 μπορεί κανείς να δει την τιμή της λογαριθμικής πιθανοφάνειας. Η ευθεία γραμμή αντιστοιχεί στο πραγματικό μέγιστο και τα σημεία δείχνουν τη μέγιστη τιμή για καθένα από τα 50 σετ για τις διαφορετικές επιλογές του m. Παρατηρεί και πάλι κανείς ότι για μικρό m τα μέγιστα που έχουμε βρει μπορεί να διαφέρουν αρκετά από το πραγματικό, ενώ για $m = 10000$ αυτή η διαφοροποίηση είναι ελάχιστη και σχεδόν αμελητέα.



Γράφημα 6.3: Οι εκτιμήσεις των α και β για τις 50 επαναλήψεις του αλγόριθμου random search με διαφορετικό αριθμό σημείων $m = 50, 500, 5000, 10000$.



Γράφημα 6.4: Οι τιμές της λογαριθμικής πιθανοφάνειας για τις 50 επαναλήψεις του αλγόριθμου random search με διαφορετικό αριθμό σημείων m =50,500, 5000, 10000.

6.6 Local Search

΄Ενα από τα μειονεκτήματα του αλγόριθμου random search είναι πως κάθε φορά ψάχνουμε όλον τον παραμετρικό χώρο, δηλαδή τα σημεία επειδή επιλέγονται τυχαία μέσα σε όλο τον παραμετρικό χώρο (η τουλάχιστον σε ένα μεγάλο κομμάτι του που επειδή πιστεύουμε πως εκεί είναι η λύση διερευνούμε). Διαισθητικά θα μπορούσε να παρατηρήσει κανείς πως κάτι τέτοιο δεν είναι η καλύτερη προσέγγιση και θα μπορούσε να βελτιωθεί αν αντί να ψάχνουμε σε ολόκληρο τον παραμετρικό χώρο ψάξουμε στην γειτονιά (δηλαδή σε σημεία κοντά) στο μέχρι τώρα μέγιστο.

Επομένως, αν συμβολίσουμε με $f(x)$ την αντικειμενική συνάρτηση που θέλουμε να μεγιστοποιήσουμε, τότε ο αλγόριθμος μπορεί να περιγραφεί ως εξής:

Αλγόριθμος:

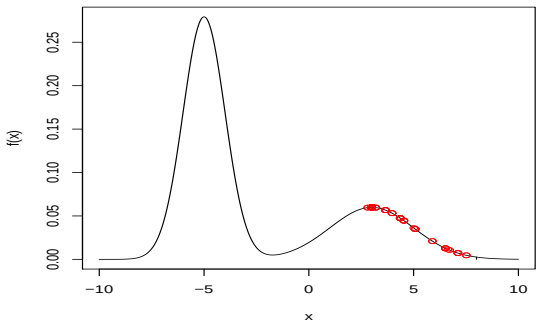
1. Ξεκίνα από ένα τυχαίο σημείο x
2. Διάλεξε ένα καινούριο σημείο x^* στη γειτονιά του x.
3. If $f(x^*) > f(x)$ θέσε $x = x^*$ και συνέχισε.
4. Σταμάτα όταν δεν μπορείς αν βρεις ένα καλύτερο σημείο μετά από μια μεγάλη σειρά σειρά επαναλήψεων

Ο παραπάνω αλγόριθμος αν και σχετικά απλοϊκός έχει πολύ καλές ιδιότητες σύγκλισης στο πραγματικό μέγιστο. Δύο σημεία που πρέπει να διευκρινιστούν είναι τι σημαίνει γειτονιά του x και πότε θα σταματήσουμε. Για το πρώτο εννοούμε σημεία κοντά στο x, το πόσο κοντά εξαρτάται από τη φύση του προβλήματος, καθώς

αν ορίσουμε τη γειτονιά να είναι ένα πολύ μικρό διάστημα γύρω από το x τότε ο αλγόριθμος θα προχωράει με πολύ μικρά βήματα και έτσι θα είναι αργός, ενώ αν ορίσουμε τη γειτονιά πολύ μεγάλη τότε πιθανότατα θα δυσκολευόμαστε να βρούμε καινούρια σημεία με μεγαλύτερη αντικειμενική συνάρτηση. Επομένως χρειάζεται εμπειρία και γνώση σχετικά με τη συνάρτηση ώστε να μπορέσουμε να ορίσουμε σωστά τη "γειτονιά". Σε κάθε περίπτωση ο αλγόριθμος συγκλίνει. Σχετικά με τον τερματισμό του αλγόριθμου μπορεί κανείς να χρησιμοποιήσει κριτήρια, όπως αν μετά από 100 προσπάθειες δεν βρεθεί καλύτερη τιμή σταμάτα ή αν θέλει να διασφαλιστεί περισσότεροι και υπολογιστικά είναι εφικτό να μεγαλώσει και άλλο τον αριθμό των επαναλήψεων αυτών.

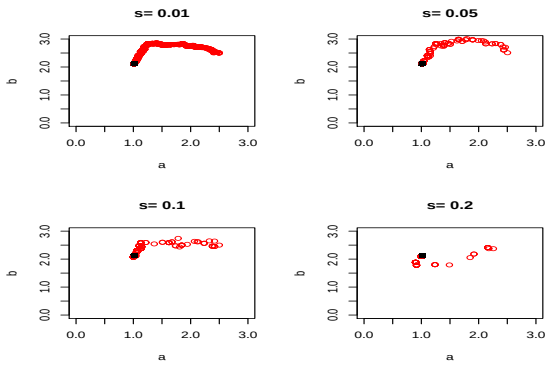
΄Ενα από τα μειονεκτήματα του αλγόριθμου είναι πως μπορεί να παγιδευτεί σε περιοχές τοπικών μέγιστων. Αυτό μπορείτε να το δείτε στο γράφημα 6.6. Η συνάρτηση που θέλουμε να μεγιστοποιήσουμε είναι δίκορφη και έχει μια παράμετρο (για λόγους ευκολίας). Αν ξεκινήσουμε από την τιμή $x = 8$ παρατηρείστε ότι σε εκείνη την κορφή υπάρχει τοπικό μέγιστο και όχι ολικό. Επομένως αν τα επόμενα σημεία είναι πολύ κοντά δεν πρόκειται ποτέ να περάσουμε στην απέναντι κορυφή και ο αλγόριθμός θα αποτύχει. Στο γράφημα μπορείτε αν δείτε τα διαδοχικά σημεία, ο αλγόριθμος θα σταματήσει μόλις φτάσει στην κορυφή αλλά εκεί υπάρχει τοπικό μέγιστο.

Πως θα μπορούσαμε αν αντιμετωπίσουμε κάτι τέτοιο; ΄Ενας απλός τρόπος είναι να δοκιμάσουμε με άλλες αρχικές τιμές, τότε είναι πολύ πιθανό να ξεκινήσουμε από σημείο στην σωστή κορυφή και να μην αντιμετωπίσουμε προβλήματα. Εναλλακτικά θα μπορούσαμε να παίρνουμε σημεία σε μεγάλη γειτονία που θα μας επέτρεπαν να "περάσουμε" από τη μια περιοχή στην άλλη. Κάτι τέτοιο θα δουλέψει μόνο όταν οι κορυφές είναι σχετικά κοντά. Μια διαφορετική λύση θα δούμε σε λίγο και αφορά τον αλγόριθμο simulated annealing, όπου η ιδέα είναι πως επιτρέπεται με κάποια ελεγχόμενη πιθανότητα να μετακινηθούμε προς σημεία με μικρότερη τιμή της αντικειμενικής συνάρτησης και άρα μπορούμε να περάσουμε από τη μια κορυφή στην άλλη.



Γράφημα 6.5: Παράδειγμα με δικορφη συνάρτηση

Για το προηγούμενο παράδειγμα με την πιθανοφάνειας της κατανομής Γάμμα έχουμε χρησιμοποιήσει τον αλγόριθμό local search για τα ίδια δεδομένα. Στο γράφημα 6.6 μπορεί να δει κανείς την ιστορία του αλγορίθμου, δηλαδή τα διαδοχικά σημεία από τα οποία πέρασε. Και στις 4 περιπτώσεις έχουμε την ίδια αρχική τιμή, αυτό που αλλάζει είναι το πως επιλέγουμε τα γειτονικά σημεία. Και στις 4 περιπτώσεις τα γειτονικά σημεία επιλέγονται με τη χρήση τυχαίων μεταβλητών από την κανονική κατανομή με κέντρο το σημείο στο οποίο βρισκόμαστε και τυπική απόκλιση με τιμές $\sigma = 0.01, 0.05, 0.1, 0.2$ αντίστοιχα. Όσο μεγαλώνει η τυπική απόκλιση τόσο διευρύνουμε τη γειτονιά στην οποία ψάχνουμε γειτονικά σημεία. Στο γράφημα βλέπουμε μόνο σημεία ως προς τα οποία η κίνηση έγινε και όχι αυτά που απορρίφθηκαν. Έτσι για μικρή τυπική απόκλιση οι κινήσεις ήταν προς σημεία πολύ κοντά και για αυτό χρειαστήκαμε πολύ περισσότερες κινήσεις. Μεγαλώνοντας την τυπική απόκλιση μπορούσαμε να μετακινούμαστε προς πιο μακρινά σημεία και άρα να φτάσουμε πιο γρήγορα (με λιγότερα σημεία) στο στόχο μας. Προσοχή όμως όταν έχουμε μεγάλη διακύμανση ενδέχεται να απορρίπτουμε πολλά σημεία καθώς αυτά είναι μακρινά και άρα όχι απαραίτητα καλύτερα. για το παράδειγμα μας οι για τυπική απόκλιση $\sigma = 0.2$ χρειάστηκε να δοκιμάσουμε πολύ περισσότερα σημεία από ότι στην περίπτωση $\sigma = 0.1$. Επομένως χρήση μεγάλης "γειτονιάς" δεν είναι πάντα η καλύτερη λύση και ουσιαστικά όταν αυτή η "γειτονιά" γίνει πάρα πολύ μεγάλη ο αλγόριθμος ταυτίζεται με τον random search.



Γράφημα 6.6: Η διαδρομή του αλγόριθμου local search όταν ξεκινήσαμε από το σημείο (2.5,2.5) για διαφορετικές τιμές της τυπικής απόκλισης, δηλαδή του πόσο μακριά μπορούσαμε να πάμε σε κάθε βήμα

6.7 Simulated Annealing

Ο αλγόριθμος simulated annealing όπως έχει ήδη ειπωθεί προσπαθεί να βελτιώσει τα προβλήματα των Local search αλγορίθμων επιτρέποντας να μετακινηθούμε προς σημεία με χειρότερη τιμή της αντικειμενικής συνάρτησης. Αυτό γίνεται με κάποια πιθανότητα η οποία όμως όσο περνάει η ώρα κι έχουμε κάνει περισσότερες επα-

ναλήψεις μικραίνει. Δηλαδή στα αρχικά βήματα δεχόμαστε μια μετακίνηση σε χειρότερο σημείο πιο συχνά.

Το όνομα του αλγόριθμου προέρχεται από τη φυσική και αφορά μεθόδους κατεργασίας μετάλλων. Ο αλγόριθμος simulated annealing σχετίζεται με τον αλγόριθμο Metropolis-Hastings και στην πραγματικότητα βασίζεται στις ιδιότητες των Μαρκοβιανών αλυσίδων και της σύγκλισης τους.

Ο αλγόριθμος χρησιμοποιείται και όταν ο παραμετρικός χώρος είναι διακριτός κι επομένως μέθοδοι που βασίζονται σε παραγώγους δεν μπορούν να χρησιμοποιηθούν (π.χ. το πρόβλημα του πωλητή που πρέπει να επισκεφτεί πολλές πόλεις).

Ο αλγόριθμος έχει μια παράμετρο που ονομάζεται θερμοκρασία, κατά αναλογία με την προέλευση του αλγόριθμου από τη φυσική. Ουσιαστικά η παράμετρος αυτή καθορίζει το ρυθμό που μεταβάλλεται η πιθανότητα να δεχτούμε μετακίνηση προς χαμηλότερο σημείο και κατ' αναλογία με τη φυσική προέλευση του αλγόριθμου καθορίζει την ταχύτητα με την όποια "κρυώνει" ο αλγόριθμος.

6.7.1 Ο αλγόριθμος

Τα βήματα του αλγόριθμου μπορούν να περιγραφούν ως εξής:

1. Προετοιμασία

   - Διάλεξε ένα τυχαίο σημείο x_0 από όπου θα ξεκινήσει ο αλγόριθμος.
   - Διάλεξε τις τιμές της αρχικής και τελικής θερμοκρασίας, έστω $T_0, T_f > 0$
   - Διάλεξε τη συνάρτηση με την οποία μεταβάλλεται η θερμοκρασία
   - Θέσε $x := x_0$ και $T := T_0$

   Τώρα είμαστε έτοιμοι για τα βασικά βήματα του αλγορίθμου

2. Διάλεξε ένα καινούριο σημείο x^* στη γειτονιά του x.

3. Αν $f(x^*) \geq f(x)$ θέσε $x = x^*$ και συνέχισε.
   αλλιώς

   - Πάρε $u \sim U(0,1)$ (προσομοίωσε μια ομοιόμορφη τυχαία μεταβλητή)
   - Αν $u < exp\left(\frac{f(x^*)-f(x)}{T}\right)$ τότε $x := x^*$ αλλιώς $x := x$

4. Μείωσε τη θερμοκρασία T σύμφωνα με την επιλεχθείσα συνάρτηση

5. Σταμάτα όταν δεν μπορείς αν βρεις ένα καλύτερο σημείο μετά από μια μεγάλη σειρά επαναλήψεων

Μπορεί κανείς να παρατηρήσεις τα εξής ενδιαφέροντα

- Όταν μετακινηθούμε σε σημείο με καλύτερη τιμή της αντικειμενικής συνάρτησης πάντα αποδεχόμαστε την κίνηση. Στην αντίθετη περίπτωση υπάρχει μια πιθανότητα που καθορίζεται από το πόσο διαφέρει η τιμή της αντικειμενικής συνάρτησης στο νέο σημείο και από τη θερμοκρασία.

- Ο τρόπος με τον οποίο αλλάζουμε τη θερμοκρασία είναι σημαντικός για τη σύγκλιση και εύρεση του βέλτιστου. αν η θερμοκρασία αλλάζει πολύ γρήγορα τότε σύντομα θα πάψουμε να δεχόμαστε μετακινήσεις προς τα πίσω, από την άλλη αν αλλάζει πολύ αργά θα έχουμε καθυστερήσεις. Επομένως η επιλογή είναι κρίσιμη για τις ιδιότητες του αλγόριθμου.

- Ο αλγόριθμος συγκλίνει στο μέγιστο μετά από έναν αριθμό επαναλήψεων δηλαδή δεν μπορούμε να μετακινηθούμε άλλο

- ΄Ενα ενδιαφέρον ερώτημα είναι πως θα ανιχνεύσουμε ότι ο αλγόριθμος μας σταμάτησε να κινείται και άρα πρέπει να σταματήσουμε. Εμπειρικά μπορούμε να χρησιμοποιήσουμε ως κριτήριο ότι δεν μετακινηθήκαμε για έναν αριθμό (πχ 100) επαναλήψεων.

- Μερικές φορές σε αρκετά προβλήματα δεν είναι απλό κι εύκολο να καθορίσουμε ποια είναι η γειτονιά της τρέχουσας λύσης που βρισκόμαστε. Αυτό απαιτεί εμπειρία και καλή γνώση του προβλήματος. Μάλιστα σε διακριτά προβλήματα θα πρέπει κανείς να ορίσει σαφώς ποιες μετακινήσεις επιτρέπονται (πχ στο πρόβλημα του πωλητή, ποιες πόλεις είναι προσβάσιμες από την πόλη που βρισκόμαστε αυτή τη στιγμή).

6.7.2 Η θερμοκρασία

Προκειμένου ο αλγόριθμος να συγκινεί πραγματικά στο μέγιστο θα πρέπει η συνάρτηση που θα χρησιμοποιηθεί για τη θερμοκρασία να είναι φθίνουσα δηλαδή η θερμοκρασία να μικραίνει όσο περνάει η ώρα. Αυτό πρακτικά σημαίνει πως όσο περνάει η ώρα κι έχουμε κάνει περισσότερες επαναλήψεις τότε η πιθανότητα να δεχτούμε μια μετακίνηση προς χειρότερο σημείο να είναι μικρότερη.

Μερικές συναρτήσεις που έχουν χρησιμοποιηθεί στην πράξη είναι οι ακόλουθες:

$$T_{new} = T_{old}(1 - \epsilon)$$

για κάποια μικρή τιμή του ϵ΄και

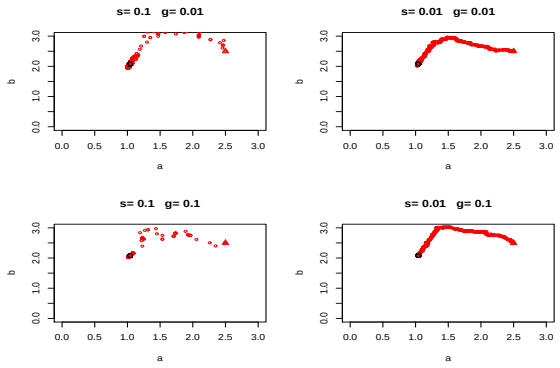
$$T_{new} = \frac{T_{old}}{1 + \beta T_{old}}$$

όπου β είναι μια κατάλληλη μικρή τιμή.

6.7.3 Παράδειγμα

Επανερχόμαστε στο πρόβλημα με τη μεγιστοποίηση της πιθανοφάνειας της Γάμμα κατανομής από ένα δείγμα που χρησιμοποιήσαμε και προηγουμένως.

Θα χρησιμοποιήσουμε τον αλγόριθμο simulated annealing. Αρχικές τιμές για τα α και β χρησιμοποιήσαμε τις τιμές 2.5 και 2.5 αντίστοιχα. Επίσης το σχέδιο ψύξης της θερμοκρασίας είναι η συνάρτηση χρησιμοποιώντας τιμές $\gamma = 0.01, 0.1$ αντίστοιχα. Τέλος τα καινούρια σημεία στη γειτονιά της μέχρι τώρα τιμής τα διαλέξαμε προσομοιώνοντας u_i $i = 1, 2$ από κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση $\sigma = 0.01, 0.1$. Οι καινούριες τιμές προκύπτουν ως $\alpha^{(t+1)} = \alpha^{(t)} + u_1$ και $\beta^{(t+1)} = \beta^{(t)} + u_2$.



Γράφημα 6.7: Διάφορες υλοποιήσεις του αλγόριθμου Simulated annealing για διαφορετικές τιμές γ και σ.

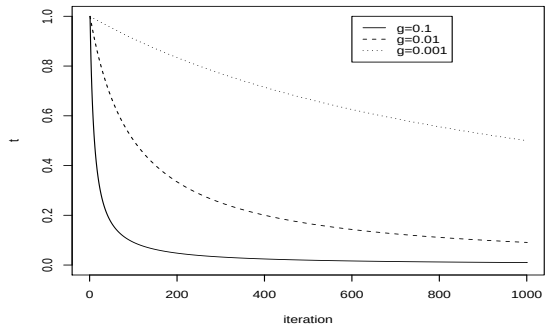
Στο γράφημα 6.7 μπορεί κανείς αν δει την ιστορία του αλγόριθμου για 500 επαναλήψεις. Οι αρχικές τιμές είναι και στις 4 περιπτώσεις οι ίδιες και οι παράμετροι που χρησιμοποιήθηκαν εμφανίζονται πάνω από κάθε γράφημα. Μπορεί κανείς να παρατηρήσει ότι όταν το σ είναι μικρό, δηλαδή επιτρέπουμε πολύ μικρά βήματα στον αλγόριθμο ο αλγόριθμος χρειάζεται περισσότερες επαναλήψεις για να βρει το μέγιστο. Από την άλλη όταν μεγαλώσει η τυπική απόκλιση μπορούμε να κάνουμε μεγαλύτερα βήματα κι έτσι φτάνουμε πιο γρήγορα στο μέγιστο.

Κάτι αντίστοιχο συμβαίνει με την επιλογή της παραμέτρου γ. Στο γράφημα 6.8 μπορεί κανείς να δεί τρεις διαφορετικές στρατηγικές για τη θερμοκρασία με τη χρήση της συνάρτησης $t_{new} = {}_{old}/(1 + \gamma_o ld)$ για τιμές του $\gamma = 0.001, 0.01, 0.1$. μπορεί κανείς να δει ότι η τιμή του μικραίνει πολύ πιο γρήγορα όσο μεγαλύτερη τιμή έχει το γ. μπορεί κανείς να δεί ότι για $\gamma = 0.1$ ο αλγόριθμος συγκλίνει πολύ πιο γρήγορα. Βέβαια μεγάλες τιμές του γ μετατρέπουν τη μέθοδο simulated annealing ίδια με τον local search αλγόριθμο.

6.8 Ο αλγόριθμος ΕΜ

΄Ενα από τα μειονεκτήματα των κλασικών αριθμητικών μεθόδων όταν αυτές χρησιμοποιούνται για την επίλυση στατιστικών προβλημάτων είναι πως δεν λαμβάνουν υπόψη τους τυχόν στατιστικές ιδιότητες, όπως για παράδειγμα περιορισμούς στο πεδίο ορισμού. Ο αλγόριθμος ΕΜ αποτελεί στην ουσία αριθμητική μέθοδο που επειδή όμως έχει κατασκευαστεί με στατιστικές τεχνικές επιτρέπει και προσφέρει αρκετή πληροφορία για την επίλυση στατιστικών προβλημάτων Ο αλγόριθμος ΕΜ χρησιμοποιείται για την εύρεση ΕΜΠ σε πολλά προβλήματα που αφορούν όμως σύγχρονες πολύπλοκες στατιστικές μεθοδολογίες.

Η βασική ιδέα του αλγορίθμου ΕΜ είναι πως εκτός από δεδομένα που έχουμε παρατηρήσει υπάρχουν και κάποια άλλα δεδομένα τα οποία αφενός δεν έχουμε παρατηρήσει αλλά αν τα είχαμε το πρόβλημα της εύρεσης ΕΜΠ θα ήταν πολύ πιο απλό.



Γράφημα 6.8: Διαφορετικές συναρτήσεις για τη θερμοκρασία της μορφής $_{new} = t_{old}/(1 + \gamma_o ld)$.

Επομένως υποθέτουμε πως υπάρχουν χαμένα (missing) δεδομένα. Είναι σημαντικό να πούμε πως η ιδέα των missing δεδομένων είναι πολύ πιο γενική από ότι χρησιμοποιείται στη στατιστική. Η τεχνική του να συμπληρώνουμε τα παρατηρούμενα δεδομένα (observed data) με μη παρατηρούμενα δεδομένα (missing data) ονομάζεται συμπλήρωση δεδομένων (data augmentation). Τα δεδομένα που προκύπτουν από το συνδυασμό παρατηρούμενων και μη παρατηρούμενων ονομάζονται πλήρη δεδομένα (complete data). Θα πρέπει να τονιστεί πως η τεχνική της συμπλήρωσης δεδομένων αποτελεί τη βάση και για άλλες πολύ δημοφιλείς τεχνικές όπως οι τεχνικές Markov Chain Monte Carlo (MCMC) και για αυτό υπάρχει μεγάλη συγγένεια ανάμεσα στον ΕΜ και τη μέθοδο MCMC, όπως θα δούμε και στη συνέχεια.

Ιστορικά ο αλγόριθμος ΕΜ πρωτοπαρουσιάστηκε στη γενική του μορφή το 1977 (Dempster *et. al.*, 1977). Διάφορες παραλλαγές του αλγορίθμου σε συγκεκριμένες εφαρμογές έχουν χρησιμοποιηθεί ήδη από τη δεκαετία του 1960 και προηγουμένως. Στην πραγματικότητα πολλοί ερευνητές σε διάφορα αντικείμενα είχαν χρησιμοποιήσει επαναληπτικούς αλγορίθμους για την επίλυση των προβλημάτων τους οι οποίοι αργότερα αποδείχτηκε ότι μπορούν να ιδωθούν σαν εφαρμογές του αλγορίθμου ΕΜ. Ο αλγόριθμος έγινε ιδιαίτερα δημοφιλής τα τελευταία χρόνια λόγω της χρήσης υπολογιστών για την επίλυση ολοένα και μεγαλύτερων στατιστικών προβλημάτων.

6.8.1 Παραδείγματα "missing data"

Όπως είπαμε και προηγουμένως η βασική ιδέα του αλγορίθμου ΕΜ είναι να συμπληρώσουμε τα παρατηρούμενα δεδομένα με κάποια μη παρατηρούμενα έτσι ώστε τα πλήρη δεδομένα να επιτρέπουν εύκολη συμπερασματολογία. Επομένως χρειάζεται να διευκρινίσουμε τι εννοούμε ως μη παρατηρούμενα δεδομένα. Η ιδέα είναι πολύ γενικότερη από την απλή ιδέα των missing data στη στατιστική, κάποιες παρατηρήσεις που για κάποιο λόγο δεν τις έχουμε. Η ιδέα έχει να κάνει με το γενικότερο μοντέλο δημιουργίας δεδομένων, δηλαδή υπάρχει με βάση το μοντέλο που χρησιμοποιούμε ένας μηχανισμός που δημιουργεί δεδομένα τα οποία όμως δεν

είμαστε σε θέση να παρατηρήσουμε. Μερικά παραδείγματα ακολουθούν:

- Το πρώτο παράδειγμα αναφέρεται στην απλή περίπτωση που κάποιες μεταβλητές από κάποιες παρατηρήσεις δεν έχουν παρατηρηθεί, δηλαδή ο πίνακας δεδομένων δεν είναι πλήρης. Αυτή είναι και η πιο κλασική και διαδεδομένη περίπτωση στη στατιστική καθώς τα περισσότερα πραγματικά σετ δεδομένων περιέχουν τέτοιου είδους μη παρατηρούμενα δεδομένα.

- Λογοκριμένες (Censored) παρατηρήσεις. Τέτοιου είδους δεδομένα είναι ιδιαίτερα διαδεδομένα στην ανάλυση επιβίωσης (survival analysis καθώς λόγω περιορισμών στον τρόπο συλλογής δεδομένων για κάποιες παρατηρήσεις /ασθενείς ξέρουμε πως ζούσαν μέχρι κάποια χρονική στιγμή αλλά μετά δεν έχουμε κάποια πληροφορία, δηλαδή γνωρίζουμε πως η τιμή της τυχαίας μεταβλητής που αφορά το χρόνο επιβίωσης είναι μεγαλύτερη από κάποια τιμή αλλά δεν ξέρουμε την ακριβή της τιμή. Για παρόμοιους λόγους που αφορούν, για παράδειγμα, την ακρίβεια των οργάνων μέτρησης, μπορεί να γνωρίζουμε πως κάποια παρατήρηση είναι μικρότερη από κάποια τιμή αλλά δεν ξέρουμε ακριβώς την τιμή της.

- Περικομμένες (Truncated) παρατηρήσεις. ΄Εχουμε μιλήσει για τέτοιου είδους δεδομένα προηγουμένως. Αφορά την περίπτωση που κάποιες συγκεκριμένες τιμές δεν μπορούν να παρατηρηθούν και εμφανίζονται επομένως με μηδενική συχνότητα εμφάνισης.

- Ομαδοποιημένα δεδομένα. Αυτή η περίπτωση αφορά πολλές έρευνες με χρήση ερωτηματολογίων όπου οι ερωτήσεις αν και αφορούν συνεχείς τυχαίες μεταβλητές έχουν κατηγοριοποιηθεί για διάφορους λόγους (απόρρητα δεδομένα, μη ακριβείς απαντήσεις κλπ). Χαρακτηριστικά τέτοια παραδείγματα είναι το εισόδημα και η ηλικία τα οποία αν και συνεχείς τυχαίες μεταβλητές συνήθως στα ερωτηματολόγια εμφανίζονται διακριτοποιημένες σε κατηγορίες. Σε αυτή την περίπτωση γνωρίζουμε το διάστημα μέσα στο οποιό πέφτει κάποια παρατήρηση αλλά δεν είμαστε σε θέση να ξέρουμε την ακριβή τιμή της παρατήρησης.

- Μίξεις κατανομών. Πολλά μοντέλα και πολλές κατανομές προκύπτουν από απλούστερα μοντέλα (κατανομές) μέσα από τη διαδικασία της μίξης. Τέτοια παραδείγματα είναι η κατανομή t-Student, η αρνητική διωνυμική κατανομή και πολλές άλλες. Σε αυτή την περίπτωση ουσιαστικά υποθέτουμε πως για την παρατήρηση x_i ξέρουμε πως ακολουθεί κάποια κατανομή $f(x_i \mid \theta_i)$ αλλά η παράμετρος της είναι και αυτή μια τυχαία μεταβλητή, δηλαδή έχουμε διαφορετική παράμετρο για κάθε παρατήρηση, και υποθέτουμε πως η παράμετρος ακολουθεί κάποια κατανομή, δηλαδή $\theta_i \sim g(\phi)$. Για παράδειγμα η αρνητική διωνυμική προκύπτει αν υποθέσουμε πως η παράμετρος της κατανομής Poisson ακολουθεί μια Γάμμα κατανομή. Τα μοντέλα μεικτών επιδράσεων (mixed effects models) που είναι ιδιαίτερα δημοφιλή σήμερα είναι ουσιαστικά μοντέλα μίξης κατανομών

- Συνελίξεις κατανομών (Convolutions). Σε αρκετές εφαρμογές η τυχαία μεταβλητή που παρατηρούμε είναι το άθροισμα δύο επιμέρους τυχαίων μεταβλητών, εμείς όμως παρατηρούμε μόνο το άθροισμα και όχι τις επιμέρους τιμές.

- Τυχαία αθροίσματα (random sums). Τέτοια μοντέλα είναι πολύ διαδεδομένα στην αναλογιστική επιστήμη. Αφορούν μια τυχαία μεταβλητή Y που προκύπτει ως

$$Y = X_1 + X_2 + \ldots + X_N$$

  όπου X_i είναι ισόνομες τυχαίες μεταβλητές των οποίων όμως το πλήθος N δεν είναι σταθερό αλλά τυχαία μεταβλητή που ακολουθεί κάποια διακριτή κατανομή (πχ την κατανομή Poisson). Για παράδειγμα το ύψος των αποζημιώσεων που μια ασφαλιστική εταιρεία θα καταβάλει ακολουθεί ένα τέτοιο μοντέλο. Σε πολλές περιπτώσεις παρατηρούμε μόνο το Y και όχι τις υπόλοιπες τυχαίες μεταβλητές.

- Hidden Markov Models. Αφορά μοντέλα για εξαρτημένα δεδομένα. Παρατηρούμε μια χρονολογική σειρά, η τιμή όμως κάθε χρονική στιγμή εξαρτάται από μια μη παρατηρήσιμη κατάσταση (state), εμείς παρατηρούμε μόνο την τιμή και όχι τις καταστάσεις σε κάθε χρονική στιγμή.

Η παραπάνω λίστα σε καμιά περίπτωση δεν εξαντλεί τις δυνατές περιπτώσεις missing data που μπορεί να προκύψουν. Ο αλγόριθμος ΕΜ μπορεί να δουλέψει και με αρκετά πιο πολύπλοκες δομές από αυτές που αναφέραμε πιο πάνω αλλά και με συνδυασμό τους κλπ. Αρκεί να μπορέσει κάποιος να εκφράσει το μοντέλο κατάλληλα ώστε αυτή η τεχνική να μπορεί να εφαρμοστεί.

6.8.2 Η βασική ιδέα

Η βασική ιδέα του αλγορίθμου είναι η εξής:

΄Εστω πως μας ενδιαφέρει να μεγιστοποιήσουμε την πιθανοφάνεια $L(\theta \mid X)$ των παρατηρηθέντων δεδομένων X. Η ιδέα είναι να συμπληρώσουμε τα δεδομένα X με κάποια πρόσθετα μη παρατηρηθέντα (missing) δεδομένα Z έτσι ώστε η πιθανοφάνεια των πλήρων (complete) δεδομένων $Y = (X, Z)$ να είναι πιο εύκολο να μεγιστοποιηθεί. Ο αλγόριθμος ΕΜ είναι επαναληπτικός κι έχει δύο βήματα. Στο πρώτο βήμα εκτιμάμε τα μη παρατηρούμενα δεδομένα Z χρησιμοποιώντας τα X και τις τιμές των παραμέτρων μέχρι τώρα (Ε- βήμα) και στη συνέχεια στο Μ-βήμα μεγιστοποιούμε την πιθανοφάνεια των πλήρων δεδομένων Y αντικαθιστώντας τα Z που δεν έχουμε παρατηρήσει με τις αναμενόμενες τιμές τους από το Ε-βήμα. Δηλαδή, με απλά λόγια, χρησιμοποιούμε τα missing data για να απλοποιήσουμε το πρόβλημα. ΄Ετσι σε κάθε επανάληψη εκτιμάμε τις τιμές των missing data με βάση την πληροφορία που υπάρχει μέχρι εκείνη τη στιγμή, δηλαδή τα παρατηρηθέντα δεδομένα και τις τιμές των παραμέτρων μέχρι εκείνη την επανάληψη.

΄Εχει αποδειχτεί πως σε κάθε επανάληψη η πιθανοφάνεια αυξάνει και πως ο αλγόριθμος συγκλίνει σε κάποια τιμή. Η ταχύτητα σύγκλισης εξαρτάται από τη δομή που έχουμε χρησιμοποιήσει και θα το εξετάσουμε παρακάτω.

Το όνομα του ο αλγόριθμος το χρωστά στα δύο του βήματα: Expectation - Maximization. Είναι πολύ βασικό να ορίσουμε κατάλληλα τα πλήρη δεδομένα

ώστε να είναι αποτελεσματικός ο αλγόριθμος. Αυτό γενικα δεν είναι εύκολο και χρειάζεται αρκετή πείρα αλλά και καλή γνώση του μηχανισμού που παράγει τα δεδομένα.

Ας δούμε λίγο πιο τεχνικά τα ακριβή βήματα του αλγορίθμου

Για την τιμή των παραμέτρων, έστω $\theta^{(r)}$ μετά την r επανάληψη Ορίζουμε τη συνάρτηση

$$\begin{aligned} Q(\theta, \theta^{(r)}) &= \int_Z \log L(\theta \mid X, Z) P(Z \mid \theta^{(r)}, X) dZ \\ &= E(\log L(\theta \mid X, Z)) \end{aligned}$$

όπου $P(Z \mid \theta^{(r)}, X)$ είναι η δεσμευμένη κατανομή των μη παρατηρούμενων δεδομένων Z για δοθείσες τιμές των παραμέτρων και των δεδομένων X. Τα βήματα του αλγορίθμου είναι τα εξής

1. *Ε-βήμα* Υπολόγισε τη συνάρτηση $Q(\theta, \theta^{(r)})$
2. *Μ-βήμα* Μεγιστοποίησε τη συνάρτηση $Q(\theta, \theta^{(r)})$ ως προς θ.

Μπορεί κανείς να δει πως το Ε-βήμα δεν είναι τίποτα άλλο από τον υπολογισμό της αναμενόμενης τιμής της πιθανοφάνειας των πλήρων δεδομένων ως προς τη δεσμευμένη όμως κατανομή των missing data. Ομοίως στο Μ-βήμα μεγιστοποιούμε την πιθανοφάνεια των πλήρων δεδομένων χρησιμοποιώντας την αναμενόμενη τιμή από το Ε-βήμα.

6.8.3 Πλεονεκτήματα και μειονεκτήματα

Πλεονεκτήματα

- Όπως όλες οι επαναληπτικές μέθοδοι έτσι και ο αλγόριθμος ΕΜ βασίζεται σε καλές αρχικές τιμές. Το σημαντικό πλεονέκτημα του αλγορίθμου είναι πως αν οι αρχικές τιμές είναι μέσα στο αποδεκτό πεδίο ορισμού τότε σε κάθε επανάληψη είμαστε σίγουροι πως οι τιμές που θα πάρουμε θα είναι και αυτές αποδεκτές, κάτι που δεν μπορεί να εγγυηθεί από άλλες μεθόδους χωρίς να ληφθεί κατάλληλη πρόνοια. Κάτι τέτοιο είναι πολύ σημαντικό καθώς προφυλάσσει από μη αποδεκτές λύσεις, κάτι που συμβαίνει αρκετά συχνά σε προβλήματα με μεγάλο αριθμό παραμέτρων.
- Ο αλγόριθμος είναι συνήθως πολύ εύκολος να προγραμματιστεί ακόμα και σε απλά πακέτα. Δεδομένου πως δεν βασίζεται σε παραγώγους και αντιστροφές πινάκων όπως για παράδειγμα η μέθοδος Newton-Raphson είναι αρκετά πιο εύχρηστος.
- Όπως έχουμε ήδη αναφέρει ο αλγόριθμος ΕΜ θεμελιώθηκε πάνω σε καθαρά στατιστική επιχειρηματολογία και επομένως προσφέρει πολύ χρήσιμη στατιστική ερμηνεία. Ο αλγόριθμος είναι μια ξεκάθαρα στατιστική τεχνική και για αυτό σε πολλά παραδείγματα χρήσης του προσφέρει μια ενδιαφέρουσα από στατιστική άποψη γνώση. Σε μερικά προβλήματα με μεγάλο αριθμό παραμέτρων η καθαρά στατιστική οπτική του αλγορίθμου είναι χρήσιμη για την καλύτερη συμπεριφορά του.

- Πολλές φορές κάποια ενδιάμεσα αποτελέσματα του αλγορίθμου έχουν τη δική τους σημασία και ενδιαφέρον και μπορούν να χρησιμοποιηθούν για περαιτέρω στατιστική ανάλυση. Για παράδειγμα το Ε-βήμα του αλγορίθμου εκτιμά τα missing data. Πολλές εφαρμογές στηρίζονται στο να αντικαθιστούν αυτά τα δεδομένα με τις εκτιμήσεις από τον αλγόριθμο.

Μειονεκτήματα

- Αργή σύγκλιση. Ο αλγόριθμος ΕΜ συγκλίνει στη λύση πιο αργά από ότι άλλοι αλγόριθμοι. Για παράδειγμα η σύγκλιση του είναι γραμμική ενώ η σύγκλιση της μεθόδου Newton-Raphson είναι τετραγωνική. Στην πράξη αυτό σημαίνει πως χρειαζόμαστε περισσότερες επαναλήψεις μέχρι να βρούμε τη λύση. Θα πρέπει να τονιστεί πως αυτό δεν σημαίνει απαραίτητα και περισσότερο χρόνο καθώς οι πράξεις που απαιτούνται σε κάθε επανάληψη μπορεί να είναι πολύ πιο απλές και άρα να χρειάζονται λιγότερο χρόνο. Θυμηθείτε γα παράδειγμα πως στη μέθοδο Newton-Raphson πρέπει να αντιστρέφουμε σε κάθε βήμα έναν πίνακα $p \times p$.

  Μια ενδιαφέρουσα παρατήρηση είναι πως η ταχύτητα σύγκλισης του ΕΜ έχει να κάνει με τη δομή που χρησιμοποιούμε για να τον κατασκευάσουμε δηλαδή με τη μέθοδο συμπλήρωσης δεδομένων (data augmentation). Όσο μεγαλύτερη είναι η πληροφορία που θεωρούμε ως missing τόσο πιο αργός είναι ο αλγόριθμος.

  Επίσης θυμηθείτε πως το γεγονός πως ο αλγόριθμος προχωρά με μικρότερα βήματα έχει το πλεονέκτημα πως δεν οδηγούμαστε σε προβλήματα μη σύγκλισης όπως με άλλες μεθόδους, όπου μπορεί ξαφνικά ο αλγόριθμος να έχει παράξενη συμπεριφορά. Δηλαδή από όσο μακριά και να ξεκινήσουμε ο αλγόριθμός θα βρει τη λύση ενώ σε άλλες μεθόδους και ειδικά σε προβλήματα με πολλές παραμέτρους αν δεν ξεκινήσουμε κοντά στη λύση ο αλγόριθμος μπορεί να μη συγκλίνει ποτέ.

- Η βασική ιδιότητα του αλγορίθμου να αυξάνει την πιθανοφάνεια σε κάθε βήμα δεν σημαίνει απαραίτητα πως στο τέλος του αλγορίθμου έχε βρεθεί το ολικό μέγιστο και όχι κάποιο τοπικό μέγιστο. Το σημείο στο οποίο θα συγκλίνει ο αλγόριθμος εξαρτάται και από τις αρχικές τιμές. Συνεπώς για να είμαστε σίγουροι πως έχει βρεθεί όντως το ολικό μέγιστο πρέπει να επαναλάβουμε τον αλγόριθμο με διαφορετικές αρχικές τιμές. Βέβαια σε πολλές περιπτώσεις, όπου η μορφή της πιθανοφάνειας μας εξασφαλίζει την ύπαρξη ενός μεγίστου, κάτι τέτοιο δεν χρειάζεται και οι αρχικές τιμές συνήθως παίζουν απλά το ρόλο της ταχύτερης σύγκλισης. Τονίζεται πως παρόμοια προβλήματα παρουσιάζουν όλες οι επαναληπτικές μέθοδοι όπως είδαμε και στο παράδειγμα με την περικομμένη κατανομή Poisson προηγουμένως.

- Η λύση εξαρτάται όπως και σε κάθε επαναληπτικό αλγόριθμο στις αρχικές τιμές. Δεδομένης της πιο αργής σύγκλισης του ΕΜ κακές αρχικές τιμές ίσως

οδηγήσουν σε πολύ περισσότερες επαναλήψεις και συνεπώς η ανάγκη για καλύτερες αρχικές τιμές είναι μεγαλύτερη.

- Ο ΕΜ δεν χρησιμοποιεί τις δεύτερες παραγώγους κι έτσι τα τυπικά σφάλματα δεν προκύπτουν κατά τη διάρκεια εκτέλεσης του αλγορίθμου. Τα τυπικά σφάλματα μπορούν να βρεθούν με άλλους τρόπους που θα δούμε σε λίγο.

6.8.4 Χρησιμα Στοιχεία

Θα πρέπει να αναφέρουμε επίσης και μερικά άλλα ενδιαφέροντα χαρακτηριστικά του αλγορίθμου. Αυτά είναι:

- Συνήθως αν και ο αλγόριθμος χρειάζεται αρκετές επαναλήψεις για να συγκλίνει (δηλαδή να εκπληρωθεί η συνθήκη τερματισμού του αλγορίθμου), κάποιες λίγες επαναλήψεις αρκούν για να φτάσουμε πολύ κοντά στη λύση. Αυτή η ιδιότητα μπορεί να συνδυαστεί με την ιδέα πως άλλες μέθοδοι έχουν πολύ γρηγορότερη σύγκλιση αρκεί να έχουμε αρχικές τιμές κοντά στη λύση. Έτσι μερικές επαναλήψεις τυ ΕΜ μας φέρνουν κοντά στη λύση και μετά κάποιος άλλος αλγόριθμος που ξεκινά πια από πολύ καλές αρχικές τιμές βρίσκει τη λύση γρηγορότερα. Δηλαδή ο ΕΜ μπορεί να συνδυαστεί με άλλες μεθόδους για ακόμα καλύτερη συμπεριφορά.

- Η πιθανοφάνεια σε κάθε επανάληψη μεγαλώνει οι παράμετροι όμως μπορεί να αλλάζουν ελάχιστα, σχεδόν ανεπαίσθητα. Συνεπώς παίζει μεγάλο ρόλο το κριτήριο το οποίο θα χρησιμοποιήσουμε για να σταματήσουμε τον αλγόριθμο, αν αυτό θα βασίζεται στην αύξηση της πιθανοφάνειας ή στη σχετική αλλαγή της τιμής των παραμέτρων. Δυστυχώς, όπως θα δούμε και στη συνέχεια, η επιλογή του κριτηρίου τερματισμού δεν είναι εύκολη υπόθεση και δεν υπάρχει γενικά αποδεκτό κριτήριο.

- Όπως είπαμε και προηγουμένως η ταχύτητα του ΕΜ εξαρτάται και από το ποσό της χαμένης πληροφορίας (missing information). Δεδομένου πως για κάποιο πρόβλημα μπορούμε να υποθέσουμε διαφορετικές ιδέες data augmentation αυτό θα οδηγήσει σε διαφορετικούς αλγορίθμους με διαφορετικές επιδόσεις. Σε έναν αλγόριθμο όπου έχουμε υποθέσει αρκετή χαμένη πληροφορία (πχ το πλήθος των missing data είναι μεγάλο σε σχέση με τα observed data) τότε ο αλγόριθμος μπορεί να είναι πολύ αργός. Συνεπώς η επιλογή του data augmentation είναι πολύ σημαντική, χρειάζεται όμως αρκετή πείρα και γνώση του προβλήματος ώστε να επιλεγεί η καταλληλότερη δομή.

6.8.5 Κριτήρια Τερματισμού

Ας δούμε μερικά κριτήρια τερματισμού του ΕΜ που χρησιμοποιούνται στην πράξη. Αργότερα στα παραδείγματα θα επανέλθουμε στο θέμα. Τα κριτήρια χωρίζονται σε δύο κατηγορίες:

- Κριτήρια βασισμένα στην πρόοδο της πιθανοφάνειας

Σταματάμε τις επαναλήψεις όταν

$$\left|\frac{L^{(r+1)} - L^{(r)}}{L^{(r+1)}}\right| \leq tol$$

όπου tol είναι μια μικρή τιμή (πχ 10^{-10}) και $L^{(r)}$ είναι η λογαριθμική πιθανοφάνεια μετά την r επανάληψη. Ουσιαστικά το κριτήριο λέει να σταματήσουμε όταν η λογαριθμική πιθανοφάνεια δεν αλλάζει πια από επανάληψη σε επανάληψη. Υπάρχουν δύο σημαντικά σημεία που πρέπει κανείς να έχει υπόψη του: το ότι δεν αλλάζει η πιθανοφάνεια δεν σημαίνει πως οι τιμές των παραμέτρων δεν αλλάζουν, αν η πιθανοφάνεια είναι σχεδόν επίπεδη σε κάποιο σημείο του παραμετρικού χώρου τότε οι παράμετροι μπορεί να αλλάζουν και μάλιστα δραματικά. Επίσης το ότι η πιθανοφάνεια αλλάζει πολύ λίγο δεν σημαίνει πως βρέθηκε απαραίτητα το μέγιστο καθώς μπορεί να σταματήσουμε αλλά ύστερα από λίγες επαναλήψεις ο αλγόριθμος να ξέφευγε από την περιοχή αυτή. Για αυτό και είναι καλή ιδέα να χρησιμοποιείται μάλλον αυστηρό κριτήριο τερματισμού ιδιαίτερα σε πολύπλοκα προβλήματα με μεγάλο αριθμό παραμέτρων.

- Κριτήρια βασισμένα στην αλλαγή των τιμών των παραμέτρων

  Σταματάμε τις επαναλήψεις όταν

  $$max_j \left(|\theta_j^{(r+1)} - \theta_j^{(r)}|\right) \leq tol$$

  όπου $\theta_j^{(r)}$ είναι η τιμή της παραμέτρου θ_j μετά την r επανάληψη. Δηλαδή το κριτήριο ικανοποιείται όταν όλες οι παράμεροι αλλάζουν από μια επανάληψη στην επόμενη λιγότερο από μια μικρή ποσότητα tol. Εναλλακτικά μπορεί κανείς να χρησιμοποιήσει κριτήρια που μετράνε τη μέση διαφορά από επανάληψη σε επανάληψη η κάποια άλλη απόσταση. Για παράδειγμα, μπορεί κανείς να σταματήσει τις επαναλήψεις όταν

  $$\sum_{j=1}^{p} \left(\theta_j^{(r+1)} - \theta_j^{(r)}\right)^2 \leq tol$$

  δηλαδή το άθροισμα τετραγωνικών αποκλίσεων για όλες τις παραμέτρους σε δύο διαδοχικές επαναλήψεις είναι μικρότερο από κάποια πολύ μικρή τιμή.

Είναι πολύ βασικό να τονίσουμε πως και τα δυο κριτήρια στην πραγματικότητα μετράνε έλλειψη περαιτέρω προόδου του αλγορίθμου παρά πραγματική σύγκλιση!

6.8.6 Ο αλγόριθμος ΕΜ στην εκθετική οικογένεια

Είναι πολύ ενδιαφέρον να δούμε την περίπτωση του αλγορίθμου όταν η κατανομή των πλήρων δεδομένων (complete data) ανήκει στην εκθετική οικογένεια κατανομών. Θα πρέπει να πούμε πως αυτό συμβαίνει σε πολλές από τις περιπτώσεις πραγματικών εφαρμογών και για αυτό είναι πολύ χρήσιμο στην πράξη.

Όταν λοιπόν η κατανομή των πλήρων δεδομένων (complete data) ανήκει στην εκθετική οικογένεια κατανομών τότε για το Μ-βήμα χρειαζόμαστε απλά τα επαρκή στατιστικά. Αυτό σημαίνει επίσης πως στο Ε-βήμα έχουμε να εκτιμήσουμε μόνο τις αναμενόμενες τιμές των επαρκών στατιστικών κάτι που κάνει την όλη διαδικασία πολύ ευκολότερη καθώς το Ε-βήμα περιέχει συνήθως απλές αναμενόμενες τιμές και μάλιστα απλά την αναμενόμενη τιμή του επαρκούς στατιστικού και όχι για κάθε παρατήρηση ξεχωριστά. Αυτό είναι κατά πολύ ευκολότερο και οδηγεί σε αλγόριθμο πολύ πιο γρήγορο στην εκτέλεση του. Το αποτέλεσμα αυτό προκύπτει απλά γιατί η λογαριθμική πιθανοφάνεια μιας κατανομής από την εκθετική οικογένεια είναι συνάρτηση απλών συναρτήσεων των δεδομένων. Αν ανατρέξει κανείς πίσω στο παράδειγμα 6.1 με την κατανομή Poisson θα δεί πως το μόνο που χρειάζεται κανείς είναι το $\sum x_i$ που είναι μια απλή συνάρτηση.

6.8.7 Τυπικά σφάλματα

Όπως είπαμε και πριν, ο αλγόριθμος ΕΜ δεν χρησιμοποιεί τις δεύτερες παραγώγους κι έτσι τα τυπικά σφάλματα δεν προκύπτουν κατά τη διάρκεια εκτέλεσης του αλγορίθμου. Μπορούν βέβαια να υπολογιστούν οι παράγωγοι αυτές μετά το τέλος του αλγορίθμου. Εναλλακτικά μπορεί κανείς να χρησιμοποιήσει τη δομή του αλγορίθμου για να βρει τα τυπικά σφάλματα.

Βασικό εργαλείο είναι η αρχή της χαμένης πληροφορίας (missing information principle) η οποία δηλώνει πως ισχύει:

$$I_{obs} = I_{com} - I_{mis}$$

όπου I_{obs} είναι η παρατηρούμενη πληροφορία, I_{com} η πληροφορία των πλήρων δεδομένων και I_{mis} η χαμένη πληροφορία (missing Information). Η παραπάνω σχέση απλά υποδεικνύει πως η πληροφορία που παρατηρούμε είναι αυτή που περιέχεται στα πλήρη δεδομένα και αφαιρέσουμε από αυτήν την πληροφορία των missing data.

Αυτό μπορεί να εκφραστεί ως

$$-\frac{d^2\ell(\theta \mid Y)}{d\theta_i d\theta_j} = \left[-\frac{d^2Q(\theta,\phi)}{d\theta_i d\theta_j}\right]_{\phi=\theta} - \left[-\frac{d^2H(\theta,\phi)}{d\theta_i d\theta_j}\right]_{\phi=\theta}$$

Αν δεν υπήρχαν missing data τότε ο πρώτος όρος στο δεξί μέλος θα ήταν απλά ο πίνακας πληροφορίας. Μια χρήσιμη σχέση για τους υπολογισμούς είναι η εξής:

$$-\left[\frac{d^2H(\theta,\phi)}{d\theta_i d\theta_j}\right] = Var\left[\frac{d\ell(\theta \mid Y,Z)}{d\theta}\right]$$

η οποία στην περίπτωση που δεν έχει απλή μορφή μπορεί να προσεγγιστεί σχετικά εύκολα με Monte Carlo.

Εκτός από την παραπάνω ιδέα για την εύρεση τυπικών σφαλμάτων μπορεί κανείς να χρησιμοποιήσει και άλλους τρόπους που βασίζονται είτε στην ιστορία των επαναλήψεων είτε ακόμα και σε bootstrap. Κάτι τέτοιο είναι σχετικά απλό καθώς ο αλγόριθμος ΕΜ σε κάθε bootstrap επανάληψη ξεκινά από καλές αρχικές τιμές και επομένως συγκλίνει σχετικά γρήγορα.

6.9 Παραδείγματα

6.9.1 Το κλασικό γενετικό πρόβλημα

Το παράδειγμα αυτό αποτελεί μια απλή και κλασική εφαρμογή του ΕΜ. Ας υποθέσουμε πως έχουμε 197 ζώα και θέλουμε να τα κατατάξουμε σε 4 κατηγορίες με βάση κάποιο θεωρητικό μοντέλο σχετικά με τις γενετικές τους σχέσεις. Τα δεδομένα μας σχετικά με τις 4 αυτές κατηγορίες είναι τα εξής

$$X = (x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$$

με θεωρητικές πιθανότητες για κάθε κελί

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4) = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right)$$

Σκοπός μας είναι να βρούμε την ΕΜΠ για το θ. Θα χρησιμοποιήσουμε τον ΕΜ και αργότερα θα δούμε και τη μέθοδο Newton-Raphson για να κάνουμε τις απαραίτητες συγκρίσεις.

Ένα τρόπος συμπλήρωσης των δεδομένων (data augmentation) στο παράδειγμα μας είναι να σπάσουμε το πρώτο κελί (x_1) σε δυο κελιά. Ο λόγος που μας οδηγεί σε αυτό είναι πως η πιθανότητα του κελιού είναι ένα άθροισμα το οποίο δημιουργεί τα προβλήματα κατά τη μεγιστοποίηση. Επομένως ελπίζουμε έτσι πως τα πλήρη δεδομένα θα έχουν πιο απλή πιθανοφάνεια την οποία και θα μπορούμε να μεγιστοποιήσουμε πιο εύκολα.

Έτσι τα πλήρη δεδομένα θα έχουν τη μορφή $Y = (y_0, y_1, y_2, y_3, y_4)$, όπου $y_i = x_i$, για $i = 2, 3, 4$ και $y_0 + y_1 = x_1$, και ομοίως οι πιθανότητες για τα κελιά y_0 και y_1 είναι αντίστοιχα $1/2$ και $\theta/4$.

Η παρατηρούμενη πιθανοφάνεια δίνεται ως

$$L(\theta \mid X) \propto (2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4}$$

ενώ η πιθανοφάνεια των πλήρων δεδομένων ως

$$L(\theta \mid Y) \propto (1-\theta)^{y_2+y_3}\theta^{y_4+y_1}$$

που είναι αρκετά πιο απλή και επομένως βρίσκουμε εύκολα πως η λογαριθμική πιθανοφάνεια για τα πλήρη δεδομένα είναι

$$\log L(\theta \mid Y) \propto (y_2+y_3)\log(1-\theta) + (y_4+y_1)\log\theta$$

Επίσης παρατηρείστε πως με βάση τον τρόπο που ορίσαμε τα πλήρη δεδομένα, υποθέτουμε πως η δεσμευμένη κατανομή $y_1 \mid \theta, X$ είναι η Διωνυμική με παραμέτρους $n = 125$ και $p = \frac{\theta}{\theta+2}$ (παρατηρείστε πως $\frac{\theta}{\theta+2} = \frac{\theta/4}{\theta/4+1/2}$). Από την πιθανοφάνεια των πλήρων δεδομένων μπορεί κανείς εύκολα να δει ότι

$$\hat{\theta} = \frac{y_1 + y_4}{y_1 + y_2 + y_3 + y_4}$$

αλλά στην περίπτωση μας δεν γνωρίζουμε το y_1. Με βάση λοιπόν αυτά που είδαμε προηγουμένως (θα δούμε σε λίγο την ακριβή μαθηματική απόδειξη) θα πρέπει να

εκτιμάμε σε κάθε επανάληψη την αναμενόμενη τιμή του y_1 χρησιμοποιώντας την ως τώρα πληροφορία και να ανανεώνουμε το θ με βάση τον παραπάνω τύπο. Ο αλγόριθμος έχει την παρακάτω μορφή:

Ε-βήμα: Υπολόγισε

$$E(y_1 \mid \theta, X) = \frac{125\theta}{\theta + 2} = t$$

που είναι απλά η αναμενόμενη τιμή της διωνυμικής κατανομής που είδαμε

Μ-βήμα: Ανανέωσε το θ ως

$$\theta^{(new)} = \frac{t + y_4}{t + y_2 + y_3 + y_4}$$

Για το παράδειγμα μας και με αρχικά τιμή $\theta^{(0)} = 0.40$ ο αλγόριθμος συγκλίνει μετά από 8 επαναλήψεις με κριτήριο τερματισμού να σταματήσουμε όταν δύο διαδοχικές τιμές του θ δεν διαφέρουν περισσότερο από 10^{-6}. Οι διαδοχικές τιμές είναι:

$$(0.4, 0.5906643, 0.6218892, 0.6261642, 0.6267342,$$

$$0.6268099, 0.626820, 0.6268213, 0.6268215).$$

Συνεπώς για τα δεδομένα μας έχουμε $\hat{\theta} = 0.6268215$.

Ας δούμε με περισσότερη λεπτομέρεια αυτό που είδαμε πριν διαισθητικά. Από τη θεωρία του ΕΜ στο Ε-βήμα χρειάζεται να υπολογίσουμε την ποσότητα

$$\begin{aligned} Q(\theta, \theta^{(r)}) &= \int_Z \log L(\theta \mid X, Z) P(Z \mid \theta^{(r)}, X) dZ = E(\log L(\theta \mid X, Z)) \\ &= constant + E\left[(y_2 + y_3)\log(1-\theta) + (y_1 + y_4)\log\theta\right] \\ &= constant + (y_2 + y_3)\log(1-\theta) + (E\left[y_1\right] + y_4)\log\theta \end{aligned}$$

αφού μόνο το y_1 είναι τυχάια μεταλητή. Η αναμενόμενη τιμή υπολογίζεται ως προς τη δεσμευμένη κατανομή του $y_1 \mid X, \theta^{(r)}$ (που όπως είδαμε είναι η διωνυμική κατανομή).

Το Μ-βήμα μεγιστοποιεί αυτή τη συνάρτηση ως προς θ που σύμφωνα με τα προηγούμενα προκύπτει πως ειναι:

$$\hat{\theta} = \frac{E\left[y_1\right] + y_4}{E\left[y_1\right] + y_2 + y_3 + y_4}$$

δηλαδή τα missing data έχουν αντικατασταθεί από την αναμενόμενη τους τιμή. Παρατηρείτε πως ουσιαστικά αυτό που δεν γνωρίζουμε στο Ε-βήμα για να υπολογίσουμε τη συνάρτηση που μας ενδιαφέρει είναι απλά η ποσότητα $E\left[y_1\right]$ και άρα αυτή η ποσότητα χρειάζεται να υπολογιστεί στο Ε-βήμα.

Ας δούμε για λόγους σύγκρισης και τη μέθοδο Newton-Raphson για το ίδιο παράδειγμα. Η μέθοδος χρησιμοποιεί τον επαναληπτικό τύπο

$$\theta^{(new)} = \theta - \frac{\frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}}{-\frac{x_1}{(2+\theta)^2} - \frac{x_2+x_3}{(1-\theta)^2} - \frac{x_4}{\theta^2}}$$

Ξεκινώντας από την ίδια αρχική τιμή και με το ίδιο κριτήριο τερματισμού μετά από 4 επαναλήψεις βρίσκουμε

$$(0.6170669, 0.6269629, 0.6268215, 0.6268215),$$

καταλήγοντας, όπως αναμενόταν, στο ίδιο $\hat{\theta}$

Ο παρατηρούμενος πίνακας πληροφορίας του Fisher είναι

$$J(\theta) = \frac{x_1}{(2+\theta)^2} + \frac{x_2 + x_3}{(1-\theta)^2} + \frac{x_4}{\theta^2}$$

ενώ ο αναμενόμενος

$$I(\theta) = \frac{n}{4(2+\theta)} + \frac{2n}{4(1-\theta)^2} + \frac{n}{4\theta^2}$$

Συνεπώς η μέθοδος scoring θα χρησιμοποιούσε τον τύπο

$$\theta^{(new)} = \theta - \frac{\frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}}{-\frac{n}{4(2+\theta)} - \frac{2n}{4(1-\theta)^2} - \frac{n}{4\theta^2}}$$

Στο Γράφημα 6.9 μπορεί κανείς να δεί τη λογαριθμική πιθανοφάνεια για το παράδειγμα μας (λείπει η σταθερά).

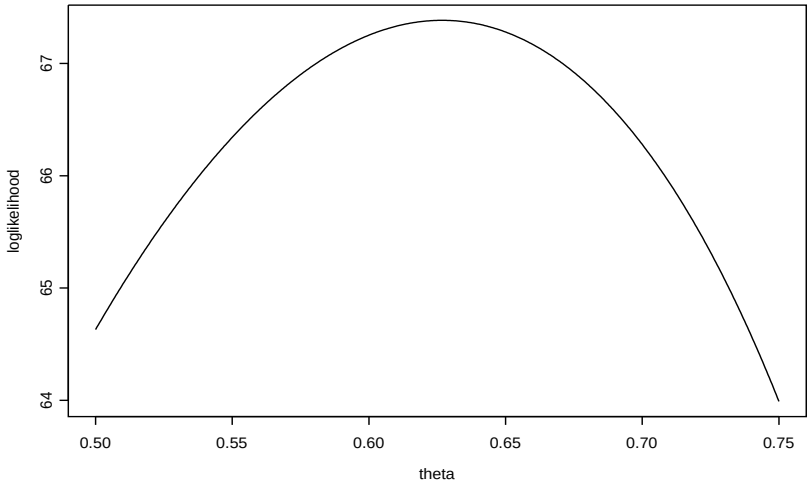
Παρατηρούμε πως η μέθοδος Newton-Raphson συγκλίνει γρηγορότερα αλλά θα μπορούσε να δώσει λύση έξω από το διάστημα $(0, 1)$. Επίσης παρατηρείστε πως απαιτεί περισσότερες πράξεις σε κάθε επανάληψη κι επομένως το κέρδος των λιγότερων επαναλήψεων δεν οδηγεί απαραίτητα σε κέρδος από άποψη χρόνου.

6.9.2 Περικομμένη κατανομή Poisson

Άς ξαναγυρίσουμε στο παράδειγμα με την περικομμένη κατανομή Poisson. Έστω πως n_i δηλώνει τη συχνότητα της i τιμής. Γνωρίζουμε επίσης πως το συνολικό μέγεθος του δείγματος είναι $\sum_{i=1}^{7} n_i = n = 1200$. Για να χρησιμοποιήσουμε τον αλγόριθμο ΕΜ θεωρούμε ως missing data τη συχνότητα της τιμής 0, επομένως δεν έχουμε παρατηρήσει την n_0. Συνεπώς τα πλήρη δεδομένα στην περίπτωση αυτή θα ήταν τα δεδομένα που περιείχαν όλες τις τιμές και του 0 συμπεριλαμβανομένου. Η κατανομή τους τότε δεν θα ήταν η περικομμένη κατανομή Poisson αλλά η απλή κατανομή Poisson και επομένως η μεγιστοποίηση της πιθανοφάνειας των πλήρων δεδομένων είναι πολύ απλή. Με βάση τη βασική ιδέα του ΕΜ στο Ε-βήμα χρειαζόμαστε την αναμενόμενη τιμή των δεδομένων που δεν έχουμε παρατηρήσει και στη συνέχεια εκτιμάμε το λ από τη μέση τιμή των πλήρων δεδομένων. Ο αλγόριθμος περιγράφεται ως:

Έστω αρχικές τιμές για τα n_0 και λ. Τότε
E-step Βρες

$$n_0^{(new)} = (n_0 + n)exp(-\lambda).$$



Γράφημα 6.9: Η λογαριθμική πιθανοφάνεια για το παράδειγμα μας (λείπει η σταθερά) σαν συνάρτηση του θ.

M-step: Ανανέωσε το λ ως

$$\lambda^{(new)} = \frac{\sum_{i=1}^{n} x_i}{n + n_0^{(new)}}$$

Σταμάτα τις επαναλήψεις αν κάποιο κριτήριο τερματισμού ικανοποιείται.

Χρειαζόμαστε κάποια καλή αρχική τιμή για το n_0. Από τη θεωρία της κατανομής Poisson γνωρίζουμε πως $n_1/n_0 = \lambda$ κι επομένως αν έχουμε μια αρχική τιμή για το λ μπορούμε να βρούμε μια καλή αρχική τιμή και για το n_0. Για παράδειγμα στην περικομμένη κατανομή Poisson ισχύει πως $n_2/n_1 = \lambda/2$ κι επομένως αρχική τιμή για το λ μπορεί να βρεθεί πολύ εύκολα.

Μερικά ενδιαφέροντα σημεία σχετικά με το παράδειγμα είναι τα ακόλουθα

- Ο EM χρειάζεται περισσότερες επαναλήψεις από ότι ο Newton-Raphson αλλά και πάλι οι υπολογισμοί είναι σχετικά απλούστεροι μέσα σε κάθε επανάληψη.

- Ο Newton-Raphson, όπως είδαμε μπορεί να οδηγηθεί σε μη επιτρεπτή λύση, κάτι τέτοιο δεν πρόκειται να συμβεί ποτέ με τον EM αρκεί η αρχική τιμή να είναι στα επιτρεπτά όρια.

- Η ιδέα μπορεί εύκολα να γενικευτεί στην περίπτωση που έχουμε περισσότερες από μια συχνότητες να μην παρατηρούνται. Δηλαδή για παράδειγμα θα μπορούσαμε να μην έχουμε παρατηρήσει τιμές ίσες με το 0 αλλά και ίσες με το 1. Δηλαδή ο αλγόριθμος με μικρές τροποποιήσεις λειτουργεί για περικομμένα δεδομένα κάθε μορφής αλλά και από κάθε κατανομή.

Στο Γράφημα 6.10 μπορούμε να δούμε την ιστορία του αλγορίθμου για 2 διαφορετικές αρχικές τιμές (0.2 και 2) αντίστοιχα. Στο δεξί γράφημα βλέπουμε την τιμή της παραμέτρου και στο δεξί την τιμή της λογαριθμικής πιθανοφάνειας. Παρατηρείστε πως για αρχική τιμή $\lambda = 2$ ο αλγόριθμος συγκλίνει πολύ πιο γρήγορα. Με τη μέθοδο Newton-Raphson για αρχική τιμή $\lambda = 0.2$ ο αλγόριθμος αποτύγχανε.

Αν θέλει κανείς να δεί το πρόβλημα χρησιμοποιώντας τους γενικούς τύπους, ουσιαστικά μπορεί να δεί πως η λογαριθμική πιθανοφάνεια για τα πλήρη δεδομένα είναι

$$\log L(\lambda|Y) \propto -\lambda \sum_{x=0}^{M} n_x + \log \lambda \sum_{x=0}^{M} x n_x$$

η οποία γράφεται ως

$$\log L(\lambda|Y) \propto -\lambda n_0 - \lambda \sum_{x=1}^{M} n_x + \log \lambda \sum_{x=1}^{M} x n_x$$

Συνεπώς στο Ε-βήμα που χρειάζομαι την αναμενόμενη τιμή της πιθανοφάνειας αυτής, ουσιαστικά χρειάζομαι μόνο την αναμενόμενη τιμή του n_0. Όμως με βάση το μοντέλο, μια παρατήρση παίρνει την τιμή 0 με πιθανότητα $e^{-\lambda}$ και δεν την παίρνει με πιθανότητα $1-e^{-\lambda}$, ενώ συνολικά έχω $n+n_0$ παρατηρήσεις. Επομένως πρόκειται για μια απλή αναμενομενη τιμή από διωνυμική κατανομή.

6.9.3 Πεπερασμένα μείγματα κατανομών (finite mixtures)

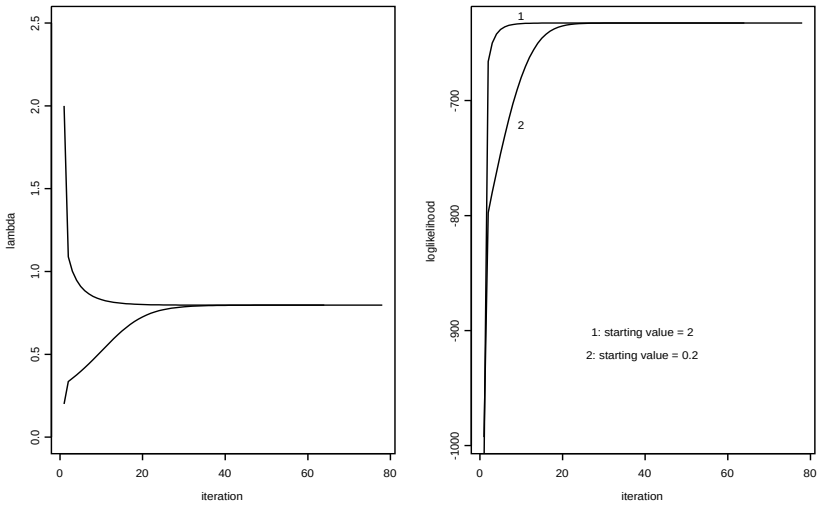
΄Εστω πως ο πληθυσμός μας αποτελείται από k υποπληθυσμούς (ομάδες). ΄Εστω πως γνωρίζουμε για κάθε ομάδα την συνάρτηση πυκνότητας πιθανότητας της. Ξέρουμε δηλαδή πως

$$\begin{aligned} &\text{ομάδα } 1 \quad f(x \mid \theta_1) \\ &\text{ομάδα } 2 \quad f(x \mid \theta_2) \\ &\qquad \dots \\ &\text{ομάδα } k \quad f(x \mid \theta_k) \end{aligned}$$

Διαλέγοντας ένα άτομο από τον πληθυσμό τυχαία τότε η συνάρτηση πυκνότητας πιθανότητας του θα είναι

$$f(x) = \sum_{j=1}^{k} p_j f(x \mid \theta_j) \tag{6.6}$$

όπου $p_j > 0$, $\sum_{j=1}^{k} p_j = 1$ και τα p_j ουσιαστικά δηλώνουν την πιθανότητα ένα τυχαία επιλεγμένο άτομο να προέρχεται από τον πληθυσμό j. Η παράμετρος θ_j δεν είναι



Γράφημα 6.10: Η ιστορία του ΕΜ για διαφορετικές αρχικές τιμές, στο δεξί γράφημα βλέπουμε την τιμή της παραμέτρου και στο δεξί την τιμή της λογαριθμικής πιθανοφάνειας

απαραίτητο να είναι μία αλλά θα μπορούσε να είναι και διάνυσμα παραμέτρων. Για παράδειγμα αν η κατανομή είναι η κανονική, τότε $\theta = (\mu, \sigma^2)$.

Η εξίσωση (6.6) ορίζει τα πεπερασμένα μείγματα της κατανομής $f(\cdot)$. Τέτοια μείγματα βρίσκουν πολλές εφαρμογές σε ποικιλία προβλημάτων στατιστικής όπως για παράδειγμα στην ανάλυση κατά συστάδες ή στη μοντελοποίηση μη ομοιογενών πληθυσμών. Ιστορικά ο αλγόριθμος ΕΜ εφαρμόστηκε σε τέτοια μοντέλα αρκετά πριν τη συστηματική μελέτη του το 1977 και αποτελεί μια πολύ ενδιαφέρουσα εφαρμογή με αρκετά ενδιαφέρουσα στατιστική ερμηνεία. Θα δούμε δύο εφαρμογές του αλλά είναι σαφές πως η ιδέα γενικεύεται σε πολλές κατευθύνσεις.

Πεπερασμένα μείγματα της κατανομής Poisson

Η κατανομή Poisson παίζει σημαντικό ρόλο ανάμεσα σε όλες τις διακριτές κατανομές και χρησιμοποιείται ευρέως σε πολλές εφαρμογές. Πολλά εναλλακτικά διακριτά μοντέλα σχετίζονται άμεσα με την κατανομή Poisson. Για αυτό και ο ρόλος της για την ανάλυση διακριτών δεδομένων είναι κυρίαρχος. Πεπερασμένα μείγματα της κατανομής Poisson έχουν κι αυτά σημαντικό ρόλο στη βιβλιογραφία. Η συνάρτηση πιθανότητας δίνεται από τον τύπο:

$$f(x_i) = \sum_{j=1}^{k} p_j f(x_i \mid \theta_j) = \sum_{j=1}^{k} p_j \frac{exp(-\theta_j)\theta_j^{x_i}}{x_i!}$$

όπου $p_j > 0$, $j = 1, \ldots, k$ και $\sum_{j=1}^{k} p_j = 1$. Το μοντέλο αυτό υποθέτει πως ο πληθυσμός έχει k υποπληθυσμούς όπου κάθε υποπληθυσμός ακολουθεί μια κατανομή Poisson αλλά με διαφορετικές παραμέτρους.

Από ένα δείγμα μεγέθους n, έστω $X = (x_1, x_2, \ldots, x_n)$ θέλουμε να βρούμε τις ΕΜΠ για τις παραμέτρους του μοντέλου, δηλαδή για τα $(p_1, \ldots, p_{k-1}, \theta_1, \ldots, \theta_k)$. Η λογαριθμική πιθανοφάνεια είναι

$$\ell(\theta \mid \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} p_j f(x_i \mid \theta_j) \right)$$

που επειδή έχει το άθροισμα μέσα στο λογάριθμο δεν είναι εύκολο να μεγιστοποιηθεί. Διάφορες αριθμητικές μέθοδοι μπορούν να χρησιμοποιηθούν αλλά επειδή υπάρχουν περιορισμοί τόσο στα p_j όσο και στα θ_j κάτι τέτοιο δεν είναι εύκολο. Θα χρησιμοποιήσουμε τον ΕΜ για να βρούμε τις ΕΜΠ.

Ας δούμε λοιπόν πως μπορούμε να επιτύχουμε την συμπλήρωση των δεδομένων μας ώστε να διευκολύνουμε τον αλγόριθμο.

΄Εστω τυχαίες μεταβλητές Z_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$ με τιμές $Z_{ij} = 1$ αν η i παρατήρηση ανήκει στη j ομάδα και 0 αλλιώς. Αν είχαμε παρατηρήσει τα Z_{ij} τοτε η εκτίμηση θα ήταν εύκολη καθώς για κάθε παρατήρηση θα ξέραμε σε ποιά ομάδα ανήκει και άρα το πρόβλημα θα ήταν απλά να εκτιμήσουμε τις παραμέτρους πολλών απλών κατανομών Poisson.

Συνεπώς στην περίπτωση μας τα Z_{ij} είναι τα missing data και στο Ε-βήμα θα τα εκτιμήσουμε με την δεσμευμένη αναμενόμενη τους τιμή, ενώ στο Μ-βήμα

θα χρησιμοποιήσουμε αυτές τις εκτιμήσεις για να ανανεώσουμε τις παραμέτρους μας. Αποδεικνύεται ότι στο Μ-βήμα χρειάζεται απλά να υπολογίσουμε σταθμικούς μέσους και διακυμάνσεις χρησιμοποιώντας ως σταθμίσεις τα αποτελέσματα από το Ε-βήμα.

Δηλαδή συμπληρώνουμε τα παρατηρούμενα δεδομένα X με τα missing Z. Για τα πλήρη δεδομένα $Y = (X, Z)$ η λογαριθμική πιθανοφάνεια γίνεται

$$\ell(\theta \mid X, Z) = \sum_{i=1}^{n} \sum_{j=1}^{k} [(z_{ij} \log p_j) + z_{ij} \left(-\theta_j + x_i \log \theta_j - \log(x_i!)\right)]$$

Παρατηρείστε πως τώρα είναι αρκετά εύκολο να βρούμε τις ΕΜΠ από αυτή την πιθανοφάνεια. Για παράδειγμα παίρνοντας την πρώτη παράγωγο ώς προς θ_1 και εξισώνοντας τη με το 0 βρίσκουμε εύκολα πως

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}$$

Επίσης μπορεί κάποιος να παρατηρήσει πως η αναμενόμενη τιμή της $\ell(\theta \mid X, Z)$ παίρνει την σχετικά απλή μορφή

$$\begin{aligned} E\left[\ell(\theta \mid X, Z)\right] &= \sum_{i=1}^{n} \sum_{j=1}^{k} E[z_{ij} \mid \theta, X] \log p_j \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{k} E[z_{ij} \mid \theta, X] \left(-\theta_j + x_i \log \theta_j - \log(x_i!)\right) \end{aligned}$$

Συνεπώς για τον υπολογισμό της αναμενόμενης τιμής της πιθανοφάνειας των πλήρων δεδομένων και τη μεγιστοποίηση της χρειαζόμαστε την εύρεση της $E[z_{ij} \mid \theta, X]$. Όμως εξ' ορισμού η τυχαία μεταβλητή Z_{ij} είναι μια Bernoulli τυχαία μεταβλητή. Δεδομένου ότι ενδιαφερόμαστε για τη δεσμευμένη αναμενόμενη τιμή αυτή θα είναι ίδια με την πιθανότητα $P(Z_{ij} = 1 \mid \theta, X)$, δηλαδή τη δεσμευμένη πιθανότητα η παρατήρηση να ανήκει στη j ομάδα. Για το M-βήμα χρειάζεται να μεγιστοποιήσουμε την αναμενόμενη τιμή της λογαριθμικής πιθανοφάνειας των πλήρων δεδομένων και συνεπώς, όπως είδαμε πριν κάτι τέτοιο είναι σχετικά απλό. Επίσης μπορεί να παρατηρήσει κανείς πως αν θέλουμε την από κοινού κατανομή των $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ik})$ αυτή είναι μια πολυωνυμική κατανομή (multinomial) με αντίστοιχες πιθανότητες αυτές που είδαμε πριν για κάθε Z_{ij}.

Με βάση τα παραπάνω τα βήματα του αλγορίθμου μπορούν να περιγραφούν ως:

΄Εστω $p_1, \ldots, p_k$ και $\theta_1, \ldots, \theta_k$ οι τιμές των παραμέτρων από την προηγούμενη επανάληψη.

Ε-βήμα: Υπολόγισε

$$w_{ij} = \frac{p_j f(x_i \mid \theta_j)}{\sum_{j=1}^{k} p_j f(x_i \mid \theta_j)}$$

Μπορεί εύκολα να δει κάποιος πως τα w_{ij} δεν είναι παρά οι εκ των υστέρων πιθανότητες η i παρατήρηση να ανήκει στη j ομάδα. Συνεπώς για μερικές εφαρμογές όπως η ανάλυση σε ομάδες αυτή η ποσότητα έχει από μόνη της ενδιαφέρον για στατιστική συμπερασματολογία (για την κατάταξη των παρατηρήσεων σε ομάδες)

Μ-βήμα: Βρες τις καινούριες εκτιμήσεις ως

$$p_j^{(new)} = \frac{\sum_{i=1}^{n} w_{ij}}{n} \text{ and } \theta_j^{(new)} = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} w_{ij}}$$

Σταμάτα τις επαναλήψεις αν κάποιο κριτήριο τερματισμού ικανοποιείται αλλιώς συνέχισε πηγαίνοντας πίσω στο Ε-βήμα.

Η παραπάνω συμπλήρωση δεδομένων μπορεί να γενικευτεί σε πολλές κατευθύνσεις. Για παράδειγμα ισχύει για πεπερασμένα μείγματα οποιασδήποτε κατανομής αλλά ακόμα και για μείγματα με διαφορετικές κατανομές για κάθε υποπληθυσμό. Επίσης η ιδέα χρησιμοποιείται εξίσου και για Μπευζιανές προσεγγίσεις εκτίμησης των παραμέτρων.

Παράδειγμα: Τα δεδομένα αφορούν τον αριθμό των ανθρωποκτονιών στην ελλάδα για το έτος 1997 στους 54 νομούς της χώρας. Έστω x_i ο αριθμός των ανθρωποκτονιών στον i νομό και t_i ο πληθυσμός του νομού σε εκατομύρια κατοίκους. Υποθέτουμε πως ο πληθυσμός είναι ανομοιογενής, υπάρχουν δηλαδή νομοί με μεγαλύτερη εγκληματικότητα από άλλους. Ας υποθέσουμε ότι υπάρχουν 4 διαφορετικές ομάδες νομών ανάλογα με την εγκληματικότητα τους. Τότε, αφού δεν γνωρίζουμε ποιος νομός ανήκει σε ποια ομάδα είναι λογικό να υποθέσουμε ένα πεπερασμένο μείγμα κατανομών Poisson ως την κατανομή του πληθυσμού όλων των νομών. Δηλαδή υποθέτουμε πως

$$f(x_i \mid t_i, \theta) = \sum_{j=1}^{4} p_j \frac{(\lambda_j t_i)^{x_i} exp(-\lambda_j t_i)}{x_i!}$$

Το μοντέλο είναι ελαφρώς διαφορετικό γιατί τώρα χρησιμοποιούμε και την πληροφορία για το μέγεθος του πληθυσμού κάθε νομού t_i. Θέλουμε να εκτιμήσουμε τα $(p_1, p_2, p_3, \lambda_1, \lambda_2, \lambda_3, \lambda_4)$. Τα βήματα του αλγορίθμου διαφέρουν ελάχιστα από αυτά που είδαμε πρίν και συγκεκριμένα τώρα ο αλγόριθμος γίνεται

Ε-βήμα: Υπολόγισε

$$w_{ij} = \frac{p_j f(x_i \mid \lambda_j t_i)}{\sum_{j=1}^{k} p_j f(x_i \mid \lambda_j t_i)}$$

Μ-βήμα: Βρες τις καινούριες εκτιμήσεις ως

$$p_j^{(new)} = \frac{\sum_{i=1}^{n} w_{ij}}{n} \text{ and } \lambda_j^{(new)} = \frac{\sum_{i=1}^{n} w_{ij} x_i t_i}{\sum_{i=1}^{n} w_{ij} t_i}$$

initial values	estimated parameters	Loglikelihood
$p = (0.25, 0.25, 0.25, 0.25)$	$\hat{p} = (0.1357, 0.3750, 0.3913, 0.0979)$	-114.8306
$\lambda = (0.01, 17.20, 32.70, 228.6)$	$\hat{\lambda} = (0, 24.0308, 34.6209, 70.8688)$	
$p = (0.25, 0.25, 0.25, 0.25)$	$\hat{p} = (0.0805, 0.0527, 0.7052, 0.1615)$	-115.4153
$\lambda = (1, 2, 5, 10)$	$\hat{\lambda} = (0, 0, 26.7921, 62.6875)$	

Πίνακας 6.2: Αποτελέσματα ξεκινώντας από διαφορετικές τιμές αλλά με το ίδιο κριτήριο τερματισμού (σταματήσαμε τις επαναλήψεις όταν η σχετική διαφορά στην πιθανοφάνεια ήταν μικρότερη από 10^{-9}

Νομός	Weights			
Αθήνα	0	0.9989	0.0011	0
Θεσσαλονίκη	0	0.2021	0.7971	0
Θεσπρωτία	0.4280	0.3499	0.2135	0.0085
Αιτωλοακαρνανία	0	0.4700	0.5259	0.0041
Αργολίδα	0	0.3791	0.5644	0.0564
...		...		

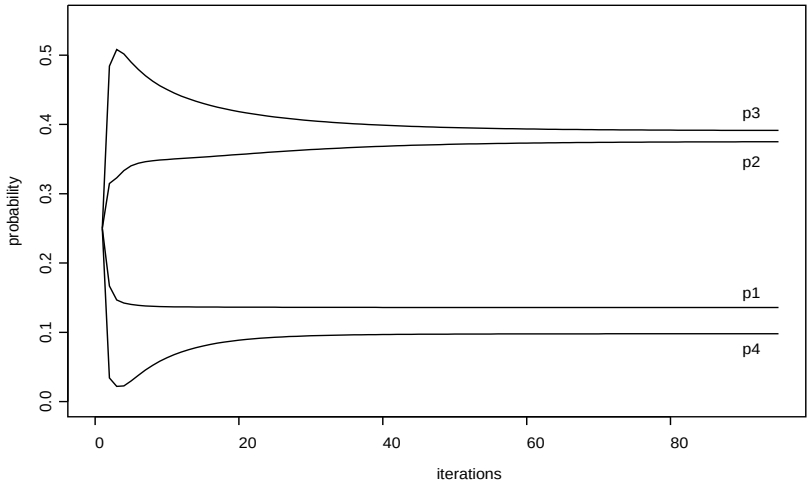
Πίνακας 6.3: Οι τιμές w_{ij} μετά το τέλος του ΕΜ από το πρώτο σετ αρχικών τιμών (δηλαδή με βάση τις ΕΜΠ).

Στον πίνακα 6.2 μπορεί κανείς να δει τα αποτελέσματα ξεκινώντας από διαφορετικές τιμές αλλά με το ίδιο κριτήριο τερματισμού (σταματήσαμε τις επαναλήψεις όταν η σχετική διαφορά στην πιθανοφάνεια ήταν μικρότερη από 10^{-9}). Αυτό που μπορεί εύκολα να παρατηρήσει κανείς είναι πως οι διαφορετικές αρχικές τιμές οδηγούν και σε διαφορετική λύση! Θα πρέπει επομένως κανείς να δοκιμάσει να ξεκινήσει με περισσότερες αρχικές τιμές ώστε να είναι πιο σίγουρος πως έχει βρει πραγματικά το μέγιστο και όχι κάποιο τοπικό μέγιστο. Οι πραγματικές ΕΜΠ αντιστοιχούν σε αυτές της πρώτης ομάδας αρχικών τιμών. Παρατηρείστε πως η δεύτερη ομάδα οδηγεί ουσιαστικά σε μια λύση με 3 και όχι 4 ομάδες καθώς παρατηρείστε πως οι εκτιμηθείσες παράμετροι για τις 2 πρώτες ομάδες ταυτίζονται.

Με βάση τα αποτελέσματα του πίνακα βρέθηκαν 4 ομάδες από νομούς. Η πρώτη ομάδα $p_1 = 0.08$ $\lambda_1 = 0$ δηλαδή περιέχει περίπου το 8% των νομών και αυτοί οι νομοί παρουσιάζουν μηδενική εγκληματικότητα. Οι άλλες ομάδες έχουν ολοένα μεγαλύτερη εγκληματικότητα (ουσιαστικά οι παράμετροι λ_j αναφέρονται σε ανθρωποκτονίες ανά εκατομμύριο κατοίκων). Όπως είπαμε και πριν το ενδιαφέρον της εφαρμογής είναι να κατατάξουμε τους νομούς σε ομάδες. Αυτό γίνεται με τη χρήση των w_{ij} μετά το τέλος του αλγορίθμου, μετά δηλαδή από την τελευταία επανάληψη. Στον πίνακα 6.3 μπορεί κανείς να δει ένα μικρό απόσπασμα από αυτά τα βάρη. Έτσι η Αθήνα ανήκει στη δεύτερη ομάδα με πιθανότητα 99% και στην τρίτη με πιθανότητα 1%. Η Θεσσαλονίκη ανήκε με μεγάλη πιθανότητα στην 3η ομάδα κλπ.

Στα Γράφηματα 6.11 και 6.12 μπορεί κανείς αν δει την ιστορία των επαναλήψεων

του αλγορίθμου για το πρώτο σετ αρχικών τιμών τόσο για τα p_j όσο και για τα λ_j. Ομοίως στο Γράφημα 6.14 μπορεί να δει την πιθανοφάνεια αλλά και το κριτήριο τερματισμού ως προς τις επαναλήψεις. Με λίγες επαναλήψεις έχουμε φτάσει πολύ κοντά στο μέγιστο και ύστερα ανάλογα και με το κριτήριο τερματισμού χρειαζόμαστε αρκετές επαναλήψεις για να φτάσουμε στο μέγιστο. Δηλαδή ο αλγόριθμος φαίνεται να χάνει σταδιακά την ενέργεια του και να προχωρά όλο και πιο αργά. Στο Γράφημα **??** έχουμε το ίδιο γράφημα αλλά για το δεύτερο σετ αρχικών τιμών. Η πιθανοφάνεια μεγαλώνει σε κάθε επανάληψη, κάτι που γνωρίζουμε από τη θεωρία αλλά τώρα με πολύ πιο αργούς ρυθμούς, ενώ το κριτήριο δεν έχει μονότονη συμπεριφορά, κάτι τέτοιο δεν αποκλείεται να συμβεί ειδικά αν οι αρχικές τιμές είναι κακές. Βασικά ο αλγόριθμος μπερδεύεται εξαιτίας των 2 πολύ ομοίων ομάδων που βρίσκει και από κάποιο σημείο και μετά γίνεται αργός.

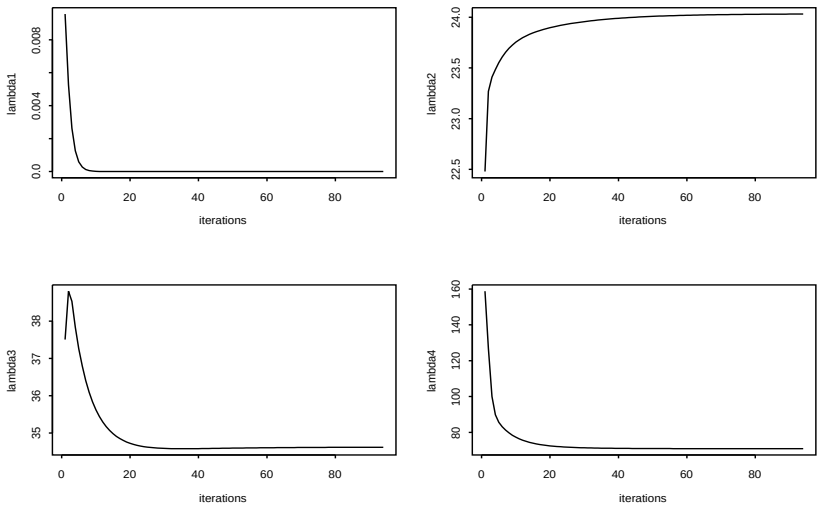


Γράφημα 6.11: Η ιστορία των επαναλήψεων για τα p_j (πρώτο σετ αρχικών τιμών).

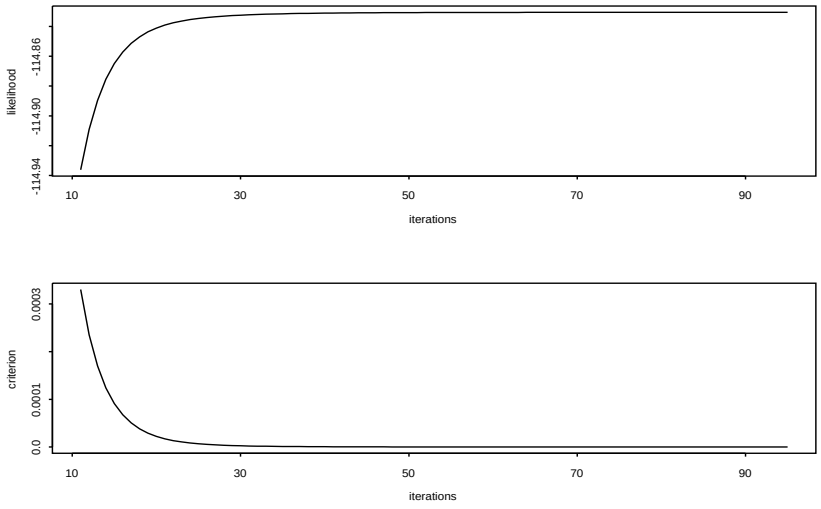
6.9.4 Πεπερασμένα μείγματα της πολυμεταβλητής κανονικής κατανομής

Η πολυμεταβλητή κανονική κατανομή αποτελεί τη βάση για πληθώρα από στατιστικά μοντέλα και τεχνικές και επομένως ο ρόλος της στη στατιστική είναι πάρα πολύ σημαντικός. Πεπερασμένα μείγματα της κατανομής αυτής αποτελούν το θεμέλιο για μεθόδους όπως η ανάλυση σε ομάδες. Ο αλγόριθμος ΕΜ αποτέλεσε την κινητήρια δύναμη για την πρακτική εφαρμογή της μεθόδου σε πραγματικά δεδομένα. Ας δούμε σύντομα τον αλγόριθμο.

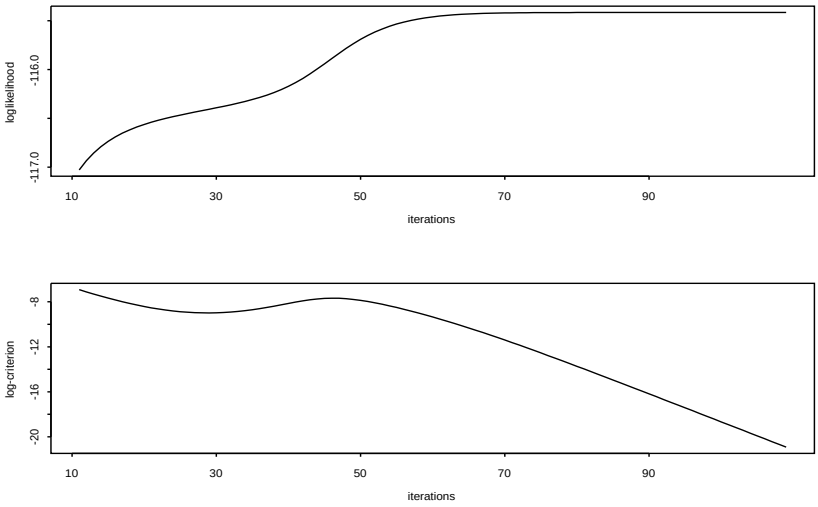
Ας υποθέσουμε ότι έχουμε k ομάδες, και οι παρατηρήσεις μέσα στη j ομάδα



Γράφημα 6.12: Η ιστορία των επαναλήψεων για τα λ_j (πρώτο σετ αρχικών τιμών).



Γράφημα 6.13: Η ιστορία των επαναλήψεων για την πιθανοφάνεια και το κριτήριο τερματισμού $|L^{k+1}/L^k - 1|$ (πρώτο σετ αρχικών τιμών)



Γράφημα 6.14: Η ιστορία των επαναλήψεων για την πιθανοφάνεια και το κριτήριο τερματισμού $|L^{k+1}/L^k - 1|$ (δεύτερο σετ αρχικών τιμών)

ακολουθούν πολυμεταβλητή κανονική κατανομή με διάνυσμα μέσων μ_j και πίνακα διακύμανσης $\mathbf{\Sigma}_j$. Τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας του τυχαίου διανύσματος $\mathbf{x}$ δίνεται ως:

$$f(\mathbf{x}) = \sum_{j=1}^{k} p_j (2\pi)^{-m/2} |\mathbf{\Sigma}_j|^{-1/2} exp\left(-\frac{1}{2}(\mathbf{x}-\mu_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\mu_j)^T\right),$$

όπου m είναι η διάσταση των δεδομένων. Στην πράξη ενδιαφέρουν 2 διαφορετικές περιπτώσεις:

- Ομοσκεδαστικό μοντέλο: Όλοι οι πίνακες διακύμανσης είναι ίδιοι $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_k = \mathbf{\Sigma}$
- Ετεροσκεδαστικό μοντέλο: Όλοι οι πίνακες διακύμανσης δεν είναι ίδιοι, δηλαδή, $\mathbf{\Sigma}_i \neq \mathbf{\Sigma}_j$, για όλα τα $i \neq j$

Όπως και πριν απευθείας μεγιστοποίηση της πιθανοφάνειας δεν είναι καθόλου απλή δουλειά για αυτό καταφεύγουμε στον αλγόριθμο ΕΜ. Και πάλι χρησιμοποιούμε την ίδια συμπλήρωση δεδομένων ορίζοντας τις ψευδομεταβλητές Z_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$ που παίρνουν τιμές 0 και 1 όπως τις ορίσαμε και προηγουμένως

Για την ετεροσκεδαστική περίπτωση ο αλγόριθμος τώρα παίρνει την εξής μορφή: Έστω αρχικές τιμές $p_1, \ldots, p_{k-1}$, $\mu_1, \ldots, \mu_k$, $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_k$.

E-βήμα Υπολόγισε

$$\begin{aligned} w_{ij} &= \frac{p_j f(\mathbf{x}_i \mid \theta_j)}{\sum_{j=1}^{k} p_j f(\mathbf{x}_i \mid \theta_j)} \\ &= \frac{p_j (2\pi)^{-p/2} |\mathbf{\Sigma}_j|^{-1/2} exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j)^T\right)}{\sum_{j=1}^{k} p_j (2\pi)^{-p/2} |\mathbf{\Sigma}_j|^{-1/2} exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j)^T\right)} \end{aligned}$$

M-βήμα Βρες τις νέες εκτιμήσεις

$$\begin{aligned} p_j^{(new)} &= \frac{\sum_{i=1}^{n} w_{ij}}{n}, \\ \mu_j^{(new)} &= \frac{\sum_{i=1}^{n} w_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} w_{ij}}, \\ \mathbf{\Sigma}_j^{(new)} &= \frac{\sum_{i=1}^{n} w_{ij}\left(\mathbf{x}_i - \mu_j^{(new)}\right)\left(\mathbf{x}_i - \mu_j^{(new)}\right)^T}{\sum_{i=1}^{n} w_{ij}} \end{aligned}$$

Σταμάτα τις επαναλήψεις αν κάποιο κριτήριο τερματισμού ικανοποιείται αλλιώς συνέχισε πηγαίνοντας πίσω στο Ε-βήμα.

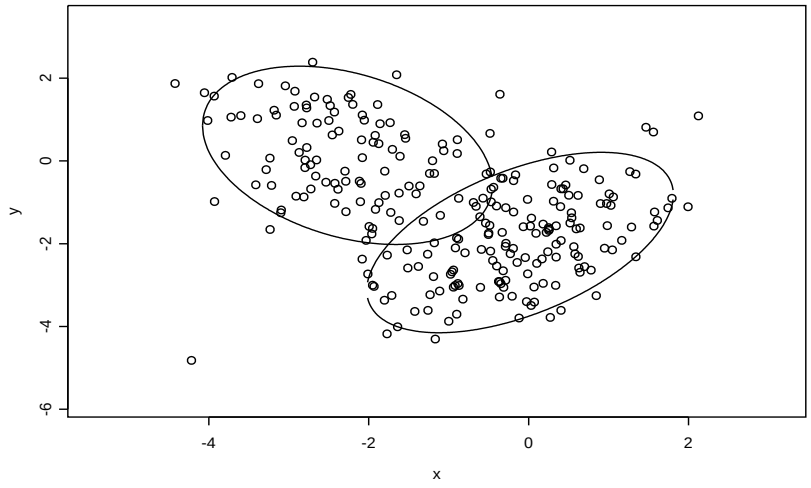
Μπορεί κάποιος να παρατηρήσει πως και πάλι στο Μ-βήμα δεν έχουμε παρά σταθμισμένες εκδόσεις των κλασικών εκτιμητριών για το διάνυσμα των μέσων και τον πίνακα διακύμανσης. Στην περίπτωση του ομοσκεδαστικού μοντέλου η μόνη διαφορά είναι στο Μ-βήμα το οποίο γίνεται

$$\mathbf{\Sigma}^{(new)} = \sum_{j=1}^{k} p_j^{(new)} \mathbf{\Sigma}_j^{(new)}$$

και θυμίζει πολύ το σταθμικό πίνακα διακύμανσης στην πολυμεταβλητή ανάλυση διακύμανσης.

Παράδειγμα: Τα δεδομένα που μπορείτε να δείτε στο γράφημα 6.15 είναι προσομοιωμένα από ένα μείγμα 2 πολυμεταβλητών κατανομών με διαφορετικούς πίνακες διακύμανσης. Στη συνέχεια χρησιμοποιήσαμε τον αλγόριθμο ΕΜ για να εκτιμήσουμε τις παραμέτρους (πίνακας 6.4). Από τα γραφήματα μπορεί κανείς να δεί πως στην περίπτωση ίσων πινάκων (γράφημα 6.16) οι ελλείψεις έχουν το ίδιο μέγεθος και προσανατολισμό κάτι που δεν συμβαίνει στο γράφημα 6.15.

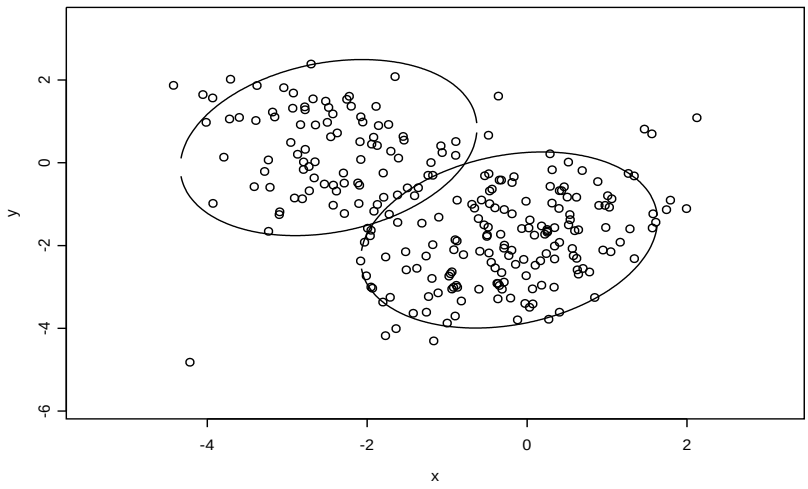
Για να έχετε μια εικόνα της δυναμικής του αλγορίθμου, για την ετεροσκεδαστική περίπτωση μπορεί κανείς να δεί στο γράφημα 6.17 τα ελλειψοειδή σε κάθε επανάληψη του αλγορίθμου. Ουσιαστικά, πολύ γρήγορα οι ελλείψεις προσεγγίζουν την



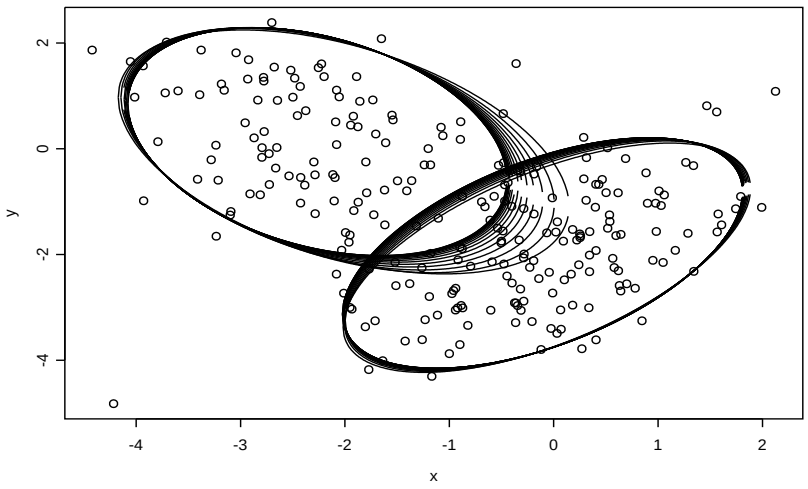
Γράφημα 6.15: Οι δύο ομάδες που προσαρμόσθηκαν στα δεδομένα. Τα ελλειψοειδή αντιστοιχούν σε περιοχές εμπιστοσύνης 95%. Κάθε ομάδα έχει διαφορετικό πίνακα συνδιακύμανης.

Άνισες Διακυμάνσεις	Ίσες Διακυμάνσεις
$p = (0.5865899, 0.4134101)$	$p = (0.6572255, 0.3427745)$
$\mathbf{\Sigma}_1 = \begin{bmatrix} 0.9568578 & 0.5763747 \\ 0.5763747 & 1.2372805 \end{bmatrix}$	$\mathbf{\Sigma} = \begin{bmatrix} 0.8974587 & 0.2312592 \\ 0.2312592 & 1.1782524 \end{bmatrix}$
$\mathbf{\Sigma}_2 = \begin{bmatrix} 0.8568740 & -0.3376073 \\ -0.3376073 & 1.2079469 \end{bmatrix}$	
$\mu_1 = (-0.1073719, -1.9717564)$	$\mu_1 = (-0.2259167, -1.8661791)$
$\mu_2 = (-2.2650514, 0.1351531)$	$\mu_2 = (-2.4823908, 0.3668937)$
loglikelihood $= -832.0454$	loglikelihood $= -846.2068$

Πίνακας 6.4: Οι εκτιμηθέισες ομάδες για τα 2 μοντέλα.



Γράφημα 6.16: Οι δύο ομάδες που προσαρμόσθηκαν στα δεδομένα. Τα ελλειψοειδή αντιστοιχούν σε περιοχές εμπιστοσύνης 95%. Έχουμε υποθέσει ίδιο πίνακα συνδιακύμανης για τις δύο ομάδες.



Γράφημα 6.17: 95% περιοχές εμπιστοσύνης για τις δύο ομάδες υποθέτωντας διαφορετικούς πίνακες διακύμανσεις. Κάθε έλλειψη αντιστοιχεί σε μια επανάληψη του αλγορίθμου.

τελική τιμή. Η επικάλυψη των δύο ομάδων ουσιαστικά σημαίνει πως για τις παρατηρήσεις που βρίσκονται μέσα στην τομή τα w_{ij} δεν είναι κοντά στο 0 ή το 1 και πως δεν είναι εύκολο να ξεχωρίσουμε σε ποιά ομάδα ανήκουν οι παρατηρήσεις αυτές.

6.9.5 Πεπερασμένα μείγματα διαφορετικών κατανομών

Ένας τρόπος για τη δημιουργία πολύπλοκων μοντέλων για τη μοντελοποίηση δεδομένων είναι ο συνδυασμός στη μορφή μειγμάτων από δύο διαφορετικές κατανομές. Κάτι τέτοιο εκτός από ευελιξία προσφέρει και τη δυνατότητα επιλογής μοντέλου, δηλαδή να βρούμε ποιό από διάφορα μοντέλα προσαρμόζει καλύτερα στα δεδομένα μας.

Στην προκειμένη περίπτωση ας υποθέσουμε ένα μοντέλο που έχει δύο υποπληθυσμούς, και συνάρτηση πυκνότητας πιθανότητας

$$f(x) = p\frac{1}{\theta}exp(-x/\theta) + (1-p)\frac{x^3\beta^4}{6}exp(-\beta x)$$

$x > 0$, $\theta, \beta > 0$, $0 \leq p \leq 1$. Δηλαδή έχουμε δύο συνιστώσες εκ των οποίων η πρώτη ακολουθεί μια εκθετική κατανομή ενώ η δεύτερη μια Γάμμα κατανομή με παραμέτρους $\alpha = 4$ και β.

Χρησιμοποιώντας τη συνηθισμένη τακτική για μίξεις κατανομών ας ορίσουμε τις λανθάνουσες μεταβλητές Z_i, $i = 1, \ldots, n$ που παίρνουν την τιμή 1 αν η παρατήρηση προέρχεται από τον εκθετικό πληθυσμό και 0 αν προέρχεται από τον πληθυσμό που ακολουθεί τη Γάμμα κατανομή. Επειδή έχουμε μόνο 2 υποπληθυμούς αρκεί να ορίσουμε μια μόνο λανθάνουσα μεταβλητή για κάθε παρατήρηση. Τα πλήρη δεδομένα θα έχουν τη μορφή $Y_i = (X_i, Z_i)$. Η λογαριθμική πιθανοφάνεια των πλήρων δεδομένων θα είναι

$$\begin{aligned} L(Y \mid data) &= \sum_{i=1}^{n} \left(z_i \log p + (1-z_i)\log(1-p)\right) \\ &- \sum_{i=1}^{n} z_i \left(\log\theta + \frac{x_i}{\theta}\right) \\ &+ \sum_{i=1}^{n} (1-z_i)\left(3\log x_i + 4\log\beta - \log 6 - \beta x_i\right) \end{aligned}$$

Συνεπώς χρειαζόμαστε στο Ε-βήμα να βρούμε τις αναμενόμενες τιμές των Z_i. Αυτό μπορεί να γίνει με τον κλασικό τρόπο στην περίπτωση μίξεων κατανομών ως εξής

- Ξεκινάμε με αρχικές τιμές για τις παραμέτρους. Αν συμβολίσουμε με $\phi = (\theta, \beta, p)$ το διάνυσμα που περιέχει όλες τις παραμέτρους, τότε η αρχική τιμή είναι η $\phi^{(0)}$. Ο τρόπος που θα επιλέξουμε τις αρχικές τιμές δεν είναι καθόλου απλός. Στην πράξη πρέπει να ξεκινήσουμε τον αλγόριθμο από διάφορες αρχικές τιμές για να είμαστε σίγουροι πως βρήκαμε το μέγιστο. Όπως θα δούμε στο συγκεκριμένο παράδειγμα η κατάσταση δεν είναι απλή και ο αλγόριθμος συγκλίνει μάλλον αργά. Έτσι ο αλγόριθμος μετά από r επαναλήψεις θα έχει τη μορφή

- Ε-βήμα: Υπολόγισε τις τιμές

$$w_i = E(Z_i \mid data, \phi^{(r)}) = \frac{p\frac{1}{\theta}exp(-x/\theta)}{p\frac{1}{\theta}exp(-x/\theta) + (1-p)\frac{x^3\beta^4}{6}exp(-\beta x)}$$

  όπου για κάθε παράμετρο χρησιμοποιούμε την τιμή που έχουμε βρεί από την προηγούμενη επανάληψη

- Μ-βήμα: Βρες την καινούριες εκτιμήτριες ως

$$\begin{aligned} p &= \frac{\sum_{i=1}^{n} w_i}{n} \\ \theta &= \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \\ \beta &= \frac{4\sum_{i=1}^{n}(1-w_i)}{\sum_{i=1}^{n}(1-w_i)x_i} \end{aligned}$$

  Οι τύποι προκύπτουν εύκολα από την πιθανοφάνεια των πλήρων δεδομένων. Παρατηρήστε πως αυτή γράφεται ουσιαστικά γράφεται ως τρία ξεχωριστά αθροίσματα και συνεπώς για κάθε παράμετρο μεγιστοποιούμε ξεχωριστά.

- Αν κάποιο κριτήριο σύγκλισης ικανοποιείται τότε σταμάτα τις επαναλήψεις αλλιώς πήγαινε πίσω στο Ε-βήμα.

Παράδειγμα. Έστω οι εξής 20 παρατηρήσεις:

5.492427, 7.987797, 5.265835, 10.468300, 7.248185
10.017489, 7.766692, 7.849390, 13.472998, 18.825573
9.819108, 12.291602, 8.262161, 6.305280, 3.014540
9.923024, 5.964323, 4.760085, 7.371590, 10.136333

Θέλουμε να προσαρμόσουμε το παραπάνω μοντέλο με τη χρήση του ΕΜ που μόλις περιγράψαμε. Στον πίνακα 6.5 μπορεί κανείς να δει τις εκτιμήσεις μαζί με τις διακυμάνσεις και τα διαστήματα εμπιστοσύνης βασισμένα στα ποσοστιαία σημεία από $B = 1000$ bootstrap επαναλήψεις.

Ειδικά για τα διαστήματα εμπιστοσύνης του p θα πρέπει να είναι κανείς πολύ προσεκτικός. Το κριτήριο σύγκλισης που χρησιμοποιήθηκε ήταν η σχετική αύξηση της πιθανοφάνειας, ο αλγόριθμος σταματούσε όταν ήταν μικρότερη από 10^{-8}. Αν χρησιμοποιηθεί πιο αυστηρό κριτήριο τότε τα αποτελέσματα θα αλλάξουν ελαφρώς. Η λογαριθμική πιθανοφάνεια του μοντέλου είναι -52.60567. Τα αποτελέσματα ουσιαστικά υποδεικνύουν ότι τα δεδομένα έχουν προέλθει από το Γάμμα κομμάτι

	εκτίμηση	διακύμανση	99%΄διάστημα εμπιστοσύνης
β	0.4644608	0.003643122	(0.352,0.700)
θ	9.075905	9.799054	(0.7603,18.8166)
p	0.0000005	0.002585328	(0.00000005,0.35128)

Πίνακας 6.5: Αποτελέσματα για τα δεδομένα

του μοντέλου, δηλαδή πως αρκούσε η κατανομή Γάμμα για να τα περιγράψει και πως αν θέλουμε να επιλέξουμε ανάμεσα στις 2 κατανομές θα προτιμήσουμε τη Γάμμα με μεγάλη ευκολία.

6.9.6 Αρνητική Διωνυμική Κατανομή

Είδαμε προηγουμένως τα πεπερασμένα μείγματα της κατανομής Poisson. Στην ουσία υποθέτουμε πως η παράμετρος της κατανομής Poisson δεν είναι σταθερά αλλά μια τυχαία μεταβλητή. Αν υποθέσουμε για την παράμετρο της κατανομής Poisson ότι ακολουθεί την κατανομή Γάμμα τότε προκύπτει η Αρνητική Διωνυμική κατανομή, που είναι και αυτή πολύ σημαντική για τη μοντελοποίηση διακριτών δεδομένων και συνήθως χρησιμοποιείται σε περιπτώσεις που η απλή κατανομή Poisson αποτυγχάνει. Από τον τρόπο γέννησης που μόλις αναφέραμε φαίνεται ξεκάθαρα η βαθιά σχέση ανάμεσα στις 2 κατανομές. Το μοντέλο που τις συνδέει είναι το εξής:

$$\begin{aligned} X \mid \theta &\sim Poisson(\theta) \\ \theta &\sim Gamma(\alpha, \beta), \end{aligned}$$

και η συνάρτηση πιθανότητας της αρνητικής διωνυμικής είναι η

$$P(x) = \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \left(\frac{\beta}{1+\beta}\right)^{\alpha} \left(\frac{1}{1+\beta}\right)^{x}, \quad x = 0, 1, \ldots, \quad \alpha, \beta > 0$$

Είναι ενδιαφέρον πως η γέννηση της αρνητικής διωνυμικής σαν μείγμα της κατανομής Poisson προσφέρει ένα ενδιαφέρον σχήμα συμπλήρωσης δεδομένων για τη χρήση του αλγορίθμου ΕΜ. Για παράδειγμα αν είχαμε παρατηρήσει τις τιμές των θ_i τότε η εκτίμηση θα ήταν πολύ απλή αφού ουσιαστικά θα είχαμε μόνο να εκτιμήσουμε τις παραμέτρους της κατανομής Γάμμα.

Επίσης είναι γνωστό πως η εκ των υστέρων κατανομή του θ_i είναι Γάμμα. Αυτό προκύπτει αρκετά εύκολα χρησιμοποιώντας το θεώρημα του Bayes και συγκεκριμένα

$$\begin{aligned} f(\theta \mid x) &= \frac{P(x \mid \theta) f(\theta)}{P(x)} \\ &= \frac{\frac{\exp(-\theta)\theta^x}{x!} \frac{\theta^{\alpha-1}\beta^{\alpha}\exp(-\beta\theta)}{\Gamma(\alpha)}}{\frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \left(\frac{\beta}{1+\beta}\right)^{\alpha} \left(\frac{1}{1+\beta}\right)^{x}} \\ &= \frac{\theta^{\alpha+x-1}(1+\beta)^{\alpha+x}}{\Gamma(\alpha+x)} \exp(-\theta(1+\beta)) \end{aligned}$$

Κάνοντας τις απλοποιήσεις μπορεί εύκολα κανείς να δει ότι η κατανομή είναι μια Γάμμα κατανομή και συγκεκριμένα ότι

$$\theta_i \mid x_i \sim Gamma(\alpha + x_i, \beta + 1).$$

Επομένως στην περίπτωση αυτή τα complete data ειναι τα ζεύγη $X_i, \theta_i)$ για ολα τα $i = 1, \ldots, n$. Αν τα θ_i είχαν παρατηρηθεί τότε άπλά θα έπρεπε κάποιος να μεγιστοποιήσει την πιθανοφάνειας από μια γάμμα κατανομή. Τα επαρκη στατιστικά για κάτι τέτοιο ειναι οι ποσότητες $\sum \theta_i$ και $\sum \log \theta_i$ κι επομένως για τον ΕΜ αλγόριθμο χρειαζόμαστε τις ποσότητες $E(\theta_i|x_i)$ και $E(\log \theta_i|x_i)$ από μια Γάμμα κατανομή. Αν και η απλή μέση τιμή ειναι μάλλον απλή ας δούμε εν συντομία τι συμβαίνει με την αναμενόμενη τιμή του λογαρίθμου.

΄Εστω η κατανομή Γαμμα με σππ που δίνεται στην (6.1). Για να βρούμε την αναμενόμενη τιμή του λογαρίθμου δηλαδή την $E(\log X)$ αρκεί να παρατηρήσει κανείς ότι

$$\frac{\Gamma(\alpha)}{\lambda^\alpha} = \int_0^\infty x^{\alpha-1} exp(-\lambda x) dx$$

Οπότε παραγωγίζοντας και στις δυο πλευρές της ισότητας ως προς α βρίσκουμε ότι

$$\frac{\Gamma'(\alpha)\lambda^\alpha - \lambda^\alpha \log(\lambda)\Gamma(\alpha)}{(\lambda^\alpha)^2} = \frac{\Gamma\left[\psi(\alpha) - \log(\lambda)\right]}{\lambda^\alpha} = \int_0^\infty \log x \;\; x^{\alpha-1} exp(-\lambda x) dx$$

όπου $\Psi(\alpha)$ είναι η δίγαμμα συνάρτηση που ορίζεται ως

$$\Psi(\alpha) = \frac{\partial \log \Gamma(\alpha)}{\partial \alpha} = \frac{1}{\Gamma(\alpha)} \frac{\partial \Gamma(\alpha)}{\partial \alpha}$$

Επομένως μπορεί εύκολα κανείς να δει ότι ισχύει

$$E(\log(X)) = \Psi(\alpha) - \log(\lambda)$$

Ο αλγόριθμος ΕΜ γίνεται επομένως:
Ε-βήμα: Υπολόγισε τις ψευδοτιμές

$$t_i = E(\theta_i|x_i) = \frac{x_i + \alpha^{old}}{1 + \beta^{old}} \;\text{ and }\; s_i = \Psi(\alpha^{old} + x_i) - \log(\beta^{old} + 1)$$

για $i = 1, \ldots, n$, όπου $\Psi(\cdot)$ είναι η δίγαμμα συνάρτηση . Αυτές οι αναμενόμενες τιμές προκύπτουν ως αναμενόμενες τιμές μιας γάμμα τυχαίας μεταβλητής καθώς είδαμε πως η δεσμευμένη κατανομή του θ_i είναι Γάμμα.
Μ-βήμα: Βρες τα καινούρια α και β ως $\beta^{new} = \alpha^{old}/\bar{t}$ και

$$\alpha^{new} = \alpha^{old} - \frac{\Psi(\alpha^{old}) + \log \beta^{new} - \bar{s}}{\Psi_3(\alpha^{old})}$$

όπου $\Psi_3(x)$ είναι η τρίγαμμα συνάρτηση

Ουσιαστικά στο Μ-βήμα έχουμε μια επανάληψη του αλγορίθμου Newton-Raphson για την κατανομή Γάμμα.

Αυτό που είναι ενδιαφέρον στον παραπάνω αλγόριθμο είναι πως στο Μ-βήμα δεν έχουμε κάποια εκτιμήτρια σε κλειστή μορφή αλλά στην ουσία χρησιμοποιούμε κάποια αριθμητική μέθοδο. Αυτό είναι χρήσιμο σε πολλές εφαρμογές καθώς ο ΕΜ δουλεύει ακόμα και σε αυτή την περίπτωση. Ο αλγόριθμος σε αυτή την περίπτωση ονομάζεται ECM (expectation- conditional maximization) καθώς στο Μ-βήμα έχουμε ουσιαστικά μια δεσμευμένη ως προς κάποια άλλη τιμή μεγιστοποίηση (στην περίπτωση μας θεωρούμε το β γνωστό που μόλις εκτιμήσαμε).

Τέλος θα πρέπει να αναφέρουμε πως επειδή ο ΕΜ έχει την ιδιότητα να αυξάνει την πιθανοφάνεια σε κάθε επανάληψη μερικές φορές αρκεί στο Μ-βήμα όχι να μεγιστοποιήσουμε την πιθανοφάνεια των πλήρων δεδομένων αλλά να σιγουρευτούμε πως η νέα εκτιμήτρια οδηγεί σε μεγαλύτερη πιθανοφάνεια από ότι η προηγούμενη επανάληψη. ΄Ενας τέτοιος αλγόριθμος ονομάζεται γενικευμένος ΕΜ (GEM Generalized EM).

6.9.7 Ομαδοποιημένα δεδομένα

Πολύ συχνά, παρά το γεγονός ότι τα δεδομένα είναι συνεχή παρουσιάζονται ή ακόμα συγκεντρώνονται με τη μορφή συχνοτήτων σε μικρά διαστήματα. Για παράδειγμα σε πολλές έρευνες αν και η ηλικία είναι μια συνεχής τυχαία μεταβλητή καταγράφεται σαν κατηγορική σε διαστήματα όπως $15-14$, $25-34$ κλπ. Τα δεδομένα που βλέπετε αφορούν την ομαδοποίηση 1000 παρατηρήσεων από κανονική κατανομή με γνωστή παράμετρο σ^2 και αφορούν τους χρόνους που μεσολάβησαν ανάμεσα σε 2 διαδοχικές τηλεφωνικές κλήσεις σε ένα τηλεφωνικό κέντρο (αφού έχει προηγηθεί τυποποίηση τους).

διάστημα	συχνότητα
≤ -3.5	1
$(-3.5, -3]$	2
$(-3, -2.5]$	6
$(-2.5, -2]$	18
$(-2, -1.5]$	41
$(-1.5, -1]$	93
$(-1, -0.5]$	145
$(-0.5, 0]$	182
$(0, 0.5]$	181
$(0.5, 1]$	156
$(1, 1.5]$	109
$(1.5, 2]$	43
$(2, 2.5]$	13
$(2.5, 3]$	9
$(3, 3.5]$	1
> 3.5	1

Αυτός ο τρόπος παρουσίασης και συλλογής δεδομένων είναι πολύ διαδεδομένος. Μπορεί κανείς να υπολογίσει περιγραφικά στατιστικά μέτρα από τέτοια δεδομένα χρησιμοποιώντας τους κατάλληλους τύπους για ομαδοποιημένα δεδομένα. Αυτό που δεν είναι γενικά γνωστό είναι πως αυτοί οι τύποι για ομαδοποιημένα δεδομένα σε καμιά περίπτωση δεν οδηγούν σε καλές εκτιμήτριες. Ας υποθέσουμε πως

τα δεδομένα προέρχονται από μια κανονική κατανομή με μέση τιμή μ και τυπική απόκλιση σ. Η εκτίμηση των παραμέτρων με τη μέθοδο μεγίστης πιθανοφάνειας δεν είναι καθόλου εύκολη υπόθεση. Για παράδειγμα η πιθανότητα του πρώτου κελιού είναι $P(x \leq -3.5) = \Phi(\frac{-3.5-\mu}{\sigma})$, ενώ για το δεύτερο κελί έχουμε πως η πιθανότητα είναι $P(-3.5 \leq x \leq -3) = \Phi(\frac{-3-\mu}{\sigma}) - \Phi(\frac{-3.5-\mu}{\sigma})$ όπου $\Phi(x)$ συμβολίζουμε την αθροιστική κατανομή της τυποποιημένης κανονικής κατανομής. Άρα η πιθανοφάνεια θα είναι

$$L(\mu, \sigma) = \prod_{i=1}^{k} p_i^{n_i}$$

όπου n_i είναι η συχνότητα του i κελιού και p_i η πιθανότητα του κελιού αυτού. Συνεπώς η μεγιστοποίηση της πιθανοφάνειας τέτοιων δεδομένων είναι μάλλον δύσκολη.

Το γεγονός όμως πως τα δεδομένα είναι ομαδοποιημένα είναι μια περίπτωση που έχουμε missing data, αντί να παρατηρήσουμε τα πλήρη δεδομένα παρατηρούμε μόνο αν ανήκουν ή όχι σε κάποιο διάστημα. Επομένως ο αλγόριθμος ΕΜ μπορεί να μας βοηθήσει.

Τα δεδομένα που έχουμε παρατηρήσει εμείς είναι τιμές $_i$, $i = 1, \ldots, 1000$ με τιμές από έως και 16 που αντιστοιχούν στα 16 κελιά που έχουμε. Αντιθέτως τα πλήρη δεδομένα είναι τιμές $_i$. Για τα πλήρη δεδομένα γνωρίζουμε πως η πλήρης πιθανοφάνεια θα είναι μια πιθανοφάνεια από την κανονική κατανομή κι επομένως μπορούμε να εκτιμήσουμε τις παραμέτρους απλά ως

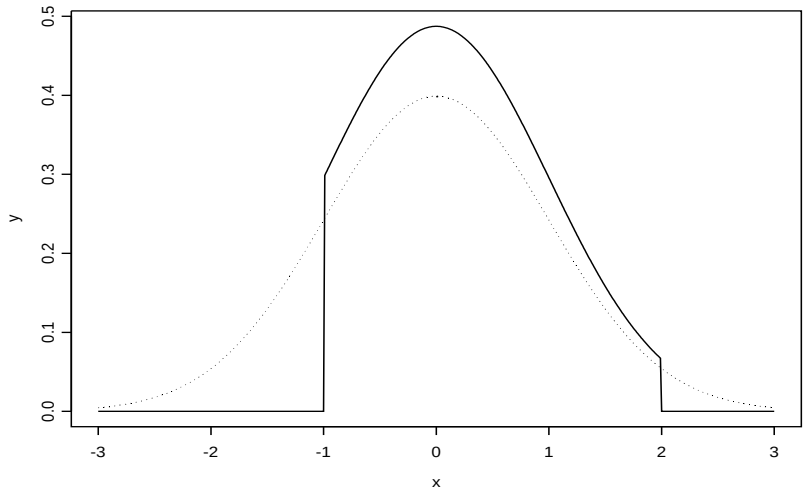
$$\hat{\mu} = \frac{\sum_{i=1}^{n} Y_i}{n}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} Y_i^2}{n} - \left[\frac{\sum_{i=1}^{n} Y_i}{n}\right]^2$$

Συνεπώς για τον ΕΜ αλγόριθμο χρειαζόμαστε στο Ε-βήμα τις αναμενόμενες τιμές των επαρκών στατιστικών $\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i^2$.

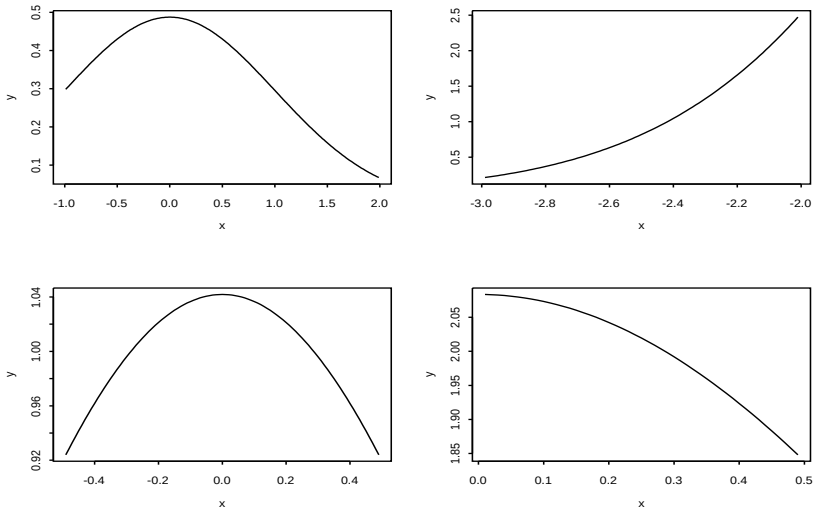
Αν τα όρια του κάθε κελιού τα συμβολίσουμε με x_L, x_U αντίστοιχα τότε μέσα σε κάθε κελί οι παρατηρήσεις μας ακολουθούν μια περικομμένη κανονική κατανομή, περικομμένη στο διάστημα $(x_L, x_U]$. Αν $f(x|\mu, \sigma)$ είναι η συνάρτηση πυκνότητας πιθανότητας μιας $N(\mu, \sigma^2)$ κατανομής τότε η συνάρτηση πυκνότητας πιθανότητας μιας περικομμένη κανονικής κατανομής στο διάστημα $(x_L, x_U]$ είναι η

$$f_T(x) = \begin{cases} 0 & x < x_L \\ \frac{f(x|\mu,\sigma)}{\int_{x_L}^{x_U} f(y|\mu,\sigma)dy} & x \in x_L, x_U) \\ 0 & x > x_U \end{cases}$$

Στο γράφημα 6.18 μπορεί κανείς να δει την τυποποιημένη κανονική κατανομή και την αντίστοιχη περικομμένη στο διάστημα $(-1, 2)$. Στο γράφημα 6.19 μπορεί να δει την περικομμένη τυποποιημένη κανονική κατανομή σε διάφορα διαστήματα. Παρατηρείστε πόσο διαφορετικά σχήματα μπορεί να πάρει που δεν θυμίζουν καθόλου την απλή κανονική κατανομή που ξέρουμε.



Γράφημα 6.18: Η τυποποιημένη κανονική κατανομή περικομμένη στο διάστημα $(-1, 2)$ και η απλή τυποποιημένη κανονική κατανομή (διακεκομμένη γραμμή)



Γράφημα 6.19: Η τυποποιημένη κανονική κατανομή περικομμένη σε διάφορα διαστήματα. Παρατηρείστε πόσο διαφορετικά σχήματα μπορεί να πάρει. Τα διαστήματα ειναι $(-1, 2), (-3, -2), (-0.5, 0.5), (0, 0.5)$.

Για το Ε-βήμα λοιπόν χρειαζόμαστε τις δύο πρώτες ροπές της περικομμένης κανονικής κατανομής. Αυτές δίνονται από τους τύπους

$$E(Y_T) = \mu + \sigma \left[\frac{\phi(\frac{x_L-\mu}{\sigma}) - \phi(\frac{x_U-\mu}{\sigma})}{\Phi(\frac{x_U-\mu}{\sigma}) - \Phi(\frac{x_L-\mu}{\sigma})} \right]$$

και

$$Var(Y_T) = 1 + \left[\frac{x_L\phi(\frac{x_L-\mu}{\sigma}) - x_U\phi(\frac{x_U-\mu}{\sigma})}{\Phi(\frac{x_U-\mu}{\sigma}) - \Phi(\frac{x_L-\mu}{\sigma})} \right] - \left[\frac{\phi(\frac{x_L-\mu}{\sigma}) - \phi(\frac{x_U-\mu}{\sigma})}{\Phi(\frac{x_U-\mu}{\sigma}) - \Phi(\frac{x_L-\mu}{\sigma})} \right]^2,$$

όπου $phi(\cdot)$ και $|Phi(\cdot)$ είναι η συνάρτηση πυκνότητας πιθανότητας και η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής αντίστοιχα. Ισχύει πως $E(Y_T^2) = Var(Y_T) + E(Y^T)^2$.

Επομένως ο αλγόριθμος ΕΜ μπορεί να περιγραφεί ως εξής:

Ε-Βήμα: Για κάθε κελί υπολόγισε τις ποσότητες $t_i = E(Y_T)$ και $r_i = E(Y_T^2)$ χρησιμοποιώντας τα όρια κάθε κελιού στου τύπους που δόθηκαν παραπάνω και τις τιμές των μ, σ που έχουμε μέχρι τώρα

Μ-Βήμα: Ενημέρωσε τις νέες παραμέτρους

$$\hat{\mu}_{new} = \frac{\sum_{i=1}^{n} t_i}{n}, \quad \hat{\sigma}^2_{new} = \frac{\sum_{i=1}^{n} r_i}{n} - \hat{\mu}^2_{new}$$

Για τα δεδομένα μας και με αρχικές τιμές $\mu = 0, \sigma = 1$ ο αλγόριθμος συγκλίνει πολύ γρήγορα μετά από 4 μόλις επαναλήψεις

μ	σ	$\log L$
0	1	−1970.2950
0.1089	1.05652	−1961.8893
0.1112	1.0589	−1961.8825
0.1112	1.0590	−1961.8825
0.1112	1.0590	−1961.8825

6.10 Περικομμένα - Λογοκριμένα δεδομένα

Εκτός από την περίπτωση ομαδοποιημένων δεδομένων πολλές φορές εμφανίζεται η περίπτωση όπου κάποιες παρατηρήσεις δεν έχουν παρατηρηθεί με ακρίβεια αλλά απλά γνωρίζουμε ότι ανήκουν σε κάποιο διάστημα. Κάτι τέτοιο προκύπτει πολύ συχνά στην ανάλυση επιβίωσης όπου για κάποιους ασθενείς ξέρουμε ότι επιβίωσαν μέχρι κάποια χρονική στιγμή αλλά μετά δεν γνωρίζουμε κάτι (είτε γιατί τέλειωσε η έρευνα, είτε γιατί τα άτομα αποσύρθηκαν από αυτήν). Σε αυτή την περίπτωση λέμε πως έχουμε λογοκριμένα δεδομένα (censored data).

Συνήθως σε αυτή την περίπτωση, παρατηρήσεις μεγαλύτερες από κάποια τιμή απλά τις περιλαμβάνουμε σε ένα διάστημα χάνοντας την πληροφορία που υπάρχει στην ουρά της κατανομής. Ας δουλέψουμε με το εξής παράδειγμα:

Τα δεδομένα που βλέπετε αφορούν 20 παρατηρήσεις από εκθετική κατανομή και αφορούν τους χρόνους επιβίωσης πειραματόζωων σε ένα νέο φάρμακο, σε μήνες. Οι παρατηρήσεις ήταν

$$\begin{aligned}
&0.05511269,\ 0.07437246,\ 0.11098159,\ 0.13552999\\
&0.20371014,\ 0.22090690,\ 0.23699706,\ 0.27435875,\\
&0.28669966,\ 0.47155521,\ 0.96822690,\ 0.97200651,\\
&1.04514368,\ 1.37989648,\ 1.49109121,\ 1.53370336
\end{aligned}$$

ενώ για τις υπόλοιπες 4 παρατηρήσεις έχει απλά καταγραφεί ότι είχαν τιμή μεγαλύτερη από 1.70. Θέλουμε να εκτιμήσουμε την παράμετρο θ της εκθετικής κατανομής.

Στην περίπτωση μας τα δεδομένα είναι γκρουπαρισμένα μόνο στην τελευταία ομάδα. Δηλαδή για τις 16 παρατηρήσεις ξέρουμε ακριβώς τις τιμές τους και μόνο για τις 4 τελευταίες έχουμε πως ανήκουν στο διάστημα $(1.70, \infty)$. Το πρόβλημα αφορά την εκτίμηση της παραμέτρου θ της εκθετικής κατανομής όταν έχουμε grouped δεδομένα και ουσιαστικά είναι μια πολύ απλή περίπτωση καθώς στην περίπτωση που θα μας απασχολήσει μόνο κάποιες παρατηρήσεις είναι ομαδοποιημένες ενώ οι υπόλοιπες δίνονται στην πλήρη τους μορφή.

΄Εστω η εκθετική κατανομή με συνάρτηση πυκνότητας πιθανότητας $f(x) = \theta exp(-\theta x)$, $x, \theta > 0$. Από ένα δείγμα $(X_1, X_2, \ldots, X_n)$ μεγέθους n ξέρουμε (ή τουλάχιστον μπορούμε πολύ εύκολα) να βρούμε πως η εκτιμήτρια μεγίστης πιθανοφάνειας είναι η $\hat{\theta} = 1/\bar{x} = n/\sum x_i$. Ανατρέχοντας στη βασική λογική του EM, αυτό που χρειαζόμαστε είναι να βρούμε την αναμενόμενη τιμή για τις 4 παρατηρήσεις που δεν ξέρουμε την ακριβή τους τιμή. Για αυτές τις παρατηρήσεις γνωρίζουμε μόνο ότι ανήκουν στο διάστημα $[1.70, \infty)$. (Επειδή οι παρατηρήσεις προέρχονται από συνεχή κατανομή δεν έχει σημασία αν το διάστημα είναι ανοικτό ή κλειστό, θα το συμβολίσουμε κλειστό χωρίς αυτό να δημιουργεί κανένα πρόβλημα).

Για να συζητήσουμε το πρόβλημα πιο γενικά, ας υποθέσουμε πως έχουμε ένα δείγμα μεγέθους n. Από τις n γνωρίζουμε πλήρως τις τιμές των m παρατηρήσεων, τις οποίες συμβολίζουμε ως $(X_1, X_2, \ldots, X_m)$ ενώ για τις υπόλοιπες k παρατηρήσεις ξέρουμε πως ανήκουν στο διάστημα (t, ∞). Η περίπτωση που μας απασχολεί έχει $n = 20$, $m = 16$, $k = 4$ και $t = 1.70$. ΄Εστω λοιπόν οι λανθάνουσες μεταβλητές $Z_1, \ldots, Z_k$ που αντιστοιχούν στις μη πλήρως παρατηρηθείσες παρατηρήσεις. Τα πλήρη δεδομένα θα είναι το διάνυσμα $Y = (X_1, \ldots, X_m, Z_1, \ldots, Z_k)$.

Η λογαριθμική πιθανοφάνεια των πλήρων δεδομένων, βασισμένοι στο γεγονός πως τα πλήρη δεδομένα ακολουθούν εκθετική κατανομή, θα είναι η

$$L(Y \mid \theta) = n \log \theta - \theta \left(\sum_{i=1}^{m} X_i + \sum_{i=1}^{k} Z_i \right)$$

και επομένως στο Μ-βήμα θα πρέπει να υπολογίζουμε την ποσότητα

$$\hat{\theta}^{(r)} = \frac{n}{\sum_{i=1}^{m} X_i + \sum_{i=1}^{k} Z_i}$$

Το μόνο που δεν γνωρίζουμε είναι η τιμή των Z_i την οποία θα πρέπει να υπολογίσουμε στο Ε-βήμα. Συγκεκριμένα αυτό που ξέρουμε είναι πως δοθείσας της πληροφορίας πως τα Z_i ανήκουν στο διάστημα $[t, \infty)$ αυτά ακολουθούν και πάλι εκθετική κατανομή περιορισμένη (truncated) στο διάστημα $[t, \infty)$, δηλαδή

$$f(z_i \mid data) = \theta exp(-\theta z_i) I(z_i \geq t), \quad \theta > 0$$

Αυτό που χρειαζόμαστε είναι η αναμενόμενη τιμή $E(Z_i \mid data)$, η οποία μπορεί να βρεθεί πολύ εύκολα ότι είναι

$$E(Z_i \mid data) = t + \frac{1}{\theta}$$

(Γενικά για κάθε εκθετική κατανομή περικομμένη στο διάστημα $[t, \infty)$ η αναμενόμενη τιμή είναι $E(Z_i) = t + \frac{1}{\theta}$ Αυτό προκύπτει εύκολα είτε λύνοντας το ολοκλήρωμα ή ισοδύναμα χρησιμοποιώντας την ιδιότητα της εκθετικής κατανομής περί έλλειψης μνήμης). Συνεπώς ο ΕΜ αλγόριθμος που πρεπει να χρησιμοποιηθεί είναι ο εξής:

- Ξεκινάμε με αρχική τιμή $\theta^{(0)}$. Μια απλή τέτοια αρχική τιμή είναι η $\theta^{(0)} = \bar{x}^{-1}$, δηλαδή ο αντίστροφος της μέσης τιμής για τα δεδομένα που έχουμε τις πραγματικές τιμές τους. Έτσι ο αλγόριθμος μετά από r επαναλήψεις θα έχει τη μορφή
- Ε-βήμα: Για τις τιμές $Z_1, \ldots, Z_k$ υπολόγισε την αναμενόμενη τους τιμή (που θα είναι ίδια για όλες)

$$s = E(Z_i \mid data, \theta^{(r)}) = t + \frac{1}{\theta^{(r)}}, \quad i = 1, \ldots, k.$$

- Μ-βήμα: Βρες την καινούρια εκτιμήτρια ως

$$\hat{\theta}^{(r+1)} = \frac{n}{\sum_{i=1}^{m} X_i + ks}$$

- Αν κάποιο κριτήριο σύγκλισης ικανοποιείται τότε σταμάτα τις επαναλήψεις αλλιώς πήγαινε πίσω στο Ε-βήμα.

Είναι πολύ ενδιαφέρον πως για το μοντέλο που συζητάμε η πιθανοφάνεια των παρατηρηθέντων δεδομένων είναι πολύ εύκολο να γραφτεί αλλά και να μεγιστοποιηθεί χωρίς τη χρήση του ΕΜ αλγορίθμου. Συγκεκριμένα, η συνεισφορά στην πιθανοφάνεια των k παρατηρήσεων τις οποίες έχουμε σε ένα διάστημα είναι $P(X \geq t) = exp(-\theta t)$. Επομένως η παρατηρηθείσα πιθανοφάνεια είναι η

$$L(X \mid \theta) = m \log \theta - \theta \sum_{i=1}^{m} X_i - kt\theta$$

και επομένως η εκτιμήτρια μεγίστης πιθανοφάνειας είναι η

$$\hat{\theta}_{ML} = \frac{m}{\sum_{i=1}^{m} X_i + kt}$$

Μπορεί εύκολα να επαληθεύσει κανείς πως ο αλγόριθμος που είδαμε πιο πριν συγκλίνει σε αυτή την τιμή. Επομένως για το τυπικό σφάλμα αρκεί κανείς να βρει τη δεύτερη παράγωγο της πιθανοφάνειας $L(X \mid \theta)$ που δώσαμε πιο πάνω ή να χρησιμοποιήσει parametric bootstrap. Για την περίπτωση αυτή τα δείγματα, μεγέθους n, θα προέρχονται από την εκθετική κατανομή με τιμή την τιμή που βρήκαμε από το δείγμα, όμως θα πρέπει να υπάρξει μέριμνα ώστε παρατηρήσεις με τιμή πάνω από t να θεωρούνται ως μια ομάδα για την οποία απλά έχουμε τη συχνότητά της.

Για τα δεδομένα βρίσκουμε πως $\hat{\theta}_{ML} = 0.9840$. Η λογαριθμική πιθανοφάνεια όταν μεγιστοποιήσαμε την πιθανοφάνεια είναι -16.2581980. Χρησιμοποιώντας parametric bootstrap με $B = 1000$ επαναλήψεις εκτιμήσαμε τη διακύμανση της εκτιμήτριας ως 0.0628. Χρησιμοποιώντας την πιθανοφάνεια βρήκαμε πως η διακύμανση της εκτιμήτριας είναι 0.0605 ($Var(\hat{\theta}) = \theta^2/m$).

Από τα παραπάνω μπορούμε να γενικεύσουμε σε πολλά επίπεδα. Κατά αρχάς μπορούν όλα τα δεδομένα να είναι ομαδοποιημένα και όχι μόνο σε κάποιο διάστημα. Έπειτα μπορεί να έχουμε οποιαδήποτε κατανομή, αν και άλλες κατανομές οδηγούν κατά κανόνα σε αρκετά πιο περίπλοκους τύπους. Επίσης τα δεδομένα μπορεί να είναι διακριτά και όχι συνεχή ή ακόμα και πολυδιάστατα.

Τέλος πρέπει να γίνει σαφές πως το παράδειγμα μας είναι μια τυπική περίπτωση δεδομένων με censoring που συναντάμε συχνά σε βιοστατιστικές εφαρμογές. Στην ουσία το μόνο που ξέρουμε για τις 4 τελευταίες παρατηρήσεις είναι πως ξεπερνούν την τιμή 1.70, σαν να σταμάτησε σε αυτή την τιμή το πείραμα. Θα μπορούσαμε να γενικεύσουμε όταν το censoring δεν εμφανίζεται σε τιμές μεγαλύτερες του 1.70 αλλά σε διάδορους ασθενείς και διάφορες τιμές. Ουσιαστικά έχουμε μια νέα δίτιμη μεταβλητή που μας δείχνει αν η παρατήρηση είναι λογοκριμένη ή όχι. Ο αλγόριθμος θα δουλέψει όπως τον περιγράψαμε. Στο Ε-βήμα θα εκτιμησουμε την αναμενόμενη τιμή των λογοκριμένων παρατηρήσεων με τη χρήση της αναμενόμενη τιμής από μια περικομμένη εκθετική κατανομή, ενώ το Μ-βήμα θα μετατραπεί κατάλληλα.

6.11 Παραλλαγές του αλγορίθμου EM

Μιλήσαμε προηγουμένως για δύο παραλλαγές του EM, συγκεκριμένα για τον ECM και τον GEM. Και οι 2 αυτές παραλλαγές αφορούσαν προβλήματα κατά το Μ-βήμα. Στην πράξη μερικές φορές το πρόβλημα υπάρχει στο Ε-βήμα καθώς οι αναμενόμενες τιμές που χρειάζονται δεν υπάρχουν σε κλειστή μορφή. Οι δύο διαφορετικές κατευθύνσεις τότε είναι είτε να υπολογίσουμε τα ολοκληρώματα που αφορούν τις αναμενόμενες τιμές με κάποια αριθμητική μέθοδο είτε να προχωρήσουμε σε Monte Carlo ολοκλήρωση δηλαδή να υπολογίσουμε τις αναμενόμενες τιμές με προσομοίωση. Θα ασχοληθούμε με τη δεύτερη αυτή προσέγγιση καθώς είναι πιο εύκολη στην υλοποίηση της με απλά στατιστικά πακέτα.

Υπάρχουν δύο διαφορετικές μέθοδοι. Αυτές είναι:

- Stochastic EM (SEM) : αυτός ο αλγόριθμος σε κάθε Ε-βήμα αντί να υπολογίζει την αναμενόμενη τιμή απλά χρησιμοποιεί μια τυχαία μεταβλητή που γεννάμε από τη δεσμευμένη κατανομή των missing data για δοθείσες τιμές των παραμέτρων (τις μέχρι εκείνη την επανάληψη τιμές) και τα πραγματικά δεδομένα. Η προσομοίωση από τη δεσμευμένη αυτή κατανομή μπορεί να επιτευχθεί είτε απλά προσομοιώνοντας από μια γνωστή κατανομή είτε χρησιμοποιώντας κάποιες πιο προχωρημένες τεχνικές όπως MCMC. Για παράδειγμα στο γεντικό πρόβλημα της ενότητας 6.9.1 είχαμε δει πως η δεσμευμένη κατανομή του x_1 ήταν η διωνυμική. Επομένως το Ε- βήμα του SEM αλγορίθμου απλά απαιτεί να γεννήσουμε μια τιμή από αυτή την κατανομή. Το Μ-βήμα παραμένει όπως είναι. Δηλαδή ο αλγόριθμος γίνεται

  Ε-βήμα: Προσομοιώσε μια τιμή t από τη $Binomial(125, \frac{\theta}{\theta+2})$ κατανομή

  Μ-βήμα: Ανανέωσε το θ όπως και πριν, δηλαδή

  $$\theta^{(new)} = \frac{t + x_4}{t + x_2 + x_3 + x_4}$$

  Ο αλγόριθμος SEM έχει κάποιες καλές ιδιότητες. Συγκλίνει στην περιοχή του μεγίστου. Δεν αυξάνει την πιθανοφάνεια σε κάθε επανάληψη αλλά έχει μια ξεκάθαρη τάση προς το μέγιστο και όταν φτάσει στην περιοχή του απλά κυμαίνεται γύρω από αυτό. Αυτό επιτρέπει να εκτιμήσουμε το τυπικό σφάλμα των εκτιμητριών. Αυτό που είναι πολύ σημαντικό για τον αλγόριθμο είναι πως επειδή έχει ένα στοχαστικό μέρος σε μερικές περιπτώσεις μπορεί να ξεπεράσει τοπικά μέγιστα και να συνεχίσει προς το ολικό μέγιστο κάτι που ποτέ δεν μπορεί να καταφέρει ο απλός ΕΜ.

- Monte Carlo EM (MCEM): Ο αλγόριθμος υπολογίζει το Ε-βήμα με Monte Carlo ολοκλήρωση. Με στατιστικούς όρους εκτιμά την αναμενόμενη τιμή παίρνοντας ένα μεγάλο δείγμα από την κατανομή και χρησιμοποιώντας τη δειγματική μέση τιμή ως εκτιμήτρια. Μοιάζει με τον SEM στο ότι και οι δυο προσομοιώνουν από τη δεσμευμένη κατανομή στο Ε-βήμα αλλά ο MCEM προσομοιώνει πολλές τιμές από όπου κι εκτιμά τη μέση τιμή ενώ ο SEM μόνο μια τιμή.

  Δηλαδή ο MCEM χρησιμοποιεί στο Ε-βήμα το γεγονός πως η αναμενόμενη τιμή προσεγγίζεται από την αντίστοιχη δειγματική δηλαδή

  $$E(z \mid data, \theta^{(r)}) \simeq \frac{1}{M} \sum_{i=1}^{M} z^{(i)}$$

  όπου $z^{(i)}$ είναι μια τυχαία μεταβλητή από την $f(z \mid data, \theta^{(r)})$. Για $M = 1$ έχουμε τον αλγόριθμο SEM

  Επομένως για το γενετικό πρόβλημα ο MCEM θα πάρει τη μορφή:

  Ε-βήμα: Προσομοιώσε M τιμές $t_i, i = 1, \ldots, M$ από τη $Binomial(125, \frac{\theta}{\theta+2})$ κατανομή. Υπολόγισε την ποσότητα

  $$\bar{t} = \frac{\sum_{i=1}^{M} t_i}{M}$$

M-βήμα: Ανανέωσε το θ όπως και πριν, δηλαδή

$$\theta^{(new)} = \frac{\bar{t} + x_4}{\bar{t} + x_2 + x_3 + x_4}$$

Αντίστοιχα για το παράδειγμα με τα πεπερασμένα μείγματα ουσιαστικά ένας αλγόριθμος MCEM υπολογίζει με Monte Carlo την αναμενόμενη τιμή των Z_{ij} και δεδομένου ότι η δεσμευμένη κατανομή είναι μια πολυωνυμική κατανομή αρκεί κανείς να προσομοιώσει, έστω m τιμές Z_i, δηλαδή ολόκληρο το διάνυσμα με 0 και 1 για την i παρατήρηση και στη συνέχεια να βρει τις σχετικές συχνότητες για καθένα από αυτά. Στη βιβλιογραφία είναι γνωστό πως αν διαλέξουμε μικρό m ο αλγόριθμος για την περίπτωση πεπερασμένων μειγμάτων μπορεί να έχει καλές ιδιότητες καθώς μπορούν έτσι να ξεπεραστούν τοπικά μέγιστα που όμως δεν είναι και ολικά μέγιστα και άρα ο αλγόριθμος δεν παγιδεύεται.

Θα πρέπει να σημειωθεί πως και στον MCEM η πιθανοφάνεια μπορεί να μην αυξάνει σε κάθε επανάληψη λόγω του Monte Carlo σφάλματος που εισάγουμε. Για να περιορίσουμε αυτό το σφάλμα θα πρέπει να διαλέξουμε ένα πολύ μεγάλο M. Από τη θεωρία για πολύ μεγάλο M η εκτιμήτρια της αναμενόμενης τιμής θα είναι πολύ κοντά στην πραγματική τιμή και επομένως ο αλγόριθμος θα μοιάζει περισσότερο με τον ΕΜ. Σε αυτή την περίπτωση όμως θα είναι πολύ αργός. Συνεπώς μια συμβιβαστική λύση είναι ξεκινήσουμε στις πρώτες επαναλήψεις με μικρό M το οποίο θα αυξάνει όσο πλησιάζουμε στο μέγιστο.

- Generalized EM: Κάτω από αυτή την επωνυμία αναφερόμαστε σε αλγορίθμους που επειδή η μεγιστοποίηση που χρειάζεται στο Μ- βήμα δεν είναι εύκολο να γίνει τότε αρκεί κατά τη διάρκεια του Μ- βήματος όχι να μεγιστοποιήσουμε την πιθανοφάνεια των πλήρων δεδομένων αλλά να βρούμε μια τιμή που μας εξασφαλίζει πως μεγαλώσαμε έστω και λίγο την πιθανοφάνεια που είχαμε από την προηγούμενη επανάληψη. Δηλαδή σιγουρευόμαστε πως ο αλγόριθμος διατηρεί τη μονοτονία του και πως σε κάθε επανάληψη η πιθανοφάνεια μεγαλώνει άσχετα με το γεγονός πως δεν μεγαλώνει όσο πιθανότατα θα ήταν εφικτό. Για παράδειγμα, στην περίπτωση της αρνητικής διωνυμικής που είδαμε προηγουμένως δεν χρειαζόμαστε τις Newton-Raphson επαναλήψεις αλλά να σιγουρευτούμε πως τα καινούρια α, β που θα βρούμε έχουν καλύτερη πιθανοφάνεια. Τέτοιο αλγόριθμοι είναι εξαιρετικά χρήσιμοι όταν η μεγιστοποίηση στο Μ-βήμα είναι δύσκολη ή χρονοβόρα.

6.11.1 MCEM για την αρνητική διωνυμική κατανομή

Στο παράδειγμα με την αρνητική διωνυμική κατανομή μια από τις αναμενόμενες τιμές χρειαζόταν τη δίγαμμα συνάρτηση. Μπορούμε να αποφύγουμε κάτι τέτοιο χρησιμοποιώντας τον αλγόριθμο MCEM. Είδαμε πως η δεσμευμένη κατανομή είναι Γάμμα και συνεπώς μπορούμε εύκολα να προσομοιώσουμε τυχαίες μεταβλητές από αυτή. Έτσι στο Ε-βήμα αρκεί να εκτιμήσουμε τις αναμενόμενες τιμές της Γάμμα που χρειαζόμαστε με προσομοίωση από την κατανομή. Δηλαδή προσομοιώνουμε

τιμές $\theta^{(j)}$, $j = 1, \ldots, m$, από τη $Gamma(\alpha + x_i, \beta + 1)$ κατανομή και στη συνέχεια υπολογίζουμε τις ποσότητες

$$t_i = E(\theta_i \mid X_i) = m^{-1} \sum_{j=1}^{m} \theta^{(j)} \quad \text{και} \quad s_i = E(\log \theta_i \mid X_i) = m^{-1} \sum_{j=1}^{m} \log \theta^{(j)}$$

Το Μ-βήμα παραμένει το ίδιο όπως και προηγουμένως.

Θα πρέπει να παρατηρήσει κανείς πως στην περίπτωση της αρνητικής διωνυμικής όπου οι αναμενόμενες τιμές μπορούν να υπολογιστούν ο MCEM δεν είναι χρήσιμος. Σε πολλά άλλα μοντέλα όμως ακόμα και η συνάρτηση πιθανότητας δεν μπορεί να γραφτεί σε κλειστή μορφή οπότε πόσο μάλλον οι αναμενόμενες τιμές. Για παράδειγμα αν αντί για την κατανομή Γάμμα χρησιμοποιήσουμε τη λογαριθμική κανονική κατανομή τα πράγματα περιπλέκονται. Πολλά μοντέλα τυχαίων επιδράσεων όμως χρησιμοποιούν τη λογαριθμική κανονική κατανομή σε συνδυασμό με την κατανομή Poisson. Με τον MCEM (και τον SEM) δεν χρειάζεται να γράψουμε σε κλειστή μορφή την πιθανοφάνεια των δεδομένων και επομένως μπορούμε να προβούμε σε εκτίμηση μεγίστης πιθανοφάνειας ακόμα και σε τόσο περίπλοκα μοντέλα.

Στο Γράφημα 6.20 μπορεί κανείς να δει μια εφαρμογή του MCEM σε πραγματικά δεδομένα από την αρνητική διωνυμική κατανομή. Οι τιμές του M είναι 1, 10, 100 ενώ στο τελευταίο γράφημα αυξάνουν κατά 10 σε κάθε επανάληψη. Όσο μικρότερη είναι η τιμή τόσο πιο πολύ τρεμοπαίζει η εκτιμήτρια. Είναι όμως πολύ σημαντικό να παρατηρήσει κανείς ότι ανεξάρτητα με την τιμή του M όλες οι περιπτώσεις πλησιάζουν γρήγορα και με την ίδια ταχύτητα το μέγιστο.

6.12 Γιατί ο αλγόριθμος συγκλίνει

Ας δούμε γιατί ο αλγόριθμος ΕΜ σε κάθε επανάληψη μεγαλώνει την πιθανοφάνεια, δηλαδή να δούμε γιατί ο αλγόριθμος συγκλίνει. Ας υπενθυμίσουμε λίγο τους συμβολισμούς και ας εισάγουμε κάποιους ακόμα. Τα παρατηρούμενα δεδομένα συμβολίζονται με X ενώ τα πλήρη (complete) δεδομένα με Y. Οι παράμετροι που θέλουμε να εκτιμήσουμε συμβολίζονται με το διάνυσμα $\psi = (\psi_1, \ldots, \psi_d)$. Η κατανομή των πραγματικών δεδομένων είναι $g(y; \psi)$ ενώ των πλήρων δεδομένων είναι $g_c(x; \psi)$. Τέλος η πιθανοφάνεια που θέλουμε να μεγιστοποιήσουμε είναι η $L(\psi)$ ενώ με $L_c(\psi)$ θα συμβολίσουμε την πιθανοφάνεια αντίστοιχα των πλήρων δεδομένων. Ουσιαστικά θέλουμε να δείξουμε πως

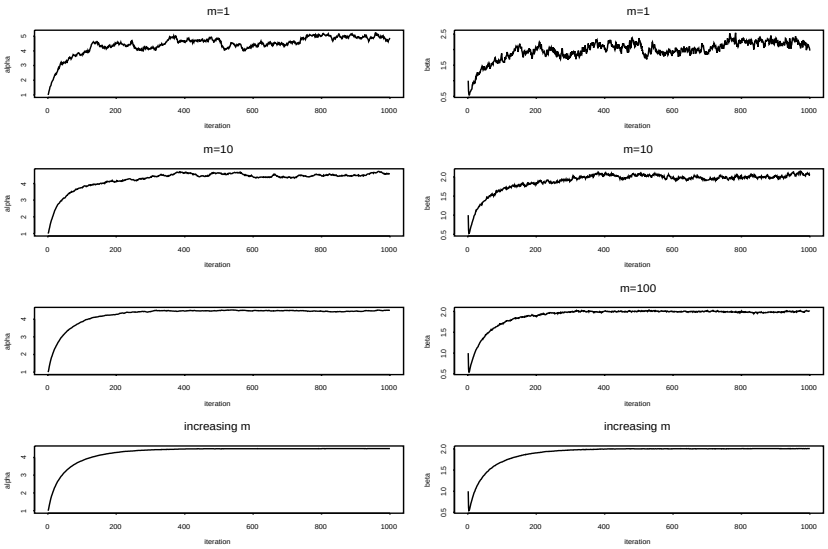
$$L(\psi^{(k+1)}) \geq L(\psi^{(k)})$$

όπου $\psi^{(k)}$ συμβολίζουμε τις παραμέτρους μετά την k επανάληψη.

Η δεσμευμένη κατανομή των πλήρων δεδομένων δοθέντος των παρατηρούμενων δεδομένων θα είναι

$$g(y \mid x; \psi) = \frac{g_c(y; \psi)}{g(x; \psi)}$$

επειδή έχουμε μια ένα προς ένα αντιστοίχιση των πλήρων δεδομένων στα πραγματικά και επομένως η από κοινού τους κατανομή θα είναι ίδια με την κατανομή των



Γράφημα 6.20: MCEM για την αρνητική διωνυμική με διαφορετικές επιλογές του M.

πλήρων δεδομένων. Έτσι η λογαριθμική πιθανοφάνεια γράφεται ως

$$\begin{aligned} \log L(\psi) &= \log g(x;\psi) \\ &= \log g_c(y;\psi) - \log g(y \mid x;\psi) \\ &= \log L_c(\psi) - \log g(y \mid x;\psi) \end{aligned}$$

Αν τώρα πάρουμε τις αναμενόμενες τιμές και στα δύο μέρη της ισότητας ως προς την δεσμευμένη κατανομή των πλήρων δεδομένων δοθέντος των πραγματικών (δηλαδή την κατανομή του $y \mid x$ αλλά ως προς την τιμή $\psi^{(k)}$ της παραμέτρου κι επομένως την εκτίμηση μέχρι την επανάληψη k) προκύπτει πως

$$\begin{aligned} \log L(\psi) &= E_{y|x;\psi^{(k)}}\left[\log g_c(y;\psi)\right] - E_{y|x;\psi^{(k)}}\left[\log g(y \mid x;\psi)\right] \\ &= Q(\psi;\psi^{(k)}) - H(\psi,\psi^{(k)}) \end{aligned}$$

όπου $Q(\psi;\psi^{(k)})$ είναι η ποσότητα που μεγιστοποιούμε σε κάθε Μ-βήμα. Επομένως η διαφορά της πιθανοφάνειας σε δύο διαδοχικές επαναλήψεις είναι

$$\begin{aligned} \log L(\psi^{(k+1)}) - \log L(\psi^{(k)}) &= \left[Q(\psi^{(k+1)};\psi^{(k)}) - Q(\psi^{(k)};\psi^{(k)})\right] \\ &- \left[H(\psi^{(k+1)};\psi^{(k)}) - H(\psi^{(k)};\psi^{(k)})\right] \end{aligned}$$

Για τον πρώτο όρο γνωρίζουμε πως από την κατασκευή του ΕΜ αλγορίθμου ότι στο

Μ-βήμα έχουμε μεγιστοποιήσει τη συνάρτηση Q και άρα ισχύει πως

$$Q(\psi^{(k+1)};\psi^{(k)}) \geq Q(\psi^{(k)};\psi^{(k)})$$

Επομένως αρκεί να δείξυμε πως

$$H(\psi^{(k+1)};\psi^{(k)}) - H(\psi^{(k)};\psi^{(k)}) \leq 0$$

Από τον ορισμό προηγουμένως της $H(\psi;\psi^{(k)})$ έχουμε πως

$$H(\psi;\psi^{(k)}) = E_{y|x;\psi^{(k)}}\left[\log g(y \mid x;\psi)\right]$$

και άρα

$$\begin{aligned} H(\psi;\psi^{(k)}) - H(\psi^{(k)};\psi^{(k)}) &= E_{y|x;\psi^{(k)}}\left[\log g(y \mid x;\psi)\right] \\ &- E_{y|x;\psi^{(k)}}\left[\log g(y \mid x;\psi^{(k)})\right] \\ &= E_{y|x;\psi^{(k)}}\left[\log g(y \mid x;\psi) - \log g(y \mid x;\psi^{(k)})\right] \\ &= E_{y|x;\psi^{(k)}}\left[\log \frac{g(y \mid x;\psi)}{g(y \mid x;\psi^{(k)})}\right] \end{aligned}$$

Γνωρίζουμε όμως πως από την ανισότητα του Jensen ισχύει πως

$$E(\log X) \leq \log E(X)$$

για κάθε τμ X.

Επομένως και στην περίπτωση μας θα είναι

$$E_{y|x;\psi^{(k)}}\left[\log \frac{g(y \mid x;\psi)}{g(y \mid x;\psi^{(k)})}\right] \leq \log E_{y|x;\psi^{(k)}}\left[\frac{g(y \mid x;\psi)}{g(y \mid x;\psi^{(k)})}\right] = 0$$

γιατί

$$E_{y|x;\psi^{(k)}}\left[\frac{g(y \mid x;\psi)}{g(y \mid x;\psi^{(k)})}\right] = \int \frac{g(y \mid x;\psi)}{g(y \mid x;\psi^{(k)})} g(y \mid x;\psi^{(k)})dy = \int g(y \mid x;\psi)dy = 1$$

κι επομένως ο λογάριθμος θα είναι 0. Συνεπώς η ποσότητα είναι πάντα 0 και άρα η πιθανοφάνεια αυξάνεται σε κάθε επανάληψη.

6.13 Επίλογος

Θα τελειώσουμε αυτή τη σύντομη περιγραφή του αλγορίθμου ΕΜ με μερικές παρατηρήσεις

- Αυτό που είναι πολύ σημαντικό με τον ΕΜ είναι πως δεν είναι μια απλή αριθμητική μέθοδος αλλά επιτρέπει πολύ σημαντική στατιστική διαίσθηση κατά την εκτίμηση και ερμηνεία των αποτελεσμάτων.

- Ο αλγόριθμος ΕΜ χρησιμοποιείται και για Μπευζιανές εκτιμήσεις. Σύμφωνα με την Μπευζιανή προσέγγιση και με συγκεκριμένη συνάρτηση ζημιάς ενδιαφερόμαστε να βρούμε την κορυφή της εκ των υστέρων κατανομής (posterior). Ο ΕΜ έχει χρησιμοποιηθεί και για αυτά τα προβλήματα καθώς αν κανείς χρησιμοποιήσει μη πληροφοριακή prior τοτε ουσιαστικά η posterior είναι η πιθανοφάνεια και άρα η κορυφή της ισοδυναμεί με την εύρεση τυ μεγίστου της πιθανοφάνειας.

- Ο αλγόριθμος με μικρές αλλαγές μπορεί να χρησιμοποιηθεί και για άλλες μεθόδους εκτίμησης όπως penalized maximum likelihood. Συνεπώς δεν είναι σωστό πως περιορίζεται αποκλειστικά σε προβλήματα μέγιστης πιθανοφάνειας.

- Υπάρχουν πολλές ομοιότητες του ΕΜ με την ιδιαίτερα δημοφιλή μέθοδο του MCMC στη Μπευζιανή στατιστική. Συγκεκριμένα η ιδέα της συμπλήρωσης των δεδομένων είναι κοινή και στις 2 μεθόδους και σε πολλά προβλήματα χρησιμοποιούν ακριβώς τις ίδιες τεχνικές. Ο ΕΜ έχει δύο βήματα που δεν είναι στοχαστικά, ενώ στο MCMC όλα τα βήματα είναι στοχαστικά. Οι παραλλαγές όμως του ΕΜ χρησιμοποιούν στοχαστικά βήματα που μοιάζουν πολύ με αυτά του MCMC. Συνεπώς υπάρχει μια σημαντική συνάφεια στις προσεγγίσεις.

- Ο ΕΜ μπορεί να χρησιμοποιηθεί συμπληρωματικά με άλλες αριθμητικές μεθόδους για ακόμα καλύτερα αποτελέσματα. Για παράδειγμα η ευκολία με την οποία πλησιάζει το μέγιστο ακόμα και αν ξεκινήσει από πολύ μακριά μπορεί να χρησιμοποιηθεί συμπληρωματικά με άλλες μεθόδους που αν βρεθούν κοντά στο μέγιστο το βρίσκουν σε λίγες επαναλήψεις. Επίσης το ίδιο το Μ-βήμα μπορεί να περιέχει κάποιες αριθμητικές μεθόδους μεγιστοποίησης.

- Αν και δεν αναφέρθηκαν εδώ, υπάρχουν τεχνικές για να κάνει κανείς πιο γρήγορο τον ΕΜ. Μια πρώτη ιδέα είναι να βρει κανείς καλύτερο data augmentation. Εναλλακτικά χρειάζονται καλύτερες αρχικές τιμές, ή κριτήρια που να σταματάνε πιο γρήγορα τον αλγόριθμο όταν αυτός έχει φτάσει στο μέγιστο.

6.14 Ασκήσεις

1. Η διμεταβλητή κατανομή Poisson αποτελεί μια γενίκευση της απλής κατανομής Poisson σε 2 διαστάσεις. Η από κοινού συνάρτηση πιθανότητας των τυχαιων μεταβλητών X και Y δίνεται από τον τύπο

$$\begin{aligned} P_{X,Y}(x,y) &= P(X=x, Y=y) = \\ & e^{-(\theta_1+\theta_2+\theta_3)} \frac{\theta_1^x}{x!}\frac{\theta_2^y}{y!} \sum_{i=0}^{min(x,y)} \left(\begin{array}{c} x \\ i \end{array}\right)\left(\begin{array}{c} y \\ i \end{array}\right) i! \left(\frac{\theta_3}{\theta_1\theta_2}\right)^i . \end{aligned}$$

   Οι περιθώριες κατανομές είναι κατανομές Poisson με μέση τιμή $\theta_1+\theta_3$ και $\theta_2+\theta_3$ αντίστοιχα για τις τυχαίες μεταβλητές X και Y ενώ η παράμετρος θ_3 είναι η συναδιακύμανση των X και Y. Ένας ενδιαφέρον τρόπος για να

προκύψει η κατανομή αυτή είναι ο εξής: Ας υποθέσουμε πως έχουμε 3 τυχαίες μεταβλητές X_1, X_2, X_3 που είναι ανεξάρτητες και καθε μια ακολουθεί κατανομή Poisson με παράμετρο θ_i, $i = 1, 2, 3$ αντίστοιχα. Τότε οι τυχαίες μεταβλητές $X = X_1 + X_3$ και $Y = X_2 + X_3$ ακολουθούν τη διμεταβλητή κατανομή Poisson.

- Χρησιμοποιείστε την παραπάνω ιδιότητα γέννησης της κατανομής για να κατασκευάσετε έναν αλγόριθμο ΕΜ για την εκτίμηση των παραμετρων
- Χρησιμοποιείστε τον αλγόριθμο Newton-Raphson για την εκτίμηση των παραμέτρων με τη μέθοδο μεγίστης πιθανοφάνειας.
- Συγκρίνετε τους 2 αλγορίθμους. Τι παρατηρείτε;

2. Έστω μια τυχαία μεταβλητή X που ακολουθεί την κατανομή Poisson με παράμετρο θ και μια τυχαία μεταβλητή Y που ακολουθεί τη διωνυμική κατανομή με παραμέτρους n, p. Έστω η τυχαία μεταβλητή $Z = X + Y$.

   - Βρείτε τη συνάρτηση πιθανότητας της Z.
   - Περιγράψτε έναν αλγόριθμο ΕΜ για την εκτίμηση με τη μέθοδο μεγίστης πιθανοφάνειας των παραμέτρων της κατανομής από ένα δείγμα $Z_1, Z_2, \ldots, Z_N$ όπου όμως η παράμετρος n είναι γνωστή.
   - Ποιά η σχέση του αλγορίθμου με τον αλγόριθμο που κατασκευάσατε για τη διμεταβλητή κατανομή Poisson του προηγούμενου ερωτήματος;

3. Έστω το ακόλουθο πείραμα. Για μια σειρά από λαμπτήρες, καταγράφεται ο χρόνος ζωής τους. Για μια άλλη παρτίδα όμως καταγράφεται μόνο αν αυτοί είναι ακόμα εν λειτουργία μετά από t ώρες και όχι ο ακριβής χρόνος ζωής τους. Έστω πως για N λαμπτήρες έχουμε τον ακριβή χρόνο ζωής τους και για M λαμπτήρες την πληροφορία αν αυτοί ακόμα λειτουργούσαν ή όχι μετά από t ώρες.

   - Αν υποθέσουμε πως ο χρόνος ζωής των λαμπτήρων ακολουθεί την εκθετική κατανομή τότε να εκτιμηθεί χρησιμοποιώντας έναν αλγόριθμο ΕΜ η παράμερος της εκθετικής κατανομής.
   - Αν τώρα υποθέσουμε ότι η κατανομή δεν είναι εκθετική αλλά ομοιόμορφη στο διάστημα $(0, \theta)$ κατασκευάστε και πάλι τον αλγόριθμο. Προσομοιώστε μάλιστα δεδομένα ώστε να ελέγξετε αν ο αλγόριθμος τρέχει σωστά. Τι παρατηρείτε;

4. Ας υποθέσουμε πως έχουμε k πληθυσμούς όπου κάθε ένας από αυτούς ακολουθεί μια εκθετική κατανομή με παράμετρο θ_j, $\quad j = 1, \ldots, k$, δηλαδή

$$f(x) = \sum_{j=1}^{k} p_j \theta_j exp(-\theta_j x)$$

   Δυστυχώς όμως το k το πλήθος δηλαδή των υποπληθυσμών δεν είναι γνωστό και πρέπει να εκτιμηθεί με τη μέθοδο μέγιστης πιθανοφάνειας. Πως μπορεί ένα αλγόριθμος ΕΜ να εκτιμήσει το k από ένα δείγμα $X_1, X_2, \ldots, X_n$;

5. Η κατανομή Birnbaum-Saunders χρησιμοποιείται για να περιγράψει κατανομές επιβίωσης ιδιαίτερα σε θέματα αξιοπιστίας. Η συνάρτηση πυκνότητας πιθανότητας της έχει τη μορφή

$$\begin{aligned} f(x;\alpha,\beta) &= \frac{1}{2\alpha\beta}\left(\frac{x}{\beta}\right)^{-1/2}\left[1+\left(\frac{x}{\beta}\right)^{-1}\right] \\ &\times \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2\alpha^2}\left(\frac{x}{\beta}-2+\frac{\beta}{x}\right)\right], \quad x,\alpha,\beta>0 \end{aligned}$$

   - Κατασκευάστε έναν αλγόριθμο Newton-Rahson για να μεγιστοποιήσετε την πιθανοφάνεια ενός τυχαίου δείγματος μεγέθους n από την κατανομή αυτή.
   - Η κατανομή μπορεί να γραφτεί ως ένα μείγμα δύο αντίστροφων κανονικών κατανομών (Inverse Gaussian) με παραμέτρους $(\beta, \alpha^{-2}\beta)$ και $(\beta, \alpha^{-2}\beta^{-1})$ αντίστοιχα. Χρησιμοποιείστε αυτή την ιδιότητα για να κατασκευάσετε έναν αλγόριθμο ΕΜ για να εκτιμήσετε τις παραμέτρους του προηγούμενου δείγματος με τη μέθοδο μεγίστης πιθανοφάνειας.
   - Ποια από τις δύο μεθόδους προτιμάτε και γιατί; Ποια είναι πιο γρήγορη;

6. ΄Ενα από τα πιο χαρακτηριστικά παραδείγματα εφαρμογής του αλγόριθμου ΕΜ είναι η απλή περίπτωση που κάποια δεδομένα δεν έχουν καταγραφεί. Ας υποθέσουμε τις παρακάτω παρατηρήσεις που προέρχονται από μια διμεταβλητή κανονική κατανομή. Κατασκευάστε έναν αλγόριθμο ΕΜ για να εκτιμήσετε τις παραμέτρους της διμεταβλητής κανονικής κατανομής με από κοινού συνάρτηση πυκνότητας πιθανότητας

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right.\right. \\ &+ \left.\left.2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\} \end{aligned}$$

   Οι παράμετροι που θέλουμε να εκτιμήσουμε είναι $(\mu_x, \mu_y, \sigma_x\sigma_y, \rho)$ Τα δεδομένα είναι (με * συμβολίζουμε τις τιμές που δεν έχουμε)

x	8	11	16	18	6	4	20	25	9	13
y	10	14	16	15	20	4	18	22	$*$	$*$

7. Σε αρκετές περιπτώσεις τα δεδομένα παρουσιάζουν για κάποιες τιμές πολύ μεγαλύτερες συχνότητες από ότι ένα απλό μοντέλο υποθέτει. Ειδικά στην περίπτωση που μας ενδιαφέρει ο αριθμός των παιδιών μιας οικογένειας έχει παρατηρηθεί το φαινόμενο η τιμή 1 να εμφανίζεται πολύ περισσότερες φορές από ότι θα περίμενε κανείς. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί inflated μοντέλα που έχουν τη μορφή

$$P_{inf}(X=x) = \begin{cases} (1-\omega)P(X=x) & x=0,2,3,\ldots \\ (1-\omega)P(X=x)+\omega & x=1 \end{cases}$$

Στην περίπτωση που μας απασχολεί ας υποθέσουμε πως το αρχικό μοντέλο $P(X = x)$ είναι μια μίξη από δύο κατανομές, μια Poisson με παράμετρο λ και μια γεωμετρική θ. δηλαδή το μοντέλο έχει την πλήρη μορφή

$$P_{inf}(X = x) = \begin{cases} (1-\omega)(\pi P_1(X = x) + (1-\pi)P_2(X = x)) & x = 0, 2, 3, \ldots \\ (1-\omega)(\pi P_1(X = x) + (1-\pi)P_2(X = x)) + \omega & x = 1 \end{cases}$$

όπου

$$P_1(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \;\; \lambda \geq 0, x = 0, 1, \ldots$$

$$P_2(X = x) = \theta(1-\theta)^x, \;\; 0 \leq \theta \leq 1, x = 0, 1, \ldots$$

Τα δεδομένα αφορούν τον αριθμό των παιδιών ανά νοικοκυριό. Οι συχνότητες ήταν οι εξής:

x	0	1	2	3	4	5	6	7	8
n_x	365	509	179	94	33	15	3	1	1

όπου n_x είναι η συχνότητα της τιμής x. Για τα δεδομένα που σας δίνονται να εκτιμήσετε με τη μέθοδο μεγίστης πιθανοφάνειας τις παραμέτρους του μοντέλου και τα τυπικά τους σφάλματα.

8. Η κατανομή που ακολουθεί ονομάζεται κατανομή του Lindlay. Η συνάρτηση πυκνότητας πιθανότητας της δίνεται από τον τύπο

$$f(x) = \frac{p^2}{p+1}(x+1)\exp(-xp), \;\; x, p > 0$$

Είναι ενδιαφέρον να παρατηρήσεις κανείς πως η κατανομή αυτή μπορεί να γραφτεί ως ένα μείγμα μιας Γάμμα κατανομής και μιας εκθετικής ως εξής

$$\pi Gamma(2, p) + (1-\pi)Expo(p)$$

όπου $\pi = \frac{p}{p+1}$

Χρησιμοποιείστε αυτή την αναπαράσταση της κατανομής για να κατασκευάσετε έναν αλγόριθμο ΕΜ για την εκτίμηση της παραμέτρου p.

9. Τα δεδομένα που ακολουθούν αφορούν τους χρόνους επιβίωσης (σε ώρες) ενός εξαρτήματος όταν αυτό δοκιμάστηκε σε 4 διαφορετικές θερμοκρασίες

βαθμών Κελσίου.

θερμοκρασία			
150	200	250	300
72∗	69	54	49
71	65	50	44
65	64	52	42
72∗	72∗	48	49
67	64	40	35
56	70	50	38
70	55	48∗	39
72∗	57	40	41
68	60	45	40∗
70	60∗	56	39

Οι αστερίσκοι υποδηλώνουν πως οι παρατηρήσεις είναι λογοκριμένες (censored) δηλαδή ότι το πείραμα σταμάτησε σε αυτό το σημείο και άρα γνωρίζουμε πως ο πραγματικός χρόνος ζωής είναι τουλάχιστον αυτός. Ένας ερευνητής θέλει να προσαρμόσει ένα απλό γραμμικό μοντέλο της μορφής

$$t_i = \alpha + \beta x_i + \epsilon_i$$

όπου $x_i = 1/t_i$, t_i είναι η θερμοκρασία για την i παρατήρηση, και $\epsilon_i \sim N(0, \sigma^2)$ Δηλαδή πρόκειται για ένα τυπικό γραμμικό μοντέλο μόνο που έχουμε λογοκριμένες παρατηρήσεις. Κατασκευάστε έναν αλγόριθμο ΕΜ για να εκτιμήσετε τις παραμέτρους του μοντέλου

10. Ένας πληθυσμός έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x) = p\frac{1}{\sigma_1\sqrt{2\pi}}exp\left(-\frac{x^2}{2\sigma_1^2}\right) + (1-p)\frac{1}{\sigma_2\sqrt{2\pi}}exp\left(-\frac{x^2}{2\sigma_2^2}\right)$$

δηλαδή ένα μείγμα δύο κανονικών κατανομών με μέση τιμή 0 και διακύμανση σ_i^2, $i = 1, 2$ αντίστοιχα. Έστω ένα τυχαίο δείγμα $X_1, \ldots, X_n$ από αυτή την κατανομή. Ενδιαφερόμαστε για την κατασκευή ενός αλγορίθμου ΕΜ για την εκτίμηση των παραμέτρων με τη μέθοδο μεγίστης πιθανοφάνειας.

(a) Περιγράψτε με λεπτομέρεια ποία είναι τα πλήρη δεδομένα (complete data)

(b) Ποιά είναι η complete likelihood ;

(c) Περιγράψτε λεπτομερώς τα βήματα του αλγορίθμου καθώς και τον τρόπο με τον οποίο καταλήξατε σε αυτά .

(d) Τι θα άλλαζε αν χρησιμοποιούσαμε τον αλγόριθμο MCEM. Περιγράψτε με λεπτομέρεια τα βήματα ;

11. Ας υποθέσουμε πως έχουμε k πληθυσμούς όπου κάθε ένας από αυτούς ακολουθεί μια αντίστροφη κανονική κατανομή με παράμετρο δ_j, $j = 1, \ldots, k$, δηλαδή

$$f(x \mid \delta_j) = \frac{\delta_j}{\sqrt{2\pi}}exp(\delta_j)x^{-3/2}exp\left(-\frac{1}{2}\left(\frac{\delta_j^2}{x} + x\right)\right), \quad x, \delta_j > 0.$$

κι επομένως

$$f(x) = \sum_{j=1}^{k} p_j f(x \mid \delta_j).$$

όπου $p_j > 0$ και $\sum p_j = 1$.

- Κατασκευάστε έναν αλγόριθμο Newton-Rahson για να μεγιστοποιήσετε την πιθανοφάνεια ενός τυχαίου δείγματος μεγέθους n από την κατανομή αυτή, υποθέτωντας πως $k = 3$.
- Χρησιμοποιήστε τον αλγόριθμο simulated annealing για να εκτιμήσετε τις παραμέτρους για το ίδιο δείγμα.
- Περιγράψτε έναν αλγόριθμο ΕΜ για να εκτιμήσετε τις παραμέτρους της κατανομής αυτής υποθέτωντας ότι k γνωστό.
- Δυστυχώς όμως το k το πλήθος δηλαδή των υποπληθυσμών δεν είναι γνωστό και πρέπει να εκτιμηθεί με τη μέθοδο μέγιστης πιθανοφάνειας. Πως μπορεί ένα αλγόριθμος ΕΜ να εκτιμήσει το k;
- Περιγράψτε έναν αλγοριθμο MCEM για το ίδιο πρόβλημα

12. 'Εστω παρατηρήσεις από μια κατανομή Γάμμα με παραμέτρους $\alpha = 2$ και β δηλαδή

$$f(x) = \frac{x\beta^2}{\Gamma(2)} exp(-\beta x), \quad x, \beta > 0.$$

΄Ενα τυχαιο δείγμα από την κατανομή ήταν το ακόλουθο

$1*$	0.5115795	0.2286870	0.6034591	$1*$
0.2986166	0.6870007	0.5977033	0.7093217	$1*$
0.4955427	0.3500113	0.5261613	0.3384634	0.4048628
0.4007785	0.3488920	0.4605230	0.2427288	0.3455211

όπου το $*$ σημαίνει πως η παρατήρηση είναι censored. Εκτιμήστε την παράμετρο του μοντέλου για αυτό το δείγμα με οποιαδήποτε μέθοδο θέλετε.

13. Η διωνυμική κατανομή έχει συνάρτηση πιθανότητας

$$P(X = x) = \left(\begin{array}{c} N \\ x \end{array} \right) p^x (1-p)^{N-x}, \quad 0 \leq p \leq 1, x = 0, 1, \ldots, N.$$

Στα δεδομένα $X_1, \ldots, X_n$ που ένας ερευνητής έχει στα χέρια του, η τιμή 0 δεν υπάρχει, δηλαδή για κάποιους λόγους δέν μπορεί να παρατηρηθεί κι επομένως τα δεδομένα προέρχονται από μια περικομμένη στο 0 διωνυμική κατανομή. Αναπτύξτε έναν αλγόριθμο ΕΜ για την ευρεση της εκτιμήτριας μεγίστης πιθανοφάνειας του p, έστω $\hat{p}_{ML}$, θεωρώντας την τιμή του N γνωστή. Συγκεκριμένα θα πρέπει να απαντήσετε με ακρίβεια τα ακολουθα

α. Ποιά ειναι η λογαριθμική πιθανοφάνεια των δεδομένων;

β. Ποιά ειναι τα πλήρη δεδομένα (complete data) που θα χρησιμοποιήσετε για να φτιαξετε τον αλγόριθμο ΕΜ;

γ. Ποιά ειναι η λογαριθμική πιθανοφάνεια των πλήρων δεδομένων (complete likelihood) ;

δ. Περιγράψτε πλήρως τα βήματα του ΕΜ αλγοριθμου ωστε να μπορεί ο αναγνώστης να υλοποιήσει πλήρως τον αλγόριθμο.

Βιβλιογραφία

Dempster, A.P., Laird, N.M. and Rubin, D., (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society,* B 39, 1-38.

Το πρώτο άρθρο το οποίο παρουσίασε τον αλγόριθμο στη γενική του μορφή. Πριν από αυτό ο αλγόριθμος είχε χρησιμοποιηθεί σε πολλές εφαρμογές. Αυτό το άρθρο ομαδοποίησε αυτές τις εφαρμογές κάτω από ένα πολύ γενικό πρίσμα. Περιέχει πολλά παραδείγματα και τη θεωρία του αλγορίθμου.

McLachlan, G. and Krishnan, N. (1997). *The EM Algorithm and Extensions*. Wiley and Sons, NY.

Το μοναδικό μέχρι στιγμής βιβλίο αφιερωμένο στον αλγόριθμο. Περιλαμβάνει τα αοτελέσματα σχετικά με τον αλγόριθμο μέχρι το 1997, όπως και πολλά παραδείγματα και πρακτικά θέματα σχετικά με τον αλγόριθμο.

McLachlan, G. and Peel, N. (2001). *Finite mixture models*. Wiley and Sons NY.

Το βιβλίο ασχολείται με πεπερασμένα μείγματα κατανομών. Καθώς ο αλγόριθμος ΕΜ αποτελεί σημαντικότατο εργαλείο για αυτά τα μοντέλα υπάρει εκτενής αναφορά σε αυτόν.

Meng, X.L. and Rubin, D. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, **80**, 267-278.

Το άρθρο αυτό εισήγαγε την παραλλαγή του αλγορίθμου, τον αλγόριθμο ECM.

Meng, X.L. and Van Dyk, D. (1997). The EM Algorithm: An Old Folk Song Sung to a Fast New Tune (with discussion). *Journal of the Royal Statistical Society*, **B 59**, 511-567.

Θα μπορούσε να πεί κανείς πως το άρθρο αυτό είναι η συνέχεια του άρθρου των Dempster etal (1977) είκοσι χρόνια μετά. Περιγράφονται οι εξελίξεις και η πρόοδος που σημειώθηκε αυτά τα 20 χρόνια που μεσολάβησαν και δίνεται έμφαση στη στατιστική ερμηνεία του αλγορίθμου.

Morgan, B.J,T. (2000) *Applied Stochastic Modelling*. Arnold Publishers, NY.

Το βιβλίο περιλαμβάνει μια πλήρη αναφορά σε πρακτικά θέματα εφαρμογής Στατιστικών μοντέλων στην πράξη. Ένα μεγάλο μέρος αφιερώνεται στα προβλήματα μεγιστοποίησης που προκύπτουν και για ατό παρουσιάζονται διάφορες αριθμητικές μέθοδοι συμπεριλαμβανομένου και του αλγορίθμου ΕΜ.